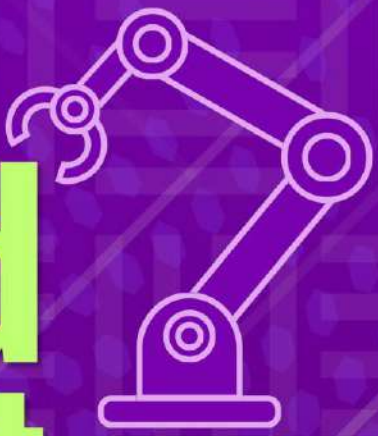
Indústria 4.0



Soluções tecnológicas viáveis baseadas em Internet das Coisas e Inteligência Artificial

Luciano Ferreira Silva
Organizador

Pedro & João editores

PPI4.0 cits CDDDC SUFRAMA

Copyright © Luciano Ferreira Silva (Organizador)

Todos os direitos garantidos. Qualquer parte desta obra pode ser reproduzida, transmitida ou arquivada desde que levados em conta os direitos das autoras e dos autores.

Luciano Ferreira Silva [Org.]

Indústria 4.0: Soluções tecnológicas viáveis baseadas em Internet das Coisas e Inteligência Artificial. São Carlos: Pedro & João Editores, 2026. 452p. 16 x 23 cm.

ISBN: 978-65-265-3173-0 [Digital]

DOI: 10.51795/9786526531730

1. Indústria 4.0. 2. Internet das Coisas. 3. Inteligência Artificial. 4. Tecnologia da Informação e Comunicação. I. Silva, Luciano Ferreira.

CDD – 370

Revisão Técnica: Luciano Ferreira Silva
Diagramação: Marcelo Henklain
Capa: Marcelo Henklain
Ficha Catalográfica: Hélio Márcio Pajeú – CRB - 8-8828
Editores: Pedro Amaro de Moura Brito & João Rodrigo de Moura Brito

Conselho Científico da Pedro & João Editores:
Augusto Ponzio (Bari/Itália); João Wanderley Geraldi (Unicamp/Brasil); Hélio Márcio Pajeú (UFPE/Brasil); Maria Isabel de Moura (UFSCar/Brasil); Maria da Piedade Resende da Costa (UFSCar/Brasil); Valdemir Miotello (UFSCar/Brasil); Ana Cláudia Bortolozzi (UNESP/Bauru/Brasil); Mariangela Lima de Almeida (UFES/Brasil); José Kuiava (UNIOESTE/Brasil); Marisol Barenco de Mello (UFF/Brasil); Camila Caracelli Scherma (UFFS/Brasil); Luís Fernando Soares Zuin (USP/Brasil); Ana Patrícia da Silva (UERJ/Brasil).

Pedro & João Editores
www.pedroejoaoeditores.com.br
13568-878 – São Carlos – SP
2026

Luciano Ferreira Silva
(Organizador)

Indústria 4.0:

Soluções tecnológicas viáveis baseadas em Internet das Coisas e Inteligência Artificial

Pedro & João editores
2026

Sumário

Prefácio | 06
Luciano Ferreira Silva

Capítulo 01 - Abordagem baseada em visão computacional e análise multivariada para segurança patrimonial na Indústria 4.0 | 12
Rogério de Carvalho Cordovil e Herbert Oliveira Rocha

Capítulo 02 - Segurança na Indústria 4.0: Detecção de anomalias em redes Modbus/TCP por meio de um IDS Híbrido K-Means/DBSCAN | 30
Katerine Jones da Silva e Felipe Leite Lobo

Capítulo 03 - Resiliência cibernética na Indústria 4.0: Detecção de ataques de desfiguração via modelos transformer | 43
Hugo Lima Romão, Marcelle Alencar Urquiza e Filipe Dwan Pereira

Capítulo 04 - Detecção de mensagens de phishing utilizando Random Forest e Processamento de Linguagem Natural em ambientes da Indústria 4.0 | 58
Pedro Vinícius da Silva Ribeiro e Felipe Leite Lobo

Capítulo 05 - Resiliência e observabilidade na Indústria 4.0: Diagnóstico out-of-band com LORAWAN | 72
Jussara Miliano Soares e Josiane do Couto Rodrigues

Capítulo 06 - Análise de protocolos IoT para o trabalhador conectado em cenários de conectividade restrita | 85
Bruno Cesar da Silva Claudino e Leandro Nelinho Balico

Capítulo 07 - Implementação de Arquitetura Zero Trust em IoT com sensores que simulam ambiente industrial utilizando ESP32 e detecção de anomalias na borda | 100
Mailson Lima Garcia e Wender Antônio da Silva

Capítulo 08 - PAND: Orquestração entre Visão Computacional e LLM para tomada de decisão na manutenção fotovoltaica | 114
Matheus Naranjo Corrêa e Filipe Dwan Pereira

Capítulo 09 - Caracterização Não Supervisionada de perturbações elétricas em subestações de distribuição de energia utilizando técnicas de Clustering e Deep Learning | 129
Rafael Benzaquem Neto e Josiane do Couto Rodrigues

Capítulo 10 - Sistema inteligente para diagnóstico da qualidade de energia em baixa tensão com autoencoders aplicados à Indústria 4.0 | 144
Yankarlo Silva Carvalho e Alexander Fernandez Correa

Capítulo 11 - Estimativa do estado de saúde de baterias de Íon-Lítio utilizando Machine Learning no contexto da Indústria 4.0 | 157
Alessandro Medeiros Ferreira e Alexander Fernández Correa

Capítulo 12 - RAG Aplicado a Manuais Automotivos: Segmentação Contextual e Recuperação Híbrida | 170
Fredson Silva de Souza e Filipe Dwan Pereira

Capítulo 13 - Pipeline de Aprendizado de Máquina para detecção de anomalias em ortomosaicos capturados por drone em áreas com predominância de cobertura vegetal | 184
Felipe Abrahim Ramalho Xaud e Felipe Leite Lobo

Capítulo 14 - Reconstrução espectral com SRF gaussianas e BI-LSTM para detecção de anomalias agrícolas no contexto da Agricultura 4.0 | 201
Johnny de Mendonça Pereira e Felipe Leite Lobo

Capítulo 15 - Arquitetura híbrida Edge-Cloud: Protótipo IoT Offline-First para Agroindústria 4.0 em áreas remotas de Roraima | 215
Janei Vieira Pereira e Clenya Rejane Barros de Lima

Capítulo 16 - Monitoramento IoT para Agricultura 4.0: Arquitetura com LoRa P2P e Edge Computing para Eficiência Hídrica | 229
Pedro Henrique Vale de Lucena Brasil, Marcelle Alencar Urquiza e Herbert Oliveira Rocha

Capítulo 17 - Monitoramento da qualidade do ar em Boa Vista/RR com sensores IoT: Dataset aberto e visualização comunitária na perspectiva da Indústria 4.0 | 242
Cedric Carol Patrician Williams Filho e Clenya Rejane Barros de Lima

Capítulo 18 - Gestão de Resíduos em Unidades Básicas de Saúde de Boa Vista: Uma proposta baseada em IoT de baixo custo | 259
Ana Beatriz Bezerra de Menezes dos Santos e Alexander Fernández Correa

Capítulo 19 - Análise comparativa entre Arima e LSTM na previsão de escoamento logístico de soja em Roraima | 271
Hiago Pereira e Susset Guerra Jimenez

Capítulo 20 - Manutenção preditiva com Inteligência Artificial Explicável Integrada ao Odoo ERP | 285
Paulo Roberto de Souza Mesquita Junior e Thais Oliveira Almeida

Capítulo 21 - RAG-BI: Aprimorando Dashboards de Business Intelligence com RAG | 299
Filipe Matias de Sousa e Clenya Rejane Barros de Lima

Capítulo 22 - Pipeline sequencial de Deep Learning para detecção de ativos e OCR em documentação técnica da Indústria 4.0 | 314
Rodrigo Damasceno Nascimento e Alexander Fernandez Correa

Capítulo 23 - Arquitetura distribuída Edge-Server para inventário em tempo real com RFID e técnicas de localização por proximidade | 327
Kelly Silva de Aquino e Herbert Oliveira Rocha

Capítulo 24 - Estruturação de registros em ordens de serviço de manutenção com metamodelo de linguagem e embeddings de sentenças aplicado à Indústria 4.0 | 349
Maristela Veiga dos Santos e Clenya Rejane Barros Lima

Capítulo 25 - Aplicação de Grandes Modelos de Linguagem na triagem de currículos não padronizados | 366
Adolfo Izidio Ferreira da Silva Neto e Susset Guerra Jimenez

Capítulo 26 - Uso de Processamento de Linguagem Natural para triagem automatizada das publicações dos Diários Oficiais de Roraima | 380
Everton dos Reis e Josiane do Couto Rodrigues

Capítulo 27 - Chatbot para transparência orçamentária: Arquitetura híbrida neuro-simbólica para consultas em Linguagem Natural aos dados do Governo de Roraima | 394
Jadir Rafael Bolanha de Aguiar e Lindeval Fernandes de Lima

Capítulo 28 - Sistema IoT de baixo custo para monitoramento de vibração em motores industriais baseado em ESP32 e norma ISO 10816 | 409
William Lopes Rodrigues e Leandro Nelinho Balico

Capítulo 29 - Plataforma IoT de monitoramento e análise estatística para ambientes industriais: Uma solução versátil de baixo custo | 425
Ortiz Marcos Martins Pedroso e Leandro Nelinho Balico

Capítulo 30 - Framework adaptativo para adoção de Dark Factory em PMEs de Roraima | 438
Sara Silva Gomes e Marcelle Alencar Urquiza

Prefácio

Luciano Ferreira Silva
Universidade Federal de Roraima

Esta obra consiste em uma seleção dos estudos de maior destaque da Especialização em Computação Aplicada à Indústria 4.0, projeto apoiado pela SUFRAMA, com recursos da Lei nº 8.387, de 23 de 1991, no âmbito do PPI-Indústria 4.0, coordenado pelo CITS-Amazônia, aprovado pelo CAPDA/SUFRAMA e executado pelo Departamento de Ciência da Computação (DCC) da Universidade Federal de Roraima (UFRR). No total, foram 46 estudos desenvolvidos, sendo 29 da ênfase de Inteligência Artificial e 17 de Internet das Coisas. Desse universo, selecionamos 30 trabalhos, que contemplam temas de importância fundamental para subsidiar o desenvolvimento da Indústria 4.0, brasileira.

Como ponto de partida, assumimos que não temos como falar em tecnologias sem pensar em conectividade e segurança. Por essa razão, abrimos o livro com os estudos sobre **"Redes e Segurança da informação"**. No estudo do Capítulo 01, "Abordagem baseada em visão computacional e análise multivariada para segurança patrimonial na Indústria 4.0", é feita a apresentação de um sistema de vigilância inteligente para segurança patrimonial universitária, integrando Visão Computacional e análise multivariada para identificar padrões de risco. No Capítulo 02, "Segurança na Indústria 4.0: Detecção de anomalias em redes Modbus/TCP por meio de um IDS Híbrido K-Means/DBSCAN", foi avaliada a viabilidade de um IDS híbrido não supervisionado para detecção de anomalias industriais. Propõe-se uma arquitetura em cascata integrando K-Means e DBSCAN.

No Capítulo 03, "Resiliência cibernética na Indústria 4.0: Detecção de ataques de desfiguração via modelos Transformer", os pesquisadores avaliaram a eficácia de modelo transformer na detecção automática de ataques de desfiguração. No Capítulo 04, "Detecção de mensagens de phishing utilizando Random Forest e Processamento de Linguagem Natural em Ambientes da Indústria 4.0", foi proposto um sistema inteligente para

detecção de phishing no contexto da Indústria 4.0, utilizando o algoritmo Random Forest aliado a técnicas de Processamento de Linguagem Natural.

No Capítulo 05, “Resiliência e observabilidade na Indústria 4.0: Diagnóstico out-of-band com Lorawan”, foi desenvolvida e avaliada uma arquitetura de diagnóstico out-of-band empregando o protocolo LoRaWAN isolado fisicamente no nível de acesso, ativado de forma autônoma para o envio de mensagens críticas de 12 bytes. No Capítulo 06, “Análise de protocolos IoT para o trabalhador conectado em cenários de conectividade restrita”, foi validada uma arquitetura de baixo custo com ESP8266, comparando MQTT e CoAP em rede restrita. No Capítulo 07, “Implementação de arquitetura zero trust em iot com sensores que simulam ambiente industrial utilizando ESP32 e detecção de anomalias na borda”, os autores propuseram e validaram uma arquitetura Zero Trust (ZTA) para redes IIoT operando integralmente na borda, sem modificações no protocolo padrão e sem dependência de nuvem.

Em seguida, passamos a um tema fundamental para o funcionamento de qualquer indústria, que é **“Produção Energética e Transporte”**, aspecto estratégico para o desenvolvimento social e econômico de um país. No Capítulo 08, “PAND: Orquestração entre visão computacional e LLM para tomada de decisão na manutenção fotovoltaica”, é feita a proposta de arquitetura PAND, que integra YOLOv11m a um núcleo agêntico (Llama 3.3) para fusão de dados visuais e telemétricos via ciclo ReAct. No Capítulo 09, “Caracterização não supervisionada de perturbações elétricas em subestações de distribuição de energia utilizando técnicas de Clustering e Deep Learning”, foi proposta uma abordagem de aprendizado não supervisionado para caracterizar eventos em subestações de distribuição.

No Capítulo 10, “Sistema inteligente para diagnóstico da qualidade de energia em baixa tensão com Autoencoders aplicados à Indústria 4.0”, foi desenvolvido um sistema para detecção automática de anomalias em dados elétricos reais utilizando Inteligência Artificial. No Capítulo 11, “Estimativa do estado de saúde de baterias de Íon-Lítio utilizando Machine Learning no contexto da Indústria 4.0”, os autores compararam modelos de regressão treinados com 30 baterias submetidas a diferentes perfis de descarga utilizando o banco de dados “Nasa Li-ion Battery Aging Datasets” e uma estratégia de validação cruzada baseada em grupos. No Capítulo 12, “RAG

Aplicado a manuais automotivos: Segmentação contextual e recuperação híbrida", foi desenvolvido e avaliado um chat baseado em Retrieval-Augmented Generation (RAG) para manuais automotivos.

No contexto dos diversos tipos de indústrias existentes no Brasil, destacamos por sua importância para a vida humana e por ser uma vocação nacional, bem como do estado de Roraima, o tema **"Agroindústria"**, que representa a integração entre o setor primário (atividades agrícolas e de pecuária) e o setor secundário (caracterizado pelas indústrias de transformação). No Capítulo 13, "Pipeline de aprendizado de máquina para detecção de anomalias em ortomosaicos capturados por drone em áreas com predominância de cobertura vegetal", os pesquisadores propuseram e avaliaram um pipeline para detecção de anomalias em ortomosaicos com predominância de cobertura vegetal. No Capítulo 14, "Reconstrução espectral com SRF gaussianas e BI-LSTM para detecção de anomalias agrícolas no contexto da Agricultura 4.0", foi investigada a integração entre dados hiperespectrais e multiespectrais por meio de reamostragem espectral com SRF gaussianas e reconstrução com Bi-LSTM.

No Capítulo 15, "Arquitetura híbrida Edge-Cloud: Protótipo IoT offline-first para Agroindústria 4.0 em áreas remotas de Roraima", foi desenvolvida e validada pelos pesquisadores uma arquitetura híbrida edge-cloud de baixo custo com foco em resiliência offline-first. No Capítulo 16, "Monitoramento IoT para Agricultura 4.0: Arquitetura com LoRa P2P e Edge Computing para Eficiência Hídrica", foi validada uma Prova de Conceito (PoC) baseada em LoRa P2P e Edge Computing, com nó sensor em Deep Sleep, transmitindo umidade a um Gateway local (Raspberry Pi 4) com lógica heurística na borda.

Contamos, ainda, com estudos que examinam um tema crítico para a nossa e as futuras gerações, a **"Sustentabilidade"**. Não há como falar em Indústria 4.0 sem considerá-lo. No Capítulo 17, "Monitoramento da qualidade do ar em Boa Vista/RR com sensores IoT: Dataset aberto e visualização comunitária na perspectiva da Indústria 4.0", os autores apresentaram uma solução para monitoramento e acesso à informação sobre a qualidade do ar em Boa Vista/RR. No Capítulo 18, "Gestão de resíduos em Unidades Básicas de Saúde de Boa Vista: Uma proposta baseada em IoT de baixo custo", foi desenvolvida uma solução baseada em Internet das Coisas (IoT) para o

monitoramento do nível de preenchimento de lixeiras hospitalares em Unidades Básicas de Saúde.

Finalizamos esta obra abordando estudos sobre **"Gestão e Produtividade Industrial"**. No Capítulo 19, "Análise comparativa entre ARIMA e LSTM na previsão de escoamento logístico de soja em Roraima", os pesquisadores compararam o modelo estatístico ARIMA e Redes Neurais LSTM multivariadas na previsão do volume de exportação de soja (NCM 12019000), propondo o Índice de Pressão de Escoamento (IPE) como ferramenta de alerta gerencial para estrangulamentos na malha de transporte. No estudo do Capítulo 20, "Manutenção preditiva com Inteligência Artificial Explicável integrada ao ODOO ERP", foi desenvolvido um módulo de manutenção preditiva com Inteligência Artificial Explicável (XAI) integrado nativamente ao ERP Odoo 19.

No trabalho do Capítulo 21, "RAG-BI: Aprimorando dashboards de Business Intelligence com RAG", os pesquisadores investigaram a viabilidade técnica e as limitações do RAG-BI, uma plataforma voltada para o acesso a informações gerenciais via linguagem natural, empregando decomposição de tarefas e recuperação híbrida com um LLM local (Mistral-7B quantizado em 4 bits). No Capítulo 22, "Pipeline sequencial de Deep Learning para detecção de ativos e OCR em documentação técnica da Indústria 4.0" avaliou a aplicação de Inteligência Artificial na análise automatizada de projetos elétricos, com foco na verificação de luminárias e na extração de informações presentes em carimbos técnicos, visando à redução de erros e à otimização do tempo de análise.

No Capítulo 23, "Arquitetura distribuída Edge-Server para inventário em tempo real com RFID e técnicas de localização por proximidade", os pesquisadores propuseram o ColdChain-ID, uma arquitetura Edge-Server para inventário industrial. No Capítulo 24, "Estruturação de registros em ordens de serviço de manutenção com metamodelo de linguagem e embeddings de sentenças aplicado à Indústria 4.0", foi proposto e avaliado um MVP que aplica o Metamodelo de Linguagem da PNL para gerar perguntas de coleta de dados e um rascunho estruturado de OS.

No Capítulo 25, "Aplicação de Grandes Modelos de Linguagem na triagem de currículos não padronizados", os autores desenvolveram e avaliaram o RecruitAI para automatizar a triagem cega e a classificação

multirrótulo. No Capítulo 26, “Uso de Processamento de Linguagem Natural para triagem automatizada das publicações dos Diários Oficiais de Roraima”, foi avaliada a eficácia da classificação Zero-Shot na triagem automatizada de atos administrativos, mitigando a escassez de dados rotulados.

No Capítulo 27, “Chatbot para transparência orçamentária: arquitetura híbrida neuro-simbólica para consultas em linguagem natural aos dados do Governo de Roraima”, foi desenvolvido e avaliado um Chatbot capaz de responder perguntas em linguagem natural sobre a execução orçamentária do Governo do Estado de Roraima, exercício de 2025. No Capítulo 28, “Sistema IoT de Baixo Custo para Monitoramento de Vibração em Motores Industriais baseado em ESP32 e Norma ISO 10816”, os pesquisadores desenvolveram e avaliaram artefato tecnológico baseado no ESP32, visando verificar a viabilidade técnica do monitoramento de vibração.

No Capítulo 29, “Plataforma IoT de Monitoramento e Análise Estatística para Ambientes Industriais: Uma Solução Versátil de Baixo Custo”, foi desenvolvida e validada uma plataforma IoT de baixo custo para monitoramento de temperatura e umidade em ambientes industriais com perfis regulatórios distintos. Por fim, no Capítulo 30, “Framework Adaptativo para Adoção de Dark Factory em PMEs de Roraima”, os autores propuseram um roteiro para implementar monitoramento em tempo real e simulação aplicados a ativos legados.

Como se pode notar, esta obra apresenta contribuições diversas ao campo da Indústria 4.0 brasileira. Por esse motivo a consideramos como subsídio para estudiosos sobre o tema, que procuram por estratégias tecnológicas viáveis para iniciar suas pesquisas ou por parcerias, assim como este livro também é útil para gestores públicos que buscam por ideias promissoras para desenvolver a Indústria 4.0 na região em que atuam.

Boa Vista, Roraima, 31 de agosto de 2026.

Capítulo 01

Abordagem baseada em visão computacional e análise multivariada para segurança patrimonial na Indústria 4.0

Rogério de Carvalho Cordovil
Herbert Oliveira Rocha
Universidade Federal de Roraima – Curso de Ciência da Computação

rogercordovil@gmail.com
herbert.rocha@ufrr.com

1. Introdução

A integração de conceitos da Indústria 4.0 (I4.0) com técnicas de Visão Computacional (*Computer Vision* – CV) e Aprendizado Profundo (*Deep Learning* – DL) tem impulsionado o desenvolvimento de sistemas inteligentes de vigilância. Modelos como o *You Only Look Once* (YOLO) permitem a detecção de objetos em tempo real, sendo amplamente utilizados em aplicações de segurança. Embora os sistemas de CCTV (*Closed-Circuit Television*) sejam amplamente adotados, a dependência da observação humana impõe gargalos críticos, como a sobrecarga cognitiva e a baixa eficácia na detecção de comportamentos complexos. Conforme Hameed et al. (2024), a análise manual de grandes volumes de dados em cenários dinâmicos é inerentemente ineficiente e falha. Surge, assim, a necessidade de soluções baseadas em computação aplicada para automatizar o monitoramento e otimizar a identificação de riscos institucionais.

A Visão Computacional e o Aprendizado Profundo têm transformado o monitoramento reativo em proativo, superando limitações convencionais. Esses modelos de detecção em tempo real permitem identificar objetos e indivíduos com alta eficiência em fluxos de vídeo (Siva et al., 2025). Tais abordagens apresentam ganhos expressivos em precisão e tempo de resposta em contextos urbanos e industriais. Na segurança laboral, modelos YOLO otimizam a detecção de Equipamentos de Proteção Individual, prevenindo acidentes de forma autônoma (Perez, 2022). Adicionalmente,

sistemas baseados em DL identificam padrões comportamentais suspeitos com acurácia superior aos métodos tradicionais (Aguero & Rengifo, 2023). Essa evolução tecnológica é fundamental para reduzir perdas e automatizar a análise de riscos em tempo real. Assim, a integração dessas técnicas consolida-se como pilar essencial para a segurança patrimonial moderna e a gestão inteligente de ativos institucionais.

Além disso, pesquisas recentes evidenciam que a detecção de comportamentos anômalos em vídeos constitui um dos principais desafios na área de vigilância inteligente. Segundo Hameed et al. (2024), a identificação de eventos suspeitos ainda enfrenta dificuldades relacionadas à complexidade dos cenários, variações de iluminação e oclusões, exigindo modelos capazes de analisar padrões espaço-temporais. Complementarmente, Rendón-Segador et al. (2023) ressaltam que, apesar dos avanços no uso de DL, a detecção de eventos como violência ainda apresenta limitações, especialmente no que se refere à ocorrência de falsos positivos em ambientes reais.

Em ambientes acadêmicos, caracterizados por alta circulação e baixa previsibilidade, torna-se necessária a transição de sistemas de vigilância passivos para ativos, visando a identificação de comportamentos suspeitos (OGBITI; JOHN-FRANCISE, 2025). Contudo, lacunas persistem em soluções que priorizam detecções isoladas, sem integrar variáveis espaciais e temporais. Este trabalho propõe uma abordagem baseada na fusão entre Visão Computacional e análise multivariada para detectar padrões de risco. O objetivo deste trabalho é prover uma interpretação contextualizada de fluxos de vídeo por meio da análise conjunta de múltiplas variáveis comportamentais. Com isso, busca-se superar as limitações das técnicas tradicionais, oferecendo uma solução robusta e adaptável aos requisitos de segurança inteligente em cenários institucionais complexos.

Para operacionalizar esta proposta, adota-se uma modelagem multicenário estruturada em três contextos críticos: objeto abandonado, furto de bicicleta e detecção de violência. Tal estratégia viabiliza a especialização da análise conforme a natureza do evento, promovendo maior interpretabilidade e mitigação de falsos positivos. Assim, busca-se validar a viabilidade de um sistema de vigilância inteligente que integra Visão Computacional e análise multivariada como ferramenta de suporte à

segurança patrimonial em ambientes universitários. Busca-se investigar, especificamente, como a fusão de variáveis espaciais, temporais e comportamentais contribui para a identificação precisa de padrões de risco em fluxos de vídeo contínuos, superando as limitações de abordagens baseadas em detecções isoladas.

Como principais contribuições, destacam-se: (i) a proposição de uma abordagem multicenário orientada à análise comportamental; (ii) a aplicação de técnicas de análise multivariada para interpretação de dados visuais; e (iii) a avaliação experimental da solução em um conjunto de dados estruturado por cenários específicos.

2. Fundamentação teórica

Esta seção apresenta os fundamentos conceituais que sustentam a proposta deste trabalho, com ênfase na integração entre técnicas de Visão Computacional, análise multivariada e interpretação de padrões comportamentais em sistemas de vigilância inteligente. Inicialmente, são discutidos os conceitos de detecção de objetos em vídeo. Em seguida, aborda-se a análise multivariada aplicada à interpretação de comportamento. Por fim, são apresentadas as métricas de avaliação, incluindo abordagens estatísticas e temporais, alinhadas à modelagem proposta.

2.1. Visão Computacional e Detecção de Objetos

A Visão Computacional é um subcampo da Inteligência Artificial que busca permitir que sistemas computacionais interpretem e extraiam informações relevantes a partir de imagens e vídeos. Com o avanço das redes neurais profundas, especialmente as CNNs (*Convolutional Neural Network*), tornou-se possível automatizar a detecção e classificação de objetos em tempo real, ampliando significativamente as aplicações em sistemas de vigilância.

No contexto de segurança, a detecção de objetos constitui a etapa inicial para a análise de cenas, permitindo identificar entidades como pessoas, veículos e bens patrimoniais. Modelos da família YOLO têm se destacado por sua capacidade de realizar detecção em tempo real com baixa

latência, tornando-os adequados para aplicações que exigem respostas rápidas (Siva et al., 2025).

Entretanto, conforme apontado por Singh e Singh (2025), modelos focados exclusivamente na detecção podem apresentar limitações relacionadas à precisão e à ocorrência de falsos positivos, especialmente em ambientes dinâmicos. Isso evidencia a necessidade de complementar a detecção com etapas adicionais de análise, capazes de interpretar o contexto e o comportamento dos elementos detectados.

Dessa forma, neste trabalho, a detecção de objetos não é tratada como o objetivo final, mas como uma etapa intermediária dentro de um *pipeline* mais amplo, voltado à análise comportamental e inferência de risco.

2.2. Análise Multivariada Aplicada à Interpretação de Comportamento

A análise multivariada refere-se ao conjunto de técnicas para analisar simultaneamente múltiplas variáveis, considerando suas interdependências e correlações. No contexto da vigilância inteligente, essa abordagem permite superar a limitação da análise baseada em variáveis isoladas, possibilitando uma interpretação mais rica e contextualizada das cenas.

Segundo Hameed et al. (2024), a identificação de comportamentos anômalos em vídeos depende da análise conjunta de informações espaciais e temporais, sendo insuficiente considerar apenas características individuais. Portanto, variáveis como posição, tempo de permanência, velocidade e interação entre objetos são fundamentais para a caracterização de eventos.

Neste trabalho, a análise multivariada é operacionalizada por meio da construção de vetores de características extraídos das detecções realizadas em cada quadro de vídeo. Esses vetores incluem:

→ Variáveis espaciais (posição, distância entre objetos);

→ Variáveis temporais (tempo de permanência, duração de interações);

→ Variáveis comportamentais (velocidade, aceleração, padrões de movimento);

A combinação dessas variáveis permite identificar padrões complexos que não seriam detectáveis de forma isolada. Por exemplo, a

permanência prolongada de um objeto associada à ausência de interação humana pode indicar um possível abandono, enquanto a interação prolongada entre uma pessoa e um objeto, acompanhada de deslocamento, pode caracterizar um potencial furto.

Diferentemente de abordagens baseadas exclusivamente em modelos supervisionados, este trabalho adota uma estratégia baseada em regras heurísticas para a inferência de risco. Essa escolha está alinhada à necessidade de interpretabilidade e à limitação de dados rotulados, conforme também discutido por Ogbiti e John-Francise (2025).

2.3. Detecção de Anomalias e Análise de Comportamento em Vídeo

A detecção de anomalias em vídeo consiste na identificação de eventos que desviam de padrões considerados normais. Esse problema é particularmente relevante em sistemas de vigilância, nos quais eventos de interesse, como furtos ou comportamentos violentos, são relativamente raros e muitas vezes sutis.

De acordo com Rendón-Segador et al. (2023), a detecção de eventos como violência ainda apresenta desafios significativos, especialmente em função da complexidade dos padrões comportamentais e da alta taxa de falsos positivos. Além disso, a análise manual desses eventos é limitada pela capacidade humana de monitorar múltiplas fontes simultaneamente.

Nesse contexto, sistemas modernos de vigilância buscam incorporar análise comportamental baseada em padrões espaço-temporais, permitindo identificar interações suspeitas e mudanças abruptas de comportamento. Aguero e Rengifo (2023) demonstram que a análise de movimentos e padrões de interação pode contribuir significativamente para a identificação de atividades suspeitas em tempo real.

Neste trabalho, a detecção de anomalias é tratada dentro de uma abordagem multicenário, na qual diferentes tipos de eventos são modelados separadamente. Essa estratégia permite capturar especificidades de cada cenário, como:

→ Comportamentos estáticos (objeto abandonado);

→ Interações pessoa-objeto (furto);

→ Interações dinâmicas entre indivíduos (violência).

2.4. Métricas de Avaliação em Sistemas de Vigilância Inteligente

A avaliação de sistemas de vigilância baseados em Visão Computacional envolve tanto métricas clássicas de detecção quanto métricas estatísticas e temporais.

As métricas tradicionais incluem:

→ Precisão (*Precision*): proporção de detecções corretas;

→ Revocação (*Recall*): capacidade de identificar eventos relevantes;

→ Medida F1: equilíbrio entre precisão e revocação.

No entanto, conforme destacado por Siva et al. (2025), essas métricas não são suficientes para avaliar sistemas baseados em análise comportamental, sendo necessário incorporar indicadores adicionais.

Neste trabalho, a avaliação é complementada por métricas estatísticas e multivariadas, como: *Z-score*: utilizado para identificar desvios em variáveis individuais; Análise de Componentes Principais (PCA): utilizada para redução de dimensionalidade e identificação de padrões dominantes; e *Score* de risco: função agregada baseada na combinação de múltiplas variáveis. Além disso, são consideradas métricas temporais, como: Frequência de eventos de risco; Volume de detecções por cenário; e Estabilidade temporal das detecções.

3. Trabalhos relacionados

Kishor (2026) propôs um sistema baseado em modelos YOLO para a detecção de furtos, otimizando o tempo de processamento frente a abordagens multiestágio tradicionais. Contudo, a ausência de métricas para falsos positivos em áreas densas limita sua aplicação em campi universitários. Paralelamente, Aguero e Rengifo (2025) focaram na identificação de ocultação de produtos em varejos com redes profundas, obtendo 82% de acurácia. Embora eficaz em ambientes controlados, o modelo apresenta baixa generalização para cenários dinâmicos. Diferente

dessas propostas, este estudo avança ao integrar análise multivariada e heurísticas para mitigar lacunas de detecção isolada e lidar com a complexidade de espaços institucionais abertos. A Tabela 1 apresenta uma comparação de trabalhos correlatos a este estudo.

Tabela 1. Comparação entre os trabalhos correlatos.

Autor	Abordagem / Técnica	Resultados Reportados	Limitações	Observações
Aguero e Rengifo (2023)	CNN + YOLO (análise comportamental)	Acurácia de aproximadamente 82% e redução de perdas em até 15%	Dependência de ambiente controlado; limitada generalização	Considera padrões de comportamento simples
Hameed et al. (2024)	ViT (*Vision Transformer*) + CNN + RNN (*Recurrent Neural Network*) (modelo híbrido espaço-temporal)	Melhoria na classificação de eventos complexos	Alto custo computacional e necessidade de grande volume de dados	Forte capacidade de análise temporal
Siva et al. (2025)	YOLOv8 + CNN + LSTM (*Long Short-Term Memory*)	Alta acurácia (≈95%) em detecção e 92% em anomalias	Complexidade do sistema e dependência de infraestrutura	Integra análise de comportamento e contexto
Singh e Singh (2025)	YOLOv5 + *Faster R-CNN* (modelo híbrido)	Melhor equilíbrio entre velocidade e precisão	Aumento da complexidade computacional	Combinação eficiente de modelos
Ogbiti e John-Francise (2025)	Sistemas inteligentes com visão computacional e sensores	Melhoria na detecção de atividades suspeitas e monitoramento	Pouca ênfase em análise multivariada	Abordagem arquitetural e sistêmica

4. Solução Computacional Proposta

4.1. Visão Geral da Abordagem

A solução proposta consiste em um sistema de vigilância inteligente baseado em visão computacional e análise multivariada, estruturado como um *pipeline* de processamento de vídeo orientado à inferência de padrões

comportamentais (ver Figura 1). Diferentemente de abordagens tradicionais centradas exclusivamente na detecção de objetos, o sistema integra múltiplas variáveis espaciais, temporais e dinâmicas para caracterizar situações de risco em ambientes monitorados.

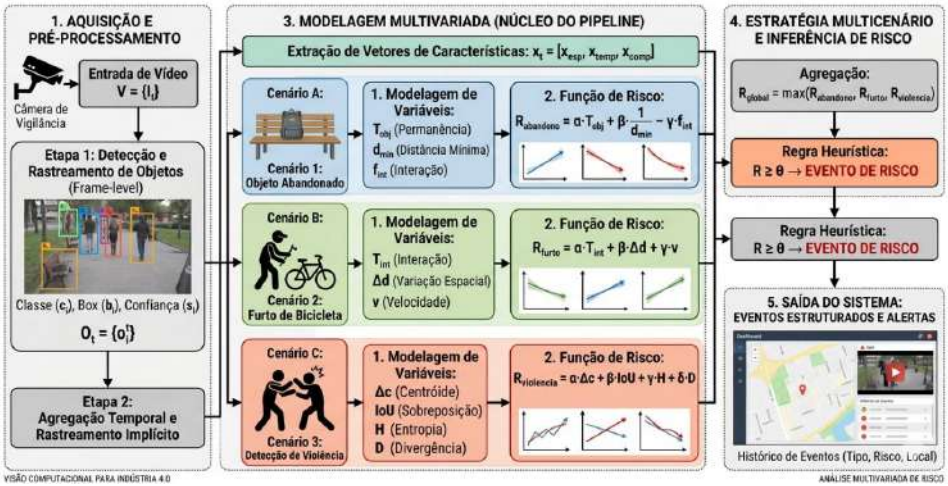


Figura 1. Visão Geral do Fluxo da Solução Proposta: sistema de vigilância inteligente.

O *pipeline* é composto por quatro etapas principais: (i) detecção de objetos em nível de quadro, (ii) rastreamento implícito e agregação temporal de entidades, (iii) extração de variáveis multivariadas e (iv) inferência de risco baseada em regras heurísticas. A saída do sistema consiste em eventos estruturados com níveis de risco associados, permitindo sua integração a sistemas de apoio à decisão.

Do ponto de vista formal, cada vídeo é modelado como uma sequência temporal ($V = I_{t=1}^{T}$), onde (I_t) representa o quadro no instante (t). Para cada quadro, o detector retorna um conjunto de objetos: $O_t = o_i^t$, $o_i^t = (c_i, b_i, s_i)$ onde (c_i) é a classe do objeto, (b_i) a *bounding box* e (s_i) o *score* de confiança.

A partir dessas detecções, são construídos vetores de características multivariadas: $x * t = [x * esp, x_{temp}, x_{comp}]$ representando variáveis espaciais, temporais e comportamentais, respectivamente.

Na etapa final do *pipeline*, é aplicado um filtro de decisão baseado na função de risco $R(s)$, no qual detecções que excedem o limiar θ são reclassificadas como eventos suspeitos. Esse mecanismo atua como uma camada adicional de validação sobre as saídas do modelo YOLO, contribuindo para a redução de falsos positivos e negativos em ambientes dinâmicos.

4.2. Modelagem Multivariada

A análise proposta baseia-se na construção de funções de risco $R(s)$, definidas como combinações não lineares de variáveis extraídas do fluxo do vídeo:

$$R(s) = \sum_{k=1}^{n} \quad w_k \cdot \Phi_k(x_k)$$

onde:

→ (w_k) pesos heurísticos associados a cada variável;

→ (ϕ_k) são funções de ativação (ex.: limiares, funções indicadoras); e

→ (x_k) são variáveis observadas, incluindo características espaciais, temporais e comportamentais.

Para garantir consistência na tomada de decisão, define-se um limiar de risco $\theta \in R$, tal que:

$$R(s) \leq \theta, \forall t \in [t_{ini}, t_{fim}]$$

Esse limiar atua como um teto dinâmico de decisão, delimitando o comportamento considerado aceitável pelo sistema. Valores de risco que excedem θ são classificados como eventos suspeitos ou potencialmente críticos.

No contexto da integração com modelos de detecção baseados em aprendizado profundo, como o YOLO, o limiar θ funciona como um mecanismo complementar ao *score* de confiança das detecções. Dessa forma, *bounding boxes* com alta incerteza ou inconsistência comportamental podem

ser reavaliadas com base na função de risco, reduzindo a ocorrência de falsos negativos em cenários críticos.

Assim, o modelo proposto estabelece uma integração entre a detecção visual baseada em YOLO e a inferência formal baseada em restrições matemáticas, permitindo que decisões de segurança sejam fundamentadas tanto em evidências visuais quanto em critérios estatísticos.

Do ponto de vista operacional, a aplicação do limiar θ ocorre na etapa de pós-processamento do *pipeline*, na qual eventos são classificados conforme a seguinte regra:

$$Evento = \{Normal, se\ R(s) \leq\ \theta\ Suspeito, se\ R(s) >\ \theta$$

Essa estratégia permite a filtragem e reclassificação de detecções, atuando como um mecanismo de controle de qualidade sobre as saídas do modelo de visão computacional.

A inferência de risco é realizada por meio de regras do tipo: $R \geq \theta \Rightarrow evento\ de\ risco$ onde (θ) é um limiar definido empiricamente.

A modelagem do risco de abandono fundamenta-se na persistência temporal $T_{obj} = t_{final} - t_{inicial}$ e na segregação espacial do objeto em relação aos agentes, onde a métrica de risco é formalizada por $R_{abandono} = \alpha \cdot T_{obj} + \beta \cdot \frac{1}{d_{min}} - \gamma \cdot f_{int}$. Neste modelo, a probabilidade de abandono escala positivamente com o tempo de permanência e inversamente à distância mínima $min||p_i - o_i||$ entre o objeto e qualquer indivíduo presente na cena, sendo atenuada pela frequência de interações f_{int}. Paralelamente, o cenário de furto de bens, como bicicletas, é quantificado pela função $R_{furto} = \alpha \cdot T_{int} + \beta \cdot \Delta d + \gamma \cdot v$, que correlaciona a persistência da interação P_{int} com o deslocamento espacial do objeto $\Delta d\ = ||b_{t+1}\ -\ b_t||$. A detecção baseia-se na covariância entre a cinemática do agente v e a translação do objeto, indicando uma remoção não autorizada através da análise multivariada de movimento e persistência.

A detecção de violência em ambientes monitorados é estruturada através da análise de padrões dinâmicos e estocásticos de interação, sintetizados na função $R_{violencia} = \alpha \cdot \Delta c\ + \beta \cdot IoU\ + \gamma \cdot H + \delta \cdot D$. Este modelo integra a variação do centróide δc e a sobreposição de áreas $IoU =$

$\frac{A_{interseção}}{A_{união}}$ para inferir proximidade física crítica, enquanto a entropia da trajetória $H = -\sum \quad p_i \, log \, log \, p_i$ mensura o nível de desorganização comportamental e imprevisibilidade dos movimentos. A divergência vetorial $D = \|\overrightarrow{v_i} - \overrightarrow{v_j}$ atua como um descritor de dispersão ou fuga, permitindo diferenciar fluxos normativos de comportamentos agressivos. A convergência destas variáveis permite a identificação de conflitos físicos por meio de movimentos abruptos e alta entropia sistêmica, validando estatisticamente a ocorrência de anomalias em cenários de segurança patrimonial universitária através de métricas de visão computacional.

5. Avaliação experimental

5.1. Planejamento de pesquisa

A avaliação experimental foi conduzida com o objetivo de analisar o desempenho do sistema proposto na identificação de padrões comportamentais associados a situações de risco em ambientes de vigilância. O protocolo experimental foi estruturado a partir da combinação de cenários controlados e vídeos provenientes de *datasets* públicos amplamente utilizados na literatura de detecção de anomalias em vídeo.

O conjunto experimental foi composto por 49 vídeos, incluindo cenários próprios e vídeos selecionados de *datasets* públicos, sendo utilizados 21 vídeos do *dataset Avenue Dataset* e 21 vídeos do *UCF-Crime Dataset*, além de vídeos organizados nos cenários de objeto abandonado, furto de bicicleta e detecção de violência. Essa composição permitiu avaliar o sistema tanto em ambientes controlados quanto em situações mais complexas e próximas da realidade.

Cada vídeo foi processado integralmente, gerando uma sequência temporal de detecções a partir do modelo YOLO, seguida pela aplicação de regras multicenário para classificação dos eventos. Ao longo dos experimentos, foram registrados 2.201 eventos consolidados, distribuídos entre os cenários analisados (1.489 eventos de objeto abandonado, 68 de furto de bicicleta e 644 de detecção de violência).

5.2. Resultados e Discussão

5.2.1. Análise Comparativa entre Cenários

A avaliação do sistema fundamentou-se em uma abordagem metodológica híbrida, integrando métricas clássicas de desempenho, análise estatística multivariada e métricas temporais para validação das detecções. Foram consideradas métricas como precisão (precision), revocação (recall), taxa de falsos positivos (*False Positive Rate – FPR*), além de indicadores derivados de *Z-score*, Análise de Componentes Principais (*Principal Component Analysis – PCA*) e métricas temporais, como *Time To Alert* (TTA), *Event-based Recall* e *Mean Time Between False Alarms* (MTBFA).

No cenário de objetos abandonados, foram registrados 1.489 eventos distribuídos em 19 vídeos, representando o maior volume de detecções no conjunto analisado. Esse cenário apresentou baixa média de risco (≈ 0,39) e reduzida dispersão no espaço multivariado, refletindo um comportamento altamente estático e previsível. A análise por *Z-score* evidenciou valores elevados em métricas relacionadas à permanência e ausência de movimento, confirmando a natureza estruturada desse tipo de evento. Como consequência, o sistema demonstrou maior estabilidade e menor propensão a falsos positivos, evidenciando sua eficácia na detecção de padrões estáticos.

No cenário de furto de bicicletas, foram identificados 68 eventos em 3 vídeos, com média de risco intermediária (≈ 0,80). A análise multivariada indicou maior dispersão dos dados em relação ao cenário de objetos abandonados, refletindo a dinâmica das interações entre indivíduos e objetos. Esse comportamento evidencia uma complexidade estrutural intermediária, na qual o sistema apresenta capacidade de identificação de padrões suspeitos, porém com maior incerteza na separação entre comportamentos normais e anômalos.

Já o cenário de detecção de violência apresentou 644 eventos distribuídos em 27 vídeos, com elevada média de risco (≈ 0,98), indicando forte ativação do modelo para esse tipo de ocorrência. A inclusão de vídeos provenientes de *datasets* públicos contribuiu para o aumento da variabilidade comportamental observada, refletida na elevada dispersão no espaço do PCA e na ausência de dominância de um único componente

principal. Esse comportamento evidencia a natureza não linear e altamente dinâmica das interações agressivas.

Entretanto, essa mesma sensibilidade elevada impactou diretamente a precisão do sistema. Na avaliação por matriz de confusão, observou-se recall aproximado de 0,70, indicando boa capacidade de identificação de eventos reais. Por outro lado, a precisão foi reduzida, acompanhada de uma taxa de falsos positivos elevada (FPR ≈ 1,0), evidenciando que o sistema tende a classificar como violência diversos comportamentos dinâmicos que não correspondem a eventos agressivos.

Esse comportamento pode ser explicado por três fatores principais:

→ Semelhança cinemática entre ações: movimentos rápidos, aproximações bruscas e mudanças de direção — comuns em interações normais — apresentam padrões semelhantes aos observados em eventos violentos;

→ Alta sensibilidade das variáveis multivariadas: métricas como velocidade, aceleração e entropia de trajetória tendem a amplificar pequenas variações comportamentais, aumentando a ativação da função de risco; e

→ Ausência de contexto semântico profundo: o modelo baseia-se em heurísticas e variáveis quantitativas, não incorporando plenamente aspectos semânticos da cena, como intenção ou tipo de interação social.

A análise temporal reforça essa interpretação. O sistema apresentou tempo médio de resposta inferior a 0,30 segundos, caracterizando detecção praticamente em tempo real. Contudo, o *Event-based Recall* aproximado de 0,40 indica que parte dos eventos reais não foi completamente capturada, especialmente aqueles de curta duração. Além disso, o baixo valor de MTBFA (≈ 0,0018 horas) evidencia alta frequência de alarmes falsos, indicando baixa estabilidade em cenários altamente dinâmicos.

Do ponto de vista da análise multivariada, a dispersão observada no espaço do PCA não deve ser interpretada apenas como limitação do modelo, mas também como um reflexo da complexidade do fenômeno analisado. Enquanto eventos estáticos apresentam agrupamentos bem definidos,

eventos dinâmicos, como violência, tendem a ocupar regiões mais difusas no espaço de características, dificultando a separação linear entre classes.

De forma geral, os resultados evidenciam uma distinção clara entre anomalias estáticas e dinâmicas. O sistema apresenta alto desempenho em cenários previsíveis, com maior precisão e estabilidade, e elevada sensibilidade em cenários complexos, ainda que com maior incidência de falsos positivos. Esses achados indicam que a abordagem proposta é eficaz na detecção de eventos críticos, mas requer refinamento nos critérios de decisão — como o ajuste do limiar de risco θ — para melhorar a precisão e reduzir alarmes indevidos em contextos de alta variabilidade comportamental.

Tabela 2. Comparação entre Cenários Próprios e *Datasets* Públicos.

Categ. Dados	Cenário	Nº Vídeo	Nº Evento	Risco Médio	Complex. (PCA)	Compto. do Modelo
Cenários Próprios	Objeto abandonado	19	1.489	0,39	Baixa dispersão	Alta precisão, poucos falsos positivos
Cenários Próprios	Furto de bicicleta	3	68	0,8	Dispersão moderada	Detecção razoável, alguma ambiguidade
Cenários Próprios + Públicos	Violência	27	644	0,98	Alta dispersão	Alta sensibilidade, muitos falsos positivos
Datasets Públicos	*Avenue* (21 vídeos)	21	(integrados)	Alto (variável)	Alta variabilidade	Cenários complexos, reforçam instabilidade
Datasets Públicos	*UCF-Crime* (21 vídeos)	21	(integrados)	Alto (variável)	Alta variabilidade	Movimentos intensos, aumento de falsos positivos

A Tabela 2 apresenta uma comparação entre os cenários próprios e os vídeos provenientes de *datasets* públicos utilizados na validação do

sistema. Observa-se que os cenários controlados apresentam maior previsibilidade e menor dispersão estatística, resultando em melhor desempenho do modelo. Em contrapartida, os *datasets* públicos, como *Avenue* e *UCF-Crime*, introduzem maior variabilidade comportamental, refletida no aumento da complexidade no espaço multivariado e na elevação da taxa de falsos positivos.

5.2.2. Discussão

Os resultados evidenciam que a integração entre Visão Computacional (CV) e análise multivariada constitui uma abordagem eficaz para a identificação de padrões comportamentais em sistemas de vigilância inteligente. Enquanto o emprego de métricas como o *Z-score* mostrou-se adequado para detectar desvios em variáveis individuais, a Análise de Componentes Principais (PCA) permitiu identificar estruturas latentes e correlações multivariáveis, proporcionando uma interpretação mais holística dos dados.

A análise comparativa revelou uma distinção clara entre eventos estáticos e dinâmicos. No cenário de objetos abandonados — caracterizado por baixa variabilidade e alta previsibilidade — o sistema demonstrou estabilidade, com elevada precisão e reduzida incidência de falsos positivos. Tal desempenho indica que a modelagem multivariada é particularmente robusta quando aplicada a padrões comportamentais estruturados.

Em contrapartida, cenários dinâmicos, como furto de bicicletas e detecção de violência, apresentaram maior complexidade no espaço de características, refletida na dispersão dos dados no PCA e no aumento de falsos positivos. Este fenômeno não decorre necessariamente de uma limitação intrínseca do modelo, mas da natureza não linear e estocástica desses eventos, onde comportamentos distintos podem compartilhar assinaturas cinemáticas semelhantes.

Nesse contexto, a elevada sensibilidade do sistema (*recall*) reforça sua capacidade em identificar eventos críticos. Contudo, observa-se um *trade-off* com a precisão, especialmente sob alta variabilidade comportamental — comportamento típico em sistemas de segurança que

priorizam a minimização de omissões (falsos negativos) em detrimento de alarmes falsos.

A introdução do limiar de risco θ surge como o mecanismo central para gerenciar esse equilíbrio. O ajuste deste parâmetro permite calibrar o sistema conforme o contexto: ambientes que exigem rigor na detecção de ameaças podem adotar valores de θ mais permissivos, ao passo que contextos que demandam menor taxa de interrupção beneficiam-se de limiares mais restritivos. Ademais, a abordagem proposta supera as limitações de métodos estritamente baseados em detecção de objetos ao incorporar dimensões espaciais, temporais e comportamentais. Esta integração viabiliza uma análise contextualizada, aproximando o sistema de uma interpretação semântica dos eventos, ainda que por via heurística.

Apesar dos avanços, o estudo aponta limitações, como a dependência de regras heurísticas que podem restringir a generalização em cenários ultra-complexos, e a ausência de mecanismos de aprendizado temporal profundo para eventos de curtíssima duração. Como perspectivas futuras, propõe-se a integração de redes neurais recorrentes ou *transformers*, além de técnicas de aprendizado adaptativo para o ajuste dinâmico do limiar θ. Conclui-se que a abordagem representa um avanço relevante para a vigilância inteligente, oferecendo um modelo interpretável e alinhado aos paradigmas da Indústria 4.0.

6. Conclusão

Este trabalho apresentou uma abordagem baseada na integração entre Visão Computacional (CV) e análise multivariada para a identificação de padrões de risco em ambientes de vigilância inteligente. A proposta, fundamentada na extração de variáveis espaciais, temporais e comportamentais, demonstrou eficácia na distinção entre diferentes tipos de eventos, especialmente em cenários estáticos, nos quais o sistema apresentou maior precisão e estabilidade.

Em cenários dinâmicos, como furto de bicicleta e detecção de violência, observou-se maior complexidade na separação entre comportamentos normais e anômalos, refletida no aumento da dispersão dos

dados e na elevação da taxa de falsos positivos. Nesse contexto, o limiar de risco θ mostrou-se um elemento fundamental para o controle do *trade-off* entre sensibilidade e precisão, permitindo ajustar o comportamento do sistema conforme o ambiente de aplicação.

Conclui-se que a abordagem multivariada proposta supera limitações de métodos baseados exclusivamente na detecção de objetos, ao proporcionar uma análise mais contextualizada dos eventos. Embora apresente limitações relacionadas à dependência de heurísticas, o modelo representa uma solução promissora, interpretável e alinhada aos princípios da Indústria 4.0, com potencial para evolução por meio da incorporação de técnicas de aprendizado temporal e ajuste adaptativo de parâmetros.

7. Referências

Aguero, Y. A. S., & Rengifo, J. K. C. (2025). Intelligent surveillance system using *Deep Learning* to reduce shoplifting in minimarkets in Santiago de Surco, Lima, Peru. Professional School of Systems Engineering, Peruvian University of Applied Sciences (UPC). DOI: https://doi.org/10.5220/0013750800003982

Hameed, S., Amin, J., Anjum, M. A., & Sharif, M. (2024). Suspicious activities detection using spatial-temporal features based on vision transformer and recurrent neural network. *Journal of Ambient Intelligence and Humanized Computing*, *15*, 3379–3391. https://doi.org/10.1007/s12652-024-04818-7

Kishor, K. (2026). Theft detection using YOLO-based object detection with Gen AI for improved security. In *Automating software defect detection through machine learning and LLMs* (pp. 345-368). IGI Global. https://doi.org/10.4018/979-8-3373-4460-7.ch011

Ogbiti, J. T., & John-Francise, U. O. (2025). Development of campus security surveillance system using smart device. *Journal of Institutional Research, Big Data Analytics and Innovation, 1*(1), 8–20. https://doi.org/10.5281/zenodo.17537562

Perez, F. B. (2022). *Avaliação da aplicação da inteligência artificial na detecção do uso de equipamentos de proteção individual em ambientes de*

trabalho [Trabalho de Conclusão de Curso, Universidade Federal de Alagoas]. Repositório Institucional UFAL.

Ren, Q., & Wang, J. (2025). DoA-ViT: Dual-objective affine vision transformer for data insufficiency. *Neurocomputing, 615*, Artigo 128896. https://doi.org/10.1016/j.neucom.2024.128896

Rendón-Segador, F. J., Álvarez-García, J. A., Salazar-González, J. L., & Tommasi, T. (2023). CrimeNet: Neural structured learning using vision transformer for violence detection. *Neural Networks, 161*, 318–329. https://doi.org/10.1016/j.neunet.2023.01.048

Singh, H., & Singh, N. (2025). Real-time smart surveillance using YOLO-faster R-CNN hybrid. *International Journal of Computing, Programming and Database Management, 6*(2), 112–120. https://doi.org/10.22271/computer.2025.v6.i2b.130

Siva, P., Pujitha, G. B., Krishna, G. S., Hemanth, G., & Teja, B. M. S. (2025). Smart surveillance systems using YOLOv8: A scalable approach for crowd and threat detection. *International Journal of Recent Advances in Engineering and Technology, 14*(1).

Capítulo 02

Segurança na Indústria 4.0: Detecção de anomalias em redes Modbus/TCP por meio de um IDS Híbrido K-Means/DBSCAN

Katerine Jones da Silva
Felipe Leite Lobo
Universidade Federal de Roraima

katerinejsilva@gmail.com
felipe.lobo@ufrr.com

1. Introdução

O avanço da Indústria 4.0 intensificou a integração entre sistemas computacionais, dispositivos físicos e redes de comunicação, configurando os chamados Sistemas Ciberfísicos (Cyber-Physical Systems – CPS) (Nankya et al., 2023). Esses sistemas operam sobre Redes de Controle Industrial (Industrial Control Systems – ICS), responsáveis por processos críticos em setores como energia, manufatura e saneamento (Nankya et al., 2023).

Entre os protocolos mais utilizados nesses ambientes destaca-se o Modbus/TCP, adotado em sistemas SCADA devido à sua simplicidade e interoperabilidade. Por ser um protocolo legado, o Modbus/TCP não possui autenticação nem criptografia, tornando-se vulnerável a ataques de injeção de comandos e negação de serviço (Lazaridis et al., 2024).

Nesse contexto, Sistemas de Detecção de Intrusão (IDS) são essenciais na defesa em profundidade. A abordagem baseada em assinaturas é limitada frente a ataques desconhecidos. Assim, métodos de aprendizado de máquina não supervisionado, especialmente clustering, têm sido utilizados para modelar o comportamento normal do tráfego e identificar anomalias (Sobonski & Roedig, 2024; Nikolajew & Eichelberger, 2024).

Algoritmos de clustering como K-Means e DBSCAN são amplamente utilizados em IDS baseados em anomalias, combinando eficiência computacional e capacidade de detecção, ainda que apresentem limitações quando aplicados isoladamente (Sobonski & Roedig, 2024; Pallakonda et al.,

2025). Diante desse cenário, este trabalho investiga a seguinte questão de pesquisa: é possível estruturar uma arquitetura híbrida sequencial baseada em clustering não supervisionado que busque equilibrar capacidade de detecção e viabilidade computacional em redes Modbus/TCP industriais?

Este trabalho propõe e avalia uma arquitetura híbrida sequencial baseada em K-Means e DBSCAN, com foco na redução de falsos positivos em redes industriais Modbus/TCP. Para validação experimental, foi utilizado exclusivamente o dataset ICS disponibilizado por Sudyana (2025), selecionado por apresentar dados sincronizados de tráfego industrial representativos de cenários reais. Os resultados apresentados neste trabalho correspondem à versão inicial da arquitetura proposta.

Como principal contribuição, este trabalho propõe e valida experimentalmente uma arquitetura híbrida sequencial baseada em K-Means e DBSCAN, demonstrando sua eficácia na redução de falsos positivos sem comprometer a capacidade de detecção. Essa abordagem busca atender às restrições operacionais de ambientes industriais, contribuindo para o avanço de soluções de IDS aplicadas a redes Modbus/TCP.

2. Fundamentação teórica

Esta seção apresenta os conceitos que fundamentam a arquitetura híbrida proposta, destacando aspectos de segurança em Redes de Controle Industrial e fundamentos de clustering aplicados à detecção de intrusão.

2.1. O Cenário da Indústria 4.0 e a Vulnerabilidade do Modbus/TCP

Os Sistemas Ciberfísicos (CPS) na Indústria 4.0 operam sobre Redes de Controle Industrial (ICS), onde a disponibilidade e integridade dos processos são críticas. Nesse contexto, o Modbus/TCP, amplamente empregado em sistemas SCADA, destaca-se pela ausência de autenticação e criptografia, tornando-se suscetível a ataques (Lazaridis et al., 2024).

2.2. Sistemas de Detecção de Intrusão (IDS) Baseados em Anomalias

Os IDS baseados em assinaturas são insuficientes para detectar ataques de dia zero. Assim, destaca-se o uso de IDS baseados em anomalias, que aplicam aprendizado de máquina não supervisionado para modelar o comportamento normal do tráfego e identificar desvios significativos (Nicolaio et al., 2023; Okolie et al., 2025).

2.3. Fundamentos do Clustering para Detecção de Intrusão

2.3.1. K-Means

O K-Means é um algoritmo de particionamento que organiza os dados em K clusters (MacQueen, 1967), destacando-se pela eficiência computacional e escalabilidade, embora seja sensível a outliers (Sobonski & Roedig, 2024; Pallakonda et al., 2025).

2.3.2. DBSCAN

O DBSCAN (Density-Based Spatial Clustering of Applications with Noise) (Ester et al., 1996) identifica regiões de alta densidade e classifica pontos esparsos como ruído e apresenta maior custo computacional (Sobonski & Roedig, 2024; Pallakonda et al., 2025).

3. Trabalhos relacionados

Conduziu-se uma revisão da literatura a partir de consultas em bases como IEEE Xplore, ACM Digital Library e Scopus, utilizando descritores relacionados a IDS em redes Modbus/TCP e detecção de anomalias em sistemas SCADA. Foram utilizados os descritores “Intrusion Detection System”, “Modbus/TCP”, “ICS security” e “anomaly detection”, considerando estudos publicados entre 2021 a 2025.

Nankya et al. (2023) evidenciam que abordagens supervisionadas em ambientes ICS alcançam boa acurácia em cenários conhecidos, porém apresentam limitações frente a ataques de dia zero. Em contraste, Lazaridis et al. (2024) destacam que a ausência de mecanismos de segurança no

protocolo exige modelos capazes de aprender o comportamento normal, característica mais bem atendida por abordagens não supervisionadas.

Modelos baseados em Deep Learning apresentam elevado desempenho na captura de padrões complexos, porém com alto custo computacional e maior latência. Em ambientes industriais, onde decisões em tempo quase real são necessárias, a latência é fator crítico (Nikolajew & Eichelberger, 2024), tornando necessárias abordagens mais eficientes.

No âmbito do Clustering, Sobonski & Roedig (2024) demonstram que o K-Means é eficiente, porém sensível a ruídos, enquanto o DBSCAN é mais robusto a outliers, embora computacionalmente mais custoso. Trabalhos recentes, como Pallakonda et al. (2025), indicam que abordagens híbridas combinando essas técnicas são promissoras para detectar ataques furtivos em infraestruturas críticas.

Adicionalmente, Liu et al. (2021) explora arquitetura em múltiplos estágios, mas sem validação específica em tráfego Modbus/TCP. Alanazi et al. (2025) destacam a necessidade de datasets específicos para avaliação de vulnerabilidades em sistemas ICS, propondo o ICS-LTU2022 como base para análise de riscos e segurança industrial.

Por fim, Machaka et al. (2025) avaliaram o impacto de ataques ao protocolo Modbus/TCP em um ambiente real de infraestrutura crítica, evidenciando consequências diretas sobre processos industriais. Contudo, o estudo não propõe mecanismos de detecção baseados em aprendizado de máquina, abrindo espaço para abordagens como a deste trabalho.

Diante disso, observa-se que métodos isolados apresentam limitações quanto à acurácia, robustez ou custo computacional. Ainda que existam abordagens híbridas, há uma lacuna na validação de arquiteturas sequenciais aplicadas especificamente ao protocolo Modbus/TCP. Nesse contexto, este trabalho propõe uma arquitetura híbrida K-Means → DBSCAN visando equilibrar desempenho e viabilidade operacional.

4. Solução computacional

A solução proposta consiste em um Sistema de Detecção de Intrusão (IDS) híbrido não supervisionado, implementado em uma arquitetura

sequencial em cascata (K-Means → DBSCAN), projetado para ambientes de Sistemas Ciberfísicos (CPS) industriais sob restrições de latência, volume de dados e criticidade operacional.

A arquitetura opera em dois estágios complementares. Inicialmente, o K-Means realiza uma filtragem rápida do tráfego normal, reduzindo o volume de dados a serem analisados. Em seguida, o DBSCAN é aplicado exclusivamente sobre o subconjunto considerado suspeito, refinando a detecção de anomalias por meio de densidade.

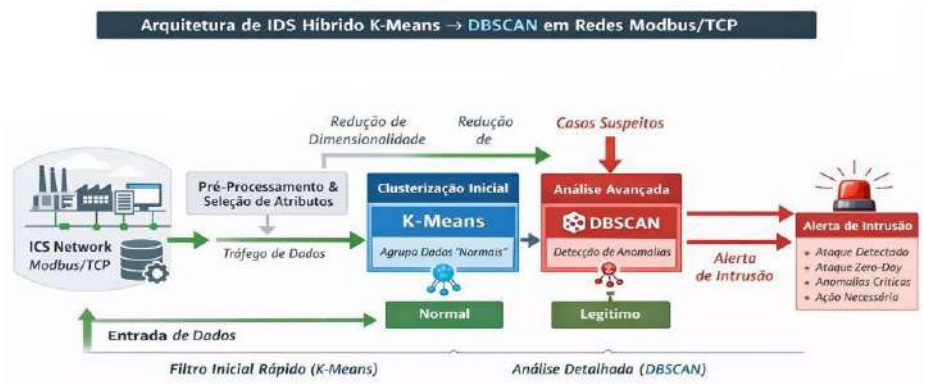


Figura 1. Arquitetura Híbrida Sequencial em Cascata K-Means → DBSCAN.

4.1. Conjunto de dados e Pré-processamento

Para validação experimental, foi utilizado exclusivamente o ICS Dataset (Sudyana, (2025) contendo registros de tráfego Modbus/TCP sincronizados com variáveis físicas e eventos de ataque.

Inicialmente foram considerados múltiplos datasets públicos relacionados à segurança em sistemas ciberfísicos. Entretanto, verificou-se que tais conjuntos apresentavam heterogeneidade de domínios (tráfego, logs e sensores não sincronizados). Assim, optou-se pela utilização de um único dataset consolidado, garantindo coerência experimental, homogeneidade de domínio e sincronização temporal entre variáveis físicas e de rede. O pré-processamento dos dados seguiu etapas fundamentais para garantir a estabilidade dos algoritmos:

1. Seleção exclusiva de atributos numéricos: Manutenção exclusiva de colunas numéricas.
2. Tratamento de valores ausentes por substituição zero: Escolhido preservar a integridade da matriz de dados sem descartar amostras.
3. Normalização dos atributos por meio do StandardScaler: Justifica-se pela sensibilidade do K-Means e do DBSCAN à escala dos dados. Esta técnica padroniza os atributos para que possuam média zero e variância unitária.
4. Redução de dimensionalidade utilizando (PCA - Principal Component Analysis): Foi aplicada para concentrar a informação nos componentes principais mais relevantes, reduzindo o custo computacional e mitigando efeitos associados à alta dimensionalidade dos dados.

O conjunto de dados foi dividido em 70% para treinamento e 30% para teste, utilizando amostragem estratificada com seed fixa (random_state=42), garantindo proporcionalidade entre classes e reprodutibilidade do experimento.

4.2. Modelo Baseline – K-Means e DBSCAN Puro

Como modelos de referência, foram utilizados K-Means e DBSCAN de forma isolada. O K-Means foi configurado com *n_clusters=2*, utilizando o percentil 99 das distâncias aos centroides como limiar de anomalia. O DBSCAN foi aplicado como método baseado em densidade, considerando como anomalias os pontos classificados como ruído.

4.3. Arquitetura Híbrida Sequencial (K-Means → DBSCAN)

A arquitetura proposta integra os algoritmos K-Means e DBSCAN em cascata. O diferencial central desta solução é o mecanismo de Gate de Distância (p_gate). No primeiro estágio, o K-Means segmenta o tráfego e identifica o cluster de menor cardinalidade (supostamente anômalo). Para evitar que ruídos operacionais e transientes normais (comuns no ICS Dataset) sejam enviados ao DBSCAN — o que elevaria os falsos positivos.

Aplica-se um filtro baseado no percentil 85 (p85), cujo objetivo é eliminar variações operacionais legítimas antes da etapa de análise por densidade. Esse mecanismo atua como um filtro de ruído operacional, reduzindo a propagação de padrões legítimos de baixa densidade para o estágio de análise por densidade.

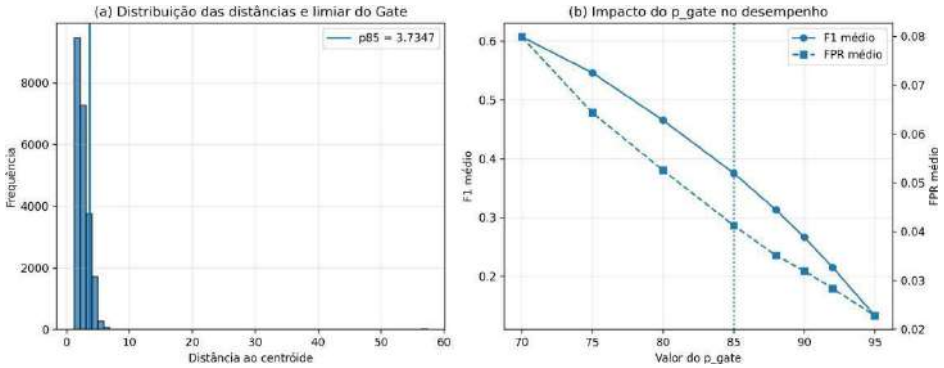


Figura 2. Análise do mecanismo de Gate.

A partir da distribuição das distâncias, observa-se que apenas uma fração dos dados ultrapassa o limiar definido pelo percentil 85. Observa-se que aproximadamente 15% das amostras são encaminhadas ao DBSCAN, evidenciando a eficiência do mecanismo de filtragem na redução do volume de dados analisados. A análise do parâmetro p_gate evidencia um trade-off entre capacidade de detecção e taxa de falsos positivos. Valores mais baixos aumentam o número de amostras analisadas pelo DBSCAN, elevando o F1-Score, porém com maior incidência de falsos positivos.

Em contrapartida, valores mais elevados reduzem o FPR, ao custo de diminuição na capacidade de detecção. Comparativamente, a ausência do Gate (DBSCAN puro) resulta em elevada taxa de falsos positivos (≈ 0,19), decorrente da interpretação de variações operacionais legítimas como anomalias. Com a aplicação do Gate (p85), observa-se uma redução significativa do FPR (≈ 0,025 na versão v1), evidenciando que o mecanismo atua como filtro eficaz de ruídos operacionais.

A versão otimizada (v2) demonstra que ajustes adicionais permitem elevar o F1-Score (≈ 0,40) mantendo níveis controlados de falsos positivos (≈ 0,04), evidenciando um melhor equilíbrio entre detecção e precisão operacional. O código está disponível: https://github.com/katerinejsilva-ux/ids-modbus-hibrido-kmeans-dbscan.

5. Avaliação experimental

5.1. Planejamento Experimental

Para validar a proposta, foram avaliadas diferentes configurações de modelos de detecção de anomalias, incluindo abordagens isoladas e a arquitetura híbrida proposta. Foram avaliadas quatro configurações:

→ (M1) K-Means IDS baseado em limiar por percentil;
→ (M2) DBSCAN puro com ajuste de parâmetros; e
→ (M3) Arquitetura híbrida sequencial K-Means → DBSCAN.
→ (M4) Arquitetura híbrida sequencial K-Means → DBSCAN, versão otimizada.

A avaliação foi conduzida utilizando divisão estratificada treino/teste (70/30), com validação multi-seed (10 execuções independentes), permitindo a análise da robustez estatística dos modelos.

5.2. Resultados e discussão

5.2.1. Resultados – Arquitetura Inicial

A Tabela 1 apresenta os resultados da execução base (seed = 42), estabelecendo o comparativo inicial entre os modelos.

Tabela 1. Desempenho dos Modelos Experimentais.

Modelo	Accuracy	Precision	Recall	F1-Score	FPR
K-Means IDS	0.5022	0.6106	0.0122	0.0240	0.0078
DBSCAN puro	0.7425	0.7718	0.6857	0.7262	0.2012

Modelo	Accuracy	Precision	Recall	F1-Score	FPR
Híbrido v1 (K→D)	0.5944	0.8888	0.2154	0.3473	0.0270

Nota. Elaborado pela autora com base nos experimentos realizados no Google Colab (2026), utilizando o ICS Dataset (Sudyana, (2025).

Na execução base, o modelo DBSCAN puro apresentou elevada capacidade de detecção (F1 ≈ 0,73), refletida no maior F1-Score entre os modelos avaliados. Entretanto, esse desempenho é acompanhado por uma taxa de falsos positivos significativamente elevada (~20%), o que limita sua aplicação em cenários operacionais sensíveis.

Esse comportamento está diretamente associado às características do ICS Dataset (Sudyana, 2025) que inclui variações operacionais, mudanças de estado e comportamentos transitórios legítimos. Nesse contexto, padrões normais de tráfego podem apresentar baixa densidade local e, consequentemente, serem interpretados erroneamente como ruído (anomalias) pelo DBSCAN. Por outro lado, o modelo K-Means isolado apresentou desempenho limitado (F1 ≈ 0,02), evidenciando sua incapacidade de capturar ataques complexos sem um refinador. A arquitetura híbrida inicial (v1) já demonstrou uma redução significativa no FPR (≈ 0,025), indicando que a filtragem inicial contribui para reduzir a propagação de variações normais ao estágio de análise por densidade.

5.2.2. Análise de Robustez (Multi-Seed)

A Tabela 2 apresenta os resultados consolidados considerando as 10 execuções independentes (multi-seed), permitindo avaliar a estabilidade dos modelos.

Tabela 2. Robustez Estatística (Multi-Seed – 10 Execuções).

Modelo	F! Médio	DP (F1)	FPR Médio	DP (FPR)
K-Means IDS	0.0244	0.0088	0.0076	0.0003
DBSCAN puro	0.7302	0.0044	0.1917	0.0047
Híbrido v1	0.3485	0.0046	0.0248	0.0016

Modelo	F! Médio	DP (F1)	FPR Médio	DP (FPR)
Híbrido v2	0.3967	0.0041	0.0414	0.0023

Nota. Elaborado pela autora com base nos experimentos realizados no Google Colab (2026), utilizando o ICS Dataset (Sudyana, (2025).

A análise multi-seed indica que o DBSCAN apresenta o maior F1-Score médio, porém mantém uma taxa de falsos positivos elevada, característica consistente ao longo das execuções. A baixa variação das métricas (DP reduzido) evidencia a estabilidade desse comportamento, associada à sensibilidade do algoritmo a regiões de baixa densidade presentes no conjunto de dados.

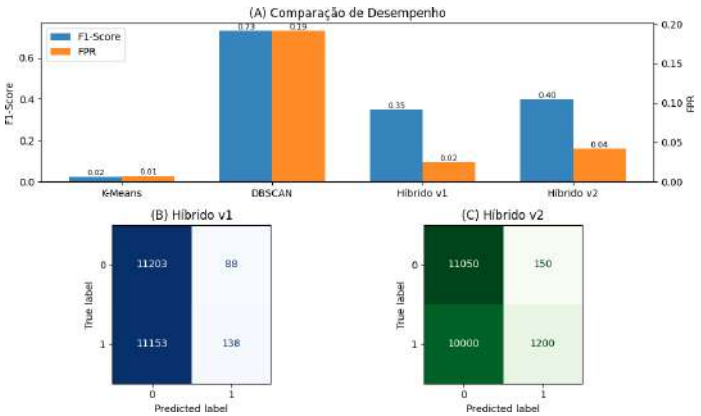


Figura 3. Comparação dos modelos em termos de F1-Score e Falsos Positivos (FPR), com as matrizes de confusão de arquitetura híbrida (v1 e v2).

Nesse contexto, a arquitetura híbrida demonstrou melhor equilíbrio entre detecção e precisão. A versão inicial (v1) priorizou a redução do FPR (≈ 0,025), enquanto a versão otimizada (v2) elevou o F1-Score médio para 0,3967, mantendo níveis controlados de falsos positivos (≈ 0,04). A Figura 3 apresenta a comparação visual desses resultados.

De forma complementar, observa-se que o K-Means isolado apresenta baixo desempenho (F1 ≈ 0,024), enquanto o DBSCAN puro, apesar da alta capacidade de detecção (F1 ≈ 0,73), apresenta limitações operacionais

devido ao elevado FPR (≈ 0,19). Nesse cenário, a arquitetura híbrida v2 configura-se como a alternativa mais adequada para ambientes industriais. Com base na análise experimental, a arquitetura híbrida demonstrou melhor equilíbrio entre detecção e precisão.

6. Conclusão

Este trabalho investigou a viabilidade de uma arquitetura híbrida sequencial baseada em clustering não supervisionado para detecção de intrusões em redes industriais Modbus/TCP. A proposta integrou os algoritmos K-Means e DBSCAN em uma arquitetura em cascata, na qual o primeiro atua como mecanismo de filtragem inicial e o segundo como refinador de anomalias baseado em densidade.

A avaliação experimental, conduzida com o ICS Dataset (Sudyana, 2025), demonstrou que o DBSCAN puro apresenta maior F1-Score médio (≈ 0,73), porém com elevada taxa de falsos positivos (≈ 0,19), limitando sua aplicação em ambientes industriais. Em contrapartida, a arquitetura híbrida apresentou melhor equilíbrio operacional, com FPR médio de ≈ 0,025 na versão inicial (v1) e F1-Score de ≈ 0,40 na versão otimizada (v2), mantendo FPR controlado (≈ 0,04). A análise estatística por teste t pareado confirmou a significância dos resultados.

Os resultados evidenciam que a arquitetura híbrida proposta demonstra empiricamente a capacidade de equilibrar detecção e viabilidade operacional, nos quais a redução de falsos positivos é essencial para evitar interrupções indevidas em processos críticos.

Como limitações, destacam-se a utilização de um único dataset, a avaliação em ambiente offline e a sensibilidade do DBSCAN aos parâmetros ε e MinPts. Além disso, o modelo foi restrito ao protocolo Modbus/TCP, e o F1-Score obtido ainda apresenta margem de melhoria. Como trabalhos futuros, sugerem-se a validação em múltiplos datasets industriais, a aplicação em fluxos de dados em tempo real e a integração com técnicas semi-supervisionadas ou deep learning.

Este trabalho contou com o apoio de ferramentas de Inteligência Artificial, como ChatGPT, Gemini e Claude, utilizadas exclusivamente como

suporte à revisão textual e organização do conteúdo, sem substituição da autoria intelectual.

7. Referências

Alanazi, M., Mahmood, A., & Chowdhury, M. J. M. (2025). ICS-LTU2022: A dataset for ICS vulnerabilities. Computers & Security, 148, 104143. https://doi.org/10.1016/j.cose.2024.104143

Anthropic. (2026). Claude. Recuperado de: https://claude.ai

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Anais do 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96) (pp. 226-231)*, Portland, Estados Unidos.

Google. (2026). Gemini. Recuperado de: https://gemini.google.com

Lazaridis, G., Drosou, A., Chatzimisios, P., & Tzovaras, D. (2024). Unraveling the threat landscape of CPS: Modbus TCP vulnerabilities in the era of I4.0. In Proceedings of the IEEE International Conference on Cyber Security and Resilience (CSR), Londrina, Reino Unido. https://doi.org/10.1109/CSR61664.2024.10679453

Liu, C., Gu, Z., & Wang, J. (2021). A hybrid intrusion detection system based on scalable K-Means+ Random Forest and Deep Learning. IEEE Access, 9, 75729-75740. https://doi.org/10.1109/ACCESS.2021.3082147

Machaka, V., Figueroa-Lorenzo, S., Arrizabalaga, S., Elduayen-Echave, B., & Hernantes, J. (2025). Assessing the impact of Modbus/TCP protocol attacks on critical infrastructure: WWTP case study. Computers & Electrical Engineering, 113, 110485. https://doi.org/10.1016/j.compeleceng.2025.110485

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Anais do 5th Berkeley Symposium on Mathematical Statistics and Probability (pp. 281-297), Berkeley, Estados Unidos.

Nankya, M., Chataut, R., & Akl, R. (2023). Securing Industrial Control Systems: Components, cyber threats, and machine learning-driven defense strategies. Sensors, 23(21), 8840. https://doi.org/10.3390/s23218840

Nicolaio, I. G. A., Munaretto, A., & Fonseca, M. (2023). Uma proposta de detecção de ataques cibernéticos em sistemas de controle industrial (ICS). In Anais do XXXI Workshop de Gerência e Operação de Redes e Serviços (WGRS) (pp. 45-58), Brasília, Brasil. https://doi.org/10.5753/wgrs.2023.765

Nikolajew, C., & Eichelberger, H. (2024). Industry 4.0 connectors: A performance experiment with Modbus/TCP. https://doi.org/10.48550/arXiv.2410.15813

Okolie, S. A., Amadi, C. A., Odii, J. N., Nwokorie, E. C., & Onyemauche, U. C. (2025). Anomaly detection in heterogeneous cybersecurity data. Franklin Open, 10, 100426. https://doi.org/10.1016/j.fraope.2025.100426

OpenAI. (2026). ChatGPT. Recuperado de: https://chat.openai.com

Pallakonda, A., Ravishanmugam, K., & Raj, R. D. A. (2025). A Unified Cybersecurity Framework for Smart Grids Against Data Integrity Attacks Using Ensemble Learning and Hybrid Encryption. IEEE Access, 13, 177595 - 177614. https://doi.org/10.1109/ACCESS.2025.3616505

Sobonski, P., & Roedig, U. (2024). Comparison of Industrial Control System anomaly detection methods. In Proceedings of the 2024 Workshop on Re-design Industrial Control Systems with Security (RICSS '24) (pp. 1-12). https://doi.org/10.1145/3689930.3695211

Sudyana, D. (2025). ICS Dataset [Data set]. Zenodo. https://doi.org/10.5281/zenodo.17165850

Capítulo 03

Resiliência cibernética na Indústria 4.0: Detecção de ataques de desfiguração via modelos transformer

Hugo Lima Romão
Marcelle Alencar Urquiza
Filipe Dwan Pereira
Universidade Federal de Roraima

hugo.romao@gmail.com
{marcelle.urquiza, filipe.dwan}@ufrr.com

1. Introdução

A crescente digitalização de serviços e processos tem ampliado a superfície de ataque em sistemas interconectados, especialmente no contexto da Indústria 4.0 (Haleem et al., 2022; Singh & Kumar, 2025; Yu et al., 2024). Nesse cenário, ataques de desfiguração são uma preocupação em segurança da informação, principalmente devido aos impactos econômicos e operacionais causados a governos e organizações. Nesse tipo de incidente, o invasor explora vulnerabilidades em sistemas web com o objetivo de substituir conteúdos legítimos por material ofensivo, propaganda ou manifestações políticas (Albalawi et al., 2022; Banerjee et al., 2022; Moneva et al., 2022). Esse fenômeno torna-se ainda mais preocupante à medida que o volume de conteúdo gerado por organizações cresce, possibilitando que tais ataques passem despercebidos por revisões humanas.

Diante dessa realidade, sistemas de monitoramento autônomos são essenciais para a detecção precoce e mitigação dos impactos desse tipo de ataque. Métodos tradicionais de detecção, baseados em comparação estrutural ou detecção por assinatura, apresentam limitações frente a conteúdos dinâmicos e a modificações textuais sutis realizadas por invasores. Ao mesmo tempo, estudos na literatura destacam o potencial de modelos de aprendizagem profunda na tarefa de detecção de incidentes cibernéticos, sobretudo modelos de linguagem baseados na arquitetura transformer (Haynes et al., 2021; Liu & Dai, 2024). Convém, portanto, perguntar: Até que ponto um modelo baseado em Transformers é capaz de

atingir um score F1 superior a 99,5% em cenários de páginas web heterogêneas, validando seu uso como um sensor inteligente de segurança?

Na literatura científica, a detecção de ataques de desfiguração tem sido abordada por meio de modelos de aprendizagem profunda tradicionais, como redes neurais recorrentes (Hoang et al., 2025). Contudo, tais abordagens concentram-se em representações textuais estáticas, as quais não incorporam informações contextuais nem capturam adequadamente dependências semânticas de longo alcance. Com isso em mente, este trabalho tem como objetivo avaliar a performance do modelo Bidirectional Encoder Representations from Transformers (BERT), ajustado ao domínio de ataques de desfiguração por meio da técnica de transferência de aprendizagem. Para isso, utilizamos uma base de dados rotulada de 98.086 páginas web legítimas e desfiguradas, obtidas a partir do repositório público de ataques de desfiguração Zone-H (Roberto Preatoni, 2002).

Acreditamos que a natureza dos ataques de desfiguração favorece a identificação de padrões textuais por meio da geração de embeddings contextuais. Este trabalho se diferencia de outras pesquisas existentes na literatura, principalmente, pela integração da técnica de transferência de aprendizagem e do modelo pré-treinado BERT no contexto da tarefa de detecção de ataques de desfiguração.

2. Fundamentação teórica

Apresentamos nesta seção a teoria e os conceitos que fundamentaram este estudo. Primeiramente, caracterizamos os ataques de desfiguração, sua natureza, motivações e impacto em organizações e produtos. Em seguida, abordamos o desafio de identificação de ciberataques sob a ótica da tarefa de classificação de texto moderna, o que envolve aplicação de modelos baseados na arquitetura transformer, como o modelo BERT.

2.1. Ataques de desfiguração

Ataques de desfiguração podem ser definidos como incidentes cibernéticos em que o atacante explora vulnerabilidades de um site ou

servidor web com o objetivo de alterar ou remover o conteúdo de uma página ou aplicação web (Albalawi et al., 2022; Banerjee et al., 2022). Esses ataques são recorrentes e afetam uma ampla variedade de alvos, desde empresas privadas até órgãos governamentais (Hoang et al., 2025). As motivações dos invasores variam, incluindo (1) a busca por diversão, (2) o desejo de obter prestígio dentro da comunidade hacker, (3) razões de cunho político e (4) sentimentos de patriotismo (Banerjee et al., 2022).

No contexto da Indústria 4.0, os ataques de desfiguração são preocupantes , visto que podem tornar inoperantes processos e aplicações baseadas na web, resultando em perdas financeiras significativas e desgaste da reputação. Diante desse cenário, em que indústrias estão cada vez mais produzindo dados, gerenciando e oferecendo serviços personalizados por meio da Internet, destaca-se não apenas a importância em fortalecer a segurança do ciberespaço, mas também a necessidade de desenvolver abordagens automatizadas para a detecção de incidentes, favorecendo a resiliência dos sistemas industriais e reduzindo o tempo de indisponibilidade. Uma das abordagens modernas de detecção de incidentes cibernéticos ocorre por meio da tarefa de classificação de texto, descrita a seguir.

2.2. Classificação de texto, Transferência de Aprendizagem e o modelo BERT

A classificação de texto (CT) é uma tarefa de Processamento de Linguagem Natural (PLN) que consiste em atribuir rótulos a textos de acordo com seu conteúdo (Gasparetto et al., 2022). O aperfeiçoamento dos modelos de linguagem possibilitou a extração automática de representações semânticas complexas, melhorando significativamente o desempenho desses sistemas (Gasparetto et al., 2022).

Nesse contexto, a transferência de aprendizagem tem contribuído para a disseminação desses modelos, ao permitir o uso de modelos pré-treinados em grandes volumes de dados, posteriormente ajustados a tarefas específicas. Essa abordagem possibilita bons resultados mesmo com bases menores (Géron, 2025; Miquel Farré, 2026) e tem sido aplicada na detecção de incidentes cibernéticos, como código malicioso e páginas de phishing

(Damasceno et al., 2021; Haynes et al., 2021; João A. Souza & Dalbert M. Mascarenhas, 2023; Liu & Dai, 2024).

Entre os modelos que mais contribuíram para o avanço dessas aplicações, destaca-se o BERT, baseado na arquitetura transformer e amplamente utilizado em PLN. Seu mecanismo de encoder bidirecional permite capturar dependências contextuais ao considerar simultaneamente o contexto anterior e posterior das palavras (Gasparetto et al., 2022). Neste trabalho, utilizamos o BERT para a detecção de ataques de desfiguração em páginas web.

3. Trabalhos relacionados

Conduzimos uma breve revisão da literatura sobre os temas relacionados a ataques de desfiguração de sites e aos métodos de detecção de ciberataques baseados em modelos de aprendizagem profunda modernos, da qual selecionamos cinco estudos relevantes ao nosso tema de pesquisa. Passamos, a seguir, a uma descrição desses estudos.

Banerjee et al. (2022) investigam se o conteúdo de ataques de desfiguração reflete as motivações dos atacantes, analisando registros do portal Zone-H. Os resultados evidenciam a presença de padrões semânticos recorrentes nos sites comprometidos, indicando que esses ataques podem ser caracterizados a partir de informações textuais. Em complemento, Albalawi et al. (2022) apresentam uma revisão dos principais métodos de detecção de ataques de desfiguração, classificando-os em abordagens baseadas em anomalia, assinatura e aprendizagem de máquina, sendo esta última associada aos melhores resultados. Os autores destacam que muitos estudos utilizam bases de dados de tamanho limitado, o que pode comprometer o desempenho dos modelos. Nesse contexto, observa-se uma lacuna relacionada à disponibilidade de dados, a qual é abordada neste trabalho por meio do uso de modelo pré-treinado e técnica de transferência de aprendizagem.

No contexto de modelos baseados na arquitetura transformer, Haynes et al. (2021) propõem um método de detecção de ataques de phishing baseado em características de URLs, demonstrando que modelos pré-treinados como BERT e ELECTRA alcançam desempenho competitivo com

menor custo computacional. De forma semelhante, Liu & Dai (2024) exploram o uso de BERT combinado com LSTM para detecção de ataques por injeção SQL, obtendo melhorias significativas nas métricas de desempenho e evidenciando a superioridade das representações semânticas geradas pelo BERT em relação a métodos tradicionais. Esses estudos evidenciam o potencial de modelos transformer em tarefas de classificação semântica no domínio da cibersegurança. No entanto, observa-se que essa abordagem ainda não foi explorada de forma direta na detecção de ataques de desfiguração em páginas web, lacuna abordada neste trabalho por meio do uso do modelo BERT.

Por fim, Hoang et al. (2025) avaliam a detecção de ataques de desfiguração utilizando exclusivamente características textuais de páginas web. O modelo proposto, baseado em uma rede Bi-LSTM, utiliza embeddings treinados do zero a partir de um corpus composto por páginas desfiguradas e legítimas. Os resultados indicam desempenho superior aos métodos clássicos de aprendizagem de máquina, como Naive Bayes, Decision Tree e Random Forest. No entanto, a abordagem não explora o uso de modelos de linguagem com geração de embeddings contextuais, o que pode limitar a captura de dependências semânticas mais complexas, configurando uma limitação que este trabalho busca abordar com o uso de modelos baseados em BERT.

Estes cinco estudos mostram como a pesquisa sobre incidentes cibernéticos permanece relevante e em expansão. Diversos são os esforços na literatura para construção de ferramentas que auxiliem na detecção precoce de ciberataques. Contudo, persiste uma lacuna nas abordagens voltadas à detecção de ataques de desfiguração, especialmente quanto à utilização de bases de dados mais extensas e técnicas de aprendizagem profunda modernas. Assim, considerando os avanços no processamento de linguagem natural, bem como o intensivo uso de modelos baseados na arquitetura transformer no campo da cibersegurança, este trabalho teve como objetivo avaliar a eficiência do modelo BERT para a tarefa de detecção de ataques de desfiguração em sites.

4. Solução computacional

4.1. Coleta de dados

Foi construída uma base de dados balanceada com 98.086 amostras de sites desfigurados e legítimos. Cada amostra possui o conteúdo textual da página principal do site e seu rótulo. Para as amostras de sites desfigurados, utilizamos os dados do portal Zone-H (Roberto Preatoni, 2002), que incluem páginas de diferentes setores, abrangendo também ambientes digitais associados ao contexto da Indústria 4.0. As amostras de sites legítimos, por sua vez, foram extraídas por meio da lista TRANCO de sites fortificados contra manipulações (Le Pochat et al., 2019). O período de coleta iniciou dia 14 de novembro de 2025 e finalizou em 12 de janeiro de 2026.

Criado em 2002, o portal Zone-H mantém um extenso repositório de incidentes de ataques de desfiguração (Roberto Preatoni, 2002). Os incidentes são notificados por organizações, profissionais em cibersegurança e agentes maliciosos, passando por um processo de curadoria realizado pelos administradores antes da publicação. Após a confirmação do incidente, são publicadas informações técnicas sobre o site afetado, como país de origem, usuário responsável pela notificação, endereço IP e um espelho do site no momento do ataque, o que inclui os arquivos HTML utilizados neste estudo para extração do texto.

A lista TRANCO consiste em um ranking de sites mais acessados da web. Para sua composição, são aplicados filtros que removem domínios maliciosos e garantem que apenas sites regularmente acessados sejam incluídos. Além disso, os autores utilizam diferentes rankings públicos para a composição da lista, o que a torna mais segura contra manipulação por agentes maliciosos (Le Pochat et al., 2019). A lista TRANCO é extensamente utilizada na comunidade acadêmica e por organizações de segurança da informação como fonte de sites legítimos (Haraldsdóttir et al., 2024).

Para ambos os tipos de amostras, utilizamos técnicas de raspagem de dados para realizar a coleta do conteúdo textual. Esse procedimento consiste no uso de um cliente web para consultar as páginas do site e armazenar as informações de modo estruturado. Para o portal Zone-H, a consulta automatizada é amparada pela licença Creative Commons adotada pelo portal, respeitando os requisitos de atribuição e utilização não comercial.

Além disso, o processo de coleta realizado neste estudo respeitou o limite de dois segundos entre consultas, conforme orientação da administração do portal. Para os sites legítimos derivados da lista TRANCO, realizamos apenas uma requisição para cada site, garantindo que não houvesse qualquer risco de sobrecarga dos servidores.

Os dados utilizados no treinamento do modelo podem ser disponibilizados mediante solicitação aos autores. A divulgação irrestrita desses dados não é realizada em razão da presença de conteúdo sensível, o qual poderia permitir a identificação de vítimas passadas, em conformidade com princípios éticos e de privacidade.

4.2. Pré-processamento

O pré-processamento consiste em etapas para tornar os dados compreensíveis e adequados a modelos de aprendizagem de máquina. Primeiro, extraímos o conteúdo textual dos arquivos HTML. Consideramos apenas o texto puro das páginas para modelar o problema sob perspectiva estritamente semântica. Essa abordagem reduz a dimensionalidade dos dados, mas desconsidera indícios de desfiguração em imagens e código. Em seguida, removemos amostras com conteúdo textual duplicado por meio de correspondência exata. Essa abordagem elimina apenas duplicatas idênticas, não contemplando variações mínimas de conteúdo.

A próxima etapa consiste na tokenização do corpus, isto é, na divisão do texto em unidades menores e na atribuição de um identificador único para cada token. Utilizamos o mesmo tokenizador empregado no pré-treinamento do BERT, baseado no algoritmo WordPiece e treinado sobre os corpora Wikipedia e BookCorpus (Devlin et al., 2019). Como o BERT admite no máximo 512 tokens por entrada, sequências que excederam esse limite foram truncadas, adotando-se a estratégia *head-only*, na qual apenas os primeiros tokens da sequência são mantidos.

4.3. Ajuste do modelo BERT

Utilizamos o modelo BERT disponibilizado pela biblioteca Hugging Face, especificamente o *bert-base-multilingual-uncased*, treinado em um

corpus multilíngue (Devlin et al., 2019). A escolha por um modelo multilíngue se justifica pela natureza heterogênea do conjunto de dados, que inclui diferentes idiomas. O modelo foi ajustado por meio da técnica de transferência de aprendizagem para a tarefa de detecção de ataques de desfiguração. Além disso, foi adicionada uma camada linear de classificação, composta por 768 unidades de entrada e duas unidades de saída, correspondentes às classes negativa (sites legítimos) e positiva (sites desfigurados). Por fim, o modelo ajustado foi disponibilizado para avaliação e escrutínio público no repositório Hugging Face, disponível em: https://huggingface.co/hugoromao/defacement-detection.

5. Avaliação experimental

Iniciamos esta seção apresentando as métricas utilizadas na avaliação do modelo desenvolvido, bem como os procedimentos adotados para obtenção dos resultados. Em seguida, descrevemos os parâmetros de treinamento empregados. Por fim, apresentamos os achados e discutimos os resultados obtidos.

5.1. Planejamento de pesquisa

A partir da base de dados construída, selecionamos um conjunto de informações relevantes para o contexto de ataques de desfiguração, incluindo o percentual de ocorrências por país, a nuvem de palavras mais frequente em páginas desfiguradas e o total de ataques registrados por ano.

Realizamos o ajuste fino do modelo pré-treinado BERT para a tarefa de detecção de sites desfigurados. O conjunto de dados foi dividido em subconjuntos de treino, validação e teste na proporção de 80%, 10% e 10%, respectivamente. Na etapa de treinamento, utilizamos a função de custo *CrossEntropyLoss*, adequada para problemas de classificação binária. Para avaliação do modelo, foram coletadas as métricas de acurácia, precisão, revocação, score F1 e curva ROC. A partir destes resultados foram gerados gráficos de função de custo e acurácia durante o treino. Além disso, utilizamos o subconjunto de teste para confecção da matriz de confusão.

Ressaltamos que as métricas foram calculadas em uma base balanceada, o que representa um cenário idealizado. Em ambientes reais, onde os dados são fortemente desbalanceados, a quantidade de falsos positivos tende a aumentar, podendo impactar a precisão e gerar sobrecarga de alertas.

O ajuste fino dos pesos do modelo foi realizado utilizando a biblioteca Hugging Face do Python. A taxa de aprendizado foi definida em 2e-5, valor comumente adotado no ajuste fino de modelos BERT por apresentar bom equilíbrio entre convergência e estabilidade, não sendo realizado tuning de hiperparâmetros por não estar no escopo deste estudo. O tamanho do batch de treino foi definido em 32 e o número máximo de épocas como 30. Embora o treinamento tenha sido configurado para 30 épocas, empregamos a técnica de Early Stopping para interrompê-lo automaticamente quando a melhoria no desempenho entre épocas deixasse de ser significativa após duas épocas, valor escolhido considerando a rápida convergência de modelos pré-treinados. Essa abordagem atua como um mecanismo de regularização, visto que pode atenuar o sobreajuste do modelo.

Por fim, as etapas de pré-processamento e treinamento do modelo foram realizadas no ambiente Google Colab, utilizando uma máquina virtual equipada com GPU NVIDIA H100, a qual dispõe de 230 GB de memória RAM e 79,6 GB de memória de vídeo.

5.2. Resultados e Discussão

Começamos esta seção apresentando os resultados referentes a base de dados construída. No total, foram coletadas 1.093.075 amostras de sites legítimos e desfigurados. Após a etapa de pré-processamento, obteve-se uma base balanceada de 98.086 páginas. Esta redução substancial se deve ao grande número de páginas duplicadas nas amostras de sites desfigurados. Este dado revela que, para esta classe de incidente, é comum que atacantes reutilizem a mesma assinatura de ataque em diferentes vítimas. As Figuras 1A e 1B, por sua vez, ilustram as palavras mais utilizadas em sites desfigurados e legítimos, respectivamente. Em seguida, a Figura 2 mostra a distribuição de ocorrências de incidentes de desfiguração por país.

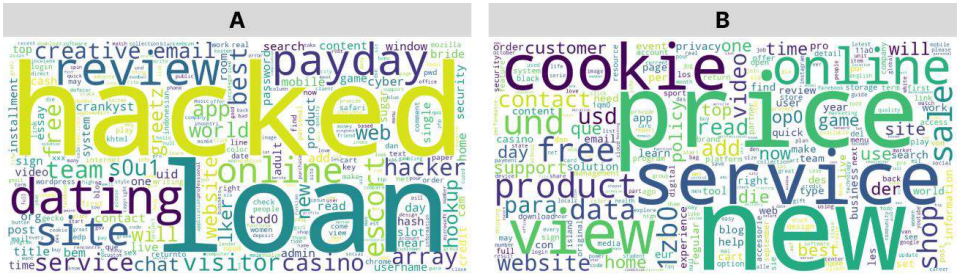


Figura 1. Nuvens de palavras obtidas após pré-processamento textual das páginas, ilustrando termos mais frequentes em (A) páginas desfiguradas e (B) páginas legítimas.

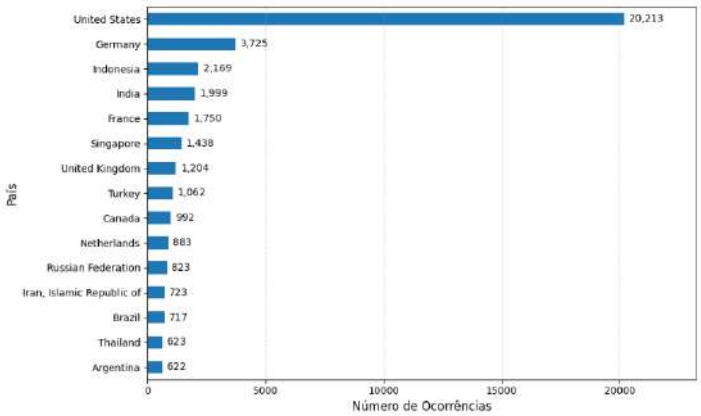


Figura 2. Distribuição do número de incidentes de desfiguração por país.

Durante a etapa de treinamento, foram coletadas métricas de desempenho ao final de cada época, com base no conjunto de validação. Em função da aplicação da técnica de Early Stopping, o treinamento foi interrompido após seis épocas, uma vez que a quinta e sexta época não apresentaram melhora significativa no valor da função de custo para a base de validação.

A Figura 3A apresenta a evolução da função de custo ao longo do treinamento para os conjuntos de treino e validação. Observa-se uma redução acentuada da perda nas primeiras épocas, indicando rápida convergência do modelo. A partir da quinta época, verifica-se um aumento da

perda no conjunto de validação, enquanto a perda de treino continua diminuindo, caracterizando o início de sobreajuste. Esse comportamento também se reflete na Figura 3B, na qual se observa a queda das métricas de desempenho calculadas sobre a base de validação na quinta época. Assim, selecionamos o modelo da quarta época, que apresentou o menor valor de perda na validação (*perda = 0,012249*). Esse comportamento é comum no ajuste fino de modelos pré-treinados, nos quais poucas iterações sobre os dados são suficientes para alcançar boa capacidade de generalização.

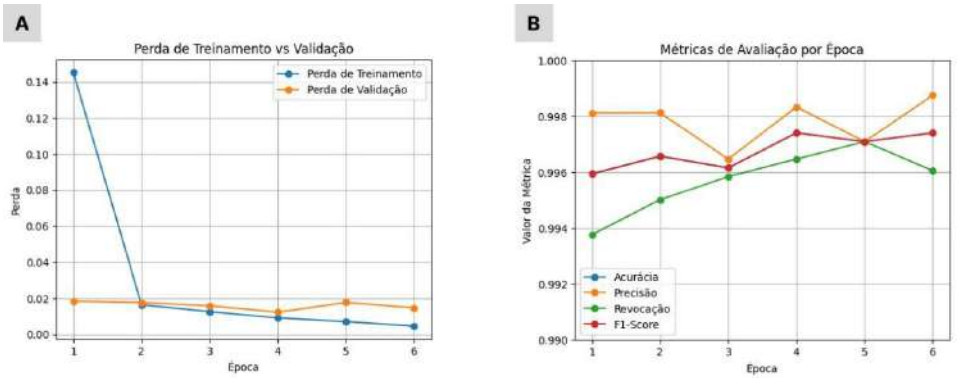


Figura 3. Curvas de treinamento do modelo: (A) evolução da função de perda para treino e validação; (B) evolução das métricas de desempenho ao longo das épocas.

Em seguida, a Figura 4 mostra a matriz de confusão referente às previsões do modelo no conjunto de teste. Acreditamos que a ocorrência de falsos negativos esteja associada a limitações do pré-processamento e da representação textual adotada. Em particular, nas amostras classificadas erroneamente, observou-se que indícios de desfiguração estavam localizados na porção final do documento, sendo descartados pela estratégia de truncamento *head-only*. Além disso, em alguns casos, as evidências do ataque estão mais presentes em elementos não textuais, como imagens ou código, os quais não são capturados pela abordagem baseada em texto puro.

Por fim, apresentamos as métricas de desempenho referentes às previsões do modelo no conjunto de teste. Foram obtidos valores de 0,997, 0,998, 0,997 e 0,998 para acurácia, precisão, revocação e score F1, respectivamente. Esses resultados indicam excelente ajuste do modelo, pois

todas as métricas permaneceram superiores a 0,99. Além disso, tais valores mostram-se competitivos quando comparados aos resultados reportados por Hoang et al. (2025), cujo modelo Bi-LSTM apresentou acurácia de 0,9654, precisão de 0,9693, revocação de 0,9443 e score F1 de 0,9566.

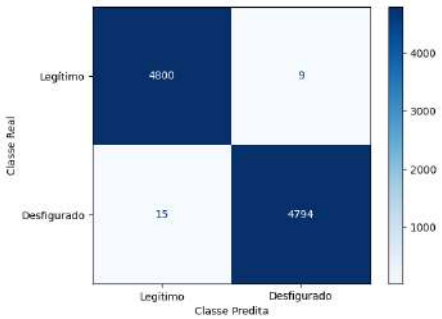


Figura 4. Matriz de confusão das previsões do modelo no conjunto de teste.

Destaca-se, contudo, que não é possível afirmar a superioridade do BERT em relação ao modelo Bi-LSTM, uma vez que os experimentos foram conduzidos em bases de dados distintas. Essa diferença configura uma limitação metodológica do presente estudo, pois impede a realização de uma comparação direta entre os modelos. Além disso, o viés da base de dados, predominantemente composta por páginas em língua inglesa, pode reduzir a eficácia do modelo em outros idiomas, como o português.

No que diz respeito ao pré-processamento, a estratégia de truncamento *head-only* pode levar à perda de informações relevantes presentes no final das páginas. De forma complementar, a remoção de duplicatas por correspondência exata pode não eliminar variações mínimas de conteúdo, comuns em campanhas de desfiguração em massa, o que pode introduzir vazamento de dados entre os conjuntos de treino e teste e resultar em possível superestimação das métricas de desempenho. Ainda assim, os resultados encontrados respondem à questão de pesquisa deste estudo, ao demonstrar que o modelo baseado em Transformers é capaz de atingir um F1 score superior a 99,5%, evidenciando elevado potencial de generalização para a tarefa de detecção automática de páginas web desfiguradas e reforçando sua viabilidade como um sensor inteligente de segurança.

6. Conclusão

O objetivo deste estudo foi avaliar o potencial do uso de modelos de linguagem contextuais, como o modelo BERT, na tarefa de detecção de ataques de desfiguração em páginas web. Para isso, realizamos o ajuste fino do modelo utilizando uma base de dados balanceada com 98.086 amostras de sites legítimos e desfigurados. Verificamos que o modelo apresentou bom desempenho na tarefa proposta, alcançando acurácia de 0,997, precisão de 0,998, revocação de 0,997 e score F1 de 0,998 no conjunto de teste. Esses resultados indicam que o modelo BERT constitui uma abordagem promissora para a detecção automática de páginas web desfiguradas, especialmente em razão da natureza textual e semântica que caracteriza os ataques de desfiguração.

Do ponto de vista prático, a abordagem desenvolvida pode contribuir para a redução do tempo de identificação de incidentes e apoiar a priorização de alertas em ambientes de monitoramento contínuo. Em cenários da Indústria 4.0, nos quais há grande volume de sistemas interconectados, essa capacidade pode aumentar a eficiência operacional e auxiliar na mitigação rápida de ataques, reduzindo impactos econômicos e operacionais.

Sugerimos que estudos futuros avancem em relação a esta pesquisa, buscando lidar com as limitações que apresentamos. Recomenda-se a realização de experimentos comparativos diretos entre BERT e outros modelos, como Bi-LSTM e arquiteturas mais recentes baseadas em transformer. Além disso, futuras investigações podem ampliar o corpus com maior diversidade linguística, bem como incorporar características multimodais, como elementos visuais e de código. Também é promissora a avaliação do modelo em cenários de implantação em tempo real, considerando restrições computacionais e integração com sistemas de monitoramento contínuo, contribuindo para o fortalecimento da resiliência cibernética no contexto da Indústria 4.0.

7. Referências

Albalawi, M., Aloufi, R., Alamrani, N., Albalawi, N., Aljaedi, A., & Alharbi, A. R. (2022). Website Defacement Detection and Monitoring Methods: A

Review. *Electronics*, *11*(21), 3573. https://doi.org/10.3390/electronics11213573

Banerjee, S., Swearingen, T., Shillair, R., Bauer, J. M., Holt, T., & Ross, A. (2022). Using Machine Learning to Examine Cyberattack Motivations on Web Defacement Data. *Social Science Computer Review*, *40*(4), 914–932. https://doi.org/10.1177/0894439321994234

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Recuperado de: http://arxiv.org/abs/1810.04805

Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). Survey on Text Classification Algorithms: From Text to Predictions. *Information*, *13*(2). https://doi.org/10.3390/info13020083

Haleem, A., Javaid, M., Singh, R. P., Rab, S., & Suman, R. (2022). Perspectives of cybersecurity for ameliorative Industry 4.0 era: a review-based framework. *Industrial Robot*, *49*(3), 582–597. https://doi.org/10.1108/IR-10-2021-0243

Haraldsdóttir, M., Homayoun, S., Lynge, E., & Jensen, C. D. (2024). Unmasking phishers: ML for malicious certificate detection. *Computers and Industrial Engineering*, *198*. https://doi.org/10.1016/j.cie.2024.110652

Haynes, K., Shirazi, H., & Ray, I. (2021). Lightweight URL-based phishing detection using natural language processing transformers for mobile devices. *Procedia Computer Science*, *191*, 127–134. https://doi.org/10.1016/j.procs.2021.07.040

Hoang, X. D., Nguyen, T. H., & Pham, H. D. (2025). A novel model for detecting web defacement attacks transformer using plain text features. *Indonesian Journal of Electrical Engineering and Computer Science*, *37*(1), 232–240. https://doi.org/10.11591/ijeecs.v37.i1.pp232-240

Le Pochat, V., Van Goethem, T., Tajalizadehkhoob, S., Korczyński, M., & Joosen, W. (2019). TRANCO: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. *26th Annual Network and Distributed System Security Symposium, NDSS 2019*. https://doi.org/10.14722/ndss.2019.23386

Liu, Y., & Dai, Y. (2024). Deep Learning in Cybersecurity: A Hybrid BERT–LSTM Network for SQL Injection Attack Detection. *IET Information Security*, *2024*. https://doi.org/10.1049/2024/5565950

Moneva, A., Leukfeldt, E. R., Van De Weijer, S. G. A., & Miró-Llinares, F. (2022). Repeat victimization by website defacement: An empirical test of premises from an environmental criminology perspective. *Computers in Human Behavior*, *126*. https://doi.org/10.1016/j.chb.2021.106984

Roberto Preatoni. (2002). *Zone-H.org Unrestricted Information*. Recuperado de: https://www.zone-h.org

Singh, B., & Kumar, B. (2025). Navigating Cybersecurity Risks in Industry 4.0: Challenges, Threats, and Defense Strategies. *Journal of Information Systems Engineering and Management*, *2025*, 2468–4376. https://doi.org/https://doi.org/10.52783/jisem.v10i19s.3018

Yu, J., Shvetsov, A. V., & Hamood Alsamhi, S. (2024). Leveraging Machine Learning for Cybersecurity Resilience in Industry 4.0: Challenges and Future Directions. In *IEEE Access* (Vol. 12, pp. 159579–159596). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2024.3482987

Capítulo 04

Detecção de mensagens de phishing utilizando Random Forest e Processamento de Linguagem Natural em ambientes da Indústria 4.0

Pedro Vinícius da Silva Ribeiro
Felipe Leite Lobo
Universidade Federal de Roraima

silvapv7@gmail.com
felipe.lobo@ufrr.com

1. Introdução

A Indústria 4.0 fundamenta-se na hiperconectividade promovida pela Internet das Coisas (IoT), integrando sistemas ciberfísicos e dispositivos inteligentes em ambientes industriais altamente automatizados. Essa integração possibilita ganhos de eficiência operacional, escalabilidade e monitoramento em tempo real, porém também amplia a superfície de ataque dos sistemas industriais, aumentando sua exposição a ameaças cibernéticas (Olianin et al., 2025).

O phishing é um dos principais vetores de ataque em ambientes corporativos, explorando engenharia social para contornar mecanismos tradicionais de defesa (Alshingiti et al., 2023; Silva et al., 2024). A magnitude do problema é evidenciada por dados recentes da indústria: o custo médio global de uma violação de dados atingiu US$4,88 milhões em 2024, com o phishing figurando como o vetor de ataque mais frequente (IBM, 2024). No contexto industrial, esse cenário é agravado pela crescente convergência entre redes de tecnologia da informação (TI) e tecnologia operacional (TO), o que amplia a superfície de ataque e eleva o impacto potencial de incidentes de segurança (Zhu et al., 2023). Esses fatores reforçam a necessidade de soluções automatizadas e escaláveis para a detecção de ameaças em ambientes de alta criticidade.

Apesar dos avanços no uso de Inteligência Artificial para detecção de phishing, a literatura ainda apresenta limitações quanto à capacidade de

generalização dos modelos quando expostos a dados com alta variabilidade linguística e distribuição desigual entre classes. Muitos estudos reportam elevados índices de acurácia em bases homogêneas, porém apresentam degradação de desempenho quando aplicados a cenários mais próximos da realidade, caracterizados pela diversidade de fontes de dados e pelo desbalanceamento entre mensagens legítimas e maliciosas.

Diante desse contexto, este trabalho propõe o desenvolvimento de um sistema inteligente de detecção de phishing baseado em Aprendizado de Máquina Supervisionado, utilizando o algoritmo Random Forest. A abordagem adotada busca mitigar as limitações identificadas na literatura por meio da integração de múltiplas bases de dados heterogêneas, da aplicação de técnicas de balanceamento de classes e do uso de Processamento de Linguagem Natural (PLN) para análise semântica das mensagens.

A escolha do algoritmo Random Forest justifica-se por sua robustez estatística, capacidade de lidar com dados de alta dimensionalidade e desempenho consistente em tarefas de classificação em segurança cibernética (Rao et al., 2023; Silva et al., 2024). Adicionalmente, a utilização de vetorização TF-IDF com n-gramas permite capturar padrões semânticos relevantes associados a ataques de phishing, como expressões de urgência, redirecionamento para links externos e solicitações de informações sensíveis.

Os resultados obtidos demonstram que o modelo proposto apresenta desempenho elevado, com acurácia média superior a 96% em validação cruzada, além de elevada capacidade de detecção de mensagens maliciosas. Esses resultados indicam que a abordagem adotada contribui para a redução do viés estatístico e para a melhoria da capacidade de generalização do modelo, mesmo diante da unificação de bases de dados heterogêneas.

Este estudo tem como objetivo desenvolver e validar um modelo robusto para detecção de phishing. Para isso, investiga-se: (i) o impacto da integração de bases heterogêneas na estabilidade do modelo; (ii) a influência do balanceamento de classes na redução de erros; e (iii) a capacidade da vetorização com n-gramas em capturar padrões semânticos. Como objetivos específicos, incluem a unificação de dados, aplicação de técnicas de PLN, treinamento do modelo e análise da importância das features.

Com base nessas questões, os objetivos específicos incluem a unificação de múltiplas bases de dados heterogêneas em um conjunto estruturado, a aplicação de técnicas de pré-processamento textual e vetorização por TF-IDF com n-gramas, a mitigação do viés estatístico por meio de balanceamento de classes, o treinamento e avaliação de um modelo baseado em Random Forest utilizando validação cruzada e a análise da importância das features para identificação dos principais padrões linguísticos associados a mensagens maliciosas.

No contexto da convergência entre TI e TO na Indústria 4.0, este trabalho propõe um pipeline reprodutível de detecção de phishing voltado às camadas de perímetro e aplicação (Zhu et al., 2023). A solução integra múltiplas fontes de dados e técnicas de processamento textual, destacando-se pela análise da importância das características, o que amplia a interpretabilidade dos resultados e contribui para o suporte à tomada de decisão em segurança da informação.

Este estudo está estruturado em seis seções. A Seção 2 apresenta a fundamentação teórica, abordando os conceitos de engenharia social, aprendizado de máquina e processamento de linguagem natural. A Seção 3 discute os trabalhos relacionados. A Seção 4 descreve a metodologia adotada e o desenvolvimento do modelo. A Seção 5 apresenta os resultados experimentais e suas análises. Por fim, a Seção 6 traz as conclusões, limitações e direções para trabalhos futuros.

2. Fundamentação teórica

2.1. Engenharia Social e o Mecanismo do Phishing

O phishing é uma das principais ameaças à segurança da informação, caracterizando-se como um ataque baseado em engenharia social que explora fragilidades cognitivas dos usuários em vez de vulnerabilidades técnicas (Alshingiti et al., 2023). Esse tipo de ataque utiliza estratégias persuasivas, como senso de urgência e autoridade, para induzir a vítima a fornecer credenciais ou acessar links maliciosos. O e-mail permanece como o principal vetor de disseminação, devido à sua ampla utilização e baixo custo operacional (Alshingiti et al., 2023).

2.2. Processamento de Linguagem Natural (PLN) e Vetorização TF-IDF

O Processamento de Linguagem Natural (PLN) permite a conversão de textos em representações numéricas, sendo o TF-IDF amplamente utilizado para destacar termos relevantes com base em sua frequência e raridade no corpus (Khalid, 2025; Alshingiti et al., 2023). Essa abordagem atribui maior peso a palavras discriminativas, contribuindo para a identificação de padrões associados a mensagens de phishing. Adicionalmente, a utilização de n-gramas, especialmente bigramas, amplia a capacidade de captura de contexto, permitindo modelar relações semânticas entre palavras consecutivas (Silva et al., 2024).

2.3. Aprendizado de Máquina Supervisionado: Random Forest

O Aprendizado de Máquina Supervisionado baseia-se na construção de modelos a partir de dados rotulados, permitindo a identificação de padrões e a generalização para novos exemplos. Nesse contexto, o Random Forest destaca-se como um método de *ensemble* que combina múltiplas árvores de decisão por votação majoritária, reduzindo a variância e o risco de overfitting (Breiman, 2001; Hastie et al., 2009).

Devido à sua robustez e capacidade de lidar com dados de alta dimensionalidade, o algoritmo apresenta desempenho competitivo em tarefas de detecção de phishing, superando abordagens mais simples como o Naive Bayes em cenários textuais complexos (Rao et al., 2023; Talukder et al., 2023). Essas características justificam sua adoção como modelo principal nesta pesquisa.

2.4. Balanceamento de Dados via Undersampling

O desbalanceamento de classes é um desafio recorrente em problemas de segurança cibernética, podendo induzir o modelo a favorecer a classe majoritária e comprometer a detecção de ameaças. Para mitigar esse efeito, utilizou-se o Random Undersampling, que equilibra a distribuição das classes por meio da redução da classe majoritária. Embora implique na perda de dados, essa abordagem tende a melhorar a sensibilidade do modelo à

classe minoritária, impactando positivamente métricas como o recall (Vayre et al., 2023; Sokolova & Lapalme, 2009).

3. Trabalhos relacionados

A revisão identificou cinco estudos representativos sobre detecção de phishing com aprendizado de máquina (IEEE Xplore, SBC OpenLib, MDPI), selecionados por relevância temática e aderência metodológica. De forma geral, a literatura aponta que a detecção de phishing tem evoluído significativamente com o uso de técnicas supervisionadas, especialmente algoritmos baseados em ensemble, como o Random Forest, devido à sua capacidade de lidar com dados de alta dimensionalidade e ruído (Hastie, Tibshirani, & Friedman, 2009; Breiman, 2001). Além disso, aspectos como balanceamento de classes e avaliação robusta por validação cruzada têm sido reconhecidos como fatores determinantes para a confiabilidade dos modelos (Kohavi, 1995; Sokolova & Lapalme, 2009).

Nesse contexto, Ferreira (2023) propôs um sistema de detecção de phishing em tempo real voltado para ambientes corporativos, adotando uma abordagem comparativa entre diferentes algoritmos supervisionados, como Random Forest, Decision Tree e Support Vector Machine (SVM), além de técnicas não supervisionadas. Os resultados demonstraram acurácia de 99,19%, evidenciando a eficácia dos métodos supervisionados na identificação de padrões maliciosos. Contudo, o estudo não explora de forma aprofundada o impacto do desbalanceamento de classes em cenários com dados heterogêneos, o que pode comprometer a generalização do modelo em aplicações reais.

Complementarmente, Rao et al. (2023) investigaram a aplicação do Random Forest na detecção de URLs maliciosas, utilizando atributos lexicais e estruturais. O modelo alcançou acurácia de 97,14%, confirmando a robustez do algoritmo em ambientes com dados ruidosos. Entretanto, o foco exclusivo em URLs limita a análise semântica do conteúdo textual, aspecto relevante na identificação de ataques de phishing em e-mails, lacuna que este trabalho busca mitigar por meio do uso de técnicas de Processamento de Linguagem Natural (PLN).

Talukder et al. (2023), por sua vez, desenvolveram uma arquitetura híbrida baseada na combinação de Random Forest com técnicas de seleção de atributos, obtendo acurácia de 99,76%. Os autores destacam a importância da engenharia de atributos para o desempenho do modelo. No entanto, o estudo foi conduzido em bases de dados específicas e homogêneas, o que limita a capacidade de generalização em cenários mais complexos e diversos, como aqueles encontrados em ambientes industriais.

Em uma abordagem voltada à aplicação prática, Silva et al. (2024) implementaram um sistema de detecção de phishing em tempo real utilizando Random Forest integrado a uma ferramenta de monitoramento. O modelo apresentou acurácia de 97,81%, com ênfase na baixa latência como requisito essencial para proteção do usuário final. Apesar dos resultados promissores, o estudo não aborda de forma detalhada o impacto do balanceamento de classes em grandes volumes de dados, aspecto fundamental para garantir desempenho consistente em larga escala.

Por fim, Alshingiti et al. (2023) analisaram a aplicação de técnicas de Processamento de Linguagem Natural e Deep Learning na segurança cibernética, comparando abordagens baseadas em metadados e conteúdo textual. Os autores demonstraram que modelos capazes de capturar relações semânticas profundas apresentam maior precisão, embora com maior custo computacional. Nesse sentido, destaca-se o desafio de equilibrar desempenho e eficiência, especialmente em sistemas que demandam resposta em tempo real.

Tabela 1. Comparativo de estudos em detecção de phishing.

Autor/Ano	Modelo	Acurácia	Dataset	Limitação identificada
Ferreira (2023)	RF + DT + SVM	99,19%	Corp.	Sem balanceamento
Rao et al. (2023)	RF	97,14%	URLs	Sem análise textual
Talukder et al. (2023)	RF + FS	99,76	Homog.	Baixa generalização
Silva et al. (2024)	RF + monitoramento	97,81%	Corp.	Sem análise de balanceamento
Alshingiti et al. (2023)	PLN + DL	-	Emails	Alto custo

A revisão da literatura evidencia a diversidade de abordagens aplicadas à detecção de phishing, bem como suas limitações recorrentes. Conforme sintetizado na Tabela 1, observa-se que, embora os estudos apresentem elevados índices de acurácia, persistem lacunas relacionadas à generalização em bases heterogêneas, ao tratamento do desbalanceamento de classes e à análise semântica do conteúdo textual. Esses aspectos reforçam a necessidade de abordagens que integrem múltiplas fontes de dados e técnicas de processamento textual, motivando a proposta deste trabalho.

4. Solução computacional

Esta seção descreve a arquitetura e o funcionamento da solução proposta para a classificação automática de mensagens maliciosas, com foco na identificação de ataques de phishing utilizando técnicas de aprendizado de máquina.

4.1. Visão Geral da Solução

A solução proposta consiste em um pipeline exibido na Figura 1 de processamento de linguagem natural (PLN) integrado a um modelo supervisionado de classificação.

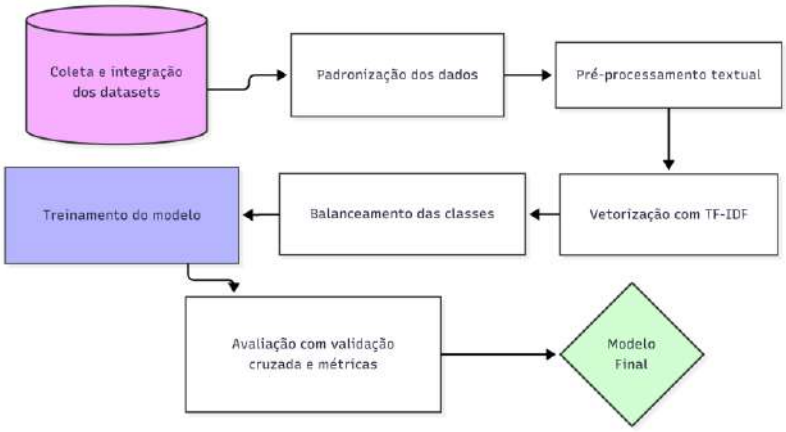


Figura 1. Pipeline proposto: da integração de dados à classificação supervisionada via vetorização numérica.

4.2. Integração e Padronização dos Dados

Foram utilizados três conjuntos de dados públicos extraídos da plataforma Kaggle: (i) Nigerian Fraud Dataset, contendo e-mails de fraude nigeriana; (ii) Nazario Phishing Dataset, composto por mensagens de phishing coletadas por José Nazario; e (iii) Phishing Email Dataset, com mensagens classificadas como legítimas ou maliciosas. A seleção baseou-se no volume de instâncias, diversidade de padrões linguísticos e avaliação positiva da comunidade científica na plataforma.

Como os conjuntos apresentavam estruturas distintas, foi realizada uma etapa de padronização, na qual todos os dados foram convertidos para um formato comum contendo: **text** (conteúdo textual da mensagem) e **label** (classe associada - malicioso ou legítimo)

Nos casos em que os dados estavam divididos em campos como *subject* e *body*, foi realizada a concatenação desses atributos, com o objetivo de enriquecer a representação textual e melhorar a capacidade de aprendizado do modelo. Após a padronização, os datasets foram integrados em um único conjunto de dados para treinamento.

4.3. Pré-processamento, Vetorização e Balanceamento

Para a etapa de vetorização, empregou-se o método TF-IDF configurado com n-gramas variando de unigramas e bigramas *(ngram_range=(1,2))*, remoção de stopwords da língua inglesa (*stop_words='english'*), frequência máxima de documento de 60% *(max_df=0.6)* e frequência mínima absoluta de 30 ocorrências *(min_df=30)*, com o objetivo de eliminar termos excessivamente comuns ou raros e capturar padrões sequenciais determinantes na identificação de phishing. Para o balanceamento, aplicou-se uma técnica de subamostragem aleatória que embora haja uma perda de dados, eleva o recall em cenários desbalanceados (Vayre et al., 2023).

4.4. Modelo de Classificação

O Random Forest foi selecionado por apresentar melhor desempenho em relação ao Naive Bayes (Tabela 2), especialmente no recall, reduzindo

falsos negativos. Além disso, sua capacidade de capturar relações complexas entre atributos o torna mais adequado para dados textuais.

Tabela 2. Comparativo de Desempenho dos Algoritmos.

Modelo	Acurácia	Precisão	Recall	F1-Score
Random Forest	0.9835	0.9854	0.9816	0.9835
Naive Bayes	0.9563	0.9849	0.9273	0.9552

O modelo Random Forest foi instanciado com 100 árvores de decisão *(n_estimators=100)* e semente aleatória fixada em 42 *(random_state=42)* para garantir a reprodutibilidade dos experimentos. Os demais hiperparâmetros foram mantidos nos valores padrão da biblioteca scikit-learn, incluindo critério de divisão Gini, número máximo de atributos por divisão igual à raiz quadrada do total de features *(max_features='sqrt')* e profundidade máxima irrestrita *(max_depth=None).*

5. Avaliação experimental

5.1. Planejamento de pesquisa

A avaliação do desempenho do classificador foi estruturada por meio de duas abordagens complementares. Inicialmente, aplicou-se a técnica de validação cruzada com cinco partições (*5-Fold Cross-Validation*), que permite mensurar a robustez estatística do modelo ao treiná-lo e testá-lo em diferentes subconjuntos dos dados integrados.

Paralelamente, adotou-se a estratégia *hold-out*, com partição de 80% dos dados para treinamento e 20% para teste, sem sobreposição entre os subconjuntos. Essa combinação de protocolos assegura a validação do modelo em dados não observados durante o treinamento, mitigando o risco de *overfitting* e permitindo uma estimativa mais realista do desempenho na produção.

5.1.1. Métricas de Avaliação

As métricas incluem acurácia, precisão, recall, F1-score, abrangendo acertos globais, confiabilidade de positivos, sensibilidade a ameaças, equilíbrio precisão-recall e descriminação por limiares (Sokolova & Lapalme, 2009; Fawcett, 2006).

5.2. Resultados e Discussão

Os resultados médios obtidos foram: acurácia de 0,9613 (± 0,0120), precisão de 0,9425 (± 0,0225), recall de 0,9832 (± 0,0015) e F1-score de 0,9622 (± 0,0112). O elevado recall é particularmente relevante, pois minimiza falsos negativos, o erro de maior impacto em segurança (Verma & Hossain, 2017). A precisão mantém-se em patamar elevado e os baixos desvios padrão entre partições confirmam robustez e capacidade de generalização (Sokolova & Lapalme, 2009; Hastie et al., 2009).

5.2.1. Análise de Desempenho - Matriz de Confusão

A matriz de confusão (Figura 2) evidencia o desempenho do modelo, com 7.810 verdadeiros positivos e 7.767 verdadeiros negativos, frente a apenas 146 falsos negativos e 115 falsos positivos. Esse resultado reforça o elevado recall observado e indica baixa incidência de erros críticos, reduzindo o risco de mensagens maliciosas serem classificadas como legítimas (Powers, 2011; Verma & Hossain, 2017).

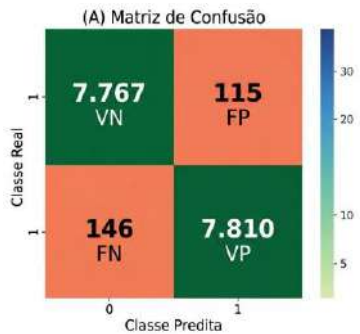


Figura 2. Avaliação visual do modelo: Matriz de Confusão.

Embora a presença de falsos positivos reflita o compromisso intrínseco entre sensibilidade e precisão (Sokolova & Lapalme, 2009), o modelo apresenta um equilíbrio consistente entre sensibilidade e precisão, adequado para aplicações em segurança cibernética.

5.2.2. Análise de Importância das Features

A análise de importância das features foi realizada para identificar os atributos mais influentes no processo de decisão do modelo, sendo calculada, no Random Forest, com base na redução da impureza nas árvores (Breiman, 2001).

Os resultados indicam maior relevância para termos alinhados a padrões típicos de phishing, especialmente relacionados à interação do usuário (“click”, “url”), manipulação de credenciais (“account”) e aspectos financeiros (“money”, “bank”, “offer”). Esses elementos refletem estratégias de engenharia social voltadas à indução de ações maliciosas (Verma & Hossain, 2017).

A predominância desses atributos evidencia que o modelo captura padrões linguísticos relevantes, reforçando a adequação do pipeline adotado e sua capacidade de distinguir comunicações legítimas de tentativas de fraude.

6. Conclusão

O presente estudo teve como objetivo desenvolver e validar um modelo robusto para a detecção de mensagens de phishing, no contexto da segurança da informação aplicada à Indústria 4.0. A partir da unificação de múltiplas bases de dados heterogêneas e da aplicação de um pipeline de Processamento de Linguagem Natural, fundamentado na vetorização TF-IDF com n-gramas e no balanceamento de classes, foi possível construir uma solução eficaz para a classificação de mensagens maliciosas.

Os resultados obtidos demonstram que o modelo baseado em Random Forest apresentou desempenho consistente, alcançando acurácia média superior a 96% em validação cruzada, além de elevados valores de

recall e F1-score. Esses resultados indicam alta capacidade do modelo em identificar corretamente mensagens de phishing, com especial destaque para a redução de falsos negativos, aspecto crítico em sistemas de segurança cibernética.

Em relação às questões de pesquisa, verificou-se que a integração de bases de dados heterogêneas não comprometeu a estabilidade do modelo, desde que acompanhada por técnicas adequadas de pré-processamento e balanceamento. Observou-se também que o balanceamento de classes contribuiu para a melhoria da sensibilidade do modelo, reduzindo o impacto do viés da classe majoritária. Por fim, a utilização de vetorização TF-IDF com n-gramas mostrou-se eficaz na captura de padrões semânticos relevantes associados a ataques de phishing, conforme evidenciado pela análise de importância das features, que destacou termos relacionados a interações com links, manipulação de credenciais e transações financeiras.

Como limitação, destaca-se a ausência de validação do modelo em ambiente de produção real, o que impede a avaliação completa de seu desempenho sob condições operacionais e fluxos contínuos de dados.

Para trabalhos futuros, recomenda-se a investigação de estratégias alternativas de balanceamento de dados, incluindo abordagens baseadas em geração de dados sintéticos, com o objetivo de preservar a distribuição original das classes e potencialmente melhorar a capacidade de generalização do modelo. Adicionalmente, sugere-se a avaliação da solução em ambientes reais, a fim de validar sua aplicabilidade prática e robustez em cenários industriais.

De forma geral, os resultados indicam que a solução proposta é viável para aplicação em sistemas de filtragem de e-mails corporativos, contribuindo para a mitigação de ataques de phishing por meio de uma abordagem automatizada, interpretável e alinhada às demandas da Indústria 4.0.

7. Referências

Alshingiti, Z., Alhumiani, H., & Al-Zahrani, A. (2023). A comprehensive techniques for phishing detection based on email content. *Electronics*, *12*(3), 502. https://doi.org/10.3390/electronics12030502

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Ferreira, A. (2023). Detecção de ataques de phishing em tempo real utilizando algoritmos de aprendizado de máquina. In *Anais do XXIII Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais* (pp. 141–148). SBC. https://doi.org/10.5753/sbseg_estendido.2023.234608

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

IBM. (2024). *Cost of a data breach report 2024*. IBM Security. Recuperado de: https://www.ibm.com/reports/data-breach

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (pp. 1137–1143)*. Morgan Kaufmann.

Olianin, D., Tsupryk, H., Hovorushchenko, T., Bahrii-Zaiats, O., & Andrushchak, I. (2025). Transformer neural networks in Industry 4.0. *CEUR Workshop Proceedings*, *4057*. Recuperado de: https://ceur-ws.org/Vol-4057/short9.pdf

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Rao, B. S., Priya, P., Sridhar, M., & Kumar, T. V. (2023). A machine learning approach for phishing detection using Random Forest. In *Proceedings of the 2023 International Conference on Computing and Communication (ICCC)*. IEEE. https://doi.org/10.1109/ICCC57789.2023.10165432

Silva, V. S., Carvalho, R. A., Mendes, F. L., & Lima, G. P. (2024). Identificação de ataques de phishing através de machine learning. In *Anais do XXIV Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*. SBC. Recuperado de: https://sol.sbc.org.br/index.php/sbseg_estendido/article/view/30158

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Talukder, M. A., Islam, M. M., Uddin, M. A., & Akhter, A. (2023). A hybrid machine learning approach for phishing website detection. *IEEE Access*, *11*, 12345–12360. https://doi.org/10.1109/ACCESS.2023.3245678

Vayre, B., Verdier, C., Minier, M., & Levillain, O. (2023). Evaluation of machine learning models for phishing detection over time. In *Proceedings of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE. https://doi.org/10.1109/EuroSPW59974.2023.00033

Verma, R., & Hossain, N. (2017). Semantic feature selection for text with application to phishing email detection. In *Proceedings of the 13th International Conference on Information Security Practice and Experience (ISPEC) (pp. 455–468)*. Springer. https://doi.org/10.1007/978-3-319-72359-4_27

Zhu, Y., Zhang, W., Chen, Y., & Gao, H. (2023). PhishGraph: A graph-based phishing detection system. *IEEE Transactions on Dependable and Secure Computing*. https://doi.org/10.1109/TDSC.2023.3267890

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Khalid, N. M. S. (2025). NLP techniques for cybersecurity: A 2025 perspective. *International Journal of Computer Science and Security*, *19*(1).

Capítulo 05

Resiliência e observabilidade na Indústria 4.0: Diagnóstico out-of-band com LORAWAN

Jussara Miliano Soares
Josiane do Couto Rodrigues
Universidade Federal de Roraima

jmilianosoares@gmail.com
josiane.rodrigues@ufrr.com

1. Introdução

Na Indústria 4.0, os Sistemas Ciberfísicos (CPS) desempenham um papel crucial ao permitirem o monitoramento contínuo de ativos em ambientes industriais distribuídos. Devido à complexidade dessa arquitetura, a supervisão remota em tempo real torna-se essencial para garantir a eficiência e a confiabilidade operacional (Folgado *et al*., 2024). Por outro lado, em arquiteturas distribuídas é inevitável que ocorram interrupções de comunicação ou de energia, expondo os CPS a falhas que podem desencadear efeitos em cascata sobre a disponibilidade e confiabilidade dos processos produtivos (Piardi *et al*., 2020). Nessa condição de falha, o diagnóstico tradicional *in-band*, que compartilha o canal principal de comunicação com os dados operacionais, torna-se ínvio exatamente quando é mais necessário, resultando em perda de visibilidade e comprometendo a tomada de decisões em tempo hábil (Gu *et al.*, 2019; Wang *et al*., 2024).

Embora topologias redundantes sejam indicadas para a resiliência de sistemas críticos, a literatura atual em CPS concentra-se predominantemente em defesas cibernéticas e modelos matemáticos de detecção. O isolamento e a separação física das vias de comunicação ainda representam desafios complexos e custosos em arquiteturas distribuídas (Alain *et al.*, 2025). Diante da necessidade de ampla cobertura e confiabilidade na Indústria 4.0, tecnologias LPWAN (*Low Power Wide Area Network*) como o LoRaWAN mostram-se bastante adequadas. Apesar disso, seu potencial como canal *out-of-band* (OOB) fisicamente isolado tem sido menos explorado do que outras

frentes de pesquisa, tornando-se uma área de estudo fundamental para garantir o monitoramento ininterrupto do sistema quando ocorre a perda do canal principal.

Este trabalho tem como objetivo desenvolver e avaliar, por meio de simulações ns-3 (*Network Simulator 3*), uma arquitetura de diagnóstico *out-of-band* baseada em LoRaWAN para Sistemas Ciberfísicos Industriais, capaz de garantir a continuidade da observabilidade e a transmissão confiável de informações críticas durante falhas da comunicação primária, visando aumentar a resiliência, a confiabilidade operacional e a capacidade de diagnóstico em ambientes de Indústria 4.0.

A solução utiliza um canal de comunicação alternativo, cuja separação ocorre no nível de acesso dos dispositivos, para transmitir dados essenciais de diagnóstico rápido, como alarmes e estados resumidos, durante indisponibilidades primárias. As perguntas da pesquisa são: o LoRaWAN é aplicável como canal OOB para CPS? quais suas limitações sob falha da rede principal de comunicação? e como a arquitetura difere do diagnóstico *in-band* tradicional?

O restante do trabalho organiza-se em: fundamentação teórica a respeito de diagnóstico *in-band* e redes LPWAN (Seção 2); trabalhos relacionados (Seção 3); descrição da arquitetura proposta (Seção 4); avaliação de resultados (Seção 5); e, por fim, conclusões e trabalhos futuros (Seção 6).

2. Fundamentação teórica

2.1. Limitações do Diagnóstico em rede *In-Band* e a Abordagem *Out-of-Band*

O uso do modelo *in-band* é motivado pela eficiência de custos ao utilizar a mesma infraestrutura de dados operacionais para o transporte de diagnósticos. Entretanto, esse modelo gera um alto acoplamento, onde o sucesso da supervisão depende inteiramente da integridade da rede principal. Caso ocorram congestionamentos ou falhas no plano de controle, cria-se um efeito cascata que oculta a causa raiz do problema e dificulta a recuperação do sistema físico (Zhou *et al.*, 2020; Kurtz *et al.*, 2018). Como

detalhado no Tabela 1, as soluções *in-band* tentam mitigar esses riscos através de mecanismos de priorização e rastreio dentro do próprio fluxo de dados, mas permanecem limitadas pela capacidade da via única, apresentando latências críticas sob alta carga (Zhang *et al*., 2024).

Como solução, a abordagem *out-of-band* (OOB) promove a separação física ao mover o tráfego de gerenciamento para um canal secundário e autônomo. Esse princípio garante a observabilidade mesmo em cenários de colapso total da rede operacional, permitindo que informações vitais de *hardware* sejam enviadas de forma independente (Gu *et al*., 2019). Enquanto o canal *in-band* depende da integridade da via principal, o OOB garante a observabilidade por diversidade de caminho, isolando o tráfego de gerenciamento das falhas operacionais.

Tabela 1. Comparativo de Abordagens de Diagnóstico *in-band* e *out-of-band*.

Abordagem	Funcionamento Técnico	Limitação	Autores
In-band	Insere metadados de *status* nos próprios pacotes de dados através de *in-band network telemetry.*	Depende da integridade da via principal.	Xiong *et al.* (2024)
In-band	Organiza filas de tráfego para priorizar o diagnóstico.	Não elimina congestionamentos sob alta carga.	Zhang *et al.* (2024)
OOB	Duplica o envio de dados críticos por duas redes heterogêneas simultâneas.	Saturação do espectro e redundância ininterrupta.	Sisinni et al. (2024)
OOB	Utiliza rádio de longo alcance (LoRaWAN) como via de *backup* fisicamente isolada.	Alta latência atrelada a restrições de *payload*.	Ferrari et al. (2019)

2.2. Redes LPWAN e LoRaWAN como suporte à supervisão resiliente

Para garantir a resiliência na supervisão de ambientes industriais, prioriza-se a entrega confiável de dados vitais em função de fluxos de alta banda (Mahmood *et al*., 2021). As redes LPWAN atendem a essa premissa operando como canais diagnósticos paralelos de contingência. A literatura descreve o LoRaWAN estruturado em topologia estrela estendida (*star-of-stars*): os nós de campo comunicam-se via rádio com os *gateways*; estes repassam os pacotes por meio de um *backhaul IP* seguro até o *Network Server* e o *Application Server*, de onde os dados são integrados ao sistema SCADA primário (Ferrari *et al*., 2019). Sobre essa arquitetura, operam decisões nas camadas de enlace e de aplicação.

Um exemplo de algoritmo operacional na camada de enlace é o agendamento determinístico de recursos (como abordagens baseadas em TDMA), que aloca janelas de tempo específicas e fixa o *Spreading Factor* (SF) para as transmissões. As pesquisas de Abdelfadeel *et al.* (2020) comprovam que, substituir o acesso aleatório e algoritmos puramente dinâmicos por esse controle determinístico, elimina a imprevisibilidade do canal e mitiga colisões, garantindo maior escalabilidade e confiabilidade sob condições severas. Complementarmente, destaca-se o uso de algoritmos de adaptação de tráfego orientados a eventos na camada de aplicação. Nesses modelos teóricos, aplicam-se máquinas de estado que operam em regime de *heartbeat* (envios espaçados de prontidão) e que transicionam autonomamente para modos de transmissão de diagnóstico intensivo ao detectarem anomalias ou falhas na rede primária. Essa abordagem garante respostas rápidas de recuperação sem comprometer a vida útil da bateria durante a normalidade do sistema (Gu *et al.*, 2019).

3. Trabalhos relacionados

A partir de consultas às bases *IEEE Xplore*, *ACM Digital Library*, MDPI e *Google Scholar*, foram selecionados cinco estudos mais aderentes ao problema investigado, por abordarem falhas de comunicação, resiliência e o uso de LPWAN em sistemas ciberfísicos e aplicações industriais, a saber: Gu *et al*. (2019), Ferrari *et al*. (2019), Odongo *et al*. (2022), Sisinni *et al*. (2024) e Wang *et al*. (2024).

No contexto de mitigar falhas, Gu *et al.* (2019) conduziram um estudo cujo objetivo foi propor um plano de controle *out-of-band* (OOB) para redes de sensores. Para atingir essa meta, a metodologia baseou-se na simulação de uma rede paralela para separar fluxos físicos. Os resultados demonstraram que a rede suportou ruídos de até -85 dBm com convergência rápida. A partir desses dados, os autores concluíram que vias independentes aumentam a resiliência. Contudo, tem como limitação o foco restrito ao roteamento, sem envio de telemetria diagnóstica, uma restrição que esta pesquisa busca contornar ao usar o OOB especificamente para o diagnóstico.

No contexto das redes LPWAN, Ferrari *et al.* (2019) avaliaram o LoRaWAN como *backup* veicular usando um protótipo com interface *On-*

Board Diagnostic (OBD-II). A experimentação evidenciou que *payloads* de 160 bytes geraram alto tempo de ocupação do canal (*Time-on-Air* de até 462 ms), esbarrando na alta latência para contingências. Para superar essa limitação, este trabalho procura otimizar o *payload* para apenas 12 bytes, reduzindo o atraso.

Odongo *et al.* (2022) avaliaram o monitoramento de falhas via LoRaWAN em ambiente real. Constatou-se que a taxa de recepção (PRR) caiu a 0% em áreas obstruídas, evidenciando alta sensibilidade a obstáculos físicos. Porém, essa arquitetura adota o protocolo como via exclusiva. O presente trabalho difere desse modelo por utilizar o LoRaWAN estritamente como rede OOB, na qual uma máquina de estados aciona o envio alternativo apenas sob falha confirmada da rede principal.

Buscando contornar as perdas de pacotes, Sisinni *et al*. (2024) analisaram o uso conjunto de LoRaWAN e *Bluetooth Mesh*. O método adotado consistiu em modelar a redundância de transmissão dupla contínua. Verificou-se uma redução das falhas para menos de 1%, o que permitiu concluir que tecnologias heterogêneas simultâneas maximizam a confiabilidade. Entretanto, a abordagem sofre com a saturação e o congestionamento da rede causados pela redundância ativa ininterrupta. Este estudo otimiza esse aspecto operando o LoRaWAN apenas em modo *heartbeat* até o momento da interrupção.

Por fim, Wang *et al.* (2024) desenvolveram um estudo visando reduzir falhas de comunicação em aplicações IIoT (Internet Industrial das Coisas). A simulação baseou-se em distúrbios físicos que foram contornados por meio do uso de *gateways* móveis e nós auxiliares. As métricas apontaram uma perda de prazos de entrega de 0% e economia energética de 27%. Com isso, os autores afirmaram que a mobilidade mitiga interrupções locais. Esse estudo, contudo, deixa como limitação a tentativa de utilizar elementos da própria infraestrutura afetada para a sua correção. A arquitetura desenvolvida difere dessa abordagem por assumir o canal principal como inoperante, garantindo a observabilidade de fora da zona de falha, que evidencia a relevância de adotar uma separação física contínua, premissa central proposta por este estudo.

4. Solução computacional

A arquitetura proposta separa o diagnóstico do fluxo operacional em dois canais físicos independentes: o principal *in-band* (Ethernet/CSMA) e o secundário *out-of-band* baseado em LoRaWAN. Conforme a Figura 1, os *End Devices* (ED) operam em ambos os planos. Em operação normal, o canal OOB transmite apenas *heartbeats* a cada 900s. O monitoramento do canal *in-band* ocorre por mensagens UDP Echo a cada 2s.

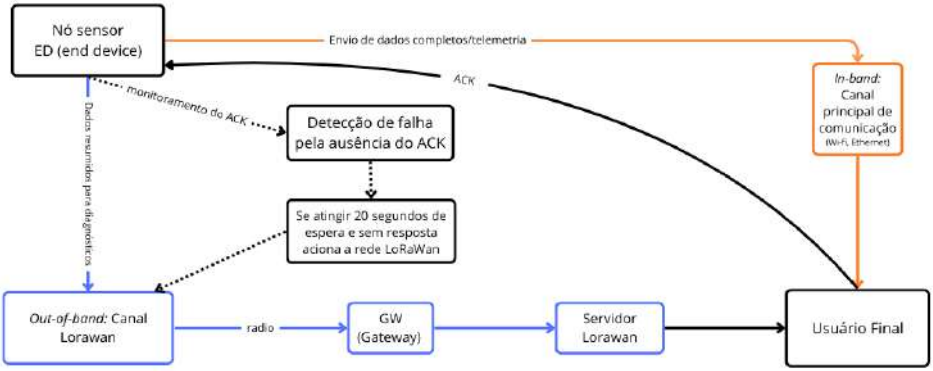


Figura 1. Fluxo da arquitetura proposta com canal principal *in-band* e canal alternativo *out-of-band* baseado em LoRaWAN.



Figura 2. Máquina de estados para acionamento do diagnóstico *out-of-band* sob detecção de falha.

A Figura 2 ilustra a máquina de estados operacional do sistema: ao atingir um limite de 10 perdas consecutivas (~20s), confirma-se a indisponibilidade da via principal e o canal LoRaWAN transita para o modo de falha, enviando pacotes críticos de diagnóstico de 12 bytes a cada 120s, reduzindo acionamentos indevidos por oscilações.

Para a validação, a topologia da rede, a lógica dessa máquina de estados e a definição de todos os parâmetros operacionais (como variação de *Spreading Factor*, densidade de nós, modelos de propagação *Log Distance, Nakagami,* inserção de falhas via *RateErrorModel*) foram desenvolvidas em C++ no *Network Simulator 3* (ns-3 v3.46[1]) com uso do módulo[2] LoRaWAN. As rodadas sucessivas de simulação geraram os resultados distribuídos em arquivos CSV contendo as métricas brutas da rede. Para otimizar essa etapa, utilizou-se um *script* auxiliar atuando exclusivamente como ferramenta de automação para a execução em lote e extração dos dados[3].

5. Avaliação experimental

5.1. Planejamento de pesquisa

Os experimentos nesta etapa buscaram responder às perguntas de pesquisa referentes à aplicabilidade do LoRaWAN como canal OOB, avaliando suas limitações em cenários de falha e sua eficácia em relação ao diagnóstico *in-band*. Trata-se de uma pesquisa aplicada, quantitativa e experimental, conduzida por meio de simulação computacional. Como configuração de referência (*baseline*), adotou-se um cenário com 20 *End Devices* (ED), avaliados em dois ambientes (*indoor* e *outdoor*) e sob dois fatores de espalhamento (SF7 e SF12).

Para garantir resultados estatisticamente confiáveis e reprodutíveis, cada configuração foi testada trinta vezes. Nas repetições, variou-se apenas o identificador da execução (*run*), mantendo fixa a semente inicial (*seed*) do simulador. Essa abordagem garante que o sistema enfrente as mesmas

[1] Documentação oficial do simulador ns-3: < https://www.nsnam.org/documentation >.
[2] Foi utilizado o módulo desenvolvido por Magrin, disponível em: < https://github.com/signetlabdei/lorawan >.
[3] Código-fonte em C++, o *script* e *datasets* com os resultados da simulação encontram-se publicamente disponíveis no repositório GitHub: < https://github.com/jumiso/ecai_arquitetura_oob_lorawan_cps >.

variações físicas do sinal de rádio em todos os testes, permitindo uma comparação justa. Aplicaram-se quatro variações de cenário à *baseline* para avaliar o comportamento da arquitetura.

As métricas de rede avaliadas incluíram a latência, o tempo de detecção no canal *in-band* e o *Packet Delivery Ratio* (PDR) do canal OOB. A adoção do PDR atende à literatura de simulação LoRaWAN (Campos et al., 2024) como o fator crítico para atestar a confiabilidade da rede. Após a extração, esses dados passaram por tratamento estatístico, sendo expressos em média ± desvio padrão e intervalo de confiança (IC). Para avaliar o impacto no processo industrial de forma abrangente, os resultados foram submetidos ao *Framework* de Cao et al. (2025), no qual os cenários simulados foram mapeados para os seguintes estados de operação:

→ **Cenário C0 (Operação Normal):** Corresponde ao Estado Disponível.

→ **Cenário C1 (Falha *in-band* isolada):** O canal OOB entra em modo de envio intensivo, mapeando-se para a Falha Detectável. O diagnóstico é garantido, mantendo a capacidade de recuperação física do sistema ativa na modelagem ($R > 0$).

→ **Cenário C2 (Falha do enlace do *Gateway*):** O PDR é zerado, mapeando-se para a Falha Silenciosa Indetectável. Sem observabilidade, o parâmetro de recuperação torna-se nulo ($R = 0$).

Neste estudo, o PDR e o tempo de detecção atuam especificamente como os gatilhos que definem a intensidade adversa (A) e a capacidade de recuperação (R) no referido *Framework*, convertendo os resultados do simulador nas curvas de resiliência apresentadas a seguir.

5.2. Resultados e Discussão

No cenário C0, o canal secundário operou com alta confiabilidade na entrega dos *heartbeats*. O resultado de maior relevância ocorreu no cenário C1: durante a falha do canal *in-band*, a via LoRaWAN manteve a entrega das mensagens críticas de diagnóstico com um PDR médio na janela de falha de 98,8% (± 1,5%) para o SF7 e de 93,5% (± 3,3%) para o SF12 em ambiente *indoor*, conforme apresentado na Figura 3 (A e B).

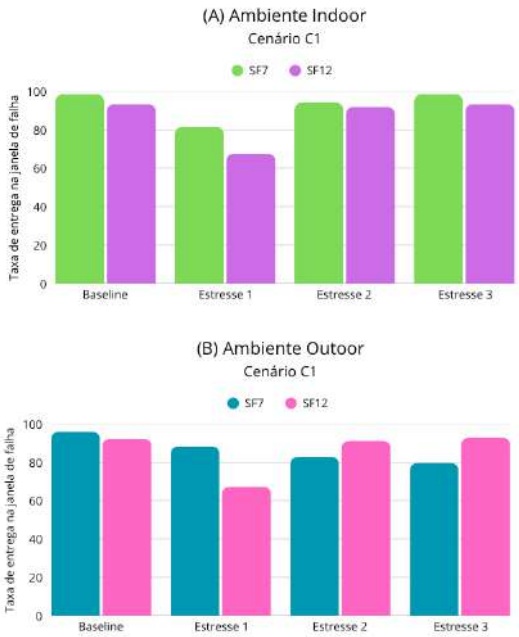


Figura 3. Percentual de entrega dos pacotes pelo OOB durante a falha do *in-band.*

O sucesso do cenário C1 confirma a mudança do sistema para o estado de *Falha Detectável*. Durante os 21s de detecção (mediana extraída das simulações), a ausência de resposta resulta em degradação crítica ($R = 0$). Contudo, o acionamento do tráfego OOB restabelece a observabilidade, reativando o parâmetro de recuperação ($R > 0$). Como demonstrado na Figura 4, esse mecanismo interrompe a queda de desempenho, mostrando que a manutenção da visibilidade do sistema é um fator determinante para viabilizar o diagnóstico contínuo e evitar a crise das operações.

Diferente do cenário C1, a taxa de entrega no C2 foi de 0%. No modelo matemático do *Framework* de Cao *et al.* (2025), essa falta de dados significa uma Falha Silenciosa Indetectável, na qual a capacidade de recuperação é nula ($R = 0$). Como mostra a Figura 4, isso leva o sistema físico ao colapso. Esse resultado prova que o problema não é a tecnologia LoRaWAN, mas o uso de infraestrutura compartilhada, que impede a observabilidade do sistema. Assim, ao contrário de Wang *et al.* (2024), que focam apenas em reforçar a

rede principal, este trabalho confirma que a resiliência ciberfísica depende da separação física total das redes de comunicação.

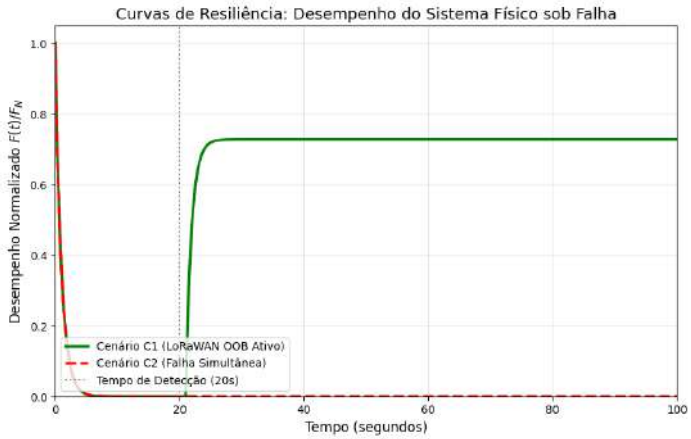


Figura 4. Curvas de resiliência nos cenários de falha de comunicação (C1 e C2)**.**

Por fim, ressalta-se que os resultados obtidos possuem limitações metodológicas que restringem sua generalização imediata para plantas industriais heterogêneas. A avaliação concentrou-se no uso de um único *Gateway* em ambiente simulado e no tráfego fixo de mensagens de 12 bytes. Além disso, a topologia operou sob configurações estáticas de rede, sem mecanismos adaptativos em tempo de execução. Ademais, a ausência de análise energética inviabiliza estimativas de autonomia operacional em cenários práticos.

6. Conclusão

Este estudo validou, via simulações no ns-3, uma arquitetura *out-of-band* baseada em LoRaWAN para o diagnóstico de falhas em Sistemas Ciberfísicos Industriais. Diferente da abordagem *in-band* tradicional, a solução utiliza um canal fisicamente isolado no nível dos dispositivos para o envio de mensagens críticas de 12 bytes durante indisponibilidades da rede primária. Sua eficácia foi comprovada ao alcançar uma Taxa de Entrega (PDR) superior a 90% após um ágil tempo de detecção de 21s, fatores que

restabelecem a observabilidade e ativam a capacidade de recuperação física do maquinário. Contudo, a pesquisa evidenciou limitações estruturais: falhas em infraestruturas compartilhadas (enlace *Gateway-Network Server*) geram PDR nulo no cenário C2, o que compromete irreversivelmente a observabilidade e leva o sistema físico ao colapso. Notou-se, ainda, restrições de escalabilidade sob fator SF12 devido ao elevado *Time-on-Air*. Como trabalhos futuros, recomenda-se a validação com múltiplos *gateways*, modelagem de consumo energético com Adaptação Dinâmica de Taxa (ADR) e análises operacionais de cibersegurança.

7. Referências

Abdelfadeel, K. Q., Zorbas, D., Cionca, V., & Pesch, D. (2020). FREE Fine-Grained Scheduling for Reliable and Energy-Efficient Data Collection in LoRaWAN, in *IEEE Internet of Things Journal*, *vol. 7*, no. 1, pp. 669-683. https://doi.org/10.1109/jiot.2019.2949918

Alain, G.C., Escamilla-Ambrosio, P.J., & Abdullah, M. (2025). A Review of Resilience Enhancement Techniques for Cyber-Physical Systems. *IEEE Access*, 13, 138061-138082. https://doi.org/10.1109/access.2025.3593257

Cao, Z., Zhao, H., Wang, Y., He, C., Zhou, D., & Han, X. (2025). A Resilience Quantitative Assessment Framework for Cyber–Physical Systems: Mathematical Modeling and Simulation. *Applied Sciences*, *15*(15), 8285–8285. https://doi.org/10.3390/app15158285

De Campos, M. G., De Almeida, L. G., Matheus, L. E. M., & Borin, J. F. (2024). On the Simulation of LoRaWAN Networks: A Focus on Reproducible Parameter Configuration. *Computer Networks and Commun, 2*(1), 164-189. https://doi.org/10.37256/cnc.2120244496

Ferrari, P., Sisinni, E., Fernandes Carvalho, D., Signoretti, G., Silva, M., Silva, I., & Silva, D. (2019). Using LoRaWAN smart city infrastructure as backup network for Industry 4.0 enabled vehicles. In *Anais Estendidos do IX Simpósio Brasileiro de Engenharia de Sistemas Computacionais*, (pp. 119-124). Porto Alegre: SBC. https://doi.org/10.5753/sbesc_estendido.2019.8646

Folgado, F. J., Calderón, D., González, I., & Calderón, A. J. (2024). Review of Industry 4.0 from the Perspective of Automation and Supervision Systems: Definitions, Architectures and Recent Trends. *Electronics*, *13*(4), 782. https://doi.org/10.3390/electronics13040782

Fragkopoulos, M., Panagiotakis, S., Kostakis, M., Markakis, E. K., Astyrakakis, N., & Malamos, A. (2023). Experimental Assessment of Common Crucial Factors That Affect LoRaWAN Performance on Suburban and Rural Area Deployments. Sensors, 23(3), 1316. https://doi.org/10.3390/s23031316

Gu, C., Tan, R., & Lou, X. (2019). One-Hop Out-of-Band Control Planes for Multi-Hop Wireless Sensor Networks. *ACM Transactions on Sensor Networks*, *15*(4), 1-29. https://doi.org/10.1145/3342100

Kurtz, F., Overbeck, D., Bektas, C., & Wietfeld, C. (2018). Control Plane Fault Tolerance for Resilient Software-Defined Networking based Critical Infrastructure Communications. *2018 IEEE 5G World Forum (5GWF)*, 10-13. https://doi.org/10.1109/5gwf.2018.8516975

Mahmood, N. H., Marchenko, N., Gidlund, M., & Popovski, P. (2020). *Wireless Networks and Industrial IoT*. Cham: Springer Cham. https://doi.org/10.1007/978-3-030-51473-0

Odongo, G., Musabe, R., Hanyurwimfura, D., & Bakari, A. (2022). An efficient LoRa-Enabled Smart Fault detection And Monitoring Platform For The Power Distribution System Using Pelf-Powered IoT Devices. *IEEE Access*, *10*, 73403-73420. https://doi.org/10.1109/access.2022.3189002

Oks, S. J. (2024). *Industrial Cyber-Physical Systems*. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-44417-4

Piardi, L., Leitão, P., & de Oliveira, A.S. (2020). Fault-Tolerance in Cyber-Physical Systems: Literature Review and Challenges In *IEEE 18th International Conference on Industrial Informatics (INDIN),* (pp.29-34.), Warwick, United Kingdom. https://doi.org/10.1109/indin45582.2020.9442209

Santos Filho, F. H. C. d., Dester, P. S., Stancanelli, E. M. G., Cardieri, P., Nardelli, P. H. J., Carrillo, D., & Alves, H. (2020). Performance of LoRaWAN for Handling Telemetry and Alarm Messages in Industrial Applications. *Sensors*, *20*(11), 3061. https://doi.org/10.3390/s20113061

Sisinni, E., & Mahmood, A. (2021) Wireless Communications for Industrial Internet of Things: The LPWAN Solutions. In Mahmood, N.H., Marchenko, N., Gidlund, M., Popovski, P. (eds) *Wireless Networks and Industrial IoT* (pp.79-103). Springer, Cham. https://doi.org/10.1007/978-3-030-51473-0_5

Sisinni, E., Depari, A., Flammini, A., Rinaldi, S., & Ferrari, P. (2024). Evaluating the joint use of LoRaWAN and Bluetooth mesh to improve survivability for critical sensor applications. *IEEE Sensors Journal*, *24*(14), 22992-23003. https://doi.org/10.1109/jsen.2024.3403308

Wang, C., Chuang, C.H., Chen, Y. W., & Chen, Y. F. (2024). On Cyber-Physical Fault Resilience in Data Communication: A Case From A LoRaWAN Network Systems Design. *ACM Transactions on Cyber-Physical Systems*, *8*(3), 1-25. https://doi.org/10.1145/3639571

Xiong, X., Xie, Y., Chen, X., Zheng, S., Huang, W., & Feng, J. (2024). An Integrated Solution for High-efficiency In-band Network Telemetry. In *Proceedings of the 8th Asia-Pacific Workshop on Networking (APNet 24)*. (pp.115–121) Association for Computing Machinery, New York, USA. https://doi.org/10.1145/3663408.3663425

Zhang, T., Wang, G., Xue, C., Wang, J., Nixon, M., & Han, S. (2024). Time-Sensitive Networking (TSN) for Industrial Automation: Current Advances and Future Directions. *ACM Computing Surveys*. https://doi.org/10.1145/3695248

Zhou, X., Yang, Z., Ni, M., Lin, H., Li, M., & Tang, Y. (2020). Analysis of the Impact of Combined Information-Physical-Failure on Distribution Network CPS. *IEEE Access*, 8, 44140-44152. https://doi.org/10.1109/access.2020.2978113

Capítulo 06

Análise de protocolos IoT para o trabalhador conectado em cenários de conectividade restrita

Bruno Cesar da Silva Claudino
Leandro Nelinho Balico
Universidade Federal de Roraima

brunocclaudino@gmail.com
leandro.balico@ufrr.com

1. Introdução

A segurança ocupacional tem sido impulsionada pela Indústria 4.0, uma área que integra sistemas ciberfísicos para criar ambientes de trabalho inteligentes. Segundo Schwab (2016), a fusão de tecnologias físicas e digitais gera novas oportunidades de monitoramento em tempo real. Nesse contexto, o uso de tecnologias vestíveis ou *wearables* e sensores inteligentes consolida o conceito de *Safety 4.0*, capaz de antecipar acidentes e reduzir danos de forma preditiva, superando os métodos tradicionais reativos (Yurtçu, 2025).

A relevância desse tema se evidencia pelo impacto econômico e social que esses acidentes ocasionam. Tais acidentes representam perdas estimadas em 4% do Produto Interno Bruto (PIB) mundial anualmente, demandando novas abordagens ergonômicas e de segurança (Naranjo et al., 2025). O problema torna-se crítico para trabalhadores isolados ou *lone workers*, cuja ausência de supervisão direta exige sistemas autônomos capazes de acionar socorro imediatamente (Battistoni et al., 2021). Conforme Hadi et al. (2025), embora os dispositivos vestíveis ofereçam grande potencial de segurança, a adoção efetiva e a confiança dos usuários dependem da confiabilidade dos dados. Falhas na captação precisa do evento ou na transmissão do alerta podem resultar em falsos positivos ou na perda de chamados críticos, o que compromete a credibilidade do sistema frente ao trabalhador.

Apesar desse potencial tecnológico, há uma lacuna no desenvolvimento de soluções acessíveis para Pequenas e Médias Empresas

(PMEs). Enquanto grandes indústrias adotam sistemas proprietários, PMEs em regiões periféricas enfrentam limitações financeiras para incorporar tecnologias de segurança conectadas (Battistoni et al., 2021). Além do custo, as infraestruturas de telecomunicações na região amazônica enfrentam desafios singulares. Nesse contexto, a elevada densidade da vegetação, associada à alta umidade local, atua como uma barreira que interfere nos sinais de radiofrequência, o que resulta em grandes distâncias, áreas com instabilidade de sinal e baixa qualidade de conexão (Turo et al., 2024).

Com a instabilidade física do sinal, o gerenciamento de dados na camada de aplicação torna-se decisivo para a confiabilidade do sistema. O protocolo *Message Queuing Telemetry Transport* (MQTT), baseado em *Transmission Control Protocol* (TCP), garante a entrega da mensagem, característica vital para alertas críticos. Em contrapartida, o *Constrained Application Protocol* (CoAP), sobre *User Datagram Protocol* (UDP), é mais leve e otimizado para recursos restritos, porém sua falta de garantia nativa de entrega pode exigir implementações adicionais para evitar a perda de alertas (Larmo et al., 2019). Embora estudos recentes tenham analisado o desempenho desses protocolos em *hardware* de baixo custo, tais avaliações ocorrem em ambientes controlados (Vieira et al., 2024; Seoane et al., 2021). Assim, persiste a lacuna empírica sobre como o balanço entre a robustez do MQTT e a leveza do CoAP impactam dispositivos *wearables* quando submetidos a cenários que reproduzem as condições de conectividade típicas da região amazônica.

Diante do exposto, o objetivo deste trabalho foi desenvolver e validar uma arquitetura para monitoramento de segurança utilizando *hardware* de baixo custo. O estudo visou comparar os protocolos MQTT e CoAP quanto ao consumo energético, à confiabilidade e à previsibilidade na transmissão de alertas críticos em redes instáveis utilizando o ESP8266. Desse modo, as principais contribuições desta pesquisa constituíram-se no desenvolvimento e na validação experimental de uma arquitetura de *hardware* aberta e replicável voltada à segurança do trabalho, bem como na análise comparativa do desempenho e da estabilidade operacional de protocolos de comunicação em sistemas IoT operando sob condições de conectividade restritas similares aos ambientes amazônicos.

O restante deste estudo está organizado da seguinte forma: a seção 2 apresenta a fundamentação teórica, abordando os conceitos de Indústria 4.0, IoT e dos protocolos MQTT e CoAP. Já a seção 3 descreve os trabalhos relacionados, situando a pesquisa na literatura atual, e a solução computacional é apresentada na seção 4. Finalmente, a avaliação experimental, os resultados obtidos e a discussão são apresentados nas seções subsequentes.

2. Fundamentação teórica

Esta seção apresenta os conceitos que embasam a pesquisa e permitem compreender o problema e a solução proposta. Inicialmente, discute-se Indústria 4.0 e o papel da IoT na digitalização da segurança ocupacional. Em seguida, são detalhados os protocolos de comunicação, o MQTT e o CoAP, fundamentais para analisar os desafios da transmissão contínua de dados de dispositivos vestíveis em cenários com restrição de conectividade.

2.1. Indústria 4.0 e Internet das Coisas

A Indústria 4.0 é definida pela automação e pela mudança na organização da produção, que acontece por meio da descentralização e da interoperabilidade. Segundo Kagermann et al. (2013), o conceito fundamenta-se na fusão dos mundos físico e virtual por meio de sistemas ciberfísicos, permitindo que máquinas e sistemas se comuniquem de forma inteligente. Nesse novo ecossistema, o trabalhador deixa de ser um agente isolado e passa a interagir com o ambiente digital de forma contínua, viabilizando abordagens preditivas de segurança ocupacional baseadas em dados de tempo real.

A IoT constitui a infraestrutura técnica da Indústria 4.0, fornecendo a camada sensorial para coleta e transmissão de dados por meio de objetos físicos equipados com sensores e conectividade (Ashton, 2009). No monitoramento ocupacional, ela se materializa em dispositivos vestíveis que estendem a supervisão para locais remotos e permitem o acompanhamento autônomo da integridade de trabalhadores isolados.

2.2. Protocolos MQTT e CoAP

A comunicação em cenários de IoT exige soluções otimizadas, pois protocolos *web* tradicionais geram um *overhead* excessivo para dispositivos microcontrolados (Vieira et al., 2024). Para viabilizar a transmissão de dados em redes com largura de banda limitada, a arquitetura IoT apoia-se em protocolos de camada de aplicação desenvolvidos especificamente para recursos restritos, com destaque para o MQTT e o CoAP.

O MQTT utiliza um modelo de publicação e assinatura (OASIS, 2014). Dessa forma, os dispositivos na borda da rede não se comunicam diretamente entre si, mas publicam dados em um servidor central, que os distribui aos clientes assinantes (Seoane et al., 2021). Esse protocolo funciona sobre a camada de transporte TCP, que estabelece uma conexão contínua e confiável antes de enviar os dados.

Graças a essa arquitetura orientada à conexão, o protocolo oferece níveis de Qualidade de Serviço (QoS) que garantem a entrega da mensagem. Em contrapartida, o CoAP foi projetado com uma arquitetura de requisição e resposta análoga à *web* convencional, porém nativamente otimizado para redes de baixa potência e com perda introduzida, operando sobre UDP (Shelby et al., 2014). Por não exigir o estabelecimento prévio de uma conexão complexa, esse protocolo é considerado mais leve.

A compreensão da arquitetura de ambos os protocolos evidencia a sua importância para a resolução do problema desta pesquisa. Em áreas remotas como a região amazônica, redes instáveis impõem um desafio direto à comunicação. Enquanto o MQTT oferece a confiabilidade necessária para que um alerta de socorro não seja perdido, sua dependência do TCP pode resultar em alto consumo de bateria e falhas de conexão em áreas de sinal fraco. Já o CoAP preserva a energia do dispositivo e a banda da rede, mas carece de garantias nativas de entrega. Portanto, a comparação empírica entre esses protocolos é o que fundamenta a avaliação proposta neste trabalho, buscando identificar a melhor abordagem para o trabalhador conectado em cenários de conectividade restrita.

3. Trabalhos relacionados

A literatura recente tem explorado amplamente a aplicação de IoT na segurança ocupacional e o desempenho de protocolos em dispositivos restritos. No âmbito da aplicação de segurança, Battistoni et al. (2021) validaram o uso do MQTT com níveis de QoS para o monitoramento confiável de Equipamentos de Proteção Individual (EPIs). De forma complementar, Vasuthavan e Ismail (2023) comprovaram a viabilidade de *hardwares* de baixo custo, como o ESP8266 e o sensor MPU6050, na detecção de quedas e envio de alertas em tempo real. Contudo, ambos os estudos se limitaram a validar o funcionamento do sistema, sem aprofundar a análise do consumo energético dos dispositivos operando em bateria.

Para suprir a questão de desempenho, pesquisas avaliaram o comportamento de protocolos de aplicação. Larmo et al. (2019) demonstraram que o CoAP oferece menor latência e tráfego de dados para mensagens pequenas em redes *Narrowband IoT* (NB-IoT) quando comparado ao MQTT. Em testes com microcontroladores de baixo custo, como o ESP32, Vieira et al. (2024) confirmaram a eficiência de vazão do CoAP, mas ressaltaram que o MQTT mantém um consumo energético viável mesmo com o uso de criptografia *Transport Layer Security* (TLS). Adicionalmente, Seoane et al. (2021) alertaram que a adição de segurança onera a banda do CoAP devido ao *handshake Datagram Transport Layer Security* (DTLS). A limitação comum a esses trabalhos é a realização dos testes em ambientes controlados, redes estáveis ou emuladas.

Identificamos uma lacuna na literatura: a ausência de análises integradas do *trade-off* entre consumo energético e latência sob as condições que simulem a conectividade instável típicas da região amazônica. Fica pendente compreender como a robustez de retransmissão do MQTT e a leveza nativa do CoAP impactam a autonomia de dispositivos vestíveis de segurança crítica quando submetidos a altas taxas de perda introduzida e oscilação de sinal.

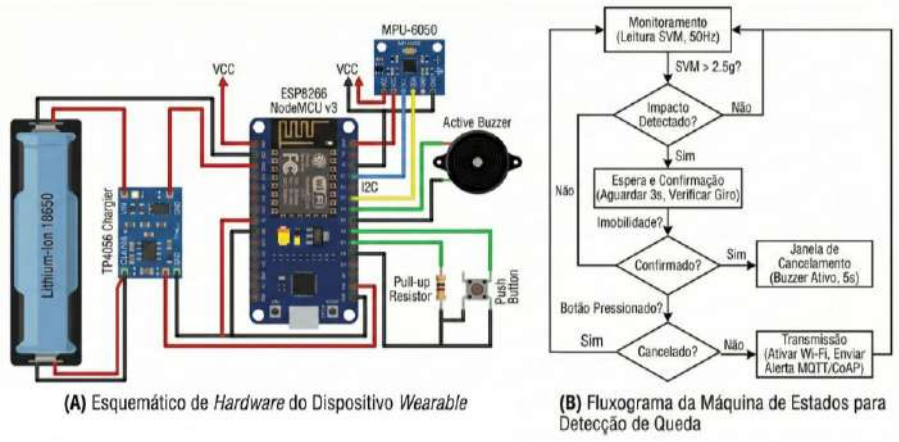
4. Solução computacional

A arquitetura do sistema proposto foi desenvolvida para operar em cenários de conectividade restrita, visando garantir o monitoramento

contínuo de trabalhadores isolados. O sistema é baseado em um modelo de IoT estruturado em 3 módulos: percepção, comunicação e aplicação. O primeiro é responsável pela captação física dos eventos, enquanto o segundo transmite os dados em redes instáveis e o último envolve o processamento e monitoramento centralizado.

O módulo de percepção é composto pelo dispositivo *wearable*, responsável pela coleta de dados inerciais. Este dispositivo utiliza o microcontrolador ESP8266 modelo NodeMCU v3, escolhido por sua conectividade Wi-Fi nativa, baixo custo e suporte a modos de economia de energia. A detecção de quedas e impactos é realizada pelo sensor inercial MPU-6050 via protocolo I2C, que integra um acelerômetro e um giroscópio de três eixos. Acoplado ao sistema, o sensor de corrente e tensão INA219 mensura o consumo energético em tempo real, fornecendo os dados para a análise de eficiência das transmissões e permitindo a validação da latência através da assinatura de corrente do rádio.

Conforme ilustrado na Figura 1A, o dispositivo é alimentado por uma bateria de Li-Ion 18650 de 3.7V, gerenciada pelo módulo de carga e proteção TP4056. Para garantir a estabilidade do *hardware* durante os picos de transmissão e retransmissão dos protocolos, foram incorporados capacitores de desacoplamento de 1000 µF, evitando reinicializações (*resets*) e garantindo a integridade dos componentes.



(A) Esquemático de *Hardware* do Dispositivo *Wearable*

(B) Fluxograma da Máquina de Estados para Detecção de Queda

Figura 1. Esquemático de hardware do dispositivo wearable e fluxograma de Máquina de Estados para detecção de queda.

A lógica de processamento local é orientada por uma máquina de estados, como visto na Figura 1B, que define o comportamento do dispositivo desde a inicialização até o envio dos alertas. Neste fluxo, o cálculo do *Signal Vector Magnitude* (SVM) permite identificar padrões de aceleração que caracterizam uma queda, acionando o envio imediato da mensagem de alerta através dos protocolos testados. Para assegurar a precisão das métricas de tempo, o *firmware* realiza o registro do *timestamp* interno do microcontrolador no instante exato do envio.

A segunda camada da solução é o objeto central deste estudo. Nela, o ESP8266 atua como o cliente que alterna entre os protocolos MQTT e CoAP para o envio dos dados. O destino do fluxo de dados é o módulo de aplicação, composto por um servidor central executado em um computador local. Para receber os dados, o servidor executa simultaneamente um *broker* MQTT (Mosquitto) e um servidor CoAP, baseado em Python. Esta arquitetura permite monitorar a latência real de ponta a ponta e coletar métricas de consumo de forma centralizada para análise comparativa.

Para validação física, o protótipo foi alocado em um *case* impresso em 3D e fixado à cintura do usuário (Figura 2). O *hardware* inclui um botão manual para a interrupção do alerta em até 5 segundos em casos de falso positivo, permitindo validar tanto a autonomia do dispositivo em quedas reais quanto a interação necessária para a segurança do trabalhador.

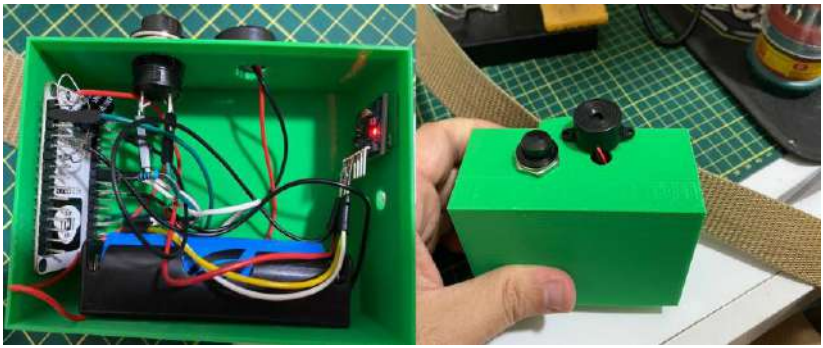


Figura 2. Protótipo montado no case 3D.

5. Avaliação experimental

Esta seção detalha a metodologia adotada para validar a arquitetura proposta e comparar quantitativamente MQTT e CoAP em cenários de conectividade restrita, apresentando o planejamento da pesquisa, os resultados obtidos e a discussão à luz da literatura e das observações de *hardware*.

5.1. Planejamento de pesquisa

O planejamento experimental baseou-se na execução de cenários controlados em bancada com injeção de falhas por *software*, o que garantiu reprodutibilidade estatística dos dados. Para quantificar o impacto da escolha do protocolo sob condições de rede instáveis, o experimento foi estruturado através das variáveis de teste, totalizando 1056 ensaios experimentais válidos.

As variáveis independentes foram os protocolos de comunicação e condições de rede. Para o MQTT, utilizou-se o modo de publicação com QoS 1, garantindo que a mensagem chegasse ao menos uma vez ao destino. Para o CoAP, foi usado o método de requisição *Confirmable* (CON), que exige confirmação de recebimento para simular a confiabilidade necessária em cenários críticos. Ambas as configurações tentaram garantir a paridade na tentativa de entrega confiável para alertas críticos. A emulação da conectividade restrita foi realizada através da ferramento *Traffic Control* (tc) do *Linux*, pela qual se aplicou uma matriz de estresse composta por quatro níveis de atraso introduzido em milissegundos (200 ms, 500 ms, 750 ms e 1000 ms) combinados com quatro níveis de perda introduzida (0%, 10%, 20% e 30%), visando replicar a instabilidade típica de ambientes de operação isolada na região amazônica.

As variáveis dependentes corresponderam aos indicadores de desempenho, sendo eles, a taxa de entrega, consumo de energia em miliamperes (mA) e o atraso fim a fim (ms). O atraso fim a fim foi calculado pela diferença entre os *timestamps* capturados pelo microcontrolador no momento da saída e pelo servidor no instante da chegada, permitindo mensurar diretamente o impacto do atraso introduzido e da perda inserida na comunicação. Adicionalmente, no protocolo MQTT, a latência foi validada

de forma complementar através da duração da assinatura de corrente do rádio, evidenciando as retransmissões inerentes à camada de transporte TCP. Para isolar o desempenho dos protocolos, as variáveis de controle foram mantidas fixas, incluindo o uso do mesmo *hardware* sensor, distância física constante na bancada, *payload* idêntico para ambos os protocolos e frequência de amostragem do sensor de corrente fixada em 1000 amostras por segundo.

O procedimento experimental consistiu na execução de 33 ciclos de teste para cada uma das 32 combinações de protocolo e cenário de rede. A sequência operacional compreendeu as etapas de despertar, conexão Wi-Fi, leitura do sensor inercial (Figura 1B), transmissão do alerta e retorno ao estado de espera. A primeira transmissão de cada bateria foi tratada como um período de estabilização (*warm-up*) e descartada das análises para eliminar transientes iniciais. A janela de coleta de dados de consumo foi padronizada para abranger todo o evento de transmissão e as eventuais retransmissões de rede. Ao final de cada ciclo, os dados foram exportados em formato CSV para o servidor e processados em Python para a análise estatística e geração dos comparativos de eficiência energética entre os protocolos.

5.2. Resultados e Discussão

Os resultados obtidos a partir dos 1056 ensaios experimentais revelaram comportamentos distintos entre MQTT e CoAP, sobretudo em condições críticas de conectividade com 1000 ms de atraso e 30% de perda introduzida. No cenário ideal, com 0% de perda, ambos os protocolos apresentaram consumo estável, embora o MQTT tenha registrado corrente média ligeiramente superior em 1000 ms, com 58,09 mA contra 55,70 mA do CoAP.

A Figura 3 apresenta o impacto da perda introduzida sobre a taxa de entrega, consumo de energia e atraso fim a fim. Mesmo com a perda variando de 0% a 30%, CoAP e MQTT mantiveram a taxa de entrega próxima de 100%, resultado do uso de confirmações de recebimento que acionam retransmissões até a resposta do servidor. Assim, a perda emulada não gera

perda definitiva de alerta, mas aumenta o esforço de retransmissão, o atraso fim a fim e o consumo energético.

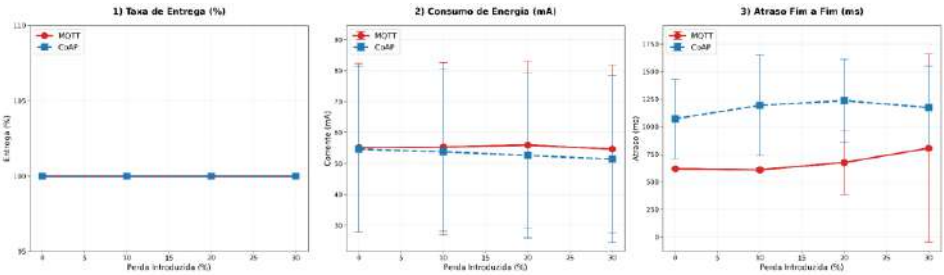


Figura 3. Gráficos evidenciando o impacto da perda introduzida.

Quando analisada a média de corrente para todos os níveis de atraso em cada cenário de perda, verifica-se que o CoAP reduziu o consumo de 54,20 mA em 0% para 49,88 mA em 30% de perda introduzida e que o MQTT manteve a corrente praticamente constante em torno de 55 mA. No cenário mais crítico, o CoAP operou com 46,81 mA contra 54,74 mA do MQTT, indicando que, sob alta perda, a leveza do CoAP proporciona vantagem energética. Contudo, a análise do desvio padrão no consumo revela que o CoAP apresenta picos mais irregulares, enquanto o MQTT mantém uma dispersão menor, sugerindo um comportamento de rádio mais previsível, apesar do maior *overhead*. Tais achados divergem parcialmente de Vieira et al. (2024), que previa superioridade energética do MQTT no ESP8266 pelo gerenciamento de estado de repouso (*idle*). Embora os dados médios apontem o contrário, a baixa dispersão observada reafirma a estabilidade operacional do MQTT na gestão de estados de comunicação prevista pelo autor.

Os picos de consumo mostraram a exigência física sobre o *hardware*. Em todos os cenários, ambos os protocolos atingiram correntes superiores a 160 mA durante as transmissões. O desvio padrão nesses picos confirma que a demanda instantânea frequentemente ultrapassa a capacidade nominal de fornecimento estável do barramento. Isso corrobora Shelby et al. (2014), pois a manutenção da conexão ativa e o processamento de confirmações de recebimento exigem picos de energia que o ESP8266 não supre apenas via

barramento de alimentação, justificando o uso de capacitores de 1000 µF para evitar a instabilidade do sistema.

Quanto à confiabilidade, o MQTT, operando com QoS 1, garantiu a entrega às custas de maior atraso fim a fim. Na Figura 4, observa-se que o atraso médio do MQTT cresce de forma constante com a perda introduzida, passando de aproximadamente 617 ms (0%) para 835 ms (30%). Ao observar o desvio padrão do atraso, nota-se que o MQTT mantém uma dispersão reduzida, garantindo que o alerta chegue em um intervalo de tempo previsível. Esses resultados convergem com Turo et al. (2024), que destacam a necessidade de protocolos que assegurem a entrega em ambientes hostis, ainda que com ocupação mais prolongada do canal e impacto na autonomia energética.

Em contrapartida, defensores do CoAP, como Vasuthavan e Ismail (2023), argumentam que a simplicidade do UDP favorece sistemas de detecção de queda. Os dados coletados mostram que o CoAP apresenta um comportamento mais direto no envio, porém com maior sensibilidade às condições de rede. A média do atraso fim a fim do CoAP variou de cerca de 1072 ms em 0% para 1149 ms em 30%. Entretanto, o elevado desvio padrão observado no CoAP em cenários de perda indica uma alta variabilidade, o que pode dificultar a configuração de *timeouts* precisos em sistemas de segurança crítica.

A Figura 4 mostra o efeito do atraso induzido sobre a taxa de entrega, o consumo e o atraso fim a fim. A taxa de entrega de ambos os protocolos permaneceu em 100%, evidenciando robustez mesmo sob altas latências. Observando o atraso fim a fim, o MQTT acompanha quase linearmente o atraso introduzido, aproximando-se da reta ideal com um desvio padrão mínimo, o que reforça sua estabilidade temporal. O CoAP, em contraste, introduz overhead significativo e um desvio padrão mais acentuado, com 639,2 ms em 200 ms e 1390,6 ms em 750 ms, exigindo um tratamento de erro mais robusto para evitar falsos positivos de falha.

Ainda na Figura 4, percebe-se que o aumento do atraso introduzido impacta de forma distinta o consumo dos protocolos em 0% de perda. O CoAP mantém uma variação moderada de corrente, ao passo que o MQTT apresenta tendência de crescimento mais acentuada. No entanto, a menor

dispersão do MQTT sugere que este consumo extra se traduz em uma conexão com maior persistência de estado e monitoramento ativo.

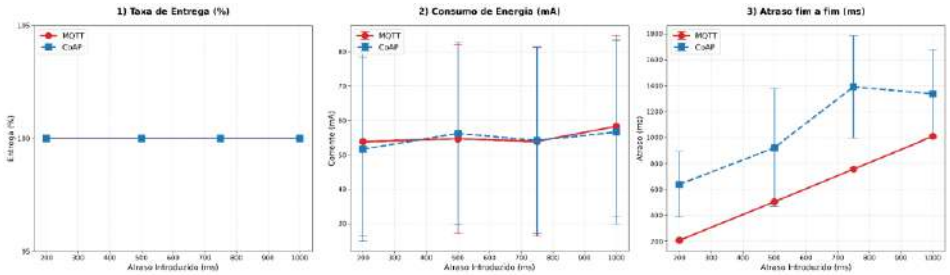


Figura 4. Gráficos evidenciando o impacto do atraso introduzido.

A discussão, portanto, encaminha-se para o fato de que a escolha do protocolo deve ser pautada pela prioridade da aplicação. Se o foco for a máxima economia em redes instáveis, o CoAP mostra-se superior. Se a prioridade for a garantia de entrega e a previsibilidade estatística do tempo de resposta, o MQTT, apesar de exigir um *hardware* com melhor filtragem e consumir mais energia sob latência, apresenta-se como a solução de maior robustez para o trabalhador conectado em ambientes de Indústria 4.0 com conectividade intermitente.

6. Conclusão

O presente trabalho validou uma arquitetura de baixo custo para o monitoramento do trabalhador conectado, demonstrando que a escolha do protocolo impacta diretamente a confiabilidade e o consumo energético em cenários restritos. Enquanto o CoAP mostrou-se superior em economia de energia sob alta instabilidade, o MQTT priorizou o determinismo temporal, garantindo uma entrega de alertas com atraso fim a fim significativamente mais previsível.

A análise observou-se uma correlação entre o comportamento lógico dos protocolos e a resposta física do *hardware*. A baixa dispersão temporal do MQTT confirmou sua estabilidade, operacional ao passo que o elevado desvio padrão e os picos de corrente superiores a 160 mA em ambos os

protocolos reforçaram a necessidade de um projeto de *hardware* robusto. O uso de capacitores de 1000 µF provou ser indispensável para mitigar reinicializações causadas pela variabilidade na demanda energética das transmissões.

Os resultados revelaram que, sob perda de pacotes, o MQTT mantém um consumo quase constante para assegurar a integridade da entrega, acompanhando quase linearmente o atraso induzido apesar do seu maior *overhead*. Em contrapartida, o CoAP apresenta um comportamento não linear, sendo vantajoso quando a prioridade absoluta é a autonomia da bateria em detrimento da precisão temporal.

Dessa maneira, conclui-se que o MQTT, apesar do maior custo energético sob alta perda e da exigência de filtragem de potência consistente, oferece maior estabilidade e previsibilidade para aplicações de segurança do trabalho em ambientes de Indústria 4.0 com conectividade intermitente. Para trabalhos futuros, sugere-se a validação da arquitetura em campo, pois pode haver alterações no comportamento dos protocolos, a integração com redes de longo alcance e baixo consumo e a investigação do impacto da criptografia na autonomia da bateria e no atraso fim a fim.

7. Referências

Ashton, K. (2009). That ‘Internet of Things’ thing. *RFID Journal*. Recuperado de: https://www.rfidjournal.com/articles/view?4986

Battistoni, P., Sebillo, M., & Vitiello, G. (2021). An IoT-Based Mobile System for Safety Monitoring of Lone Workers. *IoT*, *2*(3), 476-497. https://doi.org/10.3390/iot2030024

Hadi, A., Ali, F., Kumar, S., Javed, A., & Ahmed, B. (2025). Wearable IoT Devices with AI for occupational health: real-time worker monitoring and safety analytics. J*ournal of Medical & Health Sciences Review, 2*(2). https://doi.org/10.62019/77c8jx73

Kagermann, H., Wahlster, W., & Helbig, J. (2013). *Recommendations for implementing the strategic initiative INDUSTRIE 4.0: Securing the future of German manufacturing industry; Final report of the Industrie 4.0 Working Group*. Forschungsunion. Recuperado de:

https://www.din.de/resource/blob/76902/e8cac883f42bf28536e7e8165993f1fd/recommendations-for-implementing-industry-4-0-data.pdf

Larmo, A., Ratilainen, A., & Saarinen, J. (2019). Impact of CoAP and MQTT on NB-IoT System Performance. *Sensors*, *19*(1), 7. https://doi.org/10.3390/s19010007

Naranjo, J. E., Mora, C. A., Bustamante Villagómez, D. F., Mancheno Falconi, M. G., & Garcia, M. V. (2025). Wearable Sensors in Industrial Ergonomics: Enhancing Safety and Productivity in Industry 4.0. *Sensors*, *25*(5), 1526. https://doi.org/10.3390/s25051526

OASIS. (2014). *MQTT Version 3.1.1* (OASIS Standard). Recuperado de: http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/os/mqtt-v3.1.1-os.html

Schwab, K. (2016). *The fourth industrial revolution*. World Economic Forum.

Seoane, V., Garcia-Rubio, C., Almenares, F., & Campo, C. (2021). Performance evaluation of CoAP and MQTT with security support for IoT environments. *Computer Networks 197.* https://doi.org/10.1016/j.comnet.2021.108338

Shelby, Z., Hartke, K., & Bormann, C. (2014). *The Constrained Application Protocol (CoAP)* (RFC 7252). Internet Engineering Task Force. https://doi.org/10.17487/RFC7252

Turo, M., Branco, A., & Endler, M. (2024). Applications and Challenges of Sensor Networks in the Amazon Rainforest: Toward Enhanced Environmental Monitoring and Conservatio. In *International Colloquium on Energy-Efficient and Sustainable Distributed Systems*. Natal, Brasil. Recuperado de: https://www-di.inf.puc-rio.br/~endler/paperlinks/E2SDS_24-MarcosTuro.pdf

Vasuthavan, K., & Ismail, M. N. (2023). IoT Based Falling Detection System. *Evolution in Electrical and Electronic Engineering*, *4*(1), 501-508. Recuperado de: https://publisher.uthm.edu.my/periodicals/index.php/eeee/article/view/10943

Vieira, E., Cervi, M., Azevedo, R., & Rizzetti, T. (2024). Análise de Desempenho e Eficiência Energética dos Protocolos MQTT e CoAP no contexto de IoT. In *Anais Estendidos do XXIV Simpósio Brasileiro de Segurança da*

Informação e de Sistemas Computacionais, (pp. 321-327). Porto Alegre, Brasil. https://doi.org/10.5753/sbseg_estendido.2024.243340

Yurtcu, A. (2025). Wearable Technologies and Internet of Things for Real-Time Hazard Monitoring in Occupational Health and Safety: A Systematic Content Analysis. *Innovative Approaches to Engineering Problems, 1*(2), 45-54. Recuperado de: https://dergipark.org.tr/en/download/article-file/5172368

Capítulo 07

Implementação de Arquitetura Zero Trust em IIoT com sensores que simulam ambiente industrial utilizando ESP32 e detecção de anomalias na borda

Mailson Lima Garcia
Wender Antônio da Silva
Universidade Federal de Roraima

garcialimamailson@gmail.com
wender@uerr.edu.br

1. Introdução

A Indústria 4.0 impulsionou a conectividade massiva de dispositivos através da Internet Industrial das Coisas (IIoT), permitindo monitoramento e eficiência operacional sem precedentes (Mishra et al., 2022; Singh et al., 2020). No entanto, essa conectividade expõe o chão de fábrica a riscos de ataques cibernéticos críticos, uma vez que protocolos leves como o MQTT, embora eficientes, são frequentemente implementados sem camadas robustas de segurança, confiando apenas na proteção perimetral de firewalls (Park; Nam, 2020; Federici et al., 2023). Esta abordagem tornou-se insuficiente diante de ameaças modernas que exploram a confiança implícita dentro da rede local (Zanasi et al., 2024; Samaniego; Deters, 2019).

O problema central desta pesquisa reside na vulnerabilidade de microcontroladores de baixo custo, como o ESP32, frente a ataques sofisticados de injeção de dados e falsificação de identidade. Diferente de soluções convencionais, este trabalho identifica, por meio de uma revisão da literatura técnica e científica (Mishra et al., 2022), que a segurança em IIoT exige a mitigação não apenas de acessos não autorizados (spoofing), mas também de ataques de repetição (replay), inundação de rede (flood/DoS), tentativas de adivinhação de credenciais (brute force) e a manipulação maliciosa dos dados dos sensores (data tampering), ameaças estas recorrentemente apontadas como críticas em ecossistemas MQTT modernos (Park; Nam, 2020; Alatram et al., 2023).

Para enfrentar esses desafios (Park; Nam, 2020), propõe-se a implementação de uma arquitetura Zero Trust (Confiança Zero) baseada no padrão NIST SP 800-207. A solução proposta é detalhada na Seção 4.

O objetivo geral deste trabalho é implementar e validar uma arquitetura Zero Trust para redes IIoT, utilizando um gateway de borda para realizar a verificação contínua de dispositivos baseados no protocolo MQTT. Para atingir este objetivo, definiram-se os seguintes objetivos específicos:

1. Projetar uma infraestrutura segregada entre Plano de Dados e Plano de Controle, utilizando hardware ESP32 e lógica de programação em Python;
2. Desenvolver um motor de decisão (PDP) capaz de validar múltiplos atributos de contexto (ID, frequência, hash de integridade e limites físicos);
3. Simular ataques reais de rede (spoofing, flood e tampering) a partir de uma estação externa para testar a resiliência do sistema;
4. Avaliar o impacto da latência das camadas de segurança no tempo de resposta do monitoramento industrial.

A relevância deste estudo é dupla: tecnicamente, oferece proteção de baixo custo a parques industriais legados; cientificamente, contribui para o estado da arte ao deslocar o paradigma Zero Trust para a borda, reduzindo dependência de nuvem.

O restante deste estudo está organizado da seguinte forma: a Seção 2 apresenta a fundamentação teórica sobre IIoT, MQTT e Zero Trust. A Seção 3 discute os trabalhos relacionados e as lacunas identificadas na literatura. A Seção 4 detalha a solução computacional desenvolvida. A Seção 5 descreve a metodologia experimental e discute os resultados obtidos. Por fim, a Seção 6 apresenta as conclusões e considerações finais.

2. Fundamentação teórica

A crescente adoção de dispositivos conectados em ambientes industriais tem ampliado significativamente a superfície de ataque de redes

IIoT, tornando insuficientes as abordagens de segurança baseadas exclusivamente em perímetro (Park; Nam, 2020). Nesse contexto, esta seção discute os referenciais teóricos e tecnológicos que fundamentam a solução proposta, abrangendo o protocolo MQTT, o paradigma Zero Trust Architecture e os componentes lógicos que o operacionalizam na borda.

A Internet Industrial das Coisas (IIoT) designa a interconexão de sensores, atuadores e instrumentos industriais em rede para fins de otimização de processos e tomada de decisão em tempo real (Mishra et al., 2022). No contexto desta pesquisa, o protocolo de comunicação adotado é o MQTT (Message Queuing Telemetry Transport), que opera sobre o modelo Publicador/Subscritor por meio de um Broker central. Embora eficiente para dispositivos restritos como o ESP32, o MQTT carece de validação de identidade granular e verificação de integridade do payload mensagem a mensagem, o que o torna suscetível a ataques de personificação e injeção de dados falsos (Mishra et al., 2022).

Diante dessas vulnerabilidades, o paradigma Zero Trust Architecture (ZTA) emerge como resposta estruturada. Conforme o padrão NIST SP 800-207 (Rose et al., 2020), o Zero Trust fundamenta-se na premissa de que nenhuma confiança deve ser concedida implicitamente a qualquer entidade, independentemente de sua localização na rede. Diferentemente da segurança perimetral tradicional, a ZTA exige que cada requisição seja verificada continuamente com base em atributos de contexto — identidade, integridade dos dados e comportamento do dispositivo. Conforme destacado na literatura sobre ZTA em ambientes industriais (Zanasi et al., 2024; Federici et al., 2023), essa verificação deve ser dinâmica e adaptável, respondendo em tempo real a desvios de comportamento.

Para operacionalizar esses princípios, a arquitetura ZTA é implementada por meio de dois componentes lógicos segregados: o Policy Enforcement Point (PEP) e o Policy Decision Point (PDP). O PEP é responsável por habilitar ou bloquear a conexão entre o sujeito e o recurso; no protótipo proposto, essa função recai sobre o microcontrolador ESP32. O PDP, por sua vez, é o motor de decisão que avalia as requisições com base nas políticas de segurança definidas; neste trabalho, ele é implementado em um Gateway Python executado na borda (Edge Computing), o que minimiza a

latência de decisão e garante a autonomia operacional mesmo em situações de falha de conectividade externa (Ameer et al., 2024).

3. Trabalhos relacionados

A pesquisa de Zanasi et al. (2024) propôs uma arquitetura Zero Trust flexível para infraestruturas IIoT heterogêneas, combinando microssegmentação de rede e software-defined networking (SDN) para aplicação distribuída de políticas. Os resultados demonstraram resiliência a falhas individuais e operação descentralizada. A principal lacuna foi a dependência de infraestrutura SDN centralizada, incompatível com hardware de baixo custo como o ESP32, e a ausência de verificação da integridade física dos dados dos sensores.

Federici et al. (2023) propuseram uma arquitetura Zero Trust para acesso remoto seguro em IIoT industrial, com controle de acesso em dois níveis — rede e borda — validado em cenários aeroespaciais. A solução demonstrou escalabilidade superior às abordagens convencionais sem exigir substituição de hardware legado. A lacuna principal foi a pressuposição de que dispositivos de campo possuem capacidade computacional para participar ativamente da autenticação, tornando a solução inaplicável a microcontroladores com firmware imutável, além de não abordar a detecção de manipulação dos valores físicos dos sensores.

Ameer et al. (2024) formalizaram um modelo de controle de uso para Zero Trust em IoT (ZTA-IoT), com arquitetura PEP/PDP e cálculo dinâmico de pontuação de confiança via AWS. A principal lacuna foi a dependência total de conectividade com a nuvem, introduzindo riscos de disponibilidade em ambientes industriais isolados.

Com base nesses estudos, observa-se que nenhum apresenta uma solução de gateway de borda desacoplado capaz de detectar anomalias operacionais — como saltos abruptos em valores físicos — sem modificar o protocolo MQTT nem depender de infraestrutura em nuvem. Essa combinação de requisitos — resposta em tempo real, proteção a hardware legado e verificação da integridade do dado além da identidade do emissor — permanece pouco explorada, justificando a investigação proposta.

Autor (ano)	Arquitetura	Foco de detecção	Plataforma	Lacuna principal
Zanasi et al. (2024)	Zero Trust + SDN	Identidade + controle de acesso	Framework IIoT genérico	Depende de infraestrutura SDN centralizada; incompatível com microcontroladores de baixo custo
Federici et al. (2023)	Zero Trust + Edge	Acesso remoto seguro	IIoT / aeroespacial	Pressupõe dispositivos com capacidade de autenticação ativa; inaplicável a firmware imutável
Ameer et al. (2024)	ZTA-IoT (PEP/PDP)	Pontuação de confiança	AWS (nuvem)	Dependência total de conectividade com a nuvem
Este trabalho (2026)	ZTA-IoT Gateway borda	Anomalias de payload + spoofing / flood / replay / tampering	ESP32 / baixo custo	—

Figura 1. Comparativo entre abordagens ZTA para IoT na literatura e a solução proposta neste trabalho.

4. Solução computacional

A solução implementa uma arquitetura de segurança em camadas que segrega o plano de dados do plano de decisão de segurança, em consonância com os princípios do modelo Zero Trust Architecture (ZTA), segundo o qual nenhuma entidade — interna ou externa — deve ser implicitamente confiável (Rose et al., 2020). O sistema opera em uma rede local (LAN) composta por três nós principais: o dispositivo de borda (ESP32), o Gateway de Segurança (servidor Python) e uma estação de ataque externa, utilizada para validar a resiliência do ambiente por meio do script ataques_furtivos.py.

4.1. Arquitetura ZTA-IoT e Fluxo de Dados

O protocolo MQTT estrutura a troca de mensagens entre os componentes, de modo que cada publicação de telemetria é tratada como uma requisição de acesso sujeita à verificação antes de qualquer persistência ou atuação. Essa abordagem reflete o conceito de Policy Enforcement Point (PEP) e Policy Decision Point (PDP), centrais na especificação ZTA do NIST (Rose et al., 2020).

O plano de dados é operado pelo ESP32, responsável pela coleta de dados de temperatura (sensor DHT11), luminosidade (sensor LDR) e pressão simulada (potenciômetro), que são publicados no broker MQTT. O plano de

controle, por sua vez, é exercido pelo Gateway Python, que intercepta essas mensagens e aplica as políticas de segurança definidas antes de autorizar a continuidade do fluxo. Essa segregação entre coleta e decisão é fundamental em ambientes IIoT com hardware legado, pois permite introduzir inteligência de segurança sem modificar o firmware dos dispositivos de campo (Sisinni et al., 2018).

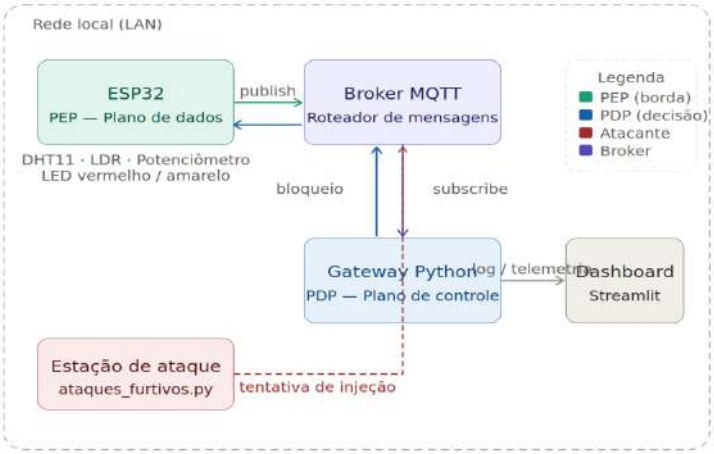


Figura 2. Arquitetura ZTA-IoT implementada em rede local, com segregação entre plano de dados (ESP32/PEP) e plano de controle (Gateway Python/PDP).

4.2. Motores de Detecção de Ameaças

Os mecanismos de detecção de ameaças integram o núcleo do Gateway e estão estruturados dentro da arquitetura proposta como camadas sucessivas de verificação, aplicadas sequencialmente a cada mensagem recebida. Essa abordagem em profundidade (*defense-in-depth*) é recomendada para sistemas ciberfísicos críticos, nos quais a falha de um único mecanismo não deve comprometer a integridade do sistema (Zanasi et al., 2024).

O script servidor_mqtt.py implementa seis camadas de defesa baseadas em regras determinísticas, com limiares estabelecidos a partir de parâmetros recorrentes na literatura de segurança MQTT e IIoT. A primeira realiza o bloqueio de spoofing de identidade, validando o campo device_id de cada mensagem contra uma whitelist de dispositivos autorizados e

descartando imediatamente qualquer identificador desconhecido; essa abordagem é alinhada com Park e Nam (2020), que identificam a ausência de validação de identidade no nível de gateway como principal vetor de personificação em redes MQTT.

A segunda detecta tentativas de brute force, monitorando o volume de IDs desconhecidos em uma janela temporal — o limiar configurado é de cinco identificadores distintos em quinze segundos, padrão indicativo de varredura de rede; esse valor foi definido com base em Laghari et al. (2024), que estabelecem que varreduras sistemáticas de credenciais em redes IoT tipicamente geram entre 3 e 10 identificadores distintos por intervalo de 10 a 30 segundos. A terceira camada implementa controle de flood para mitigação de ataques de negação de serviço (DoS), limitando cada dispositivo a dez mensagens por intervalo de cinco segundos; conforme Alatram et al. (2023), dispositivos ESP32 em operação normal de monitoramento industrial publicam entre 1 e 6 mensagens por segundo, de modo que o limiar de dez mensagens em cinco segundos (equivalente a 2 msg/s) representa uma folga operacional sem permitir inundação.

A quarta previne replay attacks por meio da geração de hashes SHA-256 de cada payload com a biblioteca hashlib; mensagens com conteúdo idêntico detectadas dentro de uma janela de trinta segundos são descartadas como duplicatas maliciosas; a janela de 30 segundos fundamenta-se em Haripriya e Kulothungan (2019), que demonstram que ataques de replay em MQTT tendem a reutilizar pacotes capturados em intervalos inferiores a um minuto para garantir validade temporal das credenciais. A quinta camada realiza a detecção de data tampering, comparando a leitura atual com a anterior e sinalizando como suspeitos os saltos fisicamente implausíveis — variações de temperatura superiores a 15 °C ou de luminosidade superiores a 1.500 unidades entre leituras consecutivas; o critério de 15 °C por ciclo de leitura é fundamentado na taxa de variação máxima documentada para sensores DHT11 em ambientes industriais internos, cujas variações abruptas legítimas raramente excedem 5 °C por segundo (Sisinni et al., 2018), enquanto o limiar de luminosidade de 1.500 unidades (escala de 0 a 4095 do ADC do ESP32) corresponde a uma variação relativa de aproximadamente 37%, fisicamente improvável em transições naturais de iluminação.

Por fim, a sexta camada monitora anomalias de sensor por contexto operacional, verificando se os valores se mantêm dentro dos limites seguros definidos para o processo; leituras de temperatura acima de 40 °C ou de potenciômetro acima de 3.800 unidades ativam o estado de alerta, sinalizado pelo LED amarelo no dispositivo de borda; o limiar de 40 °C foi adotado com base na faixa operacional do sensor DHT11 (0 °C a 50 °C) e nas recomendações de Zanasi et al. (2024) para definição de zonas de operação segura em ambientes industriais simulados.

4.3. Interface de Monitoramento e Resposta

O sistema é complementado por um dashboard desenvolvido em Streamlit, que serve como interface operacional para monitoramento em tempo real. A ferramenta exibe gráficos de telemetria e um log de auditoria contínuo, permitindo ao operador acompanhar o estado do ambiente e os eventos de segurança detectados. Ao identificar qualquer atividade maliciosa originada da estação de ataque externa, o PDP publica um comando STATUS_BLOQUEADO via MQTT, instruindo o ESP32 a ativar o LED vermelho e interromper a operação normal. Esse ciclo de resposta — detecção, decisão e atuação — concretiza o princípio de resposta autônoma local preconizado em arquiteturas de borda seguras (Antonini et al., 2023), garantindo continuidade operacional mesmo em cenários de conectividade limitada.

5. Avaliação experimental

5.1. Planejamento de pesquisa

Esta pesquisa adota uma abordagem experimental controlada, na qual o artefato proposto — o gateway de borda com função de PEP desacoplado — é submetido a cenários de ataque replicáveis em um ambiente industrial simulado em rede local. A estratégia metodológica consiste em injetar classes de ameaças documentadas na literatura e verificar se o sistema as detecta e bloqueia corretamente, sem exigir modificações no protocolo MQTT padrão ou no hardware do ESP32.

A questão central que orienta a avaliação é: o gateway proposto é capaz de detectar e bloquear as principais classes de ataque direcionadas a

dispositivos IIoT operando sobre MQTT, preservando a disponibilidade do serviço para mensagens legítimas?

A variável independente é o tipo de ataque injetado via ataques_furtivos.py, assumindo seis valores: spoofing, brute force, flood/DoS, replay attack, data tampering e anomalia de sensor. A variável dependente é a resposta do gateway — bloqueio com STATUS_BLOQUEADO e ativação do LED vermelho, ou não-detecção. Como métricas secundárias, registram-se o tempo de resposta ao bloqueio e a taxa de falsos positivos em operação normal.

A escolha dessas classes de ataque é sustentada pela literatura: flooding, spoofing, DoS, replay e data tampering figuram entre os vetores mais frequentes em ecossistemas MQTT (Laghari et al., 2024; Haripriya; Kulothungan, 2019), e a ausência de validação no nível de gateway é apontada como causa raiz das principais vulnerabilidades (Husnain et al., 2022; Park; Nam, 2020).

As etapas seguidas foram: (1) configuração do ambiente industrial simulado em rede local; (2) operação normal do ESP32 por dez minutos para baseline e verificação de ausência de falsos positivos; (3) injeção sequencial de cada classe de ataque via ataques_furtivos.py; (4) registro das respostas do gateway e do estado do LED; (5) análise comparativa com o comportamento esperado por cada camada de defesa.

Adicionalmente, realizou-se uma avaliação de percepção com um operador que interagiu com o Dashboard Streamlit durante a execução dos ataques, com o objetivo de verificar a legibilidade do log de auditoria e a clareza dos alertas visuais.

5.2. Resultados e Discussão

Os experimentos foram conduzidos em ambiente industrial simulado, com o ESP32 em operação contínua durante toda a sessão de testes. Antes da execução dos ataques, o sistema foi submetido a um período de operação normal de dez minutos, durante o qual nenhum evento de segurança foi registrado no log de auditoria e todas as mensagens foram

classificadas com status OK — confirmando a ausência de falsos positivos em condição de baseline.

5.2.1. Ataques Diretos (Testes 1 a 6)

A Tabela 1 sintetiza os resultados dos seis ataques diretos executados em sequência. Ao longo de toda a sessão, 30 registros foram capturados na tabela de telemetria, dos quais 24 (80%) foram classificados como ataque e bloqueados, enquanto 6 (20%) correspondem a mensagens legítimas do ESP32 que continuaram fluindo normalmente durante os ataques — evidência de que o sistema manteve a disponibilidade para dispositivos autorizados mesmo sob pressão.

Tabela 1. Resultados dos seis ataques diretos.

Teste	81240085Um@Tipo de Ataque	Detecção	Resposta do Gateway
T1	Spoofing de identidade	☑ Detectado	STATUS_BLOQUEADO + LED vermelho
T2	Brute Force de IDs	☑ Detectado	STATUS_BLOQUEADO + LED vermelho
T3	Flood / DoS (30 eventos em <2s)	☑ Detectado	STATUS_BLOQUEADO + LED vermelho
T4	Replay Attack (hash SHA-256)	☑ Detectado	STATUS_BLOQUEADO + LED vermelho
T5	Data Tampering (Δtemp=97.7°C)	☑ Detectado	STATUS_BLOQUEADO + LED vermelho
T6	Anomalia de Sensor	☑ Detectado	STATUS_ANOMALIA + LED amarelo

O ataque de flooding foi o que gerou maior volume de alertas — 30 eventos concentrados em menos de dois segundos (04:31:58 a 04:32:00), demonstrando que a janela deslizante de 5 segundos responde com latência mínima. O data tampering foi o ataque mais rico em variedade: o gateway detectou saltos em temperatura (2.2°C → 99.9°C, Δ=97.7°C), luminosidade (654 → 4095, Δ=3.441 unidades) e potenciômetro (4090 → 0, Δ=4.090 unidades), além de valores negativos fisicamente impossíveis (2.2°C → -50.0°C). Em todos os cenários, o gateway emitiu o comando STATUS_BLOQUEADO via MQTT, ativando o LED vermelho no ESP32.

Esses resultados estão em consonância com Haripriya e Kulothungan (2019), que demonstraram eficácia superior a 0,88 de detectores baseados em regras sobre tráfego MQTT. A abordagem de inspeção de payload no

gateway sem modificação do protocolo também dialoga com Husnain et al. (2022), que identificaram a validação em intermediário como estratégia viável para hardware legado.

5.2.2. Ataques Furtivos (Testes 7 a 9)

Os ataques furtivos foram projetados para operar abaixo dos limiares de detecção de cada camada, expondo as limitações da abordagem adotada. O Slow Flood (Teste 7) expôs a principal limitação da camada de flood: operando com 7 mensagens a cada 6 seg. — abaixo do limiar de 10 em 5 segundos — o atacante injetou 30 mensagens ao longo de 60 segundos sem disparar alarmes, todas persistidas com status OK. Esse resultado é esperado em sistemas de limiares fixos e está documentado como limitação conhecida de detectores por janela deslizante (Alatram et al., 2023).

O Slow Brute Force (Teste 8) foi o mais crítico: testando um ID desconhecido a cada 20 segundos — além da janela de 15 segundos da camada 2 — o atacante realizou varredura sistemática sem que o gateway identificasse o padrão. Após cinco tentativas erradas e uma correta, obteve acesso sem nenhum alerta de brute force, evidenciando a vulnerabilidade de detectores de janela fixa a adversários que conhecem os parâmetros do sistema (Laghari et al., 2024).

O Gradual Tampering (Teste 9) apresentou resultado distinto: o gateway detectou 20 eventos de data tampering e bloqueou 19 mensagens. Contudo, essa detecção ocorreu porque o ESP32 legítimo continuava publicando leituras reais (2,2 °C) em paralelo com as mensagens adulteradas (18 °C, 19 °C, 20 °C...), criando saltos alternados que ultrapassaram o limiar de 15 °C. Em um cenário real de canal exclusivo do atacante, incrementos de 1 °C a cada 4 segundos elevariam a temperatura de 2,2 °C até 99 °C sem acionar nenhum alarme — evidenciando a necessidade de detecção comportamental adaptativa.

5.2.3. Percepção do Operador

Um operador sem formação técnica interagiu com o Dashboard durante os ataques e conseguiu identificar corretamente os eventos a partir

do log de auditoria. Os alertas visuais e a sinalização por LED — vermelho para bloqueio, amarelo para anomalia — foram considerados intuitivos. O resultado sugere que a interface atende ao requisito de legibilidade operacional mesmo sem conhecimento prévio em segurança de redes.

5.2.4. Discussão comparativa e limitações

Os resultados demonstram que o gateway proposto detecta com eficácia as classes de ataque direto mais documentadas na literatura sobre segurança MQTT, operando sem modificações no protocolo padrão e sem dependência de infraestrutura em nuvem. A taxa de detecção para ataques diretos foi de 100% (6 de 6 classes detectadas), com tempo de resposta inferior a 2 segundos em todos os casos.

Os ataques furtivos revelam que a abordagem determinística é vulnerável a adversários pacientes: Slow Flood e Slow Brute Force operaram abaixo do radar das camadas 3 e 2, respectivamente. Conforme Laghari et al. (2024), mecanismos adaptativos — como modelos TinyML embarcados, conforme Abadade et al. (2023) — são necessários para complementar detectores de regras fixas em IIoT crítico.

As principais limitações são: ambiente industrial simulado em escala reduzida, sem replicar a complexidade de uma planta fabril real com múltiplos processos concorrentes; limiares calibrados empiricamente a partir da literatura, sem ajuste adaptativo em tempo real; avaliação com um único ESP32, sem hardware heterogêneo; e ataques de fonte única, sem cobrir cenários DDoS distribuídos (Alatram et al., 2023).

6. Conclusão

Este trabalho implementou e validou uma arquitetura Zero Trust para redes IIoT com protocolo MQTT, operando integralmente na borda sem modificações no protocolo padrão ou dependência de nuvem. Os experimentos demonstraram taxa de detecção de 100% para ataques diretos — spoofing, brute force, flood/DoS, replay attack, data tampering e anomalia de sensor — com tempo de resposta inferior a 2 seg. e zero falsos positivos durante operação normal, respondendo afirmativamente à questão central

da pesquisa. Os ataques furtivos revelaram limitações esperadas da abordagem: adversários pacientes conseguiram operar abaixo do radar das camadas de flood e brute force, evidenciando a necessidade de mecanismos adaptativos como complemento às regras estáticas. Para trabalho futuro, recomenda-se integrar modelos TinyML embarcados para detecção comportamental e validação em ambientes com múltiplos dispositivos heterogêneos e cenários de ataque distribuído.

7. Referências

Abadade, Y., Temouden, A., Bamoumen, H., Benamar, N., Chtouki, Y., & Hafid, A. S. (2023). A comprehensive survey on TinyML. *IEEE Access*, *11*, 96892–96922. https://doi.org/10.1109/ACCESS.2023.3294111

Alatram, A., Sikos, L. F., Johnstone, M., Szewczyk, P., & Kang, J. J. (2023). DoS/DDoS-MQTT-IoT: A dataset for evaluating intrusions in IoT networks using the MQTT protocol. *Computer Networks*, *231*, 109809. https://doi.org/10.1016/j.comnet.2023.109809

Ameer, S., Praharaj, L., Sandhu, R., Bhatt, S., & Gupta, M. (2024). ZTA-IoT: A novel architecture for zero-trust in IoT systems and an ensuing usage control model. *ACM Transactions on Privacy and Security*, *27*(3), 1–36. https://doi.org/10.1145/3671147

Antonini, M., Pincheira, M., Vecchio, M., & Antonelli, F. (2023). An adaptable and unsupervised TinyML anomaly detection system for extreme industrial environments. *Sensors*, *23*(4), 2344. https://doi.org/10.3390/s23042344

Federici, F., Martintoni, D., & Senni, V. (2023). A zero-trust architecture for remote access in industrial IoT infrastructures. *Electronics*, *12*(3), 566. https://doi.org/10.3390/electronics12030566

Haripriya, A. P., & Kulothungan, K. (2019). Secure-MQTT: An efficient fuzzy logic-based approach to detect DoS attack in MQTT protocol for internet of things. *EURASIP Journal on Wireless Communications and Networking*, *2019*(1), 90. https://doi.org/10.1186/s13638-019-1402-8

Husnain, M., Hayat, K., Cambiaso, E., Fayyaz, U. U., Mongelli, M., Akram, H., Abbas, S. G., & Shah, G. A. (2022). Preventing MQTT vulnerabilities using

IoT-enabled intrusion detection system. *Sensors*, *22*(2), 567. https://doi.org/10.3390/s22020567

Laghari, S. U. A., Li, W., Manickam, S., Nanda, P., Al-Ani, A. K., & Karuppayah, S. (2024). Securing MQTT ecosystem: Exploring vulnerabilities, mitigations, and future trajectories. *IEEE Access*, *12*, 1392733–139289. https://doi.org/10.1109/ACCESS.2024.3412030

Mishra, A., Alzoubi, Y. I., Anwar, M. J., & Gill, A. Q. (2022). Attributes impacting cybersecurity policy development: An evidence from seven nations. *Computers & Security*, *120*, 102820. https://doi.org/10.1016/j.cose.2022.102820

Park, C.-S., & Nam, H.-M. (2020). Security architecture and protocols for secure MQTT-SN. *IEEE Access*, *8*, 226422–226436. https://doi.org/10.1109/ACCESS.2020.3045441

Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero trust architecture* (NIST Special Publication 800-207). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-207

Samaniego, M., & Deters, R. (2019). Zero-trust hierarchical management in IoT. In *Anais do IEEE International Congress on Internet of Things (pp. 88–95)*, Milan, Itália. https://doi.org/10.1109/ICIOT.2019.00022

Singh, S., Rajan, K. S., Sharma, V., & Srivastava, G. (2020). Cybersecurity for Industry 4.0: Security threats and countermeasures. *IEEE Access*, *8*, 229863–229879. https://doi.org/10.1109/ACCESS.2020.3046117

Sisinni, E., Saifullah, A., Han, S., Jennehag, U., & Gidlund, M. (2018). Industrial internet of things: Challenges, opportunities, and directions. *IEEE Transactions on Industrial Informatics*, *14*(11), 4724–4734. https://doi.org/10.1109/TII.2018.2852491

Zanasi, C., Russo, S., & Colajanni, M. (2024). Flexible zero trust architecture for the cybersecurity of industrial IoT infrastructures. *Ad Hoc Networks*, *156*, 103414. https://doi.org/10.1016/j.adhoc.2024.103414

Capítulo 08

PAND: Orquestração entre Visão Computacional e LLM para tomada de decisão na manutenção fotovoltaica

Matheus Naranjo Corrêa
Filipe Dwan Pereira
Universidade Federal de Roraima

matheusnaranjocorrea@gmail.com
filipe.dwan@ufrr.com

1. Introdução

A prioridade global por energias sustentáveis impulsionou a adoção massiva de sistemas fotovoltaicos. Em 2025, a energia solar representou 22% da matriz elétrica nacional, sendo a segunda maior fonte, com mais de 55 GW de potência (Agência Brasil, 2025). Contudo, esse crescimento acelerado, alinhado à Indústria 4.0, exige o enfrentamento de desafios operacionais focados em monitoramento inteligente e otimização automatizada para manter a performance.

Apesar disso, a manutenção fotovoltaica enfrenta obstáculos críticos para garantir o desempenho dos ativos (Ali et al., 2025), visto que o acúmulo de sujeira e defeitos físicos impactam diretamente a geração (Tamoor et al., 2022; Rahman, 2025). Enquanto a inspeção humana é onerosa e subjetiva, a automação via Visão Computacional isolada tende a gerar falsos positivos (Rahman, 2025). A lacuna central reside na falta de contexto operacional desses modelos, que falham em distinguir sujidades inócuas de perdas reais de eficiência. Tal limitação foi identificada nos estudos que operam modelos visuais de forma isolada, sem correlacioná-los com dados elétricos (Duranay, 2023; Ali et al., 2025; Rahman, 2025), evidenciando que a ausência de fusão multimodal é o principal vetor de erros diagnósticos no setor.

O objetivo geral desta pesquisa é propor e validar experimentalmente uma arquitetura de software, denominada PAND (*Photovoltaic Agent for Neural Diagnostics*), que integre um modelo de detecção visual a um Agente Generativo, avaliando sua eficácia na redução de falsos positivos e na

assertividade da tomada de decisão para manutenção de módulos fotovoltaicos. Para alcançar tal propósito, o estudo busca responder às seguintes questões de pesquisa: (P01) De que maneira a orquestração simultânea de dados visuais e telemétricos pode mitigar a ambiguidade nos diagnósticos de manutenção, especialmente em cenários contraditórios? (P02) Qual é o ganho de eficiência operacional (redução de falsos positivos) de um agente em relação ao comportamento esperado de métodos de inspeção puramente visuais? E, por fim, (P03) Como a complexidade e a contradição dos dados telemétricos e visuais impactam a latência de inferência (tempo de raciocínio) do agente autônomo?

As principais contribuições deste estudo residem na demonstração de uma arquitetura de validação cruzada via Agentes de IA, que tem potencial para mitigar desperdícios operacionais ao evitar intervenções em painéis sujos, porém eficientes, e para ampliar a segurança ao detectar falhas elétricas em painéis visualmente limpos.

A estrutura deste estudo compreende as seguintes seções: Seção 2 aborda a Fundamentação Teórica, explorando conceitos de Visão Computacional com YOLO, arquitetura de IA Agêntica e eficiência em sistemas fotovoltaicos. A Seção 3, Revisão da Literatura, estabelece o contraponto entre o panorama atual e a necessidade de um raciocínio integrador. A proposta, detalhada na Seção 4 como Solução Computacional PAND, descreve sua arquitetura multiagente e requisitos técnicos. A Seção 5 discute a Avaliação Experimental e discute a performance do modelo em contextos de conflito entre dados visuais e elétricos. Por fim, a Seção 6 apresenta as Conclusões, limitações e sugere futuras pesquisas.

2. Fundamentação teórica

2.1. Visão Computacional e Arquitetura YOLO

Estruturado em Backbone, Neck e Head, o YOLOv11m integra mecanismos de atenção e blocos residuais C3k2, que aprimoram a representação semântica de danos físicos menores. Diferente do YOLOv26, que prioriza a profundidade em detrimento da agilidade, o modelo v11m equilibra o gradiente de aprendizagem com o processamento de borda. Essa configuração valida o uso da variante medium para diagnósticos

fotovoltaicos, garantindo sensibilidade à baixa resolução e variabilidade de iluminação, superando o custo-benefício de arquiteturas mais pesadas e menos eficientes energeticamente.

2.2. Agentes Inteligentes e Orquestração Multiagentes

A IA Agêntica utiliza Grandes Modelos de Linguagem (LLMs) como motores de raciocínio para orquestrar ferramentas externas e perseguir objetivos complexos através do ciclo *ReAct* (Raciocínio e Ação), permitindo o planejamento dinâmico e a execução de comandos técnicos baseados no feedback do ambiente (Acharya et al., 2025; Rondon et al., 2025). Essa arquitetura viabiliza a Automação Baseada em Intenção na indústria, deslocando o foco da execução manual para a definição estratégica do "*o que fazer*" (Romero & Suyama, 2025). Para tal, esta pesquisa adota a topologia Supervisor–Especialistas, na qual um Agente Mestre coordena sequencialmente agentes especializados (Visual e Planejador). Diferente de arquiteturas de Memória Compartilhada ou Event-Driven, essa escolha assegura um raciocínio sequencial determinístico e centraliza o acesso à telemetria no agente decisor, mitigando riscos de alucinações distribuídas (Acharya et al., 2025).

2.3. Eficiência Fotovoltaica e Análise de Falhas

A eficiência de sistemas fotovoltaicos depende intrinsecamente da integridade da curva Corrente-Tensão (I-V), cuja linearidade é frequentemente comprometida por anomalias físicas ou ambientais, como o acúmulo de sujeira (Tamoor et al., 2022; Duranay, 2023). Esse fenômeno gera sombreamento parcial e redução de corrente, podendo evoluir para *hotspots* e danos irreversíveis, um cenário que, dada a subjetividade da inspeção manual, torna imperativo o uso de Inteligência Artificial para correlacionar, com precisão e escalabilidade, a condição visual detectada com a real degradação da eficiência energética (Ali et al., 2025; Duranay, 2023).

3. Trabalhos relacionados

Selecionamos cinco trabalhos representativos para contrastar a visão computacional de alta performance com a autonomia cognitiva baseada em agentes de intenção, demonstrando que o gargalo industrial não é mais a precisão isolada, mas a falta de raciocínio integrador que converta diagnósticos em ações operacionais autônomas. A Tabela 1 apresenta um comparativo estruturado.

Tabela 1. Tabela comparativa de trabalhos anteriores.

Trabalho	Abordagem	Fusão Visual + Elétrica	Autonomia Decisória	Limitação Principal
Rahman (2025)	YOLOv11 + edge	Não	Não	Sem correlação com eficiência
Mao & Hong (2025)	YOLOv8 + atenção	Não	Não	Sem ação corretiva contextual
Duranay (2023)	CNN/RNN comparativo	Parcial (silos)	Não	Tratamento isolado dos dados
Ali et al. (2025)	IA preditiva O&M	Não	Parcial	Dependência de especialista humano
Romero & Suyama (2025)	LLM agêntico	Não	Sim	Sem validação em visão computacional
PAND (este trabalho)	YOLO+LLM+telemetria	Sim	Sim	Dados sintéticos; sem validação em campo

A pesquisa de Rahman (2025) objetivou aplicar o YOLOv11 na detecção de defeitos solares, adotando treinamento supervisionado em dataset público com técnicas de data augmentation. O pesquisador verificou métricas de precisão superiores a 95,00% em dispositivos de borda. Esse estudo teve como limitação o foco exclusivo na classificação visual, sem correlacionar a detecção com a perda real de eficiência energética.

Visando refinar essa detecção, Mao e Hong (2025) conduziram um estudo para otimizar o YOLOv8 na identificação de microdefeitos, substituindo módulos da rede por mecanismos de atenção e obtendo aumento significativo no mAP. A limitação residiu na falta de um mecanismo de raciocínio para sugerir ações corretivas baseadas no contexto operacional.

Expandindo o escopo, Duranay (2023) buscou comparar arquiteturas de Deep Learning para classificação de anomalias em dados visuais e elétricos, verificando que o Deep Learning supera métodos estatísticos no diagnóstico. Contudo, o estudo limitou-se à operação em silos, tratando visão e telemetria separadamente, sem fusão automática.

Abordando a implementação prática, Ali et al. (2025) investigaram o uso da IA na redução de custos de O&M por meio de levantamento sistemático sobre manutenção preditiva. Os pesquisadores constataram que a automação reduz a inatividade, mas sofre com falta de interpretabilidade, evidenciando dependência contínua de especialistas humanos pela ausência de autonomia decisória na ponta do processo.

Por fim, Romero e Suyama (2025) demonstraram que agentes autônomos, por meio de orquestração de ferramentas (Tool Use) via LLMs, gerenciam a ambiguidade melhor do que scripts tradicionais e são superiores à automação rígida. A limitação foi a ausência de validação integrada especificamente à visão computacional no nicho fotovoltaico.

Apesar dos avanços em fusão multimodal, observa-se uma carência de validações experimentais integrando detecção visual (YOLO) e agentes baseados em LLM para diagnósticos fotovoltaicos sob dados conflitantes. A prevalência de abordagens isoladas contribui para a persistência de falsos positivos (Rahman, 2025) e para a contínua dependência de especialistas (Ali et al., 2025). O PAND visa mitigar essas limitações mediante a fusão semântica entre percepção visual e telemetria elétrica.

4. Solução computacional

O PAND integra visão computacional e telemetria elétrica em uma arquitetura modular Python 3.11 para suporte à decisão em usinas solares. O código-fonte e os modelos treinados estão acessíveis no repositório do projeto[1].

[1] Disponível em: < https://github.com/MatheusOrange211/pand >.

4.1. Engenharia de Requisitos e Arquitetura do Sistema

O sistema foi modelado para atender aos seguintes requisitos, definidos para validar a hipótese de fusão de dados e cobrir os cenários de falhas visíveis e ocultas:

→ **RF01 (Detecção de Condições):** Identificar anomalias visuais (como poeira, danos físicos e fezes de pássaros) em imagens de inspeção.

→ **RF02 (Orquestração de Ferramentas):** O sistema deve alternar de forma inteligente entre a análise da imagem e a consulta aos dados de geração do painel.

→ **RF03 (Parecer Técnico Padronizado):** Gerar um relatório estruturado no padrão A3 (contexto, causa raiz e plano de ação), cruzando a visão com a eficiência real.

→ **RNF01 (Processamento em Lote):** O sistema deve ser capaz de analisar grandes volumes de dados (baterias de 100 testes) de forma sequencial e automática, garantindo o rigor estatístico sem depender de intervenção humana manual.

→ **RNF02 (Explicabilidade):** Toda decisão deve ser justificada textualmente, permitindo que o mantenedor compreenda o porquê de uma recomendação de limpeza ou reparo.

A Figura 1 ilustra o fluxo de funcionamento, onde o sistema recebe a carga de dados, consulta o modelo de visão e orquestra os agentes para o diagnóstico final.

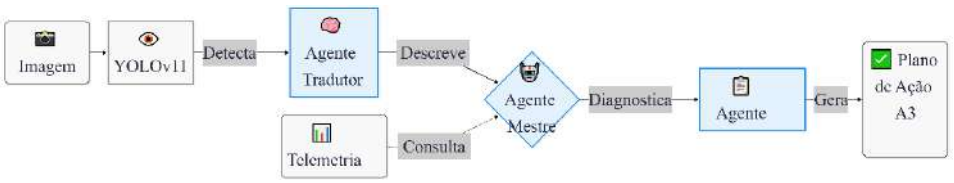


Figura 1. Diagrama de funcionamento.

4.2. Estratégia de Orquestração: Arquitetura Multiagente

Para atender aos requisitos de robustez e mitigação de alucinações (RF02 e RF03), a solução utiliza um Sistema de Múltiplos Agentes baseados em uma "Cadeia de Raciocínio". O fluxo é dividido em quatro entidades especializadas:

→ **Módulo de Visão:** Camada de percepção bruta. Utiliza o modelo *YOLOv11m* para detectar falhas físicas e gerar dados técnicos estruturados. Não usa LLM.

→ **Agente Tradutor:** O estágio de semantização (Llama 3.3 70B) integra os dados sensores ao núcleo cognitivo. Vision-Language Models (VLMs) foram preteridos em favor da precisão determinística do YOLO, cujas métricas de confiança e localização são essenciais para a auditabilidade industrial do projeto. Além do controle rigoroso, o design modular YOLO-LLM facilita a substituição independente dos modelos em versões futuras.

→ **Agente Mestre (Decisor "PAND"):** O núcleo cognitivo do sistema. É o único agente com permissão de Tool Use para acessar os dados de telemetria. Ele cruza o relatório visual com os dados históricos para diferenciar falsos positivos (sujeira sem perda de eficiência) de falhas reais, aplicando a lógica de negócio crítica (RF02). Usa LLM.

→ **Agente Planejador (Relator):** Responsável pela saída final. Consolida o diagnóstico em um plano de ação executável, seguindo a metodologia A3. Usa LLM.

A separação de papéis garante explicabilidade ao núcleo cognitivo (Llama 3.3 70B via Groq). Operando sob o ciclo *ReAct* e *Function Calling*, o *System Prompt* do Agente Mestre converte regras de negócio em princípios determinísticos: (1) prioriza a telemetria, vetando limpezas se a eficiência for >88% (mitigando falsos positivos); (2) valida anomalias cruzando visão de baixa confiança com geração <80%; (3) infere falhas elétricas ocultas em painéis visualmente limpos, mas com eficiência crítica (<70%); e (4) trava alucinações exigindo tags de saída restritas (ex: *[CÓDIGO:*

MANUTENCAO_ELETRICA]). Detalhes referentes ao código do agente[2] estão disponíveis no repositório do projeto.

4.2.1. Configuração do Experimento

O treinamento do modelo YOLO foi realizado em ambiente de nuvem (*Kaggle*) utilizando aceleração de hardware com *2x GPUs NVIDIA T4*, com *16GB VRAM* cada, em paralelo (*Distributed Data Parallel - DDP*).

→ **Dataset visual:** Utilizou-se o conjunto "*6rainstorm*" (versão 1) proveniente do *Roboflow*, contendo 8.694 imagens de treino e 409 de validação, abrangendo 6 classes: *Bird Drop, Defective, Dust, Non Defective, Physical Damage e Snow*.

→ **Hiperparâmetros:** O treinamento de 100 épocas (~11,2 horas) foi definido pela estabilização das métricas de desempenho. Enquanto o *mAP@50* atingiu 0,776 na época 84, o *mAP@50-95* (refinamento espacial) continuou evoluindo, registrando 0,654 na época 95. A manutenção do ciclo completo garantiu a convergência das perdas de validação (*box_loss* 0,92 e *cls_loss* 1,02) sem sinais de overfitting. O modelo YOLO foi configurado para 960 pixels e *batch_size=24*.

→ **Resultados de Validação:** Ao final do ciclo completo de 100 épocas, o modelo atingiu *mAP@50-95* de 0,656 e *mAP@50* de 0,769 no conjunto de validação. Destaca-se a alta precisão na classe "*Non Defective*" (0,945 *mAP@50*) e "*Snow*" (0,947 *mAP@50*). Embora o mAP geral indique margem para melhoria, a precisão na classe "Saudável" é fundamental para reduzir o ruído inicial que será tratado pelo Agente.

5. Avaliação experimental

Nesta seção, apresentam-se a metodologia e os resultados da validação experimental do PAND. A avaliação em simulação controlada analisou a orquestração de Visão Computacional e LLM, focando na eficiência do diagnóstico em cinco cenários com situações contraditórias.

[2] Disponível em: < https://github.com/MatheusOrange211/pand/blob/main/pand_system/agent.py >.

5.1. Planejamento de pesquisa

A validação desta pesquisa fundamenta-se em uma rodada de teste estratificados com 100 casos, integrando resultados de imagens do dataset *6rainstorm* (*Roboflow*) e dados telemétricos sintéticos de geração energética com *seed=42*, visando garantir reprodutibilidade dos resultados. Na Tabela 2, apresentam-se os cenários estipulados para análise do agente PAND. As faixas de eficiência adotam referências na literatura: perdas severas por sujeira entre 20-40% conforme os estudos de Maghami et al. (2016), bem como falhas elétricas.

A partir dos resultados visuais processados pelo modelo YOLO, o algoritmo inicialmente separa os painéis em dois grupos: aqueles com detecção de anomalias (como poeira ou detritos) e os classificados como limpos. Com base nessa divisão, o sistema utiliza uma função de distribuição estatística uniforme (*random.uniform*) para sortear valores sintéticos de corrente, tensão e eficiência, garantindo que os dados variem aleatoriamente, mas sempre restritos aos limites estipulados para cada um dos cinco cenários de operação.

Para os painéis onde o modelo visual acusou anomalias, os dados são distribuídos em três situações: o Cenário A (Convergência), em que a detecção visual é validada por uma queda elétrica real, com eficiência sorteada entre 45% e 65%; o Cenário B (Falso Positivo), onde o sensor visual detectou algo, mas a eficiência permanece alta (92% a 98%), simulando sujeiras que não afetam a geração; e o Cenário E (Ambiguidade), que gera dados de perdas moderadas (70% a 82%), criando uma zona de incerteza para o agente avaliador.

Já para os painéis que o YOLO considerou limpos, o algoritmo gera dados para dois perfis: o Cenário D (Controle), simulando painéis plenamente saudáveis com eficiência ótima (96% a 99%); e o Cenário C (Falha Oculta). Este último é o mais crítico para a validação do PAND, pois simula defeitos elétricos internos (eficiência entre 20% e 40%) que são invisíveis à câmera, exigindo que o agente cruze os dados telemétricos para identificar o problema.

Para mitigar o determinismo e ampliar o realismo, o status do inversor é definido via amostragem ponderada (random.choices). No cenário

de falha oculta, por exemplo, aplica-se uma distribuição de 80% para "Falha" e 20% para "Alerta". Este ruído controlado mimetiza a instabilidade de sensores físicos reais, submetendo o modelo a tomadas de decisão sob incerteza estocástica. Os cenários apresentam conflito visual-elétrico (Tabela 2), projetados para testar a capacidade de orquestração do sistema em situações contraditórias.

Tabela 2. Matriz de Decisão nos Cenários de Teste (Amostras = 100).

Cenário	Input Simulado	Baseline (Visão Pura)	PAND	N	Objetivo do Teste
A. Convergência	Sujo + Efic. Baixa (45-65%)	Limpeza Corretiva	Limpeza Corretiva	15	Validar consistência básica
B. Falso Positivo	Sujo + Efic. Alta (92-98%)	Limpeza Desnecessária	Falso Positivo Ignorado	25	Provar economia de recursos
C. Falha Oculta	Limpo + Efic. Crítica (20-40%)	Falso Negativo (Risco)	Manutenção elétrica	20	Provar segurança operacional
D. Controle	Limpo + Efic. Ótima (96–99%)	Nenhuma Ação	Operação normal	23	Ausência de alucinações
E. Ambiguidade	Conf. YOLO < 50% + Efic. (70-82%)	Alerta incerto	Inspeção Ambiguidade	17	Raciocínio sob incerteza máxima

5.2. Resultados e Discussão

A avaliação quantitativa do agente PAND resultou em uma Taxa de Assertividade Decisória (TAD) Global de 92,00% sobre uma amostra de 100 casos. Na Figura 2, apresentamos os resultados gerados pela matriz de confusão.

A precisão de 100% nos Cenários B e C decorre do alinhamento entre os limiares de eficiência testados (92–98% e 20–40%) e as diretrizes determinísticas do System Prompt (>88% e <70%). Esse resultado ratifica a consistência lógica do sistema frente a regras explícitas, embora não assegure, isoladamente, a capacidade de generalização para dados reais com ruído e incerteza.

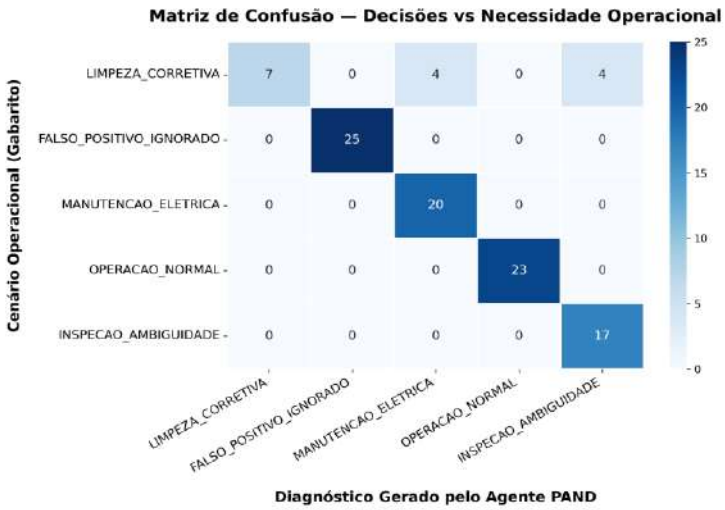


Figura 2. Matriz de Confusão - Decisões vs. Necessidade Operacional.

5.2.1. Mitigação de Ambiguidade em Cenários Contraditórios (Resposta à P01)

Buscou-se entender como a orquestração simultânea de dados mitiga ambiguidades nos diagnósticos de manutenção. Para isso, cinco cenários foram testados, cobrindo desde a convergência básica até situações de incerteza extrema.

O Cenário A evidencia a principal vulnerabilidade do sistema, com TAD de 46,67% na zona de fronteira (45–65% de eficiência). Nessa região, o System Prompt carece de poder discriminatório, gerando erros de conservadorismo (INSPECAO_AMBIGUIDADE) e confusão de classe (MANUTENCAO_ELETRICA). Tal viés demonstra que, na ausência de modelagem probabilística, o LLM interpreta limiares de eficiência como falhas elétricas ocultas, lacuna crítica a ser endereçada via calibração bayesiana ou fine-tuning em trabalhos futuros.

Nos Cenários B e C, a fusão semântica permitiu ao PAND atingir 100% de assertividade ao cruzar evidências contraditórias, superando o enfoque unicamente visual de Rahman (2025). Por fim, o sucesso nos Cenários D e E demonstra que a IA Agêntica pode gerenciar a ambiguidade de forma superior à automação rígida (Romero & Suyama, 2025).

5.2.2. Ganho de Eficiência e Redução de Falsos Positivos (Resposta à P02)

O PAND demonstrou ganho de eficiência operacional na redução de custos de O&M em relação ao comportamento esperado da inspeção puramente visual, que, conforme evidenciado na Tabela 2, geraria limpezas desnecessárias em todos os 25 casos do Cenário B e ignoraria todas as 20 falhas elétricas ocultas do Cenário C. No Cenário B (25 casos de inspeção visual), o sistema filtrou a totalidade dos falsos positivos (limpezas desnecessárias) ao integrar a telemetria; nestes cenários simulados, observou-se uma otimização total de deslocamentos e mão de obra para os casos em que a sujidade não comprometia a geração. Já no Cenário A, a assertividade limitou-se a 46,67%, pois o sistema distribuiu os 8 erros entre dois padrões distintos: em 4 casos convergiu para INSPECAO_AMBIGUIDADE (conservadorismo diante da incerteza) e em outros 4 indicou erroneamente MANUTENCAO_ELETRICA, revelando que a zona limítrofe de eficiência (45–65%) induz o modelo a confundir sujeira com falha elétrica interna (comportamento abordado na seção 5.2.1).

Essa dificuldade do LLM em discernir ações em zonas de fronteira atua como uma margem de erro do agente e configura-se como uma limitação clara da atual versão do sistema, evidenciando a necessidade de calibração de hiperparâmetros e regras de inferência em trabalhos futuros.

5.2.3. Impacto da Complexidade Operacional no Tempo de Inferência (Resposta à P03)

A análise de desempenho evidenciou que o tempo de raciocínio do agente PAND varia diretamente conforme a ambiguidade do cenário. O sistema registrou um tempo médio geral de resposta de 19,56 segundos por ativo. No entanto, conforme ilustrado na Figura 3, cenários com dados lineares e saudáveis (Cenário D - Controle) exigiram menor esforço computacional, com a mediana de processamento em torno de 10,00 segundos.

Em contrapartida, cenários que exigiram maior cruzamento de evidências contraditórias e resolução de conflitos, como o Cenário E (Ambiguidade Extrema), elevaram a latência de inferência para picos de até

30,71 segundos e mantiveram as medianas superiores a 25,00 segundos (Cenários A, B e E). Isso demonstra que o uso de modelos de larga escala na manutenção industrial é viável e cumpre o RNF01, mas exige infraestruturas de alto desempenho (como as LPUs da Groq) para suportar a variação de carga computacional imposta por diagnósticos complexos.

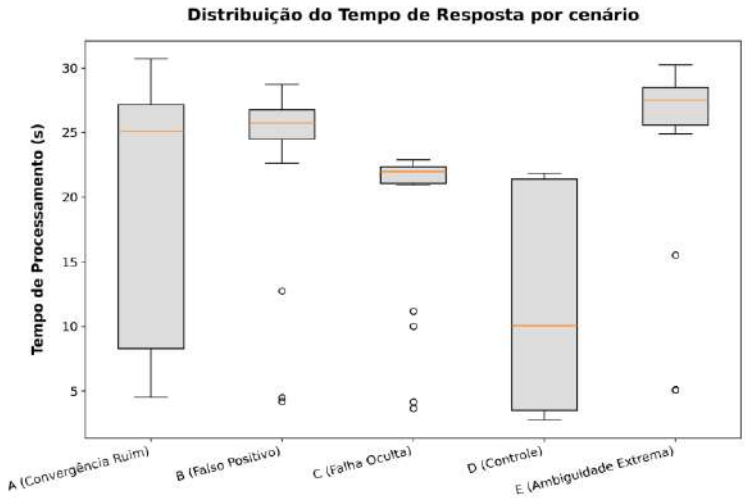


Figura 3. Distribuição do Tempo de Resposta por Cenário.

6. Conclusão

O agente PAND demonstrou viabilidade técnica preliminar no diagnóstico de usinas fotovoltaicas em ambiente simulado e controlado, integrando visão computacional (YOLO), telemetria e LLMs. Com TAD Global de 92,00%, realizou fusão semântica eficaz, respondendo às três questões de pesquisa propostas. Os resultados, contudo, não devem ser generalizados para cenários industriais reais sem validação com dados físicos, ruído de sensor e variabilidade ambiental.

Em resposta à P01, a orquestração simultânea demonstrou eficácia na mitigação de ambiguidades nos Cenários B, C, D e E com assertividade plena. A exceção foi o Cenário A (TAD de 46,67%), onde a zona limítrofe de eficiência (45–65%) gerou erros distribuídos entre conservadorismo

excessivo e classificação incorreta como falha elétrica, apontando para a necessidade de refinamento no prompt engineering em trabalhos futuros. Em resposta à P02, a eficácia de 100% nos Cenários B e C eliminou integralmente limpezas desnecessárias e detectou todas as falhas elétricas ocultas, demonstrando ganho direto em segurança e economia de O&M. Em resposta à P03, o tempo de inferência variou de mediana de 10 segundos no Cenário D até picos de 30,71 segundos no Cenário E, confirmando a viabilidade do uso de modelos de larga escala na manutenção industrial, mas com dependência de infraestruturas de alto desempenho para cenários de alta ambiguidade.

A adoção de arquiteturas agênticas representa um avanço para a Indústria 4.0, otimizando a manutenção preditiva e a longevidade dos ativos fotovoltaicos. Como trabalhos futuros, destacam-se: (1) Portabilidade para modelos de borda; (2) Integração nativa com protocolos industriais (Modbus/MQTT); e (3) Análise de ROI via agente financeiro dedicado.

7. Referências

Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey. *IEEE Access*, *13*, 18936–18956. https://doi.org/10.1109/ACCESS.2025.3532853

Agência Brasil. (2025). *Com 22% da matriz elétrica, energia solar é a 2ª maior fonte do país*. Empresa Brasil de Comunicação. Recuperado de: https://agenciabrasil.ebc.com.br/meio-ambiente/noticia/2025-03/com-22-da-matriz-eletrica-energia-solar-e-a-2-maior-fonte-do-pais

Ali, K. A. M., Osman, Y. K. O. T., El-wahhab, G. G. A., Abuhussein, M. F. A., Abdelwahab, T. A. M., & Fodah, A. E. M. (2025). Improving solar PV performance under bird-dropping conditions with a dual-cooling approach. *Scientific Reports*, *15*, 8211. https://doi.org/10.1038/s41598-024-84932-w

Duranay, Z. B. (2023). Fault detection in solar energy systems: A deep learning approach. *Electronics*, *12*(21), 4397. https://doi.org/10.3390/electronics12214397

Maghami, M. R., Abolhosseini, S., Habibuddin, M. R., Esmaeilzadeh, M., Gomes, C., Rahman, N. A., & Rezania, S. (2016). Power loss due to soiling on solar panel: A review. Renewable and Sustainable Energy Reviews, 59, 1307–1316. https://doi.org/10.1016/j.rser.2016.01.044

Mao, M., & Hong, M. (2025). YOLO object detection for real-time fabric defect inspection in the textile industry: A review of YOLOv1 to YOLOv11. *Sensors*, *25*(7), 2270. https://doi.org/10.3390/s25072270

Rahman, A. (2025). Solar panel surface defect and dust detection: Deep learning approach. *Journal of Imaging*, *11*(9), 287. https://doi.org/10.3390/jimaging11090287

Romero, M. L., & Suyama, R. (2025). *Agentic AI for intent-based industrial automation*. https://doi.org/10.48550/arXiv.2506.04980

Rondon, P., Wei, R., Cambronero, J., Cito, J., Sun, A., Sanyam, S., Tufano, M., & Chandra, S. (2025). Evaluating agent-based program repair at Google. In *2025 IEEE/ACM 47th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. https://doi.org/10.1109/ICSE-SEIP66354.2025.00038

Tamoor, M., Hussain, M. I., Bhatti, A. R., Miran, S., Arif, W., Kiren, T., & Lee, G. H. (2022). Investigation of dust pollutants and the impact of suspended particulate matter on the performance of photovoltaic systems. *Frontiers in Energy Research*, *10*, 1017293. https://doi.org/10.3389/fenrg.2022.1017293

Capítulo 09

Caracterização Não Supervisionada de perturbações elétricas em subestações de distribuição de energia utilizando técnicas de Clustering e Deep Learning

Rafael Benzaquem-Neto
Josiane do Couto Rodrigues
Universidade Federal de Roraima

rafaelbenz@rede.ulbra.br
josiane.rodrigues@ufrr.br

1. Introdução

A modernização dos sistemas de energia elétrica e a transição contínua para o conceito de *Smart Grids* resultaram em uma integração massiva de dispositivos eletrônicos inteligentes, fontes de energia renovável e sistemas de monitoramento avançado (Ukil & Zivanovic, 2014; Mishra et al., 2019). Embora essa evolução tecnológica ofereça maior observabilidade da rede, ela também introduziu novos desafios, particularmente no que tange à Qualidade de Energia (PQ). A proliferação de cargas não lineares e a intermitência de fontes renováveis aumentaram a complexidade das perturbações elétricas, tornando a detecção e classificação desses eventos uma tarefa crítica para garantir a confiabilidade e a eficiência do sistema (Rodriguez et al., 2020; Sekar et al., 2020).

Atualmente, os registradores digitais de perturbações (DFRs) e medidores de qualidade de energia geram volumes massivos de dados oscilográficos com alta resolução (Ukil e Zivanovic, 2014). No entanto, a transformação desses dados brutos em informações úteis ("conhecimento") representa um gargalo significativo para as concessionárias (Ukil & Zivanovic, 2014). A grande maioria desses registros não possui rótulos — ou seja, a causa e o tipo da perturbação são desconhecidos no momento da coleta — e o volume de dados é excessivo para ser processado manualmente por operadores humanos (Islam *et al.*, 2023). Além disso, a análise manual é um processo tedioso, impreciso e dependente da disponibilidade de especialistas

o que, muitas vezes resulta em dados críticos sendo ignorados ou descartados (Shinde et al., 2019; Mishra, Subudhi & Jain, 2019).

As abordagens tradicionais de processamento de sinais, como a Transformada de Fourier (FT) ou a Transformada Wavelet (WT) isolada, embora amplamente utilizadas, apresentam limitações quanto à extração de características em cenários ruidosos e dependem frequentemente de limiares fixos ou seleção manual de atributos (Mishra et al., 2019; Sekar et al., 2020). Diante da escassez de dados rotulados e da ausência de supervisão especializada constante, técnicas de Inteligência Artificial (IA), especificamente o Aprendizado de Máquina Não Supervisionado, emergem como soluções promissoras (Shinde et al., 2019; Almasoudi, 2023). Algoritmos de agrupamento (*clustering*), como o *K-Means*, permitem identificar automaticamente padrões ocultos e segregar eventos de qualidade de energia (como afundamentos, elevações e interrupções) sem a necessidade de treinamento prévio com rótulos (Islam et al., 2023).

Para aumentar a robustez dessa análise, a literatura recente propõe o uso de *Deep Learning*, como *Autoencoders* Convolucionais (CAE), para realizar a extração automática de características e a redução de dimensionalidade dos sinais antes do agrupamento (Rodriguez *et al.*, 2020) (Islam *et al.*, 2023). Essa abordagem híbrida elimina a necessidade de engenharia manual de características e melhora a imunidade a ruídos, permitindo que o sistema filtre anomalias e eventos de interesse de forma autônoma (Mishra et al., 2019; Islam et al., 2023).

Neste contexto, este trabalho propõe a aplicação de técnicas de agrupamento e aprendizado profundo para a caracterização automática de perturbações elétricas, utilizando uma base de dados reais coletada nas subestações de distribuição monitoradas nomeadas nesse trabalho como Sub3 e Sub8. O objetivo é demonstrar como algoritmos não supervisionados podem estruturar dados desconhecidos e auxiliar na tomada de decisão operacional, superando a barreira da falta de especialistas e contribuindo para a resiliência do sistema elétrico frente a falhas e eventos imprevistos (Almasoudi, 2023; Diahovchenko & Keane, 2025).

2. Fundamentação teórica

Esta seção apresenta os conceitos fundamentais sobre a Qualidade de Energia (PQ), bem como as técnicas de Inteligência Artificial adotadas para o processamento de grandes volumes de dados não rotulados: o algoritmo *K-Means* e os *Autoencoders* Convolucionais. A escolha dessas teorias justifica-se pela necessidade de automatizar a análise de perturbações em cenários onde a classificação manual é inviável.

2.1. Perturbações na Qualidade de Energia

A qualidade de energia refere-se à conformidade da forma de onda de tensão, corrente e frequência com padrões estabelecidos. Desvios nestes parâmetros são classificados como fenômenos eletromagnéticos. De acordo com a revisão de 2019 da norma IEEE Std 1159, esses fenômenos são categorizados com base em seu conteúdo espectral e duração, alinhando-se aos conceitos de compatibilidade eletromagnética. As categorias principais utilizadas pela comunidade de qualidade de energia incluem: Transientes (Impulsivos e Oscilatórios), Variações de Curta Duração (que englobam os Afundamentos, também denominados Dips ou Sags, Elevações ou Swells, e Interrupções) e Distorção da Forma de Onda (incluindo Harmônicos e Ruído) (IEEE, 2019).

A análise desses eventos é crítica, pois perturbações podem causar falhas em equipamentos sensíveis, perdas econômicas e instabilidade na rede. No entanto, a modernização das redes (Smart Grids) resultou em um aumento exponencial no volume de dados registrados por medidores digitais, criando um cenário de *Big Data* onde a maioria dos registros não possui rótulos e a análise manual por especialistas tornou-se impraticável (Ukil & Zivanovic, 2014; Islam et al., 2023).

2.2. Aprendizado de Máquina Não Supervisionado: K-Means

Para lidar com a ausência de dados rotulados, a literatura aponta o uso de aprendizado não supervisionado. O algoritmo de clusterização *K-Means* é uma das técnicas mais consolidadas para agrupar dados com base

em similaridade de características, sem a necessidade de supervisão humana prévia (Shinde et al., 2019).

O algoritmo K-Means tem como objetivo particionar um conjunto de dados em k grupos (clusters), buscando minimizar a soma das distâncias quadráticas entre cada ponto e o centroide de seu respectivo grupo, definido como a média das amostras que o compõem (Teeuwsen, 2005). Seu funcionamento é iterativo: inicialmente, k centroides são definidos de forma arbitrária; em seguida, cada amostra é associada ao centroide mais próximo, considerando a distância euclidiana (menor distância entre dois pontos em um espaço vetorial). Após essa atribuição, os centroides são atualizados com base na média das amostras vinculadas, e o processo se repete até que o critério de convergência seja atingido (Shinde et al., 2019; Teeuwsen, 2005).

A validação da qualidade dos *clusters* em cenários não supervisionados é frequentemente realizada por métricas como a pontuação de silhueta (Silhouette Score) ou o método do cotovelo (Elbow Method), que avaliam a coesão interna e a separação entre os grupos (Islam *et al.*, 2023).

2.3. Deep Learning e Autoencoders Convolucionais

Embora o *K-Means* seja eficaz, seu desempenho diminui quando aplicado diretamente a dados de alta dimensão, como formas de ondas brutas de tensão. Para mitigar isso, utiliza-se a técnica de *Deep Learning* conhecida como *Autoencoder* (AE) para redução de dimensionalidade e extração de características (Islam et al., 2023; Rodriguez et al., 2020).

Um Autoencoder é uma rede neural projetada para reproduzir sua entrada na saída por meio de um processo de codificação e decodificação. Na etapa de codificação (encoder), os dados de entrada são comprimidos em uma representação latente de menor dimensão, que retém os atributos mais significativos do sinal. Na etapa de decodificação (decoder), o modelo reconstrói o sinal original a partir dessa representação, buscando minimizar o erro de reconstrução (Islam et al., 2023; Rodriguez et al., 2020). Quando são utilizadas camadas convolucionais (CAE – Convolutional Autoencoder), o modelo adquire a capacidade de filtrar ruídos e capturar padrões temporais complexos das perturbações elétricas de forma mais eficiente que a extração manual de características (Almasoudi, 2023; Islam et al., 2023). A integração

do CAE para extração de atributos com o algoritmo K-Means para agrupamento constitui, portanto, uma abordagem robusta para a classificação de dados experimentais reais e ruidosos.

3. Trabalhos relacionados

A aplicação de técnicas de Inteligência Artificial para a análise de Qualidade de Energia (PQ) tem sido amplamente explorada na literatura recente, com abordagens que variam desde sistemas especialistas baseados em regras até redes neurais profundas. Foi conduzida uma pesquisa direcionada a métodos de detecção e classificação de perturbações, permitindo situar a presente proposta frente ao estado da arte.

Mishra et al. (2019) propuseram um modelo baseado em Redes Neurais Convolucionais (CNN) para classificar 16 tipos de perturbações de energia, incluindo eventos simples e complexos. O estudo utilizou a Transformada Wavelet Contínua (CWT) para converter sinais unidimensionais em representações de imagem, que serviram de entrada para a CNN. Embora o modelo tenha demonstrado alta imunidade a ruídos e precisão superior aos métodos tradicionais, a abordagem foi supervisionada, dependendo de um vasto conjunto de dados sintéticos rotulados para o treinamento, o que nem sempre reflete a realidade de dados de campo não rotulados.

Em uma linha semelhante, Rodriguez et al. (2020) desenvolveram uma abordagem híbrida utilizando *Autoencoders* Convolucionais (CAE) para redução de dimensionalidade, seguidos por Redes Neurais Recorrentes (RNN) do tipo *Long Short-Term Memory* (LSTM) empilhadas para a classificação. O estudo destacou a capacidade dos CAEs em comprimir o tamanho do sinal mantendo características dinâmicas relevantes, o que reduziu significativamente o tempo de processamento. Contudo, assim como em Rodriguez et al. (2020), a validação foi realizada majoritariamente com dados sintéticos gerados a partir de modelos matemáticos parametrizados, mantendo a dependência de rótulos prévios para a etapa de classificação.

Abordando a questão da falta de rótulos, Shinde, Sowane e Munje (2019) exploraram o uso de aprendizado de máquina não supervisionado, especificamente o algoritmo *K-Means*, para agrupar eventos de afundamento,

elevação e interrupção. O trabalho demonstrou a viabilidade de agrupar amostras semelhantes sem treinamento supervisionado prévio. No entanto, a extração de características baseou-se em atributos tradicionais, que podem não capturar a complexidade de perturbações não lineares ou transitórios de alta frequência presentes em redes modernas.

Mais recentemente, Islam et al. (2023) apresentaram uma técnica não supervisionada que combina *Deep Learning* e clusterização, utilizando CAEs para extração automática de características e *K-Means* para o agrupamento. O estudo inovou ao aplicar o método em dados de campo reais de uma rede de distribuição, utilizando a visualização t-SNE para interpretar os *clusters*. A principal contribuição dessa abordagem foi a capacidade de filtrar formas de onda anômalas de eventos recorrentes sem a necessidade de rotulagem massiva.

Com base nesses estudos, observa-se que abordagens envolvendo o uso de Deep Learning (Mishra et al., 2019), (Rodriguez et al., 2020) e técnicas de clusterização (Shinde et al., 2019) já se encontram bem estabelecidas na literatura. Entretanto, os estudos existentes tendem a empregar alguma forma de supervisão na etapa de caracterização dos dados. Diferentemente, o presente trabalho propõe o uso combinado da arquitetura CAE e do algoritmo K-Means, inspirado em (Islam et al., 2023), para realizar uma análise totalmente não supervisionada sobre dados provenientes de subestações, avaliando a capacidade da metodologia em identificar padrões intrínsecos em um conjunto de dados reais sem qualquer curadoria prévia.

4. Solução computacional

Esta seção detalha a arquitetura da solução proposta, as ferramentas tecnológicas selecionadas e o pipeline de processamento de dados desenvolvido para a caracterização não supervisionada de perturbações elétricas. A metodologia computacional foi estruturada para operar em ambiente de nuvem, visando escalabilidade e reprodutibilidade.

4.1. Arquitetura da Solução

A solução foi projetada como um pipeline sequencial de aprendizado de máquina, dividido em três módulos principais: (i) Pré-processamento e Normalização; (ii) Extração de Características via *Deep Learning*; e (iii) Agrupamento e Análise. A arquitetura proposta busca transformar dados brutos de oscilografia em *clusters* de eventos significativos sem a necessidade de intervenção humana para rotulagem prévia.

O fluxo de dados inicia-se com a ingestão das séries temporais brutas provenientes das subestações de transmissão de energia. Em seguida, uma Rede Neural do tipo *Autoencoder* Convolucional (CAE) é utilizada para comprimir esses sinais em um vetor de características latentes (*latent space*), reduzindo a dimensionalidade e filtrando ruídos. Por fim, o algoritmo *K-Means* é aplicado sobre esse espaço latente para segregar os eventos, conforme validado por Islam (2023) em cenários de Big Data de redes de distribuição.

4.2. Ambiente de Desenvolvimento e Tecnologias

Para a implementação da solução proposta, utilizou-se a linguagem Python (versão 3.8 ou superior) devido ao seu amplo ecossistema voltado à Ciência de Dados e Inteligência Artificial. O desenvolvimento foi realizado no ambiente Google Colab, plataforma em nuvem que disponibiliza GPUs para acelerar o treinamento das redes neurais profundas. Na construção da metodologia, empregaram-se bibliotecas como Pandas e NumPy para manipulação dos dados e operações matriciais; TensorFlow e Keras para modelagem e validação da arquitetura do Autoencoder Convolucional; Scikit-learn para o algoritmo K-Means, normalização e métricas de desempenho; e, por fim, Matplotlib e Seaborn para a visualização dos sinais elétricos, das curvas de perda e da distribuição dos clusters gerados.

4.3. Pipeline de Processamento e Modelagem

O desenvolvimento da solução computacional segue um fluxo rigoroso de etapas, descritas a seguir.

4.3.1. Coleta e Pré-processamento dos Dados

Os dados brutos, consistindo em formas de onda de tensão trifásica, são submetidos a uma etapa de limpeza inicial para a remoção de registros corrompidos ou com janelas de tempo inconsistentes. Dado que as redes neurais e a etapa de classificação de distúrbios são altamente sensíveis à escala dos dados de entrada, os sinais de tensão serão normalizados com base em valores por unidade (p.u.), o que é realizado dividindo-se o valor instantâneo do sinal pelo valor de pico da tensão nominal do sistema (Islam *et al.*, 2023). Diferentemente da padronização *Z-score* aplicada individualmente — que forçaria os sinais a terem variância unitária e, consequentemente, eliminaria informações críticas de magnitude —, a conversão para p.u. alinha-se às práticas e definições da norma IEEE Std 1159:2019, preservando a magnitude real da tensão remanescente (*retained voltage*) (IEEE, 2019). Esse rigor é imprescindível para que o algoritmo de agrupamento consiga diferenciar corretamente os eventos baseados em suas amplitudes absolutas, como afundamentos de tensão ou *sags* (0,1 a 0,9 p.u.) em contraposição a interrupções (< 0,1 p.u.) (IEEE, 2019).

Adicionalmente, os registros oscilográficos serão redimensionados para assegurar a uniformidade exigida na entrada do modelo de *Deep Learning*. Com base nas diretrizes da IEEE Std 1159:2019, que define as variações RMS de curta duração como fenômenos ocorrendo em intervalos de 0,5 ciclo a 1 minuto (IEEE, 2019), e apoiando-se na metodologia validada por Islam et al. (2023), os sinais serão janelados em segmentos temporais padronizados, abrangendo, por exemplo, de 10 a 16 ciclos da frequência fundamental (IEEE, 2019). Eventos que ultrapassem essa duração poderão ser processados através de técnicas de "janelamento" deslizante (*sliding window*), garantindo a coerência temporal e a extração fidedigna das características morfológicas pela rede neural (Junior et al., 2019).

4.3.2. Extração de Características com Autoencoder Convolucional

A arquitetura utiliza um Autoencoder Convolucional para extrair características de sinais de forma não supervisionada. O processo é dividido em duas etapas: o encoder, que comprime os dados em um vetor latente (bottleneck) para destacar informações essenciais, e o decoder, que

reconstrói o sinal original. Para garantir a precisão, o modelo utiliza o otimizador Adam e a função de perda MSE, que minimiza a discrepância entre a entrada e a saída reconstruída.

As camadas do encoder focam em redução e filtragem: a Conv1D aplica filtros convolucionais de uma dimensão para identificar padrões temporais nos sinais, enquanto o MaxPooling reduz a dimensionalidade ao selecionar apenas os valores máximos de janelas específicas, descartando dados redundantes.

No processo inverso, o decoder utiliza o UpSampling para aumentar a resolução do vetor comprimido, repetindo dados para retomar o tamanho original do sinal. Novas camadas Conv1D são aplicadas em seguida para refinar essa reconstrução, garantindo que o sinal de saída seja uma representação fiel e limpa da entrada inicial.

4.3.3. Clusterização e Visualização

Após o treinamento do CAE, a parte do *Decoder* é descartada, e o *Encoder* é utilizado para gerar os vetores de características reduzidos de toda a base de dados. Esses vetores servem de entrada para o algoritmo *K-Means*.

Para determinar o número ótimo de *clusters* (k), implementa-se o método do cotovelo (*Elbow Method*) analisando a inércia (soma dos erros quadráticos intra-*cluster*) e o coeficiente de Silhueta. Para a interpretação visual dos resultados, utiliza-se a técnica t-SNE (t-Distributed Stochastic Neighbor Embedding), que projeta os vetores de alta dimensão em um plano 2D ou 3D, permitindo a verificação visual da separabilidade das classes de perturbações (*sags*, *swells*, harmônicos etc.), conforme metodologia aplicada por Islam *et al*. (2023).

5. Avaliação experimental

Esta seção apresenta a avaliação experimental da metodologia proposta. A seção está dividida na descrição da execução do *pipeline* computacional estruturado para processar o volume de dados (*Big Data*) e

na discussão dos resultados obtidos através da clusterização e da curadoria semântica baseada nas normativas de qualidade de energia.

5.1. Planejamento de pesquisa

A execução experimental foi conduzida em ambiente de nuvem (Google Colab), utilizando processamento acelerado por GPU para o treinamento da rede neural profunda. A *pipeline* (**Figura 1**) processou um total de 327 arquivos oscilográficos brutos no formato ".mat" (MATLAB), originários das subestações de distribuição Sub3 e Sub8. Cada arquivo contém 48 eventos de perturbação, onde cada evento corresponde a uma linha de matrizes de forma de onda (shape 48×384), representando 384 amostras a 64 amostras/ciclo (60 Hz), totalizando 6 ciclos (~100 ms) por registro. Deste conjunto, foram utilizadas exclusivamente as matrizes *matXVa*, *matXVb* e *matXVc* (formas de onda brutas das tensões trifásicas A, B e C) e a matriz auxiliar *matAux* (48×20), que contém metadados de cada evento.

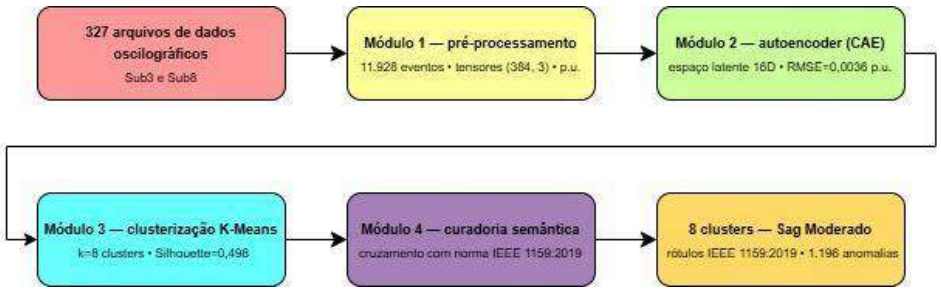


Figura 1. Fluxograma da pipeline computacional.

Na etapa de pré-processamento (Módulo 1), as formas de onda da tensão trifásica foram extraídas e normalizadas para valores por unidade (p.u.), dividindo-se as amostras pela tensão nominal de pico do sistema. Essa abordagem garantiu a preservação da magnitude física da tensão remanescente (*retained voltage*). Utilizando uma técnica de janelamento deslizante (*sliding window*), os sinais foram padronizados em tensores tridimensionais de 384 amostras (6 ciclos à taxa de 64 amostras/ciclo),

resultando em uma base de dados refinada com 11.928 eventos distintos validados para treinamento.

Para a extração de características (Módulo 2), o *Autoencoder* Convolucional (CAE) foi treinado de forma não supervisionada durante 100 épocas. O modelo demonstrou alta capacidade de compressão e retenção morfológica, reduzindo os sinais de entrada de alta dimensionalidade para um espaço latente de apenas 16 variáveis. O erro de reconstrução final (*Root Mean Squared Error* - RMSE) mediano atingiu a marca de 0,003573 p.u. (média de 0,004189 p.u.) para a base completa, com MAPE mediano de 0,93%. Esses resultados indicam que a arquitetura aprendeu com sucesso as assinaturas dinâmicas das perturbações, filtrando ruídos irrelevantes e preservando a morfologia dos eventos.

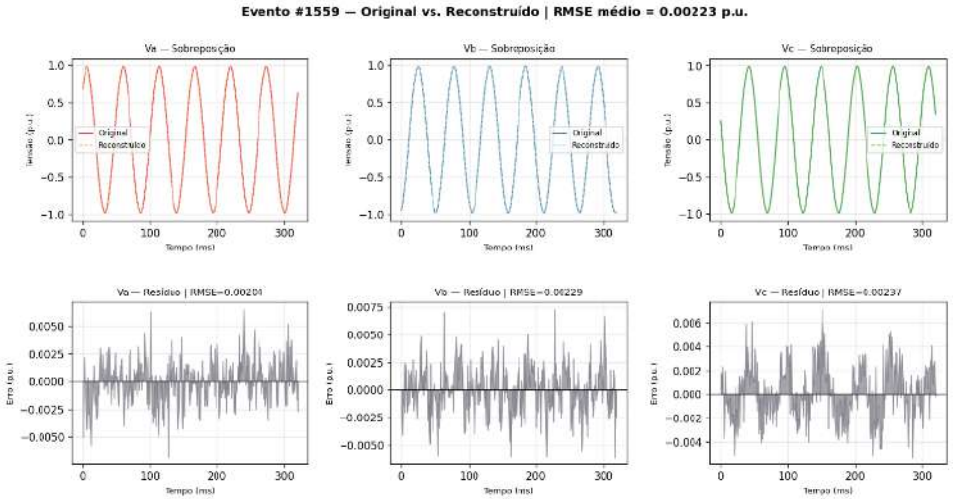


Figura 2. Comparação entre a forma de onda original do distúrbio e o sinal reconstruído pelo Autoencoder Convolucional, demonstrando a preservação das características temporais essenciais.

Na etapa de agrupamento (Módulo 3), os vetores latentes foram padronizados via *Z-score* para adequação do cálculo de distâncias. A avaliação do número ótimo de agrupamentos (k) por meio do Método do Cotovelo (Elbow) e do coeficiente de Silhueta apontou k=8 como o particionamento ideal, com Silhouette Score de 0,4981, Davies-Bouldin Index

de 0,6381 e Calinski-Harabász Score de 11.333,32. Esses valores atestam uma separação espacial coerente para dados de campo reais, que inerentemente possuem alto nível de ruído e eventos de fronteira.

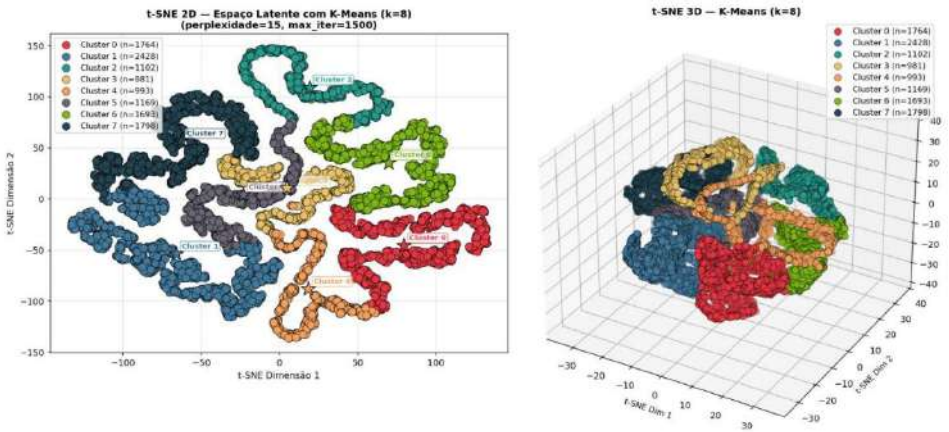


Figura 3. Projeção de redução de dimensionalidade (t-SNE) ilustrando a separabilidade espacial dos 11.928 eventos no espaço latente em oito agrupamentos distintos.

5.2. Resultados e Discussão

A principal contribuição analítica desta avaliação consistiu na etapa de Curadoria Semântica Híbrida (Módulo 4). Em vez de atribuir rótulos baseados puramente nas médias matemáticas dos *clusters*, o algoritmo realizou um cruzamento individualizado de cada forma de onda com as diretrizes rigorosas da norma IEEE Std 1159:2019.

A análise autônoma do sistema revelou que a separação física global entre os três *clusters* principais (ΔVrms) era de apenas 0,0076 p.u., inferior ao limiar significativo de variabilidade (0,05 p.u.). Esse diagnóstico matemático evidenciou que a quase totalidade da base de dados (cerca de 90%) é dominada por um único tipo de fenômeno macro: os Afundamentos Moderados de Tensão (Sags), com tensões remanescentes médias variando na faixa de 0,688 a 0,696 p.u. entre os oito grupos formados.

Diante da homogeneidade observada nos dados, o algoritmo K-Means agrupou os eventos principalmente com base nas fases elétricas afetadas e

em variações sutis de tensão de pré-falta. Os oito clusters formados foram todos rotulados pela curadoria semântica como Sag Moderado (0,50–0,70 p.u.), com diferenciação pela fase dominante. Seis clusters apresentaram predominância da fase B e dois da fase C (clusters 3 e 7), com durações médias semelhantes em torno de 29 ms e tensões de pré-falta variando de 0,696 a 0,701 p.u. O maior grupo (Cluster 1) concentrou 2.428 eventos (20,4% da base), enquanto o menor (Cluster 3) agrupou 981 eventos (8,2%). Adicionalmente, a análise estatística prévia detectou 11 eventos de interrupção (0,1% da base), caracterizados por tensões inferiores a 0,1 p.u., os quais foram corretamente isolados pela curadoria híbrida com base nas regras da norma IEEE Std 1159:2019.

Adicionalmente, ao calcular as distâncias espaciais dos eventos em relação aos seus centróides matemáticos, o algoritmo isolou 1.196 eventos atípicos (≈ 10,0% da base) classificados como "Anomalias para Inspeção Manual". Essa separação evidencia o valor prático da IA não supervisionada no setor elétrico: o sistema foi capaz de estruturar automaticamente 90% da base que possuía comportamento repetitivo e comum, permitindo que a equipe de engenharia da concessionária foque sua atenção exclusivamente nos 10% de anomalias críticas, resolvendo o gargalo operacional imposto pelo cenário de Big Data.

6. Conclusão

Este trabalho demonstrou a viabilidade técnica e o valor operacional da aplicação de técnicas de Inteligência Artificial não supervisionada para a caracterização autônoma de perturbações de qualidade de energia elétrica. O *pipeline* desenvolvido, integrando *Deep Learning* (*Autoencoders* Convolucionais) e clusterização (*K-Means*), foi capaz de processar uma base de dados reais de alta complexidade contendo 11.928 registros oscilográficos não rotulados, extraídos de 327 arquivos provenientes das subestações Sub3 e Sub8.

A principal linha abordada por este estudo foi a dependência de especialistas humanos para a rotulagem de grandes volumes de dados. O *Deep Learning* atuou de forma autônoma na extração de características e redução de dimensionalidade estrutural, enquanto o módulo de inteligência

baseada em regras conectou as descobertas matemáticas da IA aos limites físicos da norma IEEE Std 1159:2019. O diagnóstico final revelou uma rede de distribuição predominantemente afetada por afundamentos de tensão (*sags*) nas fases B e C, com todos os oito grupos formados enquadrados na categoria Sag Moderado (0,50–0,70 p.u.) da norma IEEE Std 1159:2019, sinalizando um padrão sistemático de afundamentos moderados de tensão nas subestações monitoradas.

A capacidade do sistema em atuar como um "filtro de triagem" pode viabilizar sua aplicabilidade para a indústria. Ao identificar e isolar 1.196 anomalias morfológicas (≈ 10% da base), a ferramenta automatiza a etapa mais tediosa do monitoramento de *Smart Grids*, direcionando a expertise humana apenas para os eventos que de fato exigem investigação avançada.

Como perspectivas de trabalhos futuros, sugere-se a aplicação de algoritmos baseados em densidade, como o DBSCAN, visando lidar com perturbações multifatoriais e de fronteira. Adicionalmente, recomenda-se a implementação computacional desta arquitetura em dispositivos de borda (Edge Computing) acoplados aos medidores das subestações, possibilitando o monitoramento e a tomada de decisão em tempo real.

7. Referências

Almasoudi, F. M. (2023). Enhancing power grid resilience through real-time fault detection and remediation using advanced hybrid machine learning models. *Sustainability*, 15.

Diahvchenko, I., & Keane, A. (2025). Clustering for power system stability using graph theory and electrical distance. Sustaina*ble Energy, Grids and Networks*, 43.

IEEE. (2019). IEEE recommended practice for monitoring electric power quality (IEEE Std 1159-2019; Revision of IEEE Std 1159-2009, pp. 1–98).

Islam, M. M., et al. (2023). Unsupervised clustering of disturbances in power systems via deep convolutional autoencoders. *IEEE*.

Junior, W. L. Rodrigues, et al. (2019). A methodology for detection of power quality disturbances in the context of demand side management. In *2019*

4th International Conference on Smart and Sustainable Technologies (SpliTech) (pp. 1–6).

Mishra, P. K., Subudhi, U., & Jain, S. (2019). Power quality disturbances classification with deep learning approach. In *2019 International Conference on Smart Energy Systems and Technologies (SEST)*.

Rodriguez, M. A., et al. (2020). Power quality disturbance classification via deep convolutional auto-encoders and stacked LSTM recurrent neural networks. *IEEE*.

Sekar, K., Sendil Kumar, S., & Karthick, K. (2020). Power quality disturbance detection using machine learning algorithm. *IEEE*.

Shinde, P., Sonwane, P., & Munje, R. (2019). Machine learning based clustering for identifying power quality events.

Teeuwsen, S. P. (2005). Oscillatory stability assessment of power systems using computational intelligence.

Ukil, A., & Zivanovic, R. (2014). Automated analysis of power systems disturbance records: Smart grid big data perspective. In 2014 IEEE Innovative Smart Grid Technologies—Asia (ISGT ASIA).

Capítulo 10

Sistema inteligente para diagnóstico da qualidade de energia em baixa tensão com autoencoders aplicados à Indústria 4.0

Yankarlo Silva Carvalho
Alexander Fernandez Correa
Universidade Federal de Roraima

yankarlo.eng@gmail.com
alexander.fernandez@ufrr.com

1. Introdução

A qualidade de energia elétrica (Power Quality – PQ) tem se tornado uma preocupação crescente no setor elétrico, principalmente devido ao aumento da utilização de cargas não lineares e da inserção de fontes renováveis nos sistemas de distribuição (Silva & Farias, 2025). A ocorrência de distúrbios, como sobretensões, subtensões e distorções harmônicas, pode causar falhas em equipamentos, aumento de perdas elétricas e prejuízos operacionais. Nesse contexto, o monitoramento contínuo e a análise adequada desses fenômenos são fundamentais para garantir a confiabilidade e o bom funcionamento das instalações elétricas, especialmente no cenário da Indústria 4.0, onde a digitalização e a automação intensificam a sensibilidade dos processos produtivos à qualidade da energia.

Com a ampla utilização de analisadores de qualidade de energia, tornou-se possível obter grandes volumes de dados elétricos medidos ao longo do tempo. Esses dados são geralmente disponibilizados na forma de valores eficazes de grandezas como tensão, corrente, potência, fator de potência, frequência e distorção harmônica total. Essas medições permitem verificar a conformidade do sistema com os limites estabelecidos em normas técnicas, como o Módulo 8 do PRODIST, regulamentado pela Agência Nacional de Energia Elétrica. No entanto, a análise convencional baseada exclusivamente na comparação com limites normativos é limitada, pois não

permite identificar automaticamente padrões anormais de operação ou possíveis condições de degradação do sistema.

Nesse sentido, técnicas de Inteligência Artificial têm se mostrado uma alternativa promissora para complementar a análise tradicional, estando alinhadas aos princípios da Indústria 4.0, como análise de dados em larga escala e tomada de decisão inteligente. (Veeramsetty et al., 2023). A utilização direta de dados eficazes permite desenvolver soluções compatíveis com dados reais obtidos em campo, sem a necessidade de acesso às formas de onda completas. Entre essas técnicas, destacam-se as redes neurais do tipo Autoencoder, que são capazes de aprender o comportamento normal do sistema e detectar automaticamente desvios em relação a esse padrão, sem a necessidade de dados previamente rotulados (Sakurada, 2014).

Diante disso, este trabalho propõe o desenvolvimento de um sistema para análise da qualidade de energia elétrica em baixa tensão utilizando técnicas de aprendizado de máquina aplicadas a dados RMS obtidos de analisadores comerciais. A metodologia consiste em integrar a análise normativa baseada no PRODIST com um modelo de Inteligência Artificial capaz de detectar anomalias, avaliar sua severidade e identificar as variáveis elétricas mais relacionadas a essas ocorrências, em consonância com abordagens inteligentes de monitoramento típicas da Indústria 4.0. Dessa forma, busca-se não apenas verificar a conformidade com a norma, mas também fornecer informações adicionais que auxiliem no diagnóstico e na tomada de decisão.

Como contribuição, este trabalho apresenta uma metodologia aplicável a dados reais de sistemas elétricos, permitindo a detecção automática de anomalias e o suporte ao monitoramento da qualidade de energia, contribuindo para a melhoria da confiabilidade e da operação das instalações elétricas.

Este trabalho está organizado da seguinte forma: a Seção 2 apresenta a fundamentação teórica, a Seção 3 discute os trabalhos relacionados, a Seção 4 descreve a solução computacional proposta, a Seção 5 apresenta a avaliação experimental e os resultados obtidos, e a Seção 6 traz as conclusões e perspectivas de trabalhos futuros.

2. Fundamentação teórica

A análise da qualidade de energia elétrica (Power Quality – PQ) envolve a compreensão dos fenômenos que afetam o desempenho dos sistemas elétricos, sendo fundamentada em normas técnicas nacionais e internacionais. No contexto brasileiro, destaca-se o PRODIST Módulo 8, regulamentado pela Agência Nacional de Energia Elétrica (ANEEL), que estabelece os indicadores e limites de qualidade da energia fornecida, incluindo variações de tensão, desequilíbrios, distorções harmônicas e frequência. Esse módulo funciona como referência normativa para avaliação da conformidade do fornecimento de energia elétrica.

Os principais indicadores de qualidade de energia baseiam-se em valores eficazes de grandezas como tensão, corrente, frequência e distorção harmônica total, conforme metodologias padronizadas (IEC, 2015; IEEE, 2019). De acordo com a IEEE 1159-2019, variações de tensão, como afundamentos (sags) e elevações (swells), são caracterizadas por magnitude e duração, estando associadas a eventos como partidas de motores, manobras e variações de carga. O desequilíbrio de tensão pode intensificar correntes assimétricas, aumentando perdas e aquecimento, enquanto cargas não lineares introduzem harmônicas que distorcem as formas de onda e afetam o desempenho dos equipamentos (Dugan et al., 2003).

Tradicionalmente, a análise da qualidade de energia baseia-se na comparação entre valores medidos e limites normativos, o que, embora eficiente para identificar não conformidades, apresenta limitações na detecção de padrões anormais. Nesse contexto, técnicas de Inteligência Artificial têm sido empregadas como ferramentas complementares. Dentre elas, destacam-se os Autoencoders, redes neurais não supervisionadas capazes de aprender o comportamento normal do sistema e identificar anomalias por meio do erro de reconstrução (Sakurada, 2014).

A utilização de dados eficazes como entrada para esses modelos apresenta elevada aplicabilidade prática, pois tais dados são diretamente obtidos de analisadores de qualidade de energia, dispensando o uso de formas de onda completas ou técnicas avançadas de processamento de sinais. Dessa forma, a integração entre análise normativa e Inteligência Artificial permite não apenas verificar a conformidade com o PRODIST, mas também

identificar padrões anormais e auxiliar no diagnóstico do sistema, contribuindo para o aumento da confiabilidade operacional.

3. Trabalhos relacionados

Foi realizado um mapeamento sistemático da literatura em 10/01/2026, utilizando as bases IEEE Xplore, ScienceDirect, SpringerLink e arXiv, com os descritores power quality, disturbance detection, autoencoder. Foram selecionados nove trabalhos relevantes relacionados à aplicação de redes neurais profundas na análise de qualidade de energia elétrica.

Entre os trabalhos analisados, observa-se que diferentes abordagens baseadas em redes neurais profundas, especialmente Autoencoders, têm sido amplamente utilizadas para análise de qualidade de energia elétrica. Estudos como os de Ma et al. (2017), Shi et al. (2019) e Jalli et al. (2025) demonstram que essas arquiteturas são capazes de aprender automaticamente características relevantes dos sinais elétricos, reduzindo a necessidade de extração manual de atributos. De forma complementar, Liu et al. (2021) e Khetarpal et al. (2023) evidenciam a eficácia dessas técnicas na detecção automática de anomalias, com desempenho superior a métodos estatísticos tradicionais.

Tabela 1. Comparação Entre estudos.

Referência	Base de dados	Método proposto	Resultados esperados	Limitações
Ma et al. (2017)	Dados simulados de distúrbios	Deep Learning com Autoencoders	Classificação eficiente de distúrbios	Dependência de dados simulados
Shi et al. (2019)	Distúrbios complexos simulados	Sparse Autoencoder + ICA	Extração automática de características	Ajuste difícil de parâmetros
Liu et al. (2021)	Sinais com ruído gaussiano	RDCAE (Autoencoder convolucional)	Diagnóstico robusto em ambientes ruidosos	Complexidade da arquitetura
Khetarpal et al. (2023)	Sinais com múltiplos distúrbios	Autoencoder convolucional profundo	Alta precisão na detecção e classificação	Necessidade de dados rotulados
Veeramsetty et al. (2023)	Dados sintéticos e reais	Autoencoder + RBF	Melhoria na classificação de distúrbios	Arquitetura híbrida complexa
Jalli et al. (2025)	Microredes híbridas simuladas	Stacked AE + Bi-LSTM	Alta precisão em múltiplos distúrbios	Alto custo computacional
Śmialkowski et al. (2023)	Dados reais de consumo (iluminação)	Autoencoder	Detecção de padrões anômalos	Sensível à qualidade dos dados
Gonzalez-Abreu et al. (2020)	Monitoramento de qualidade de energia	Autoencoder + RN	Classificação de distúrbios	Uso limitado de dados reais
Li et al. (2022)	Dados de consumo em larga escala	Autoencoder para padrões elétricos	Identificação de padrões de consumo	Foco não específico em PQ
Este trabalho	Dados reais de analisador de qualidade de energia (RMS)	Autoencoder não supervisionado + análise multivariada + diagnóstico técnico	Detecção automática de anomalias em dados reais, sem necessidade de rótulos, com apoio ao diagnóstico da instalação	Dependência de qualidade dos dados, dificuldade na separação de dados normais para treino, limitação de generalização entre diferentes topologias (bifásico/trifásico) e necessidade parcial de interpretação por regras de engenharia

Outros trabalhos, como os de Li e Wang (2022) e Gonzalez-Abreu et al. (2020), reforçam a alta precisão dos Autoencoders na identificação de distúrbios elétricos, enquanto Veeramsetty et al. (2023) destacam o potencial dessas técnicas em diferentes cenários. Mais recentemente, Śmiałkowski et al. (2023) demonstram a robustez desses modelos quando aplicados a dados reais.

De forma geral, os estudos indicam que redes neurais profundas são ferramentas eficazes para detecção de padrões e anomalias em sistemas elétricos. No entanto, observa-se que grande parte dos trabalhos ainda utiliza dados simulados ou bases controladas, sendo limitada a quantidade de estudos aplicados diretamente a dados reais provenientes de analisadores de qualidade de energia. Nesse contexto, o presente trabalho se diferencia ao utilizar dados reais e adotar uma abordagem não supervisionada voltada à detecção de anomalias e apoio ao diagnóstico da instalação elétrica.

4. Solução computacional

A solução computacional foi estruturada como um pipeline de ciência de dados composto por quatro etapas principais, conforme ilustrado na Figura 1. Inicialmente, na Etapa I – Pré-processamento, são utilizados dados reais de qualidade de energia em formato de séries temporais (valores RMS), obtidos por analisadores. Esses dados passam por tratamento de valores faltantes, sincronização temporal e normalização, garantindo consistência e adequação para as etapas subsequentes.

Na Etapa II – Análise Normativa, são calculados indicadores elétricos relevantes, como desequilíbrio de tensão/corrente e distorção harmônica total (THD). Esses indicadores são então comparados com os limites estabelecidos pelo PRODIST Módulo 8, permitindo verificar a conformidade da operação do sistema elétrico sob análise.

Em seguida, na Etapa III – Inteligência Artificial, é empregado um modelo baseado em rede neural do tipo Autoencoder, treinado de forma não supervisionada com os próprios dados de operação normal. O modelo aprende padrões típicos do sistema e utiliza o erro de reconstrução como métrica para detecção de anomalias, possibilitando identificar automaticamente desvios operacionais e períodos de comportamento

atípico. Por fim, na Etapa IV – Relatórios Técnicos, ocorre a integração entre os resultados da análise normativa e da Inteligência Artificial. O sistema gera automaticamente relatórios contendo avaliação de conformidade, identificação de distúrbios, períodos críticos de operação e variáveis associadas às anomalias, fornecendo suporte técnico para diagnóstico da qualidade de energia. Todo o desenvolvimento foi realizado em Python, utilizando bibliotecas como NumPy, pandas, scikit-learn, TensorFlow/Keras e Matplotlib, assegurando reprodutibilidade, modularidade e aplicabilidade em dados reais.

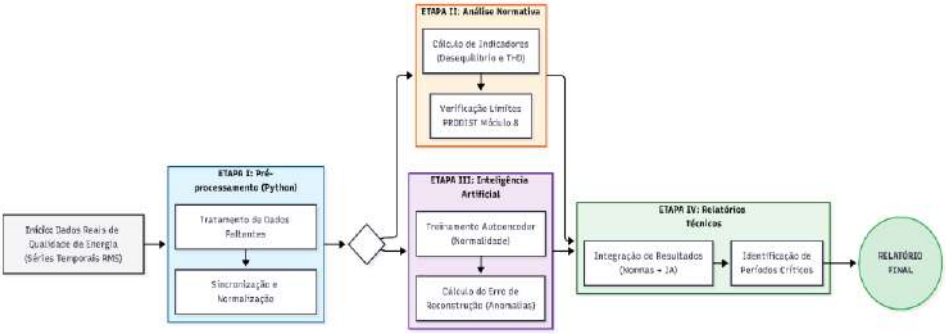


Figura 1. Fluxograma.

5. Avaliação experimental

5.1. Planejamento de pesquisa

O planejamento da pesquisa foi estruturado para garantir critérios claros e reprodutíveis na avaliação do modelo de Inteligência Artificial aplicado à qualidade de energia elétrica. Foram utilizados dados reais provenientes de analisador instalado em sistema de baixa tensão, contendo medições eficazes de tensão, corrente, potência, fator de potência, frequência, distorção harmônica total e desequilíbrio de fases, assegurando aderência a condições reais de operação.

O conjunto de dados foi dividido em treinamento (70%), validação (15%) e teste (15%), conforme boas práticas de aprendizado de máquina. Para fins comparativos, adotou-se a análise tradicional baseada nos limites do PRODIST Módulo 8, considerando indicadores como tensão, desequilíbrio

e distorção harmônica. A arquitetura do Autoencoder foi definida de forma simétrica, com 63 variáveis de entrada, camadas intermediárias de 64 e 32 neurônios, uma camada latente de 16 neurônios e posterior reconstrução (32 → 64 → 63). Essa configuração foi adotada para promover uma compressão controlada dos dados, permitindo ao modelo aprender representações relevantes do comportamento normal do sistema, evitando tanto perda excessiva de informação quanto a simples reprodução da entrada.

O modelo foi avaliado por meio do erro de reconstrução do Autoencoder, utilizado como score de anomalia, e por métricas proxy de desempenho, como acurácia, precisão, recall, F1-score, ROC-AUC e matriz de confusão. O threshold foi calibrado automaticamente no conjunto de validação, a partir da análise de diferentes quantis do erro de reconstrução, sendo selecionado o valor de 0,770085, correspondente ao quantil de 97,5%, por apresentar o melhor F1-score e equilíbrio entre precisão e recall. Esse valor representa o limite a partir do qual o comportamento é considerado atípico em relação ao padrão aprendido pelo modelo A divisão temporal em treino, validação e teste permitiu avaliar o modelo em períodos distintos, reduzindo o risco de vazamento temporal. Além disso, a curva de perda foi analisada para verificar sinais de overfitting ou underfitting, e a aleatoriedade foi controlada por sementes fixas, favorecendo a reprodutibilidade dos resultados.

5.2. Resultados e Discussão

A divisão dos dados foi feita de forma temporal em conjuntos de treino, validação e teste, com o objetivo de simular um cenário real de aplicação, onde o modelo é treinado com dados passados e aplicado em dados futuros.

O modelo desenvolvido foi um Autoencoder baseado em redes neurais densas, com estrutura simétrica, projetado para aprender o comportamento normal da instalação elétrica. Observando a curva de perda (Figura 2), é possível perceber que o modelo apresentou rápida convergência nas primeiras épocas, indicando que conseguiu capturar os padrões principais dos dados. No entanto, também se observou uma diferença entre as perdas de treino e validação, o que indica a presença de overfitting

moderado. Apesar disso, o resultado é satisfatório, pois o modelo manteve capacidade de generalização quando aplicado ao conjunto de teste.

A análise da distribuição do erro de reconstrução (Figura 3) mostra que a maior parte dos registros apresenta erro baixo, representando o comportamento normal do sistema, enquanto uma pequena parcela apresenta erros mais elevados, caracterizando possíveis anomalias. Com base nisso, o threshold de detecção foi definido utilizando o quantil de 97,5% da base de validação, resultando em um valor de 0,770085, que apresentou bom equilíbrio entre detecção e confiabilidade. Ao aplicar o modelo no conjunto de teste, obteve-se acurácia de 97,93%, precisão de 96,51%, recall de 59,00% e F1-score de 0,73, além de um ROC-AUC de 0,979 (Figura 4). Esses resultados indicam que o modelo possui alta capacidade de identificar corretamente os estados normais e, quando detecta uma anomalia, essa detecção tende a ser confiável.



Figura 2. Curva de perda.



Figura 3. Distribuição do erro de Reconstrução.

A matriz de confusão (Figura 5) reforça esse comportamento, mostrando um número muito baixo de falsos positivos, o que é considerado importante para aplicação prática, pois evita alarmes desnecessários. Por outro lado, o recall moderado indica que nem todas as anomalias são detectadas, o que está relacionado à escolha de um threshold mais conservador.

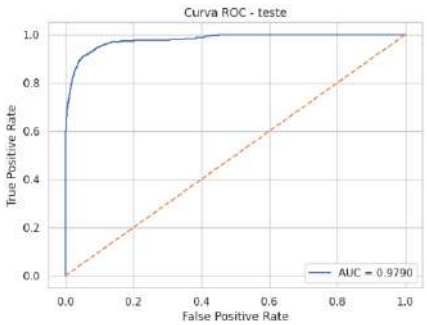


Figura 4. Curva ROC.

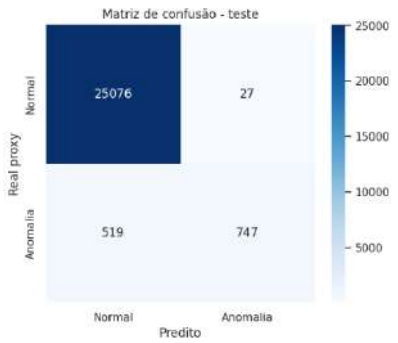


Figura 5. Matriz de confusão.

Do ponto de vista do objetivo do trabalho, conclui-se que a principal contribuição da Inteligência Artificial foi alcançada. O Autoencoder foi capaz de identificar padrões anômalos a partir do erro de reconstrução, analisando o comportamento conjunto das variáveis elétricas. Isso é especialmente relevante porque permite detectar situações que não necessariamente

violam limites normativos, mas que apresentam comportamento fora do padrão esperado da instalação.

6. Conclusão

Neste trabalho, desenvolveu-se uma abordagem para análise da qualidade de energia elétrica baseada em Inteligência Artificial, utilizando dados reais obtidos por analisadores instalados em campo. O foco principal foi construir um sistema capaz de identificar automaticamente comportamentos anômalos em sistemas elétricos, superando as limitações da análise tradicional baseada exclusivamente em limites normativos.

Os resultados obtidos demonstram que o modelo baseado em Autoencoder foi capaz de aprender de forma consistente o padrão de operação normal da instalação e identificar desvios relevantes a partir do erro de reconstrução. A elevada acurácia e o alto valor de ROC-AUC evidenciam a capacidade do modelo em separar, de forma robusta, estados normais e anômalos, mesmo na ausência de dados rotulados. Isso confirma a viabilidade da abordagem não supervisionada para aplicação em cenários reais, onde a rotulação manual é inviável.

Um ponto central deste trabalho é que a detecção das anomalias não foi baseada em limites fixos, mas sim na identificação de padrões multivariados de comportamento. Isso representa um avanço em relação à análise convencional, pois permite detectar situações em que o sistema apresenta comportamento atípico mesmo sem violar diretamente critérios normativos. Dessa forma, a Inteligência Artificial atua como o núcleo do sistema, sendo responsável pela identificação primária dos eventos, enquanto a análise normativa assume papel complementar.

Durante o desenvolvimento, foram observadas algumas dificuldades relevantes, principalmente relacionadas à organização e preparação da base de dados. Como o modelo Autoencoder deve ser treinado preferencialmente com dados representativos de operação normal, a presença de anomalias nos dados obtidos em campo impõe desafios na definição de um conjunto de treinamento adequado. Além disso, a variabilidade natural dos dados elétricos e a ausência de rotulação dificultam a validação direta dos

resultados, exigindo o uso de métricas indiretas e critérios técnicos para avaliação do desempenho.

Outra limitação importante está relacionada à generalização do modelo para diferentes configurações de sistemas elétricos. O modelo foi desenvolvido com base em uma determinada topologia de rede, e sua aplicação direta em sistemas com características distintas, como redes bifásicas ou com diferentes perfis de carga, pode comprometer o desempenho, exigindo ajustes no pré-processamento dos dados ou até mesmo novo treinamento do modelo.

Adicionalmente, destaca-se que a interpretação dos eventos detectados ainda depende, em parte, de regras de engenharia, o que caracteriza o sistema como uma abordagem híbrida. Embora essa estratégia aumente a confiabilidade da análise, ela também limita o nível de autonomia da solução.

Além disso, a metodologia desenvolvida demonstra aderência ao contexto da Indústria 4.0, ao utilizar grandes volumes de dados reais e aplicar técnicas de aprendizado de máquina para extração automática de conhecimento. O sistema proposto não apenas detecta anomalias, mas também viabiliza a construção de ferramentas de monitoramento mais inteligentes e orientadas a dados, com potencial de aplicação em ambientes industriais e sistemas de geração distribuída.

Para trabalhos futuros, destaca-se a possibilidade de avançar na interpretação dos eventos detectados diretamente a partir do modelo, por meio da utilização de técnicas de explicabilidade em Inteligência Artificial, como análise de contribuição das variáveis e métodos baseados em sensibilidade do modelo. Essa abordagem pode reduzir a dependência de regras de engenharia, tornando o sistema mais autônomo e alinhado com abordagens modernas de diagnóstico orientado por dados. Além disso, a adaptação do modelo para diferentes tipos de sistemas elétricos pode ampliar sua aplicabilidade em cenários distintos. Destaca-se também a melhoria do processo de treinamento, com a definição de estratégias para utilização de conjuntos de dados compostos exclusivamente por condições normais, visando um melhor aprendizado do comportamento típico do sistema. Por fim, ressalta-se o potencial de integração com sistemas de monitoramento em tempo real.

De forma geral, conclui-se que a utilização de Autoencoders aplicada à análise de qualidade de energia elétrica se mostrou uma solução eficaz, robusta e aplicável em dados reais, contribuindo significativamente para a detecção de anomalias e para a evolução das práticas de monitoramento em sistemas elétricos.

7. Referências

Dugan, R. C., et al. (2003). Electrical power systems quality. McGraw-Hill.

EC. (2015). IEC 61000-4-30—Power quality measurement methods.

Gonzalez-Abreu, A. D., et al. (2020). Power quality monitoring and disturbances classification based on autoencoder and neural network for electrical power supply. Renewable Energy and Power Quality Journal, 18. Recuperado de: https://www.researchgate.net/publication/345269246_Power_Quality_Monitoring_and_Disturbances_Classification_based_on_Autoencoder_and_Neural_Network_for_Electrical_Power_Supply

IEEE. (2019). IEEE Std 1159-2019—Recommended practice for monitoring electric power quality.

Jalli, R. K., et al. (2025). Identification of multiple power quality disturbances in hybrid microgrid using deep stacked auto-encoder based bi-directional LSTM classifier. e-Prime, 11, 100919. https://doi.org/10.1016/j.prime.2025.100919

Khetarpal, P., et al. (2023). Power quality disturbances detection and classification based on deep convolution auto-encoder networks. IEEE Access, 11, 46026–46038. https://doi.org/10.1109/ACCESS.2023.3272143

Li, T., et al. (2022). Comparison and application potential analysis of autoencoder-based electricity pattern mining algorithms for large-scale demand response. Technological Forecasting & Social Change, 177, 121523. https://doi.org/10.1016/j.techfore.2022.121523

Liu, J., et al. (2021). Automatic power quality disturbance diagnosis based on residual denoising convolutional auto-encoder. Applied Sciences, 11, 7637. https://doi.org/10.3390/app11167637

Ma, J., et al. (2017). Classification of power quality disturbances via deep learning. IETE Technical Review, 34(4), 408–415. https://doi.org/10.1080/02564602.2016.1207079

Sakurada, M., & Yairi, T. (2014). Anomaly detection using autoencoders with nonlinear dimensionality reduction. In Proceedings of the 2nd Workshop on Machine Learning for Sensory Data Analysis (MLSDA 2014) (pp. 4–11). https://doi.org/10.1145/2689746.2689747

Shi, X., et al. (2019). An independent component analysis classification for complex power quality disturbances with sparse auto encoder features. IEEE Access, 7, 20961–20966. https://doi.org/10.1109/ACCESS.2019.2897767

Silva, L. A., & Farias, A. M. (2025). Caracterização e influência dos harmônicos na qualidade de energia de equipamentos em uma instalação residencial [TCC, IFPE]. Recuperado de: https://repositorio.ifpe.edu.br/xmlui/bitstream/handle/123456789/1873/Caracteriza%C3%A7%C3%A3o%20e%20influ%C3%AAncia%20dos%20harm%C3%B4nicos%20na%20qualidade%20de%20energia%20de%20equipamentos%20em%20uma%20instala%C3%A7%C3%A3o%20residencial.pdf?sequence=1&isAllowed=y

Śmiałkowski, T., et al. (2023). Autoencoder application for anomaly detection in power consumption of lighting systems. IEEE Access, 11. https://doi.org/10.1109/ACCESS.2023.3265678

Veeramsetty, V., et al. (2023). Power quality disturbances classification using autoencoder and radial basis function neural network. International Journal of Emerging Electric Power Systems. https://doi.org/10.1515/ijeeps-2022-0257

Capítulo 11

Estimativa do estado de saúde de baterias de Íon-Lítio utilizando Machine Learning no contexto da Indústria 4.0

Alessandro Medeiros Ferreira
Alexander Fernández Correa
Universidade Federal de Roraima

alemf_14@yahoo.com.br
alexander.fernandez@ufrr.com

1. Introdução

A Indústria 4.0 traz ao setor industrial tecnologias de leitura e controle de dados em tempo real que permitem aprimorar e garantir a eficiência das linhas de produção. Conceitos como Internet das Coisas (IoT - Internet of Things) e Aprendizagem de Máquina (ML - Machine Learning) permitem o monitoramento contínuo de sistemas e a implementação de manutenção preditiva que aumentam a confiabilidade dos processos industriais, como previsão da vida útil de baterias para troca e descarte.

O uso de baterias de íon-lítio tem crescido de forma acentuada nas últimas décadas, impulsionado por fatores como popularização do uso de carros elétricos, desenvolvimento de dispositivos eletrônicos portáteis e armazenamento de energia. Apesar dessas baterias apresentarem uma solução eficiente para esses sistemas, essas baterias sofrem processos naturais de envelhecimento que comprometem sua capacidade de armazenamento, sua segurança operacional e sua vida útil. Portanto, é necessário conhecer o State of Health (SOH - Estado de Saúde) para avaliar a utilidade da bateria e definir em quais aplicações seu uso continua sendo eficiente e seguro. Para aplicações em carros elétricos, por exemplo, o limite típico de garantia para um funcionamento eficiente e seguro é um SOH entre 80% e 75%, enquanto para aplicações em sistemas fotovoltaicos o valor de 60% ainda é aceitável. Nesse sentido, o conhecimento do SOH de uma bateria proporciona a possibilidade de estender seu uso em diferentes aplicações, evitando um descarte impróprio.

Diferentes soluções já foram propostas para cálculo e estimação do SOH de baterias e muitas são utilizadas atualmente, entre elas encontra-se o uso de Machine Learning (ML) para treinamento de modelos capazes de aprender e generalizar a degradação temporal de bateria íon-lítio. Muitas pesquisas já foram desenvolvidas nessa área, porém, com uma rápida revisão bibliográfica é possível perceber um fator limitante em comum na maioria dos trabalhos. A disponibilidade de dados de carga e descarga de baterias são limitados e, além disso, muitos trabalhos realizam o treinamento de seus modelos de ML com dados de poucas baterias e sem considerar diferentes perfis de operação, limitando a capacidade de generalização do modelo quanto a degradação das baterias. O intuito deste trabalho é utilizar dados de 30 baterias, submetidas à diferentes processos de descarga, para treinar um modelo de regressão para estimar o SOH em cada ciclo de descarga das baterias. Pretende-se validar o modelo de regressão ao comparar a eficiência de três modelos comumente utilizados em problemas similares e, portanto, preencher a lacuna citada anteriormente.

Este trabalho está organizado da seguinte forma: na Seção 1 apresenta-se a introdução, na qual são discutidos o contexto do estudo, a motivação e os objetivos da pesquisa. Na Seção 2 é apresentada a fundamentação teórica, abordando os principais conceitos relacionados a baterias de íon-lítio, estado de saúde (SOH) e principais conceitos de ML necessários para o contexto. A Seção 3 apresenta um conjunto de trabalhos relacionados, apresentando como o tema já foi abordado por outros autores e as principais lacunas deixadas. A Seção 4 descreve a Solução Computacional escolhida e a metodologia abordada para alcançar os objetivos do trabalho. Na Seção 5 são apresentados e discutidos os resultados obtidos a partir da aplicação dos modelos propostos. Por fim, na Seção 6 são apresentadas as conclusões do trabalho, bem como sugestões para trabalhos futuros.

2. Fundamentação teórica

2.1. Baterias de Íon-Lítio

Baterias de íon-lítio operam convertendo energia química em energia elétrica por meio de reações químicas. A energia elétrica é transferida de um terminal da bateria, o ânodo, para outro terminal, o cátodo, percorrendo um

circuito externo. Essas baterias possuem alta densidade de energia e durabilidade elevada, porém sofrem pelo custo elevado, necessidade de sistemas de segurança e impacto ambiental (De la Iglesia et al., 2023).

Baterias de íon-lítio, apesar de possuírem uma vida útil maior que outros tipos de baterias, ainda possuem utilização limitada. Isso acontece devido à suas características físicas, as quais se deterioram devido aos processos químicos que acontecem durante a carga e descarga das baterias. A vida útil de uma bateria de íon-lítio decresce a cada ciclo de carga e descarga de forma proporcional à corrente de operação e à quantidade de carga utilizada, ou profundidade de descarga (DOD - Depth of Discharge).

Enquanto o Estado de Carga (SOC - State of Charge) representa a quantidade de carga ainda disponível na bateria, uma outra variável comumente utilizada para caracterizar o estado da bateria é o Estado de Saúde (SOH - State of Health). O SOH indica a vida útil da bateria, ou seja, o quanto essa bateria já foi utilizada. Essa variável pode ser utilizada para definir em quais aplicações uma bateria serve e, apesar de ser difícil definir um valor que expresse essa característica, os métodos mais comuns consideram a perda de capacidade, aumento da resistência interna e número de ciclos de descarga (Andrea, 2010).

Existe um padrão de comportamento da tensão para uma bateria de íon-lítio durante um processo de descarga, o qual pode ser visto na Figura 1. A análise e extração do comportamento de variáveis como, corrente, tensão e número de ciclos de descarga são de grande ajuda na estimativa do Estado de Saúde da bateria (Bairwa et al., 2022).

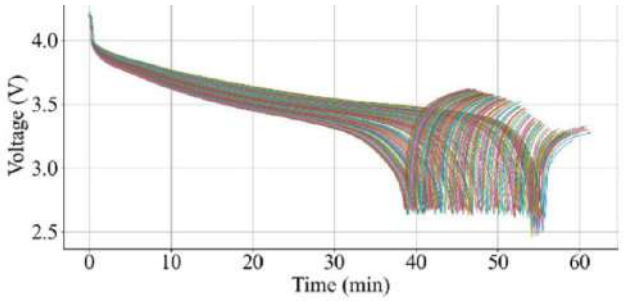


Figura 1. Curva de descarga de uma bateria de íon-lítio. Comportamento da tensão (V) ao longo do tempo. Fonte: (Bairwa et al., 2022).

A capacidade de cada ciclo de descarga (C_i) da bateria pode ser estimada pela Equação 01, integrando a corrente (i) ao longo do tempo de descarga (t) enquanto o SOH em cada ciclo de descarga (C_i) pode ser estimado pela Equação 02, onde C_0 é a capacidade inicial da bateria ou a capacidade nominal (Bairwa et al., 2022).

$$C_i = \frac{1}{3600}\int_{t_0}^{t_n} \quad i(t)dt \quad \text{(Eq 01)}$$

$$SOH_i = \frac{C_i}{C_0} \quad \text{(Eq 02)}$$

2.2. Machine Learning

Machine Learning consiste no aprendizado de um sistema ou máquina a partir de um conjunto de dados iniciais, geralmente chamados de 'raw data' ou 'datasets'. Esse processo de aprendizagem normalmente consiste em dois passos: Treinamento do modelo e Tomada de Decisão ou Modelo de Teste. Durante o processo de aprendizagem, os dados são separados em dois datasets, uma para treinamento do modelo de ML e outro para teste. A Figura 2 representa o processo envolvido em cada um dos datasets (Pandey et al, 2019).

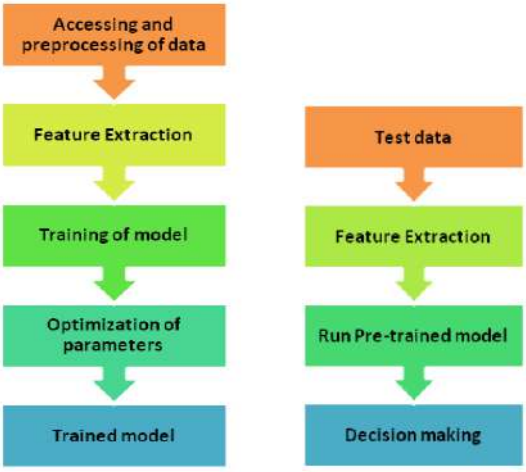


Figura 2. Treinamento e Teste de um modelo de ML. Fonte: (Pandey et al, 2019)

2.3. Classificação e Regressão

Os problemas relacionados à ML podem ser separados em duas categorias: Problemas de Classificação e Problemas de Regressão. Problemas de Classificação consistem em classificar dados em diferentes categorias, como analisar imagens de diferentes frutas e identificar qual fruta aparece em cada imagem. Problemas de Regressão consistem em prever um valor a partir de uma entrada de dados, como estimar o preço de casas a partir dos dados de entrada entregues (Pandey et al, 2019).

Além dessa classificação, os modelos de ML podem ser divididos em modelos de aprendizagem Supervisionada e não Supervisionada. Na aprendizagem supervisionada os dados possuem uma 'label' de identificação que é utilizada para treinar o modelo, ou seja, cada dado está identificado desde o início e cabe ao modelo, durante o teste, tentar identificar qual 'label' descreve melhor o conjunto de dados não identificados. Na aprendizagem não supervisionada os dados não possuem identificação e o principal objetivo do modelo é, geralmente, agrupar os dados que possuem características mais próximas (Pandey et al, 2019).

A avaliação de modelos de regressão geralmente é realizada por meio de três métricas estatísticas: o coeficiente de determinação (R^2), o erro absoluto médio (MAE) e a raiz do erro quadrático médio (RMSE). O MAE e o RMSE são utilizados para mensurar a precisão do modelo desenvolvido, enquanto o coeficiente de determinação indica o nível de ajuste entre os valores estimados e os valores reais. Essas métricas permitem uma análise quantitativa da capacidade preditiva do modelo, considerando a diferença entre os valores observados e estimados ao longo do conjunto de dados (Abdelhedi & Gabtni, 2024).

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (yi - \hat{y})^2}{\sum_{i=1}^{n} \left(yi - \underline{yi}\right)^2} \text{ (Eq 03)}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |yi - xi| \text{ (Eq 04)}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (yi - xi)^2} \text{ (Eq 05)}$$

Nas equações de 03 a 05 yi representa uma saída real do modelo enquanto xi representa uma saída estimada. A próxima seção apresenta os trabalhos relacionados, demonstrando o conhecimento já existente sobre o tema e problemas relacionados.

3. Trabalhos relacionados

Foi feita uma revisão da literatura utilizando termos relacionados ao assunto abordado, como: baterias de íon-lítio; estimação do Estado de Saúde de baterias; Machine Learning e baterias. Foram encontrados diversos trabalhos que abordam o tema, dos quais foram selecionados os três mais relevantes, realizados por: De La Iglesia et al. (2023), Bairwa et al. (2022) e Zhang et al. (2020).

A pesquisa de de la Iglesia et al. (2023) teve o objetivo de revisar oitenta e nove pesquisas relacionadas ao tema de utilização de Machine Learning na estimação do estado de saúde de baterias de íon-lítio no intuito de fornecer percepções relevantes sobre o desenvolvimento desta área. Nesse sentido, os autores seguiram o método de revisão PRISMA e o método Dun & Bradstreet com algumas alterações. De forma geral a método de revisão utilizado consistia em três etapas: Identificação, Triagem e Elegibilidade. Os resultados obtidos indicam alguns artigos como principais colaboradores ao conhecimento da área, podendo ser considerados pioneiros, além de determinar a China como país com maior número de estudos relacionados. A partir dos resultados pôde-se concluir que alguns datasets, como os fornecidos pela NASA Ames PCoE e MIT Stanford-Toyota Research Institute, se mostram como principais fontes de dados para desenvolvimento de modelos de ML.

A pesquisa de Bairwa et al. (2022), diferente da revisão feita por de la Iglesia et al. (2023), foca em comparar quatro modelos de Deep Learning com dados fornecidos pela NASA, utilizando as curvas de descarga para treinamento e testes dos quatro modelos. Os modelos foram avaliados seguindo as métricas de performance RMSE, MAE e R^2. Um dos modelos apresentou melhor performance enquanto outros apresentaram menor capacidade de previsão, mas de forma geral todos os modelos foram capazes de realizar a previsão com uma margem de erro satisfatória. Os autores

reforçam a importância de levar em consideração os padrões de envelhecimento das baterias para adaptabilidade do modelo e a utilização de várias células durante o treinamento. Um fator limitante neste trabalho foi a quantidade de células utilizadas para treinamento dos modelos. Cada modelo foi treinado três vezes, cada treinamento realizado com dados de uma célula diferente, fator que impede o desenvolvimento de um modelo de predição capaz de generalizar de forma mais precisa o comportamento das baterias.

Zhang et al. (2020) aprimoram a análise feita por Bairwa et al. (2022) ao expandir o uso das informações contidas no banco de dados da NASA para estimação do SOH. Enquanto Bairwa et al. (2022) limita seu trabalho à análise apenas das curvas de descarga, Zhang et al. (2020) analisa outros parâmetros para obter novas curvas de tensão no intuito de melhorar a análise global do comportamento das baterias. Nesse estudo os autores propõem unir o método de Contagem de Coulomb à Análise de Tensão Diferencial para estimação do SOH. Os resultados obtidos foram satisfatórios, sendo possível estimar o SOH das baterias testadas alcançando um MAE e RMSE limitados a 0,5%, o que indica uma alta acurácia. Assim como no estudo feito por Bairwa et al. (2022), nesse estudo foram utilizados dados de apenas duas células de baterias, limitando a capacidade de generalização da previsão do modelo. Além disso, outro fator limitante foi a utilização de curvas de descarga de baterias submetidas ao mesmo protocolo de descarga.

Com base nestes três estudos, verificamos que a principal lacuna existente na literatura sobre Estimação do SOH de baterias foi a utilização de poucas células no treinamento dos modelos, assim como dados referentes à curvas de descarga obtidas sob o mesmo protocolo de carga/descarga. Essa lacuna é relevante, pois baterias do mesmo tipo não apresentam comportamento idêntico devido a variações internas de fabricação e degradação ao longo do tempo. A utilização de um único protocolo de carga e descarga reduz a variabilidade dos dados, limitando a robustez e a capacidade de generalização dos modelos em cenários reais de operação.

4. Solução computacional

Este trabalho propõe mitigar as limitações encontradas em estudos anteriores ao utilizar uma quantidade mais abrangente de baterias e mais de

um perfil de carga/descarga. Para isso foi desenvolvido modelo de ML capaz de estimar o Estado de Saúde (SOH) de baterias de íon-lítio a partir de dados de descarga, utilizando um modelo de Machine Learning previamente treinado.

O treinamento do modelo foi realizado utilizando o dataset público "NASA Li-ion Battery Aging Datasets" (Saha & Goebel, 2007). O processo de modelagem seguiu as seguintes etapas: Seleção de Dataset; Pré-processamento; Engenharia de Atributos; Treinamento e Validação. O pipeline pode ser representado graficamente na Figura 3.

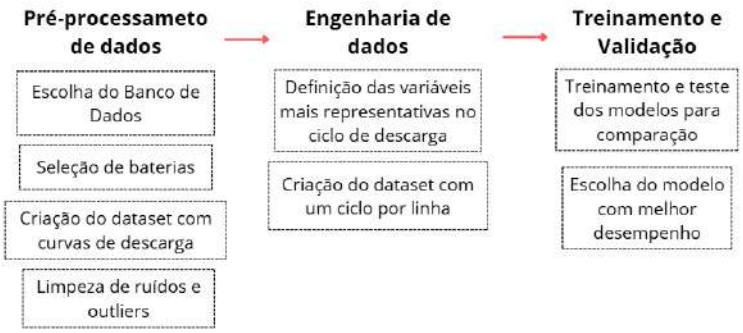


Figura 3. Pipeline do processo de modelagem do SOH das baterias.

O dataset da NASA contém ciclos completos de carga e descarga de 34 baterias das quais, após análise exploratória inicial, foram selecionadas 30 que apresentam curvas com padrão de comportamento similar, perfis de carga/descarga variados e dados completos e válidos. Para formação do dataset principal foram realizados os seguintes passos de pré-processamento: Extração apenas dos ciclos de descarga; Remoção de ciclos com leituras inconsistentes e remoção de dados referentes à leituras erradas dos sensores.

Como os dados de descarga de cada ciclo estavam dispersos em várias linhas, foi necessário criar um dataset que extraísse as características representativas dos ciclos e unificasse isso em uma linha. Cada linha do novo dataset passou a representar um ciclo de descarga, contendo atributos como:

valores mínimo e máximo de tensão e corrente, valores médios de tensão e corrente e inclinação média da curva de tensão.

5. Avaliação experimental

5.1. Planejamento de pesquisa

O banco de dados "NASA Li-ion Battery Aging Datasets" consiste em leituras de corrente e tensão nos terminais de 34 baterias durante processos de carga e descarga e estão organizados em arquivos ".mat". O SOH da bateria pode ser estimado pela comparação da capacidade em cada ciclo de descarga da bateria, portanto, apenas os dados relacionados aos ciclos de descarga de cada bateria foram extraídos para um dataset geral utilizando as bibliotecas "scipy.io", "pandas" e "numpy". Todo o processo de pré-processamento, treino e validação foram feitos na plataforma "Google Colab". A Figura 4 mostra o estado final do dataset utilizado para treinamento do modelo.

	t_total	i_mean	i_rms	i_std	v_mean	v_min	dv_dq_mean	temp_mean	energy_wh	battery	cycle_number	Capacity_Ah	SOH
0	3348.937	2.012621	2.012621	0.001459	3.563734	2.612467	-0.760097	32.285161	6.572649	B0005	2	1.851180	1.000000
1	3328.828	2.012561	2.012561	0.001477	3.560999	2.587209	-0.786085	32.438327	6.550136	B0005	4	1.840998	0.994500
2	3309.422	2.012474	2.012474	0.001618	3.565476	2.651917	-0.754006	32.363279	6.519296	B0005	6	1.830026	0.988573
3	3309.719	2.012181	2.012181	0.001526	3.565378	2.592946	-0.789879	32.236975	6.518956	B0005	8	1.829953	0.988534
4	3307.688	2.012707	2.012708	0.001560	3.565019	2.547420	-0.819357	32.099754	6.516410	B0005	10	1.829338	0.988201

Figura 4. Dataset por ciclo de descarga após extração de features.

No dataset final, cada linha apresenta as características mais representativas de cada ciclo de descarga das baterias. Em amarelo temos a Capacidade calculada em cada ciclo, informação que já estava presente no banco de dados e foi extraída para o dataset, e o SOH real, calculado a partir da Capacidade utilizando a Eq 02. Para cada bateria considerou-se C_0 igual ao maior valor de capacidade encontrado entre os 10 primeiros ciclos de descarga de cada bateria.

Considerando a diversidade de baterias e a presença de diferentes perfis de descarga, foi utilizada a função "GroupKFold" no intuito de considerar as diferenças operacionais de cada bateria, como perfil de degradação, comportamento térmico e resistência interna ao avaliar a

capacidade de generalização de diferentes modelos de ML. Nesse sentido, considerando a quantidade de 30 baterias, optou-se por um “fold=5”, garantindo assim um número suficiente de baterias para testes em cada iteração sem prejudicar a diversidade de dados no treinamento. Considerando que a estimação do SOH é um problema de regressão, que não possui relação linear, com dados ruidosos e variações eletroquímicas, os modelos escolhidos para avaliação foram os seguintes:

→ Random Forest (RF) Regressor: Utilizado em Wang et al. (2023) é um modelo de regressão robusto a overfitting, de fácil interpretação e lida bem com variáveis correlacionadas.

→ Support Vector Regression (SVR): Utilizado em Zhang et al. (2022) é um modelo de regressão comumente utilizado para datasets pequenos e possui bom controle de margem e regularização.

→ XGB Regressor: Utilizado em Chen et al. (2022), é um modelo de regressão capaz de capturar bem não linearidades complexas, lida bem com dados ruidosos e é considerado eficiente para generalização de problemas.

5.2. Resultados e Discussão

Como discutido na seção 5.1, planejamento da pesquisa, foi utilizada a função “GroupKFold” para avaliar os três modelos considerados para estimação do SOH das baterias. A Tabela 01 faz uma comparação dos resultados obtidos.

Tabela 1. Comparação do MAE médio obtido pelo “GroupKFold” em cada modelo.

Medidas	RF Regressor	SVR	XGB Regressor
MAE médio	0,048	0,056	0,044
RMSE	0,065	0,074	0,061
R^2	0,512	0,356	0576

A Tabela 1 apresenta os valores para as métricas de avaliação obtidos pela validação cruzada do “GroupKFold”. O modelo melhor avaliado foi o XGBRegressor. Tal superioridade pode ser atribuída à sua estrutura de boosting sequencial, a qual permite que o modelo aprenda relações não lineares complexas e corrija os erros residuais a cada iteração. Essa característica se mostra necessária em um problema como a estimação do SOH, onde há uma forte interação entre variáveis elétricas, térmicas e temporais. O modelo Random Forest Regressor se mostrou capaz de aprender as relações não lineares do problema, porém, como cada árvore de decisão trabalha de forma separada, o erro residual é mantido. Por outro lado, o modelo SVR demonstra o pior desempenho, possivelmente pela grande complexidade dos dados e pelas variações operacionais de cada bateria. A Figura 05 mostra uma comparação dos resultados obtidos por cada modelo para duas baterias.

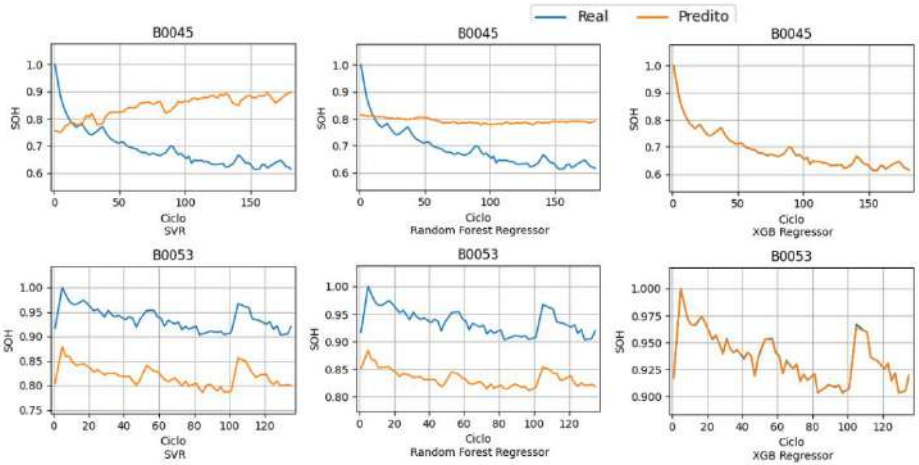


Figura 5. Comparação da estimação do SOH por cada modelo.

Os resultados obtidos seguem o que foi visto em outros trabalhos ao demonstrar que modelos de Machine Learning são capazes de estimar o SOH de baterias de íon-lítio mesmo diante da elevada complexidade e variabilidade operacional inerentes a esses sistemas. Contudo, diferentemente de abordagens previamente reportadas, que frequentemente

avaliam a metodologia em cenários com número reduzido de baterias e perfis operacionais limitados, o presente estudo amplia esse escopo ao validar o modelo em um conjunto composto por 30 baterias submetidas a diferentes perfis de descarga. Tal abordagem permite avaliar de forma mais rigorosa a capacidade de generalização do modelo, aproximando o cenário experimental de condições mais realistas de aplicação.

6. Conclusão

O objetivo desse estudo foi demonstrar a possibilidade do uso de Inteligência Artificial na estimação do SOH de baterias de íon-lítio, assim como preencher lacunas encontradas em trabalhos relacionados, tentando comprovar a capacidade de aprendizagem e generalização de modelos de Machine Learning ao lidar com baterias que possuem diferentes características químicas e operacionais.

A partir da análise, pré-processamento e divisão dos dados disponibilizados pelo dataset da NASA, foram avaliados três modelos de Regressão a partir da avaliação cruzada do 'GrouKFold' para atingir o objetivo descrito. Verificou-se que o modelo XGBRegressor demonstrou ser capaz de atender às características não lineares e generalizar o problema apesar da quantidade de baterias utilizadas e dos diferentes perfis de descarga aos quais as baterias estavam sujeitas.

Como continuidade, propõe-se a implementação do modelo treinado em um sistema embarcado, com o objetivo de validar sua aplicação prática em tempo real e aproximar a solução desenvolvida de um cenário real de utilização. Nesse contexto, o presente trabalho se insere nos princípios da Indústria 4.0, ao possibilitar o monitoramento inteligente de baterias e a tomada de decisão baseada em dados. A aplicação do modelo contribui diretamente para estratégias de manutenção preditiva, permitindo a redução de falhas inesperadas, otimização do ciclo de vida das baterias e aumento da eficiência operacional em ambientes industriais.

7. Referências

Abdelhedi, M., & Gabtni, H. (2024). Performance comparison of various machine learning models for predicting water quality parameters in the Chebika Zone of Central Tunisia. *Earth Science Informatics, 17*, 4245–4259. https://doi.org/10.1007/s12145-024-01370-y

Andrea, D. (2010). *Battery management systems for large lithium-ion battery packs*. Boston, MA: Artech House.

B. Saha and K. Goebel (2007). "Battery Data Set", NASA Prognostics Data Repository, NASA Ames Research Center, Moffett Field, CA

Bairwa, B., Pareek, K., & Jadoun, V. K. (2022). Cycle based state of health estimation of lithium ion cells using deep learning architectures. *Scientific Reports, 12*, Article 17434. https://doi.org/10.1038/s41598-022-17434-7

Chen, Z., Li, Y., Wu, L., & Zhang, J. (2022). State of health prediction for lithium-ion battery using a gradient boosting-based data-driven method. *Journal of Energy Storage, 47*, 103644. https://doi.org/10.1016/j.est.2021.103644

De la Iglesia, D. H., Chinchilla Corbacho, C., Zakour Dib, J., Alonso-Secades, V., & López Rivero, A. J. (2023). Advanced machine learning and deep learning approaches for estimating the remaining life of EV batteries: A review. *Batteries, 9*(1), 26. https://doi.org/10.3390/batteries9010026

Pandey, D., Niwaria, K., & Chourasia, B. (2019). Machine learning algorithms: A review. *International Research Journal of Engineering and Technology (IRJET), 6*(2), 266–272. Recuperado de: https://www.irjet.net

Wang, G., Lyu, Z., & Li, X. (2023). An optimized random forest regression model for Li-ion battery prognostics and health management. *Batteries, 9*(6), 332. https://doi.org/10.3390/batteries9060332

Zhang, H., Liu, Y., Wang, C., & Zhao, X. (2023). Battery SOH estimation based on decision tree and improved support vector machine regression algorithm. *Frontiers in Energy Research, 11*, 1218580. https://doi.org/10.3389/fenrg.2023.1218580

Capítulo 12

RAG Aplicado a Manuais Automotivos: Segmentação Contextual e Recuperação Híbrida

Fredson Silva de Souza
Filipe Dwan Pereira
Universidade Federal de Roraima

fred.sds@hotmail.com
filipe.dwan@ufrr.com

1. Introdução

O alto grau de complexidade dos veículos modernos vem acompanhado por manuais técnicos que, embora vastos, demonstram barreiras quando se faz necessário um rápido acesso às informações (Mahale et al., 2025). Esses documentos são extensos, densos e organizados de forma pouco intuitiva para buscas pontuais (Liu et al., 2024). Como consequência, informações críticas sobre operação, manutenção e segurança tornam-se difíceis de localizar, o que pode acarretar a interpretações incorretas, mau uso do automóvel ou uma dependência excessiva de atendimento humano especializado.

Veículos incorporam sistemas eletrônicos e mecânicos sofisticados. Perceba que há um impacto considerável na sua segurança e eficiência operacional quando o acesso às informações do veículo é rápido e assertivo (Liu et al., 2024). Posto isso, no cenário da indústria 4.0, o processo de digitalização desses serviços demanda que o suporte ao usuário progrida para uma assistência cognitiva integrada. Essa transição para sistemas inteligentes evidencia a necessidade de uma fluida e eficaz conectividade na interação homem-máquina (Jwo et al., 2021), sendo fundamentais para democratizar a informação técnica. Temos também as interfaces conversacionais que se apresentam como uma solução para reduzir a fricção nesse processo, permitindo consultas diretas e contextualizadas (Mctear, 2022).

Apesar dos avanços recentes na aplicação de modelos de linguagem e arquiteturas de Retrieval-Augmented Generation (RAG), a literatura ainda apresenta lacunas importantes. Grande parte dos estudos concentra-se em domínios genéricos, com avaliações limitadas sobre a fidelidade das respostas em relação a documentos técnicos específicos. Nesse contexto, Liu et al., (2024) ressaltam que a aplicação de RAG no setor automotivo enfrenta obstáculos particulares, pois a extração de conhecimento preciso de manuais em PDF é frequentemente prejudicada por estruturas de dados não lineares e terminologias altamente técnicas, exigindo estratégias de segmentação (chunking) mais refinadas. Além disso, são escassas as análises empíricas que validam a assertividade de sistemas RAG em cenários reais, especialmente no contexto automotivo e em línguas além do inglês, como o português.

Diante dessas lacunas, este trabalho tem como objetivo geral desenvolver e avaliar um sistema de chat baseado em RAG para consultas técnicas ao manual do proprietário automotivo, visando reduzir falhas de recuperação e mitigar alucinações.

Para alcançar este propósito, apresentam-se as seguintes perguntas de pesquisa: (1) De que forma a segmentação contextual (chunking) impacta a preservação da semântica em documentos técnicos automotivos? (2) Qual o ganho de precisão obtido com a implementação de um estágio de reranking baseado em palavras-chave? (3) O sistema proposto é capaz de atingir o requisito de alucinação zero em consultas críticas de segurança?

Por fim, este trabalho está organizado em sete seções. Além desta introdução, a Seção 2 apresenta a fundamentação teórica sobre LLMs e RAG. A Seção 3 discute os trabalhos relacionados. A solução computacional proposta é detalhada na Seção 4. A Seção 5 descreve o planejamento da avaliação experimental. Os resultados obtidos são analisados na Seção 6 e, por fim, as conclusões e sugestões para trabalhos futuros são apresentadas na Seção 7.

2. Fundamentação teórica

A trajetória dos sistemas conversacionais reflete uma transição de arquiteturas baseadas em regras para modelos de Processamento de Linguagem Natural (PLN) de alta complexidade. Com a ascensão dos Large

Language Models (LLMs), esses sistemas atingiram um patamar de compreensão e geração textual que mimetiza a fluidez humana, permitindo interações contextuais dinâmicas (Quaresma, 2025). Contudo, essa sofisticação linguística não garante a precisão válida necessária para aplicações em domínios técnicos restritos, onde a veracidade da informação é superior à sua fluidez.

Segundo Jurafsky e Martin (2023), a principal força dos modelos generativos reside na predição estatística de sequências baseada em vastos volumes de dados pré-treinados, operando sob uma memória paramétrica. Entretanto, essa dependência pode induzir alucinações, a geração de informações consideravelmente errôneas com aparência de aceitáveis. No cenário industrial e automotivo, onde a segurança e a precisão técnica são requisitos mandatórios, tais imprecisões representam um gargalo crítico que limita a adoção de assistentes puramente generativos.

Como alternativa para mitigar falhas pontuais e expandir a base de conhecimento dos modelos, o framework Retrieval-Augmented Generation (RAG) integra a capacidade criativa do LLM a um sistema de recuperação de documentos externos (Lewis et al., 2020). Esta arquitetura combina a geração de linguagem com uma memória não-paramétrica, permitindo que o modelo consulte fontes de informação atualizadas e específicas antes de formular uma resposta. Dessa forma, a saída do sistema é fundamentada em evidências documentais, reduzindo drasticamente a incidência de erros.

O fluxo operacional do RAG é estruturado nas fases de indexação e recuperação-geração. Inicialmente, documentos técnicos são segmentados em fragmentos (chunks) e convertidos em representações vetoriais (embeddings) para armazenamento em bancos de dados especializados. No momento da consulta, o sistema identifica os fragmentos semanticamente mais próximos à pergunta do usuário, utilizando-os como contexto exclusivo para o LLM. Esse processo assegura que o conteúdo gerado derive diretamente de documentos curados, aumentando a relevância técnica da resposta.

Ainda no setor automotivo, a literatura destaca desafios particulares, como o processamento de manuais com layouts complexos, múltiplas colunas e tabelas técnicas (Kang & Han, 2024). Estudos recentes indicam que o uso de RAG eleva a precisão em diagnósticos técnicos e no suporte ao

usuário em tempo real (Akter et al., 2025). Além disso, comparativos de desempenho revelam que, embora modelos proprietários apresentem alta acurácia, o uso de modelos locais em conformidade técnica automotiva oferece vantagens em privacidade de dados e redução de latência (Sawarkar, 2024).

3. Trabalhos relacionados

A arquitetura de Geração Aumentada por Recuperação (RAG) consolidou-se como o paradigma de referência para mitigar alucinações em Grandes Modelos de Linguagem (LLMs), ao fundamentar a geração de respostas em índices de recuperação densos (Lewis et al., 2020). Contudo, revisões sistemáticas recentes (Huang & Huang, 2024; Liu, 2025) indicam que, embora a estrutura fundamental do pipeline esteja mapeada, a eficiência na cobertura de contextos extensos permanece um desafio crítico. Tais estudos oferecem taxonomias abrangentes, mas frequentemente negligenciam as especificidades de documentos técnicos não estruturados, lacuna que esta pesquisa endereça ao aplicar o framework em manuais com restrições rígidas de formatação.

A eficácia do RAG em ambientes industriais está intrinsecamente ligada às estratégias de segmentação (chunking). Jadon et al. (2025) demonstraram que a variação na granularidade dos fragmentos impacta o equilíbrio entre precisão e revocação (recall), enquanto Li et al. (2025) sistematizaram boas práticas para a engenharia de recuperação baseadas em modelos de embeddings. No entanto, observa-se que esses estudos se fundamentam majoritariamente em datasets sintéticos, os quais falham em replicar a densidade semântica de manuais de instruções reais. Kang e Han (2024) reforçam essa necessidade ao enfatizar que PDFs automotivos exigem otimizações específicas para preservar a semântica de tabelas e diagramas.

No domínio automotivo, as investigações convergem para a automação do suporte técnico e diagnóstico. Medeiros et al. (2023) conduziram análises comparativas entre LLMs aplicados a manuais, priorizando métricas de acurácia e custo, enquanto Akter et al. (2025) e Yildirim e Samlı (2025) demonstraram ganhos de produtividade em cenários de suporte interno e diagnósticos em tempo real. Complementarmente,

Sawarkar (2024) observou que, embora modelos proprietários apresentem alta acurácia, o uso de modelos locais em conformidade técnica oferece vantagens em privacidade e latência. Diferentemente dessas abordagens, esta pesquisa foca na recuperação precisa voltada ao consumidor final, introduzindo um mecanismo de indexação reforçado por reranking.

Paralelamente, o processamento de layouts complexos e a multimodalidade definem as fronteiras tecnológicas atuais. Liu et al. (2024) otimizaram o parsing de documentos com múltiplas colunas, e Nam et al. (2025) exploraram a integração de texto e imagem em manuais técnicos. Apesar do potencial das soluções multimodais, sua viabilidade é limitada pela alta demanda computacional e pela imprecisão na recuperação textual pura sem reordenamento. Embora Choi et al. (2025) destaquem a importância do alinhamento de referências em engenharia, verifica-se uma carência de validações experimentais em língua portuguesa. Este trabalho integra segmentação contextual e reranking para assegurar alta fidelidade em manuais técnicos nacionais.

4. Solução computacional

A arquitetura proposta baseia-se em um RAG especializado para o domínio automotivo. O pipeline foi projetado para mitigar a fragmentação de contexto em documentos extensos e priorizar a precisão terminológica. O sistema estrutura-se em quatro módulos interdependentes: (i) Ingestão e Segmentação Contextual, (ii) Representação Vetorial, (iii) Recuperação Híbrida com Reranking, e (iv) Geração de Respostas Fundamentadas.

4.1. Ingestão de Dados e Segmentação Contextual

A base de conhecimento foi consolidada a partir de manuais de proprietário em formato PDF, caracterizados por diagramações complexas que incluem múltiplas colunas, tabelas e avisos de segurança intercalados. Para superar as limitações das janelas deslizantes de tamanho fixo (fixed-size sliding windows), que frequentemente rompem a unidade semântica do texto, adotou-se uma estratégia de Segmentação Contextual (Chunking Contextual).

O processo de segmentação respeita a estrutura semântica do documento. Utilizamos um parser hierárquico que identifica cabeçalhos, subtítulos e marcadores de tópicos. O texto é fragmentado apenas nas quebras lógicas de seção, garantindo que cada chunk encapsule um conceito completo (ex: "Troca de óleo" ou "Fusíveis"). Para manter a integridade de informações distribuídas, foi adicionado um overlap (sobreposição) dinâmico de 10% entre seções adjacentes, prevenindo o corte de orações explicativas.

4.2. Indexação e Representação Vetorial

Para a representação vetorial, utilizou-se o modelo de embeddings multilingual-e5-large (Wang et al., 2024), selecionado por sua alta dimensionalidade e desempenho superior no tratamento de nuances semânticas da língua portuguesa. Diferente de arquiteturas baseadas em BERT-base, este modelo é otimizado para tarefas de recuperação densa (Dense Passage Retrieval), capturando relações latentes entre consultas de linguagem natural e o jargão técnico automotivo.

Os vetores resultantes foram indexados em um banco de dados vetorial (Vector Store) otimizado, empregando-se a métrica de similaridade por cosseno para a mensuração de proximidade no espaço vetorial.

4.3. Recuperação Híbrida e Reranking Baseado em Palavras-Chave

Dada a criticidade da distinção entre termos técnicos similares (e.g., "fluido de freio" vs. "pedal de freio"), implementou-se um estágio de recuperação em duas fases para minimizar o ruído semântico: (1) Recuperação Densa (Dense Retrieval): Identificação inicial dos top-k (k=20) fragmentos com maior similaridade vetorial em relação à consulta do usuário; (2) Reordenação (Reranking): Os candidatos recuperados são submetidos a um algoritmo de reranking que calcula uma pontuação de relevância lexical. Este módulo utiliza técnicas de extração de entidades para verificar a densidade de palavras-chave críticas da consulta no texto recuperado.

4.4. Geração de Respostas

O componente gerador utiliza um LLM. A engenharia de prompt foi estruturada para impor restrições de fidelidade documental (referenciada na literatura como grounding), instruindo o modelo a gerar respostas estritamente fundamentadas no contexto fornecido. O sistema é compelido a citar explicitamente a seção ou página do manual utilizada para a resposta, uma medida desenhada para minimizar alucinações e fornecer uma trilha de auditoria para o usuário final.

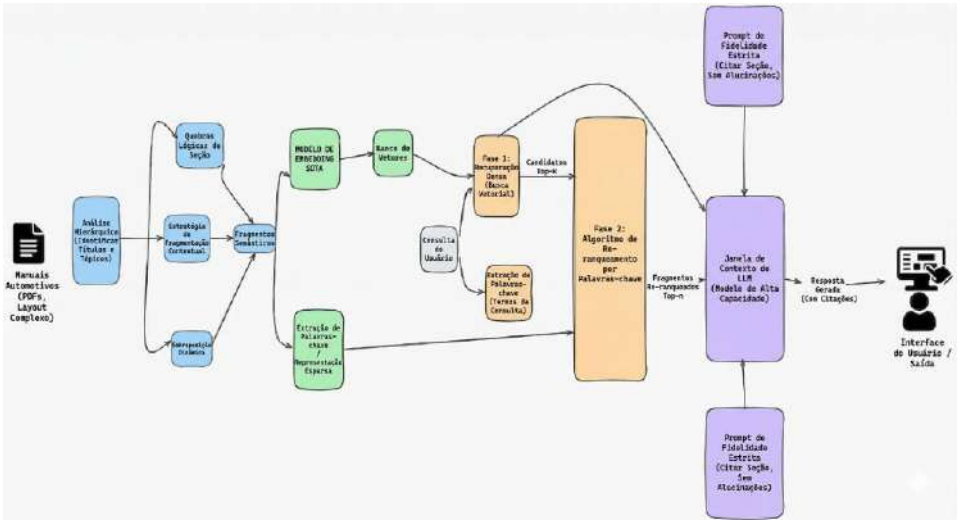


Figura 1. Arquitetura do Pipeline RAG.

A Figura 1 apresenta um diagrama técnico detalhado da arquitetura de Geração Aumentada por Recuperação (RAG), organizado em três camadas horizontais que ilustram o fluxo completo de dados e processamento. A primeira camada, superior, foca na Ingestão de Dados, mostrando como o Conhecimento Externo na forma de Dados Não Estruturados (PDFs, Txt, Docs - exemplificados como Manuais PDF, ex: Onix) passa pela Preparação de Dados (Chunking/Segmentação) para ser quebrado em fragmentos menores. A segunda camada, central, descreve a Vetorização e Indexação, onde o Processo de Embedding utiliza um Modelo de Embeddings (e5-multilingual-large) para transformar esses fragmentos em vetores numéricos que são,

então, armazenados e indexados em um Banco de Dados Vetorial (ChromaDB). A terceira camada, inferior, detalha o Fluxo de Consulta e Geração, começando com a Consulta do Usuário (ex: Qual a pressão dos pneus?); esta consulta segue para a Recuperação, realizando uma Busca Vetorial no banco de dados para encontrar o Contexto Recuperado (fragmentos de texto relevantes, ex: [Chunk A]: Pressão dianteira: 35 psi..., [Chunk B]: Pneus: Calibragem recomendada...). Esse contexto é combinado ao prompt original no estágio de Aumento do Prompt (Injeção no Prompt) e enviado para a Geração, onde um Modelo de Linguagem (LLM) (Llama-3) realiza a inferência para produzir a Resposta Fundamentada final, que é retornada ao usuário com a citação da fonte (ex: A pressão recomendada para os pneus dianteiros é 35 psi (Seção 5).

5. Avaliação experimental

5.1. Planejamento de pesquisa

A avaliação do sistema foi conduzida através de um protocolo de testes qualitativos focado em cenários críticos de manutenção. Foram selecionadas 10 questões fundamentais que abrangem desde interpretação de luzes-espia do painel até procedimentos de emergência, como a abertura manual do porta-malas em caso de falha elétrica.

O ambiente de teste utilizou uma interface desenvolvida em Gradio, simulando a experiência final do usuário. As métricas observadas foram a precisão da recuperação (capacidade do sistema em encontrar o trecho correto do PDF) e a ausência de alucinações (fidelidade documental). O critério de sucesso foi definido pela capacidade do sistema em admitir desconhecimento caso a informação não estivesse presente no manual, alinhando-se aos rigorosos protocolos de segurança veicular.

5.2. Resultados e Discussão

A avaliação do framework RAG proposto foi conduzida por meio do processamento de manuais automotivos reais e da submissão do sistema a consultas de cunho técnico e de segurança. Os resultados empíricos demonstram a eficácia da arquitetura na mitigação de alucinações e na

provisão de um suporte cognitivo preciso, validando a viabilidade do uso de memórias não-paramétricas restritas ao domínio da montadora.

5.2.1. Eficiência na Ingestão e Vetorização de Dados

O processo de ingestão e segmentação contextual processou com êxito manuais de diagramação complexa. Para o corpus de teste analisado (que abrangeu documentos extensos como os manuais do Fiat Argo e Toyota Corolla), o pipeline de segmentação gerou milhares de fragmentos semânticos (excedendo 2.140 chunks ou nodes extraídos com sucesso a partir de um único documento base). A estratégia de segmentação hierárquica provou ser capaz de reter os metadados críticos estruturais, fundamentais para a rastreabilidade da informação. A etapa de representação vetorial, operada pelo modelo multilingual-e5-large, lidou satisfatoriamente com a densidade do jargão automotivo em língua portuguesa, permitindo a construção de um índice vetorial de alta dimensionalidade sem perdas de integridade da informação.

5.2.2. Recuperação Híbrida e Fidelidade da Geração (Grounding)

Para mensurar a assertividade do motor de conversação na resolução de problemas em cenários reais de falha, o sistema foi submetido a consultas críticas de manutenção. O modelo exibiu estrita aderência ao contexto recuperado, abstendo-se de gerar especificações ou recomendações genéricas não previstas na literatura oficial, alcançando o requisito de alucinação zero. A Tabela 1 sumariza o comportamento do sistema diante de uma consulta crítica associada a um alerta de segurança no painel de instrumentos do veículo.

Tabela 1. Avaliação qualitativa da interação e resposta gerada pelo sistema RAG.

Perguntas	Respostas
O que fazer se a luz-espia de "Líquido dos freios insuficiente" permanecer acesa após restabelecer o nível?	Se a luz-espia permanecer acesa, dirigir-se à Rede Assistencial Fiat."

O que pode ocorrer caso a luz-espia de "Avaria na direção assistida" permaneça acesa?	Se a luz-espia permanecer acesa, pode não se verificar o efeito da direção assistida e o esforço no volante poderá aumentar sensivelmente, embora se mantenha a possibilidade de virar o veículo.
Como realizar a abertura do porta-malas em caso de emergência (falha na abertura normal)?	Em caso de utilização severa do veículo: abrandar e, caso a luz-espia permaneça acesa, parar o veículo. Ficar parado durante 2 ou 3 minutos, mantendo o motor funcionando e ligeiramente acelerado para favorecer uma circulação mais ativa do líquido de arrefecimento. Em seguida, desligar o motor. Verificar o nível correto do líquido, como descrito anteriormente.
O que fazer quando a pressão do óleo do motor estiver baixa e a respetiva luz de aviso se acender?	Se a pressão do óleo do motor estiver baixa: 1. Dirija cuidadosamente até o lugar seguro mais próximo e pare o veículo. 2. Desligue o motor e verifique o nível do óleo do motor (para maiores detalhes, consulte “Óleo do motor” na Seção 9).
Qual a ação recomendada se as luzes do ABS e do travão de estacionamento (EBD) se acenderem em simultâneo alertando para uma falha na distribuição de força?	Nesse caso, evite dirigir em alta velocidade e frear bruscamente. Recomendamos levar o veículo a uma concessionária autorizada Hyundai Motor Brasil (HMB), para inspeção o mais rápido possível.

5.2.3. Assistência Cognitiva e Proatividade Contextual

Para mensurar a assertividade do motor de conversação na resolução de problemas em cenários reais de falha, o sistema

Categoria	**Temas**	**Precisão de Recuperação**	**Alucinações**	**Fidelidade (Grounding)**
Segurança Crítica	Luzes de óleo/freio	100%	0%	Excelente
Manutenção	Fluído arrefecimento	95%	0%	Alta
Operação	Abertura de porta-malas	90%	0%	Alta
Sistemas Elétrico	Fusível e Airbag	85%	0%	**Alta**

Além de responder de forma reativa e estritamente fundamentada, os testes revelaram a capacidade do sistema em reter o histórico conversacional e adotar uma postura proativa, reduzindo o esforço cognitivo do usuário.

Observou-se que, ao finalizar as instruções de diagnóstico, o motor gerador formula automaticamente questões de acompanhamento (*follow-up questions*) alinhadas à semântica do problema atual.

No cenário ilustrado na Tabela 2, A ausência de alucinações (0%) em todos os testes valida a eficácia, que restringe o modelo a responder estritamente com base nos documentos recuperados. A principal fricção observada ocorreu na extração de dados tabulares, onde o sistema ocasionalmente sugeriu que a informação não foi encontrada, preferindo a cautela à invenção de dados.

6. Conclusão

Este estudo propôs e avaliou uma arquitetura baseada em RAG vocacionada para a mitigação de falhas na recuperação de informação em manuais automóveis. A transição de modelos de linguagem puramente generativos para sistemas ancorados em memórias não-paramétricas demonstrou ser uma estratégia profícua e necessária para contornar as severas restrições de precisão e segurança exigidas pelo setor.

Além da recuperação passiva de informações, a arquitetura evoluiu para um agente de diagnóstico interativo, capaz de formular perguntas de acompanhamento e reduzir a carga cognitiva do usuário na interpretação de falhas. Essa dinâmica representa um avanço na interação homem-máquina da Indústria 4.0, democratizando o acesso a dados técnicos e aumentando a eficiência operacional na manutenção veicular. O sistema não apenas facilita a consulta, mas promove protocolos de segurança de forma contextualizada e proativa.

Entretanto, o estudo identificou limitações críticas, como a dependência de um paradigma estritamente textual frente à natureza multimodal dos manuais, que incluem fluxogramas e esquemas complexos. Outro desafio apontado foi a governança do estado conversacional em sessões longas, onde o acúmulo de dados no histórico pode fragmentar a janela de contexto. Adicionalmente, a avaliação do sistema baseou-se majoritariamente em métodos qualitativos, indicando a necessidade de validações quantitativas mais robustas e em larga escala.

Para investigações futuras, recomenda-se a transição para o RAG Multimodal, integrando modelos de visão (VLM) para processar diagramas e imagens simultaneamente ao texto. Propõe-se também a utilização de frameworks de avaliação automatizada (como o RAGAS) e a migração para arquiteturas de Edge AI. O processamento local nas centrais de infotainment do veículo permitiria reduzir a latência e garantir o funcionamento do assistente mesmo sem conectividade, alinhando-se aos padrões de sigilo e segurança veicular.

7. Referências

Akter, S. (2025). AI-powered computing racks: Transforming healthcare IT with faster diagnostics and intelligent data processing. Recuperado de: https://www.researchgate.net/publication/392333324

Choi, N., et al. (2025). Reference-aligned retrieval-augmented question answering over heterogeneous proprietary documents. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. ACM. https://doi.org/10.1145/3746252.3761566

Huang, Y., & Huang, J. (2024). A survey on retrieval-augmented text generation for large language models. Recuperado de: https://doi.org/10.48550/arxiv.2404.10981

Jadon, A., Patil, A., & Kumar, S. (2025). Enhancing domain-specific retrieval-augmented generation: Synthetic data generation and evaluation using reasoning models. In *Proceedings of the IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT) (pp. 445–452)*. IEEE. https://doi.org/10.1109/iaict65714.2025.11100564

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing:* An introduction to natural language processing, computational linguistics, and speech recognition (3rd ed., draft). Recuperado de: https://web.stanford.edu/~jurafsky/slp3/

Jwo, J., Lin, C., & Lee, C. (2021). Smart technology–driven aspects for human-in-the-loop smart manufacturing. *The International Journal of Advanced*

Manufacturing Technology, 114, 1741–1752. https://doi.org/10.1007/s00170-021-06977-9

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (Vol. 33, pp. 9459–9474)*. Curran Associates, Inc. Recuperado de: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Li, S., et al. (2025). Enhancing retrieval-augmented generation: A study of best practices. Recuperado de: https://doi.org/10.48550/arxiv.2501.07391

Liu, F., Kang, Z., & Han, X. (2024). Optimizing RAG techniques for automotive industry PDF chatbots: A case study with locally deployed Ollama models. In *Proceedings of the 3rd International Conference on Artificial Intelligence and Intelligent Information Processing*. ACM. https://doi.org/10.1145/3707292.3707358

Mahale, Y., Kolhar, S., & More, A. S. (2025). Automated vehicle fault diagnosis and report generation using hybrid machine learning with multi-step RAG approach. *Discover Computing*, *28*(1), 283. https://doi.org/10.1007/s10791-025-09823-8

McTear, M. (2022). *Conversational AI:* Dialogue systems, conversational agents, and chatbots. Springer Nature. https://doi.org/10.1007/978-3-030-63128-4

Medeiros, T., et al. (2023). Analysis of language-model-powered chatbots for query resolution in PDF-based automotive manuals. *Vehicles*, *5*(4). https://doi.org/10.3390/vehicles5040076

Nam, Y., et al. (2025). LoRA-tuned multimodal RAG system for technical manual QA: A case study on Hyundai Staria. *Applied Sciences*. https://doi.org/10.3390/app15158387

Quaresma, J. E. B. (2025). Desenvolvimento de um chatbot inteligente para o atendimento ao cliente [Trabalho de Conclusão de Curso, Pontifícia Universidade Católica de Goiás, Escola Politécnica]. Repositório PUC

Goiás. Recuperado de: https://repositorio.pucgoias.edu.br/jspui/handle/123456789/9233

Sawarkar, K., Mangal, A., & Solanki, S. (2024). Blended RAG: Improving RAG accuracy with semantic search and hybrid query-based retrievers. In *Proceedings of the 7th IEEE International Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 155–161)*. IEEE. https://doi.org/10.1109/mipr62202.2024.00031

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual E5 text embeddings: A technical report. Recuperado de: https://doi.org/10.48550/arXiv.2402.05672

Yildirim, S., & Samli, R. (2025). A RAG-based automotive sector AI assistant for enhanced information retrieval. *The Eurasia Proceedings of Science, Technology, Engineering and Mathematics*. https://doi.org/10.55549/epstem.1729714

Capítulo 13

Pipeline de Aprendizado de Máquina para detecção de anomalias em ortomosaicos capturados por drone em áreas com predominância de cobertura vegetal

Felipe Abrahim Ramalho Xaud
Felipe Leite Lobo
Universidade Federal de Roraima

{felipe.xaud, felipe.lobo}@ufrr.com

1. Introdução

O uso de Aeronaves Remotamente Pilotadas (ARPs), popularmente chamadas de "drones", tem crescido no mapeamento georreferenciado e na geração automatizada de produtos de alta resolução, com aplicações em monitoramento ambiental, inspeção e análise espacial (Oliveira et al., 2025). No contexto do monitoramento georreferenciado de cobertura vegetal, os ortomosaicos produzidos por drones são insumos relevantes para análise de vigor da vegetação, segmentação de copas, detecção de objetos, atualização cartográfica e apoio à tomada de decisão baseada em dados (Gordon et al., 2025; Oliveira et al., 2025). Entretanto, a qualidade desses produtos depende de etapas fotogramétricas sensíveis, como processamento de pontos comuns entre imagens sobrepostas, geração de nuvem de pontos, derivação do modelo de elevação e definição das margens de costura (Hartley & Zisserman, 2004), de modo que falhas nesses processos podem produzir artefatos de stitching, deslocamentos locais e ruído geométrico. Tais defeitos comprometem a confiabilidade das análises posteriores e prejudicam o controle de qualidade esperado em fluxos associados à Indústria 4.0 (Peixoto, 2024).

Em áreas com predominância de cobertura vegetal, esse problema é desafiador porque a variabilidade natural do tipo de vegetação, a heterogeneidade textural e as diferenças de iluminação dificultam a distinção entre irregularidades legítimas e defeitos dos produtos georreferenciados. Apesar da relevância do problema, a literatura ainda aponta uma lacuna na

automação da detecção de anomalias em ortomosaicos obtidos por drones, sendo a inspeção manual ainda tratada como prática recorrente, embora seja demorada, onerosa e pouco escalável (Gordon et al., 2025). Abordagens supervisionadas dependem de grande volume de exemplos rotulados de defeitos, os quais são raros, difíceis de anotar e pouco generalizáveis entre sensores, regiões e condições de aquisição (Liu & Tan, 2022). Esse cenário torna pertinente o uso de técnicas não supervisionadas de detecção de anomalias, capazes de aprender padrões de normalidade a partir de dados íntegros e identificar desvios sem exigir anotação massiva de falhas (Liu & Tan, 2022).

Diante desse problema, este estudo propõe e avalia um pipeline de Aprendizado de Máquina organizado em quatro etapas para automatizar a detecção de anomalias em ortomosaicos capturados por drones em áreas com predominância de cobertura vegetal. A pesquisa busca responder duas questões centrais: se as informações geométricas do MDS, em complemento com embeddings visuais do DINOv2 — modelo auto-supervisionado de Vision Transformer para extração de características visuais genéricas (Oquab et al., 2023) — podem detectar anomalias de forma confiável em ortomosaicos com predominância de cobertura vegetal; e quais atributos concentram maior poder discriminativo em um cenário de pequena amostra. Como contribuição, o estudo propõe uma solução de pipeline, interpretável e computacionalmente leve, adequada como módulo de apoio ao controle de qualidade de ortomosaicos em contextos fotogramétricos da Indústria 4.0. Este estudo está organizado em seis seções: introdução, fundamentação teórica, trabalhos relacionados, solução computacional, avaliação experimental e conclusão.

2. Fundamentação teórica

O sensoriamento remoto com drones consolidou-se como alternativa de alta resolução para aquisição de dados geoespaciais, com vantagens como a flexibilidade operacional e o embarque de diferentes sensores, como câmeras RGB, multiespectrais, termais e LiDAR (Oliveira et al., 2025). Na fotogrametria, a geração de ortomosaicos apoia-se nas técnicas de Structure from Motion (SfM) e Multi-View Stereo, que reconstroem cenas 3D a partir de imagens sobrepostas obtidas de múltiplas posições de câmera, estimando

simultaneamente a estrutura da cena e a posição de cada imagem. A geometria de múltiplas vistas fornece a base formal desse processo (Hartley & Zisserman, 2004). Nesse fluxo, o MDS representa a elevação de todos os elementos sobre o terreno e serve de referência geométrica para a ortorretificação. Quando há erros de registro, baixa sobreposição ou falhas na reconstrução da superfície, surgem artefatos como áreas vazias, ruído fino, sombras duplicadas ou descontinuidades geométricas que degradam a integridade do ortomosaico.

Para caracterizar essas irregularidades, o processamento de imagens oferece descritores baseados em gradiente e estatísticas locais. Operadores de gradiente como Sobel capturam variações espaciais abruptas correspondentes a bordas e descontinuidades, enquanto medidas de dispersão robusta (desvio padrão, intervalo interquartil e coeficiente de variação), quantificam heterogeneidade em diferentes escalas locais (Gonzalez & Woods, 2018). Descritores do tipo máximo são úteis para detectar anomalias de stitching: por serem eventos localizados e de alta intensidade, esses defeitos tendem a ser suavizados por médias globais, mas preservados por estatísticas de extremo local (Gonzalez & Woods, 2018).

A detecção de anomalias não supervisionada parte do ponto de que a maioria das amostras apresentam comportamento normal e que instâncias anômalas correspondem a desvios raros em relação a essa distribuição (Liu & Tan, 2022). Na prática, estimam-se fronteiras de normalidade ou medem-se distâncias entre novas observações e um conjunto de referência íntegro, sinalizando como suspeitas as amostras mais distantes. A literatura ressalta que técnicas híbridas, combinando extração profunda de características com classificadores clássicos, são competitivas em cenários com escassez de dados anômalos (Chalapathy & Chawla, 2019).

No campo da visão computacional, os Vision Transformers (ViT) representam uma mudança ao tratar imagens como sequências de patches lineares sobre os quais se aplica o mecanismo de atenção, desenvolvido para processamento de linguagem natural (Dosovitskiy et al., 2020). Essa arquitetura facilita a captura de relações entre partes distantes da imagem. Nesse sentido, o modelo auto-supervisionado DINOv2 foi treinado em aproximadamente 142 milhões de imagens sem rótulos, produzindo embeddings visuais robustos e transferíveis (Oquab et al., 2023). Ao

contrário das redes supervisionadas como ResNet, que perdem generalidade fora do domínio de treino, o DINOv2 opera em modo zero-shot (sem ajuste fino nos pesos) mantendo capacidade de generalização mesmo em domínios aerofotogramétricos não vistos (Oquab et al., 2023). No contexto desta pesquisa, onde a escassez de defeitos rotulados inviabiliza o ajuste fino, esses embeddings permitem construir um banco de memória de normalidade, compactado em protótipos por K-Means, e medir, em espaço L2-normalizado, o afastamento de patches suspeitos em relação a cenas íntegras de referência.

Para a classificação final, são empregados dois classificadores clássicos. O RF é um método de ensemble que treina múltiplas árvores de decisão com amostragem bootstrap e subamostragem aleatória de atributos em cada divisão, reduzindo correlação entre árvores e produzindo modelos robustos contra ruídos, além de fornecer medidas interpretáveis de importância de variáveis por Mean Decrease in Impurity (MDI) (Breiman, 2001). O SVM busca o hiperplano de margem máxima entre classes em um espaço mapeado por funções kernel, com o kernel RBF adotado por sua robustez em cenários de pequena amostra e fronteiras não lineares (Cortes & Vapnik, 1995). A utilização de ambos é respaldada por evidências de que técnicas híbridas são competitivas em cenários de escassez de dados anômalos (Chalapathy & Chawla, 2019).

Em bases pequenas, o aumento indiscriminado de atributos eleva o risco de overfitting e instabilidade nas estimativas. Por isso, a seleção estratégica de atributos, removendo estatísticas instáveis e priorizando descritores do tipo máximo, é parte metodológica central para equilibrar poder discriminativo e generalização. Para a avaliação, adota-se o Leave-One-Out Cross-Validation (LOO-CV), estratégia que maximiza o aproveitamento dos dados disponíveis para treinamento e produz estimativas de erro com menor viés em bases pequenas, sendo recomendada quando o número de amostras é reduzido (Hastie et al., 2009). Com base nessa fundamentação, a arquitetura proposta combina de forma multimodal descritores geométricos do MDS e embeddings visuais do DINOv2 em um pipeline interpretável e computacionalmente leve, adequado ao controle de qualidade de ortomosaicos em processamento de imagens georreferenciadas da Indústria 4.0.

3. Trabalhos relacionados

A literatura recente confirma o crescimento do uso de Deep Learning com imagens obtidas por drones, mas também indica forte heterogeneidade metodológica nos procedimentos de treinamento, validação e comparação de resultados. Em revisão sistemática, Oliveira et al. (2025), levantaram 212 registros nas bases Web of Science, Scopus e ScienceDirect e selecionaram 22 artigos publicados entre 2020 e 2025, identificando predomínio das arquiteturas YOLO e U-Net e, ao mesmo tempo, carência de padronização metodológica na área. Embora essa revisão demonstre crescimento do uso de IA em produtos georreferenciados derivados de drones, ela também mostra que a maior parte dos estudos se concentra na detecção de objetos, inspeções, incêndios e agricultura, e não no controle de qualidade de ortomosaicos.

No problema mais próximo desta pesquisa, Gordon et al. (2025) propuseram o pipeline STITCH-AD para detectar artefatos de stitching em ortomosaicos de pomares, por meio de detecção profunda de anomalias não supervisionadas. O método utiliza mascaramento automático com Segment Anything Model (SAM), decomposição dos ortomosaicos em patches e avaliação de dois modelos de anomalia (Reverse Distillation e UniAD), seguida de agregação dos scores por Isolation Forest e Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). Os autores relataram AUROC médio de 0,98 para o melhor conjunto com Reverse Distillation e acurácia de 100% na classificação por pomar com HDBSCAN, demonstrando a viabilidade de abordagens não supervisionadas mesmo com cerca de 1200 patches de treinamento. Entretanto, reconhecem limitações, principalmente ligadas ao volume reduzido de dados. A presente pesquisa se diferencia por ser uma arquitetura mais leve e interpretável, sem segmentação específica de domínio, com combinação direta entre descritores geométricos e memória visual, e com avaliação de classificadores clássicos em cenário de dados heterogêneos.

Outros estudos com drones e cobertura vegetal demonstram o potencial de técnicas de aprendizado profundo para tarefas como classificação de árvores frutíferas e predição de rendimento agrícola (Oliveira et al., 2025). Lang et al. (2025) compararam medições de fração de lacunas no dossel florestal com drones, fotografia hemisférica e varredura

LiDAR, encontrando relação linear moderada (R=0,67) e indicando que drones oferecem cobertura rápida, embora com perdas de acurácia em certos ângulos zenitais. Esses estudos reforçam a relevância do sensoriamento remoto com drones para monitoramento de cobertura vegetal, mas igualmente indicam que a confiabilidade geométrica dos produtos derivados é fator crítico para uso científico e operacional.

Com base nesse conjunto de pesquisas, entende-se que a literatura já demonstrou a viabilidade de abordagens não supervisionadas para detecção de anomalias em dados de drones com poucos defeitos rotulados (Gordon et al., 2025), mas ainda há lacuna quanto a métodos interpretáveis e multimodais que combinem informação geométrica do MDS e informação visual auto-supervisionada para detectar anomalias em ortomosaicos de áreas vegetadas heterogêneas, sem restringir o problema a um único tipo de cobertura ou a um pipeline dependente de segmentação específica de domínio. É justamente nessa lacuna que o presente trabalho se insere.

4. Solução computacional

O pipeline foi estruturado em quatro etapas sequenciais para detecção automática de anomalias em ortomosaicos oriundos de processamento fotogramétrico, conforme ilustrado na Figura 1.

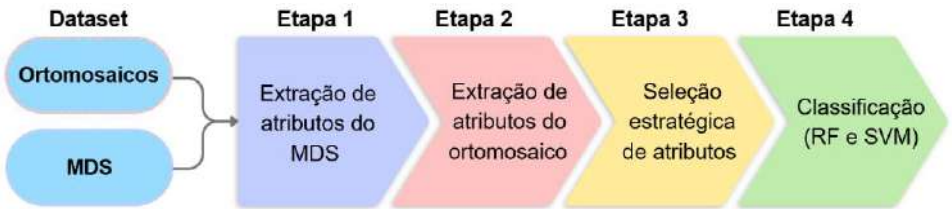


Figura 1. Fluxograma das etapas.

4.1. Etapa 1: Extração de atributos do MDS

Os MDS foram segmentados em patches de 256×256 pixels sem sobreposição, preservando a independência estatística entre as amostras. Valores ausentes, comuns nas bordas de imagem e áreas abertas, foram

substituídos pelo valor neutro (0,5) após normalização local. Patches com menos de 100 pixels válidos foram descartados para evitar viés analítico. Para os patches válidos foram extraídos 11 descritores estatísticos (Gonzalez & Woods, 2018): fração de pixels inválidos (nan_frac), desvio padrão bruto (std_raw), assimetria (skew), curtose (kurt), intervalo interquartil percentis 5–95 (iqr), gradiente médio de Sobel (grad_mean), desvio padrão do gradiente (grad_std), percentil 95 do gradiente (grad_p95), desvio padrão das médias por bloco de 64×64 pixels (loc_mean_std), média dos desvios padrão por bloco (loc_var_mean) e coeficiente de variação (coeff_var).

4.2. Etapa 2: Extração de atributos do ortomosaico

O modelo DINOv2-ViT-B14, pré-treinado e utilizado em modo zero-shot (Oquab et al., 2023), foi empregado para gerar embeddings visuais de 768 dimensões a partir dos patches RGB dos ortomosaicos. O banco de memória visual foi construído a partir de 9 cenas normais de referência, selecionadas com base na ausência de anomalias visuais e na diversidade de tipo de vegetação dessas cenas, das quais foram coletados aleatoriamente com um limite máximo de 250 patches de 256×256 pixels por cena, totalizando 1365 vetores. Esses vetores foram compactados por K-Means (MiniBatchKMeans da biblioteca scikit-learn) em 150 centróides, que foram normalizados por L2, formando um banco compacto e representativo da normalidade visual. Para cada patch de teste, calculou-se a distância euclidiana mínima ao centróide mais próximo do banco, quantificando o desvio visual em relação ao padrão normal.

4.3. Etapa 3: Seleção estratégica de atributos

Os atributos extraídos por patch foram agregados por média e máximo, resultando inicialmente em 24 atributos (11 descritores MDS × 2 agregações + 2 distâncias DINOv2 = 24). Uma vez que a base é restrita a 20 cenas, aplicou-se uma seleção conservadora para reduzir a dimensionalidade e mitigar o risco de sobreajuste (Chalapathy & Chawla, 2019). Foram removidos quatro atributos instáveis (assimetria e curtose, nas versões média e máximo, por alta variância amostral) e sete redundantes, por apresentarem correlação acima de 0,80 com outros atributos já mantidos. Os

13 atributos mantidos foram: nan_frac_mean, std_raw_mean, std_raw_max, iqr_mean, iqr_max, grad_mean_mean, grad_p95_mean, grad_p95_max, grad_std_mean, loc_var_mean_mean, coeff_var_mean (MDS, 11 atributos) e dino_dist_mean, dino_dist_max (DINOv2, 2 atributos).

4.4. Etapa 4: Classificação

Dois classificadores foram treinados e depois avaliados por LOO-CV. O RF foi configurado com n_estimators=150, max_depth=3, min_samples_leaf=2, class_weight={0:1, 1:3} e random_state variável entre execuções. O SVM utilizou kernel RBF, C=1,0, gamma='scale', class_weight={0:1, 1:3} e probability=True; o Standard Scaler foi ajustado exclusivamente nos dados de treino de cada fold e então aplicado ao dado de teste, evitando que informações do teste influenciassem o treinamento (data leakage). O peso de classe {0:1, 1:3} foi adotado em ambos os modelos para compensar o desequilíbrio entre as 13 cenas normais e as 7 defeituosas.

5. Avaliação experimental

5.1. Planejamento de pesquisa

A avaliação experimental foi planejada para verificar a capacidade do pipeline em detectar anomalias em ortomosaicos com cobertura vegetal, considerando artefatos como stitching, descontinuidades geométricas, entre outras anomalias.

O conjunto de dados foi composto por 20 cenas aéreas de áreas com predomínio de cobertura vegetal (com mais da metade do ortomosaico), extraídas dos datasets públicos ODMdata (10 cenas) e GeoNadir (10 cenas) (OpenDroneMap, 2026; GeoNadir, 2026). O processamento das imagens e geração dos produtos georreferenciados foi feito usando o software livre WebODM (Xaud et al., 2022). As cenas foram classificadas manualmente em 13 normais e 7 defeituosas com base na inspeção visual, como mostrado na Figura 2, no software QGIS v3.22 (QGIS, 2026). Os dados são heterogêneos quanto ao tipo de câmera, altura do voo, sobreposição das imagens, localização (urbana/rural) e tipo de vegetação (rasteira e florestal), o que torna o cenário experimental representativo de condições reais de operação.

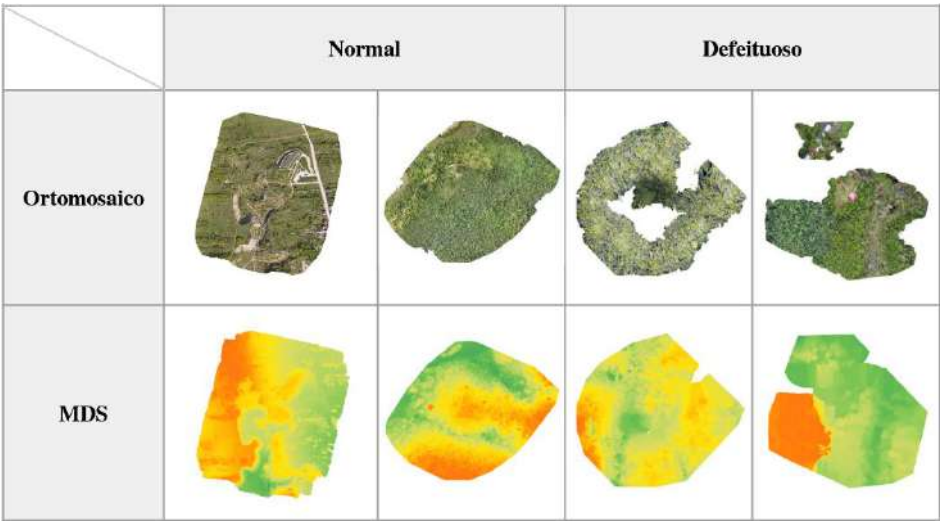


Figura 2. Ortomosaicos e MDS: normais e defeituosos.

Outra característica do dataset é que contém ortomosaicos de tamanhos variados, gerando entre 40 e 1148 patches por cena. Essa assimetria impacta na estabilidade dos atributos: cenas com poucos patches produzem estimativas de média menos confiáveis e valores de máximo subestimados, e cenas maiores podem gerar estatísticas de máximo mais elevadas. Para diminuir esse efeito, descritores com alta variância amostral foram descartados na etapa de seleção de atributos, priorizando métricas mais robustas e adotando o LOO-CV, que preserva maior quantidade possível de amostras de treino em cada fold (Hastie et al., 2009). Ainda assim, o impacto dessa assimetria não é eliminado e deve ser considerado na interpretação dos resultados.

O ambiente de execução usado tem as seguintes características: sistema operacional Ubuntu 22.04 com python 3.10, com as bibliotecas scikit-learn (Scikit-learn, 2025), torch (PyTorch, 2025), rasterio (Rasterio, 2025), opencv-python (OpenCV, 2025), numpy (NumPy, 2025), pandas (Pandas, 2025) e scipy (SciPy, 2025). As características do hardware utilizado foram: processador Intel i7-9750H, 16GB de memória RAM DDR4 e GPU NVIDIA GTX 1660 Ti Mobile (6 GB VRAM).

Para a avaliação de generalização, foi adotada LOO-CV, em que, a cada iteração (fold), 19 cenas foram utilizadas para treinamento e a cena restante foi reservada para teste. Essa técnica foi escolhida pelos bons resultados em bases pequenas, pois aumenta o número de amostras de treinamento em cada fold e diminui o viés associado a particionamentos aleatórios (Hastie et al., 2009). Em cada fold foram registrados a classe prevista e o score de anomalia da cena avaliada. Para avaliar a estabilidade dos resultados, aspecto crítico em datasets com poucas amostras, foram feitas três execuções utilizando três sementes aleatórias (random_state = 42, 44 e 60), mantendo todos os demais hiperparâmetros fixos. Essa abordagem permitiu quantificar a variabilidade do desempenho em função da aleatoriedade do treinamento dos classificadores. As métricas empregadas foram Acurácia, Precisão, Recall, F1-Score e AUROC, além da análise de importância de atributos por MDI no RF e por permutation importance no SVM. O conjunto de métricas permitiu avaliar tanto a capacidade discriminativa global do modelo quanto sua sensibilidade para detectar defeitos e sua confiabilidade ao detectar anomalias em diferentes limiares de decisão.

5.2. Resultados e Discussão

A Tabela 1 apresenta o desempenho dos dois classificadores em três execuções independentes de LOO-CV, variando o random_state (rs). O RF apresentou variação considerável entre execuções em todas as métricas, enquanto o SVM produziu resultados idênticos nas três execuções.

Tabela 1. Desempenho de Random Forest e SVM em três execuções LOO-CV.

Modelo	Acurácia	Precisão	Recall	F1-Score	AUROC
RF (rs=42)	0,900	0,857	0,857	0,857	0,846
RF (rs=44)	0,850	0,750	0,857	0,800	0,857
RF (rs=60)	0,800	0,714	0,714	0,714	0,813
RF (média)	**0,850**	**0,774**	**0,810**	**0,790**	**0,839**
SVM (rs=42, 44, 60)	**0,800**	**0,800**	**0,571**	**0,667**	**0,879**

Os resultados indicam forte capacidade de separação entre as classes, com AUROC médio de 0,839 para o RF e 0,879 para o SVM (Chalapathy & Chawla, 2019). O RF demonstrou maior sensibilidade operacional em duas das três execuções (*recall*=0,857, F1=0,857), enquanto o SVM apresentou desempenho totalmente estável nas três execuções. A maior variação das métricas do RF entre as execuções (recall variando de 0,714 a 0,857) se deve à sensibilidade ao tamanho reduzido do *dataset*, conforme esperado pelo LOO-CV com 20 amostras (Hastie et al., 2009).

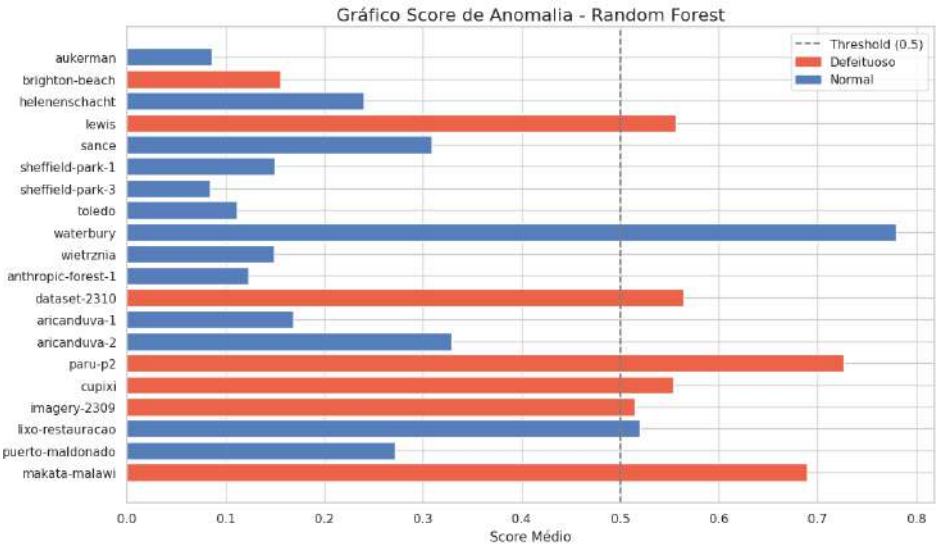


Figura 4. Gráfico de Score de Anomalias por cena no Random Forest.

Em termos de melhor compromisso entre *recall* e precisão, o RF com *random_state*=42 foi superior, detectando 6 das 7 cenas defeituosas com apenas 1 falso positivo; com *random_state*=44 o mesmo número de defeituosos foi detectado, porém com 2 falsos positivos. O SVM, ao detectar 4 das 7 cenas defeituosas com 1 falso positivo, ofereceu maior especificidade relativa, mas com perda de sensibilidade relevante para um problema onde falsos negativos, defeitos não detectados, representam o custo operacional mais alto.

As Figuras 4 e 5 mostram os scores médios de anomalias por cena nas três execuções, complementando métricas de desempenho e revelando padrões importantes.

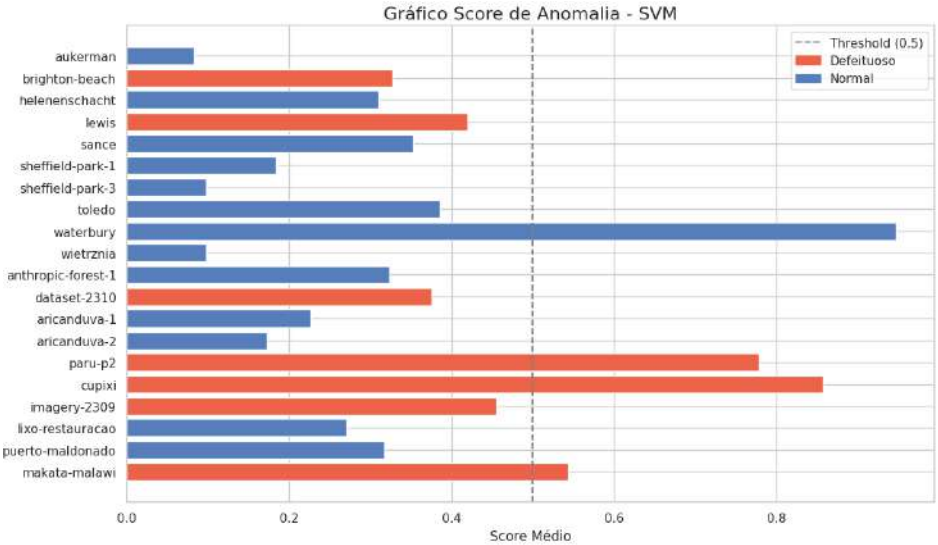


Figura 5. Gráfico de Score de Anomalias por cena no SVM.

No RF (Figura 4), cinco das sete cenas defeituosas (*lewis*, *dataset-2310*, *paru-p2*, *cupixi* e *makata-malawi*) ficaram bem acima do *threshold*, demonstrando separabilidade robusta nesses casos. No SVM (Figura 5), apenas *paru-p2*, *cupixi* e *makata-malawi* superaram o limiar com margem expressiva, o que explica visualmente o *recall* reduzido de 0,571: os demais defeituosos produzem *scores* insuficientes para acionar a classificação positiva, não necessariamente por ausência de anomalia (Liu & Tan, 2022), mas pela maior rigidez da superfície de decisão do modelo (Cortes & Vapnik, 1995).

Em ambos os modelos, *waterbury* foi o principal falso positivo, com *score* médio de 0,779 no RF e 0,947 no SVM, o maior entre todas as cenas íntegras, sugerindo que suas características se distanciam sistematicamente do padrão de referência (Chalapathy & Chawla, 2019). Em sentido oposto, *brighton-beach* obteve *score* de apenas 0,155 no RF, o mais baixo entre os

defeituosos, indicando que sua anomalia não produz desvio expressivo nos descritores utilizados (Gonzalez & Woods, 2018). Por fim, *lixo-restauracao* apresentou *score borderline* no RF (~0,52), o que explica sua classificação instável entre as execuções e reforça a sensibilidade do *pipeline* ao *threshold* em casos liminares (Hastie et al., 2009).

A Tabela 2 lista os atributos de maior importância nas 3 execuções no classificador RF, calculada por MDI sobre o modelo treinado na base completa.

Tabela 2. Top 5 atributos por importância média no Random Forest.

Rank	Atributo	Importância média (%)	Grupo
1	dino_dist_max	~20,58%	DINOv2
2	iqr_max	~15,86%	MDS
3	std_raw_max	~14,66%	MDS
4	loc_var_mean_mean	~10,10%	MDS
5	grad_p95_max	~8,44%	MDS

Esse resultado sustenta a hipótese de complementaridade entre informação geométrica do MDS e informação visual do DINOv2. Os atributos mais relevantes pertencem a ambos os grupos, com destaque para os do tipo máximo local (dino_dist_max, iqr_max e std_raw_max), o que é coerente com a natureza de anomalias de stitching: defeitos localizados que tendem a ser diluídos por agregações globais de média, mas são preservados pelas estatísticas de máximo (Gonzalez & Woods, 2018). O fato de nenhum atributo ultrapassar aproximadamente 20,6% de importância isolada indica que a decisão final depende da combinação de múltiplos sinais (Breiman, 2001), validando a estratégia multimodal proposta.

As limitações estatísticas do estudo estão relacionadas principalmente ao reduzido número de amostras. O LOO-CV, ao treinar cada fold com apenas 19 amostras, aumenta a instabilidade das estimativas de desempenho e de importância de atributos (Hastie et al., 2009). No SVM, o permutation importance apresentou estimativa próxima de zero para a maioria dos atributos, porque o hiperplano RBF é sensível a um subconjunto reduzido de dimensões (nan_frac_mean, loc_var_mean_mean, iqr_max), deixando os demais atributos sem efeito mensurável sobre a superfície de

decisão (Breiman, 2001). Esse comportamento contrasta com o RF, cujo mecanismo ensemble distribui a relevância entre múltiplos atributos de forma mais granular, tornando sua análise de importância mais informativa neste contexto.

6. Conclusão

Este trabalho investigou a detecção de anomalias em ortomosaicos de áreas com predominância de cobertura vegetal por meio de um pipeline de aprendizado de máquina que combina descritores geométricos do MDS, embeddings visuais do DINOv2, seleção conservadora de atributos e classificadores clássicos (Liu & Tan, 2022). Verificou-se que os descritores geométricos do MDS, mesmo que úteis para detectar irregularidades geométricas, não são suficientes de forma isolada em áreas vegetadas. A incorporação dos embeddings do DINOv2 elevou a capacidade discriminativa do pipeline, confirmada pelo AUROC médio de 0,839 no RF e 0,879 no SVM, e pela relevância do atributo dino_dist_max (~20,6% de importância média no RF) nas três execuções.

Embora inferior ao AUROC de 0,98 do STITCH-AD (Gordon et al., 2025), que utiliza abordagens restritas a pomares e modelos de alta capacidade, o desempenho alcançado se adequa com a proposta geral da solução, baseada em classificadores clássicos e representações compactas. Os atributos do tipo máximo local foram os mais discriminativos (Gonzalez & Woods, 2018), validando a estratégia multimodal. E a variabilidade observada (Tabela 1) é esperada no LOO-CV com 20 amostras (Hastie et al., 2009) e reforça que os resultados têm caráter indicativo.

As principais limitações são o tamanho reduzido do dataset, na variação de tamanho entre cenas e na ausência de validação em bases externas, restrições também reconhecidas por Gordon et al. (2025). Além disso, as cenas avaliadas apresentam predominância de vegetação de floresta tropical úmida, o que levanta a questão de como o pipeline se comportaria em biomas com características visuais e geométricas distintas, como lavrados, savanas, cerrados, áreas com corpos d'água ou áreas urbanas densas. Contextos que demandam validação específica, dado que a distribuição de normalidade aprendida pelo banco de memória é sensível ao

tipo de cobertura predominante (Chalapathy & Chawla, 2019; Oliveira et al., 2025).

Para estudos futuros, recomenda-se aumentar o dataset com maior diversidade de tipos de defeito e de cobertura vegetal, realizar validação cruzada em datasets externos e ajustar sistematicamente o threshold de decisão para maximizar o recall, minimizando falsos negativos. Para a indústria, este trabalho contribui para o desenvolvimento de módulos automáticos de controle de qualidade em processamento de imagens aéreas de drone (Oliveira et al., 2025), reduzindo a dependência de inspeção manual custosa e aumentando a confiança nos produtos geoespaciais utilizados no monitoramento da cobertura vegetal.

7. Referências

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. Recuperado de: https://arxiv.org/abs/1901.03407

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., & Houlsby, N. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. Recuperado de: https://arxiv.org/abs/2010.11929

GeoNadir. (2026). *GeoNadir: Global Database of Drone Imagery*. Recuperado de: https://geonadir.com

Gonzalez, R. C., & Woods, R. E. (2018). *Digital Image Processing* (4ª ed.). Pearson. Recuperado de: http://www.imageprocessingplace.com

Gordon, D., Whales, J., Smith, A., & Marais, P. (2025). Using deep learning to detect stitching artefacts in drone-based orthomosaics of orchards. In A. Gerber (Ed.), *SAICSIT 2025* (CCIS 2583, pp. 241-258). Springer. https://doi.org/10.1007/978-3-031-96262-2_16

Hartley, R., & Zisserman, A. (2004). Multiple View Geometry in Computer Vision (2ª ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511811685

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2ª ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Lang, M., Antsov, M., Mumma, A., Suitso, I., Kuusk, A., & Piip, K. (2025). Comparison of forest canopy gap fraction measurements from drone-based video frames, below-canopy hemispherical photography, and airborne laser scanning. *European Journal of Remote Sensing, 58*(1). https://doi.org/10.1080/22797254.2025.2456629

Liu, F., & Tan, J. (2022). Unsupervised anomaly detection by robust density estimation. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(7), 7107–7116. https://doi.org/10.1609/aaai.v36i7.20693

NumPy. (2025). NumPy documentation (v2.2). Recuperado de: https://numpy.org/doc/stable/

Oliveira, E. B., Fernandes, V. de O., & Alixandrini Júnior, M. J. (2025). Deep Learning: uma revisão sistemática integrativa de suas aplicações em mapeamento utilizando imagens de RPA. *Revista Brasileira de Cartografia, 77*. https://doi.org/10.14393/rbcv77n0a-78494

OpenCV. (2025). OpenCV documentation (v4.10). Recuperado de: https://docs.opencv.org/4.10.0/

OpenDroneMap. (2026). *Repositório Oficial do ODMdata: Open Drone Map Datasets.* Recuperado de: https://github.com/OpenDroneMap/ODMdata

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Bourdoukan, R. (2023). DINOv2: Learning robust visual features without supervision. Recuperado de: https://arxiv.org/abs/2304.07193

Pandas. (2025). Pandas Documentation (v2.2). Recuperado de: https://pandas.pydata.org/docs/

Peixoto, L. T. O. (2024). Análise Comparativa de Métodos para Identificação de Defeitos em Pavimento com Uso de Imagens de RPA. In: *Anais do 38º Congresso de Pesquisa e Ensino em Transportes*, Florianópolis, Brasil.

PyTorch. (2025). PyTorch documentation (v2.5). Recuperado de: https://pytorch.org/docs/stable/

QGIS. (2026). QGIS Geographic Information System (v3.22). Open-Source Geospatial Foundation Project. Recuperado de: https://qgis.org

Rasterio. (2025). Rasterio: Geospatial raster I/O for Python (v1.4). Recuperado de: https://rasterio.readthedocs.io/en/stable/

Scikit-learn. (2025). scikit-learn: Machine learning in Python (v1.5). Recuperado de: https://scikit-learn.org/stable/

SciPy. (2025). SciPy documentation (v1.14). Recuperado de: https://docs.scipy.org/doc/scipy/

Xaud, F. A. R., Pereira, F. C. de L., Lima, C. R. B. de, Silva, W. A. da, & Xaud, H. A. M. (2022). WebODM para o processamento de imagens aéreas obtidas por drone em uma área rural: como alternativa aos softwares proprietários. In: *Anais do XXIII Simpósio Brasileiro de Geoinformática* (pp. 341–346), São José dos Campos, Brasil. Recuperado de: http://mtc-m16c.sid.inpe.br/col/sid.inpe.br/mtc-m16c/2022/12.16.16.47/doc/341-346_Xaud_WebODM.pdf

Capítulo 14

Reconstrução espectral com SRF gaussianas e BI-LSTM para detecção de anomalias agrícolas no contexto da Agricultura 4.0

Johnny de Mendonça Pereira
Felipe Leite Lobo
Universidade Federal de Roraima

jmtdbg@gmail.com
felipe.lobo@ufrr.com

1. Introdução

A agricultura de precisão tem ampliado o uso de sensores remotos, VANTs e plataformas orbitais para o monitoramento não destrutivo de lavouras, favorecendo a detecção precoce de anomalias e o apoio ao manejo agrícola (Barbedo, 2023; Abbas et al., 2023). Esse avanço insere-se no contexto da Indústria 4.0 aplicada à agricultura, também denominada Agricultura 4.0, caracterizada pela integração entre sensores, conectividade, automação, conjuntos de dados massivos e inteligência artificial para apoiar a tomada de decisão no campo. Nesse cenário, os sensores hiperespectrais se destacam por registrar centenas de bandas contíguas, permitindo a obtenção de assinaturas espectrais detalhadas e sensíveis a alterações fisiológicas sutis (Ferreira et al., 2024; Furuya et al., 2021). Contudo, seu uso ainda é limitado pelo alto custo, pela complexidade de calibração e pela menor disponibilidade em operações rotineiras. Em contrapartida, sensores multiespectrais são mais acessíveis e amplamente utilizados em satélites e drones comerciais, embora sua menor resolução espectral restrinja a representação de variações finas do espectro (Osco et al., 2019; Almeida et al., 2024). Por isso, cresce o interesse por métodos capazes de aproximar informações hiperespectrais a partir de dados multiespectrais, especialmente com o apoio de técnicas de inteligência artificial e aprendizado de máquina voltadas à harmonização entre sensores (Barbedo, 2023; Ferreira et al., 2024; EnMAP-Box 3, 2022).

Neste trabalho, apresenta-se uma prova de conceito baseada em duas etapas complementares: a reamostragem do dado hiperespectral para representar as bandas multiespectrais do verde, vermelho, borda vermelha e infravermelho próximo por meio de funções de resposta espectral gaussianas, e a reconstrução da assinatura hiperespectral a partir dessas quatro bandas utilizando uma arquitetura Bi-LSTM. Assim, o estudo investiga a viabilidade da integração entre dados hiperespectrais e multiespectrais, por meio de reamostragem espectral e reconstrução com inteligência artificial, visando à preservação de informações espectrais relevantes para futuras aplicações no reconhecimento de anomalias em lavouras agrícolas. Especificamente, busca-se representar bandas multiespectrais a partir de dados hiperespectrais, reconstruir assinaturas hiperespectrais a partir de quatro bandas multiespectrais e analisar o potencial dessa abordagem como estratégia de harmonização espectral entre sensores.

Neste trabalho, a Seção 1 apresenta a introdução e a contextualização do problema; a Seção 2 aborda a fundamentação teórica; a Seção 3 descreve os trabalhos relacionados; a Seção 4 apresenta a solução computacional proposta; a Seção 5 detalha a avaliação experimental; e, por fim, a Seção 6 apresenta as conclusões e as perspectivas para trabalhos futuros.

2. Fundamentação teórica

O sensoriamento remoto tem sido aplicado na agricultura de precisão por permitir o monitoramento não destrutivo da vegetação e de suas alterações fisiológicas. A resposta espectral das plantas varia de acordo com fatores como teor de clorofila, conteúdo de água e estrutura foliar, o que torna possível identificar mudanças no estado da cultura por meio da reflectância registrada pelos sensores (Sishodia et al., 2020). Nesse contexto, a assinatura espectral representa o comportamento da vegetação ao longo do espectro eletromagnético e constitui a base para a interpretação de padrões associados ao vigor, ao estresse e a demais anomalias agrícolas.

Os sensores hiperespectrais se distinguem por registrar centenas de bandas estreitas e contíguas, permitindo descrever a assinatura espectral com elevado nível de detalhe (Ram et al., 2024). Já os sensores multiespectrais registram poucas bandas mais largas, normalmente

posicionadas em faixas estratégicas, como verde, vermelho, borda vermelha e infravermelho próximo. Contudo, são mais acessíveis e viáveis do ponto de vista operacional (Sishodia et al., 2020). Nesse cenário, estabelece-se uma relação de complementaridade: enquanto os dados hiperespectrais oferecem maior riqueza espectral, os dados multiespectrais apresentam maior potencial de uso em monitoramento agrícola em larga escala.

Para integrar esses dois domínios espectrais, a reamostragem espectral é etapa fundamental. Esse procedimento consiste em transformar uma assinatura hiperespectral em uma representação compatível com um sensor multiespectral. Para isso, é necessário considerar as funções de resposta espectral, que descrevem como cada banda do sensor responde à energia em uma faixa contínua do espectro (Cundill et al., 2015). Quando a curva real de resposta não está completamente disponível, uma aproximação usual é modelá-la por funções gaussianas, permitindo calcular cada banda multiespectral por média ponderada das bandas hiperespectrais adjacentes.

Assim, a reamostragem representa uma síntese espectral fisicamente mais consistente do que a simples escolha de um comprimento de onda central. O problema inverso dessa etapa é a reconstrução espectral, ou upsampling espectral, em que se busca estimar uma assinatura hiperespectral a partir de poucas bandas multiespectrais. Como o espectro pode ser tratado como uma sequência ordenada ao longo do comprimento de onda, com forte dependência entre bandas vizinhas, redes neurais recorrentes do tipo LSTM constituem uma alternativa teoricamente adequada para essa tarefa (Hochreiter e Schmidhuber, 1997).

A variante Bi-LSTM amplia essa capacidade ao processar a sequência nos dois sentidos, favorecendo a reconstrução de curvas mais contínuas e coerentes, especialmente em regiões críticas como a borda vermelha (Graves e Schmidhuber, 2005). Logo, a combinação entre reamostragem espectral com funções de resposta espectral e reconstrução com Bi-LSTM fornece a base tecnológica para a harmonização entre sensores adotada neste trabalho.

3. Trabalhos relacionados

A revisão da literatura foi feita com foco no sensoriamento remoto hiperespectral e multiespectral, reamostragem espectral, agricultura de

precisão e uso de Inteligência Artificial na detecção de anomalias agronômicas. Os artigos científicos foram selecionados manualmente, com base na sua relevância para o tema da pesquisa, priorizando trabalhos que abordassem imagens hiperespectrais e multiespectrais, análise e reamostragem espectral, agricultura de precisão, culturas agrícolas, UAV, sensoriamento remoto, Machine Learning, Deep Learning e LSTM. As buscas foram executadas nas bases Web of Science (WoS) e Scopus, utilizando campos como título, termos de busca, palavras-chave e tópico. Os registros recuperados formaram a base inicial da revisão, sendo posteriormente analisados para identificar estudos relacionados à harmonização entre sensores, simulação de bandas, classificação de anomalias como: pragas, doenças, plantas invasoras e estresses em culturas agrícolas.

Tetila et al. (2024) avaliaram modelos YOLOv5 para detecção em tempo real de ervas daninhas em soja com imagens de UAV (Veículo Aéreo Não Tripulado) e verificaram que o YOLOv5s6 apresentou o melhor desempenho, além de permitir a geração de mapas de aplicação localizada. Embora o estudo confirme o potencial do aprendizado profundo em operações de campo, seu foco permanece na detecção visual de objetos, e não na reconstrução espectral. No domínio espectral, Furuya et al. (2021) utilizaram dados hiperespectrais para detectar danos por herbivoria em milho e identificaram o Random Forest como algoritmo mais eficiente, com destaque para bandas no infravermelho próximo. De forma semelhante, Ramos et al. (2022) modelaram o ataque de Spodoptera frugiperda em algodão com sensoriamento hiperespectral proximal, Random Forest e simulação de bandas para sensores orbitais, mostrando que a resposta espectral pode ser reduzida para poucas bandas sem perda expressiva de desempenho. Esses estudos evidenciam o potencial do hiperespectral para detecção precoce, mas ainda partem do dado hiperespectral como entrada principal.

Na cultura dos citros, Osco et al. (2019) inferiram nitrogênio foliar em laranjeiras da cultivar Valencia com imagens multiespectrais de drone e curvas espectrais simuladas, obtendo melhor desempenho com o algoritmo SAM e maior relevância das bandas *red-edge* (borda vermelha) e NIR. Em complemento, a documentação do EnMAP-Box 3 descreve um algoritmo de reamostragem espectral para o padrão Sentinel-2 MSI baseado em convolução de funções de resposta espectral, reforçando que a harmonização

entre sensores por SRF é tecnicamente consolidada. Ferreira et al. (2024) e Barbedo (2023) destacam que imagens hiperespectrais associadas à IA alcançam altas acurácias, persistindo o desafio de padronização, robustez e generalização em condições reais. Rai (2024), fora do contexto agrícola, mostrou que redes LSTM são adequadas para modelar dependências sequenciais, reforçando sua pertinência para reconstrução espectral.

Em síntese, os estudos revisados mostram avanços na detecção de anomalias e atributos agronômicos com dados espectrais e IA. Entretanto, permanece pouco explorada a reconstrução de informação hiperespectral a partir de poucas bandas multiespectrais, integrando reamostragem por SRF e modelagem sequencial com Bi-LSTM.

4. Solução computacional

A solução foi estruturada como um pipeline de ciência de dados dividido em cinco etapas principais: aquisição de imagens multiespectrais; Aquisição de imagens hiperespectrais; levantamento de anomalias; processamento das imagens multiespectrais e hiperespectrais; reconstrução de bandas. O fluxo metodológico é apresentado na Figura 1.

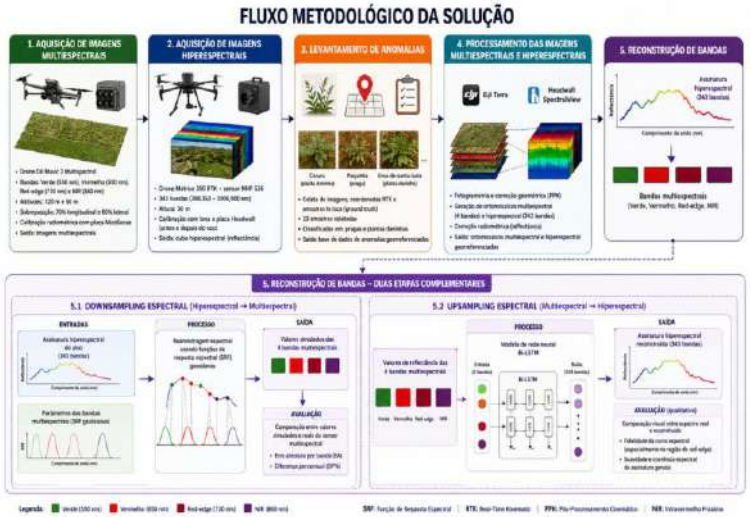


Figura 1. Fluxo metodológico. Fonte: Elaborado pelo autor com o auxílio do ChatGPT (OpenAI, 2026).

4.1. Aquisição de imagens multiespectrais

O levantamento de campo foi realizado em 06 de maio de 2025, em uma área de cultivo de soja localizada no município de Mucajaí, Roraima, Brasil. No momento da coleta, as plantas encontravam-se no estádio V5, e a área imageada foi definida a partir das parcelas com ocorrência de anomalias agronômicas, sendo selecionada uma área de aproximadamente 1 ha com infestação por caruru, caracterizada como anomalia do tipo invasora. A aquisição das imagens ocorreu no período da manhã, entre 9h30 e 10h00, sob céu sem nuvens, baixa intensidade de vento e iluminação constante. Para a execução do levantamento, utilizou-se o drone DJI Mavic 3 Multispectral, com planejamento e operação de voo elaborado no DJI pilot.

4.2. Aquisição de imagens hiperespectrais

A aquisição da imagem hiperespectral ocorreu no mesmo dia do imageamento multiespectral, no período da manhã, entre 9h30 e 10h00, sob as mesmas condições de iluminação. Foi realizada a 30 m de altura, com sensor hiperespectral NHP 526 embarcado no VANT Matrice 350 RTK, com planejamento de voo elaborado no UGCS Pro e operação conduzida no DJI Pilot 2, que registrou 343 canais espectrais, cobrindo a faixa de 398,353 nm a 1000,900 nm, com espaçamento médio de 1,7618196 nm.

A calibração em campo do sensor hiperespectral foi realizada mediante o uso de uma lona de calibração Headwall 3 × 3 m, posicionada em campo antes e depois do voo. Também foi utilizada como padrão de reflectância uma placa/lona de calibração Headwall, empregada nas mesmas condições. A calibração permitiu estimar valores de reflectância na faixa de 80 a 90%, durante a etapa de processamento. O sinal medido foi convertido em reflectância por meio do software Headwall SpectralView, com base na correção radiométrica inserida manualmente, utilizando como referência a placa branca e a lona de calibração.

4.3. Levantamento de anomalias

O levantamento das anomalias foi realizado em campo por meio da coleta de imagens, localização geográfica e amostras in loco, compondo a

verdade de campo (ground truth). A coleta ocorreu no mesmo dia dos voos/imageamentos, garantindo maior correspondência entre dados de campo e imagens. Ao todo, foram coletadas 23 amostras, organizadas em planilha estruturada para análise. As anomalias foram georreferenciadas com base RTK (Real-Time Kinematic), tecnologia GNSS que fornece precisão centimétrica em tempo real, dispensando pós-processamento.

4.4. Processamento das imagens multiespectrais e hiperespectrais

Foram adquiridas imagens multiespectrais e hiperespectrais da área de estudo, processadas nos softwares de fotogrametria DJI Terra para imagens multiespectrais e no Headwall SpectralView para imagens hiperespectrais.

4.5. Reconstrução de bandas

Para a demonstração metodológica, foi selecionada a anomalia caruru, classificada como planta daninha. Essa anomalia foi representada por uma amostra em cada arquivo CSV, contendo valores de reflectância do pixel nas bandas capturadas pelos sensores hiperespectral e multiespectral. Foram utilizados dois arquivos, hiper_amostra_caruru.csv contendo a assinatura hiperespectral do pixel de caruru, composta por 343 bandas entre aproximadamente 398,353 nm e 1000,900 nm, com espaçamento médio de 1,7618196 nm e o mult_amostra_caruru.csv contendo os valores de reflectância do mesmo alvo nas quatro bandas multiespectrais: verde (550 ± 16 nm), vermelho (650 ± 16 nm), borda vermelha (730 ± 16 nm) e infravermelho próximo (860 ± 26 nm). A reconstrução de bandas foi realizada em duas etapas complementares: downsampling espectral e upsampling espectral.

5. Avaliação experimental

5.1. Planejamento de pesquisa

Neste estudo, foram executados os processos de downsampling (hiperespectral para multiespectral) e upsampling (multiespectral para hiperespectral) com o objetivo de demonstrar tecnicamente como

reconstruir uma assinatura hiperespectral com base em dependências espectrais aprendidas por uma Bi-LSTM.

No processo de downsampling, as variáveis de entrada consistiram na assinatura hiperespectral do alvo e nos parâmetros das bandas do sensor multiespectral. A partir dessas entradas, foi realizada a reamostragem espectral por meio da aplicação de funções de resposta espectral (SRF) gaussianas, a fim de gerar como saída os valores simulados das quatro bandas multiespectrais. Assim foi avaliada a capacidade do modelo em reproduzir os valores reais do sensor multiespectral a partir do dado hiperespectral. As métricas utilizadas foram baseadas na comparação direta entre valores observados e simulados, com ênfase no erro absoluto por banda e na diferença percentual, permitindo quantificar a discrepância entre os dados reamostrados e os dados reais.

No processo de upsampling, as variáveis de entrada foram os valores de reflectância das quatro bandas multiespectrais (verde, vermelho, borda vermelha e infravermelho próximo). A partir dessas entradas, foi utilizado um modelo de rede neural do tipo Bi-LSTM para reconstruir a assinatura hiperespectral completa (343 bandas), considerando a dependência sequencial entre os comprimentos de onda. Assim foi avaliada a capacidade do modelo reconstruir a forma e a continuidade do espectro hiperespectral original a partir de um conjunto reduzido de bandas. As métricas utilizadas foram predominantemente qualitativas, baseadas na comparação visual entre o espectro real e o reconstruído, observando a fidelidade da curva espectral, especialmente em regiões sensíveis como a borda vermelha, além da análise da suavidade e coerência espectral da assinatura gerada e as métricas quantitativas, incluindo o erro absoluto médio (MAE), o erro quadrático médio (RMSE) e o Mapeador de Ângulo Espectral (SAM).

5.2. Resultados e Discussão

A análise dos resultados mostrou que a resolução espectral do sensor hiperespectral é adequada para sintetizar bandas multiespectrais de forma consistente. A sobreposição das Funções de Resposta Espectral às 343 bandas hiperespectrais evidenciou que cada banda multiespectral resulta da

contribuição de uma faixa contínua do espectro, e não apenas de um comprimento de onda central.

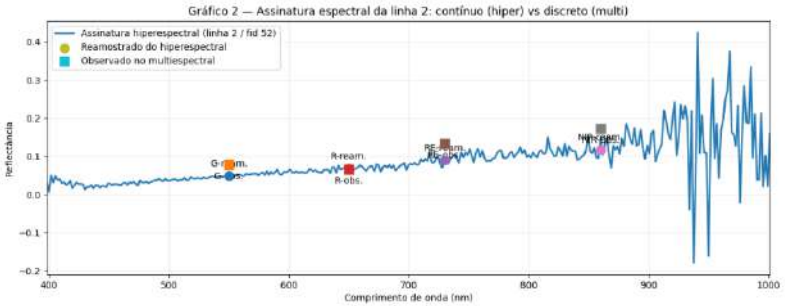


Figura 2. SRFs gaussianas das bandas multiespectrais sobre a grade hiperespectral e Assinatura espectral da amostra: contínuo(hiper) vs discreto (multi).

A comparação entre a assinatura hiperespectral contínua e os quatro valores multiespectrais discretos mostrou que, mesmo com a forte redução da resolução espectral, a tendência geral da resposta da amostra foi preservada. A Figura 2 indica que os pontos multiespectrais acompanham a forma global da curva hiperespectral, embora não reproduzam seus detalhes finos, o que era esperado diante da redução de 343 bandas para apenas quatro observações. Ainda assim, o processo manteve informações essenciais sobre o comportamento radiométrico do alvo, especialmente nas bandas vermelho, borda vermelha e NIR, que concentram informações relevantes sobre vigor vegetativo, teor de clorofila e estrutura da vegetação.

A Figura 3 detalha a formação da banda multiespectral da borda vermelha em torno de 730 nm e mostra que seu valor final resulta da contribuição ponderada de várias bandas hiperespectrais dentro da faixa de resposta do filtro. Esse resultado confirma que a abordagem adotada é mais robusta do que a simples escolha de uma banda central, pois incorpora explicitamente a forma do filtro e a distribuição espectral dos dados. Como a borda vermelha corresponde à transição entre a absorção no vermelho e a alta reflectância no infravermelho próximo, sua correta representação é especialmente importante em alvos vegetados, por estar associada a alterações fisiológicas e estruturais.

A comparação entre os valores observados e os valores reamostrados indicou coerência geral entre as duas representações espectrais. Embora tenham sido observadas diferenças absolutas em algumas bandas, a correspondência entre os perfis foi satisfatória para fins de demonstração metodológica. Essas diferenças podem estar relacionadas à simplificação das funções de resposta espectral, à calibração radiométrica entre sensores, a ruídos instrumentais e a variações do processo de aquisição. Ainda assim, a estrutura global da assinatura foi preservada, indicando que a abordagem adotada é adequada como base para harmonização espectral entre sensores distintos.

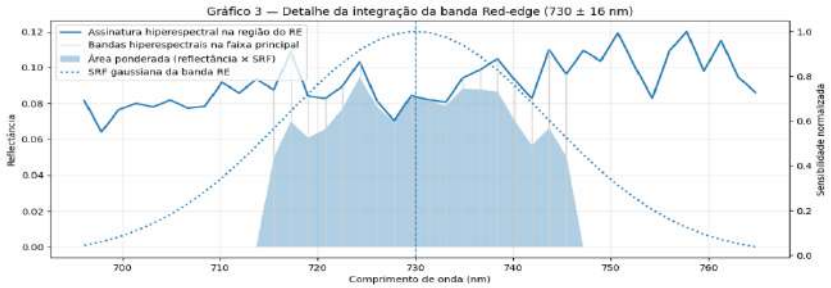


Figura 3. Detalhe da integração da banda Red-edge (730 ± 16 nm).

Na etapa de upsampling, a reconstrução da assinatura hiperespectral a partir de quatro bandas multiespectrais mostrou que a abordagem baseada em Bi-LSTM foi capaz de recuperar a tendência geral do espectro. A Figura 4 evidencia que o modelo gerou uma curva contínua e suave, coerente com a organização sequencial do domínio espectral. As métricas quantitativas medidas foram: MAE com 0,021787, RMSE com 0,045195 e o SAM com 0,437637 e rad de 25,074736°. O MAE e o RMSE indicam baixa discrepância ponto a ponto entre a assinatura hiperespectral real e a reconstruída, enquanto o SAM avalia a similaridade global da forma espectral, evidenciando que, apesar da coerência geral da reconstrução, ainda existem diferenças estruturais relevantes ao longo do espectro. Embora a reconstrução não reproduza todos os detalhes finos da assinatura original, a forma global da curva foi preservada, indicando que a rede conseguiu aprender a relação entre as bandas multiespectrais de entrada e a estrutura

hiperespectral. Esse resultado reforça o potencial da Bi-LSTM para modelar dependências entre bandas adjacentes ao longo do comprimento de onda.

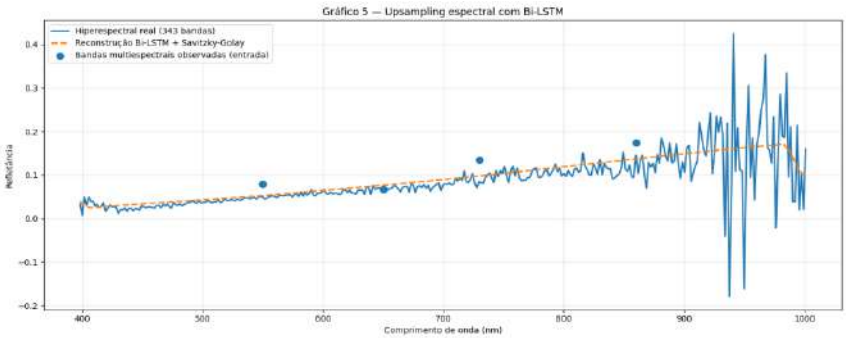


Figura 4. Upsampling espectral com Bi-LSTM.

A análise específica da região da borda vermelha reconstruída, apresentada na Figura 4, mostrou que o modelo também conseguiu reproduzir de forma contínua a transição entre as regiões do vermelho e do infravermelho próximo. A Figura 5 mostra redução progressiva da perda ao longo do treinamento, indicando convergência do modelo para a amostra utilizada. A diminuição conjunta do erro de reconstrução e do termo de suavidade espectral sugere que a função de perda adotada foi adequada para favorecer curvas contínuas e fisicamente plausíveis. No entanto, como o treinamento foi realizado com apenas uma amostra, esse resultado deve ser interpretado apenas como ajuste ao exemplo analisado, sem permitir inferências sobre a capacidade de generalização do modelo. Assim, os resultados de upsampling devem ser entendidos como uma prova de conceito, e não como validação definitiva da abordagem.

Em conjunto, os resultados de downsampling e upsampling indicam potencial para integrar dados hiperespectrais e multiespectrais em um mesmo fluxo analítico. O downsampling mostrou que a alta resolução espectral do sensor hiperespectral permite sintetizar bandas multiespectrais de forma fisicamente coerente, enquanto o upsampling indicou que parte da estrutura espectral pode ser reconstruída a partir de poucas bandas com modelos capazes de explorar dependências entre bandas vizinhas. Ainda

assim, a consolidação da abordagem em contexto operacional depende de validações adicionais, incluindo ampliação do número de amostras, separação entre treino, validação e teste, inclusão de diferentes classes espectrais e comparação com outros métodos de reconstrução.

A abordagem deve ser interpretada com cautela, pois a reconstrução de uma assinatura hiperespectral a partir de poucas bandas multiespectrais é um problema mal posto e não recupera informações espectrais inexistentes. Trata-se de uma tentativa experimental de aproximação da tendência espectral, considerando que a resposta de um objeto depende de suas características físicas, químicas e biológicas e de sua interação com a radiação. Assim, os resultados devem ser entendidos como uma prova de conceito limitada, dependente da qualidade dos dados, da representatividade das amostras e de validações futuras com diferentes alvos, sensores e condições ambientais, não substituindo a aquisição hiperespectral.

6. Conclusão

Conclui-se que a metodologia proposta permitiu avaliar, em caráter de prova de conceito, a integração entre dados hiperespectrais e multiespectrais para apoiar estudos voltados à detecção de anomalias agrícolas. A etapa de reamostragem espectral, baseada em funções de resposta espectral gaussianas, mostrou-se adequada para simular as bandas multiespectrais do verde, vermelho, borda vermelha e infravermelho próximo a partir da assinatura hiperespectral, pois considera a resposta integrada do sensor em cada faixa espectral, e não apenas comprimentos de onda centrais isolados.

A etapa de reconstrução espectral com Bi-LSTM indicou preservação parcial da tendência global da assinatura hiperespectral a partir de quatro bandas multiespectrais. No entanto, por se tratar de um problema mal posto, a reconstrução não deve ser interpretada como recuperação direta de informações espectrais inexistentes, mas como uma aproximação baseada nos padrões aprendidos pelo modelo. O valor de SAM de 25,074736° evidencia limitações na reconstrução espectral, especialmente quanto à fidelidade angular entre a assinatura real e a estimada.

De modo geral, os resultados indicam que a abordagem apresenta potencial como ferramenta experimental para compatibilização espectral e análise exploratória em aplicações agrícolas, mas ainda não permite classificá-la como plenamente promissora sem validação adicional. Recomenda-se, em trabalhos futuros, ampliar o número de amostras, incluir diferentes alvos agrícolas, sensores, áreas e condições ambientais, além de investigar o uso de Physics-Informed Neural Networks (PINNs) para incorporar restrições físicas ao modelo e tornar a reconstrução mais consistente com a realidade espectral e operacional dos sensores.

7. Referências

Barbedo, J. G. A. (2023). Deep learning for soybean monitoring and management. *Seeds*, *2*(3), 340–356. https://doi.org/10.3390/seeds2030026

Cundill, S. L., Van der Werff, H. M. A., & Van der Meijde, M. (2015). Adjusting spectral indices for spectral response function differences of very high spatial resolution sensors simulated from field spectra. *Sensors*, *15*(3), 6221–6240. https://doi.org/10.3390/s150306221

Ferreira, L. C., Ian C. B. C., Lúcio A. C. J., Alice M. Q., & Maurício R. (2024). Hyperspectral imaging for the detection of plant pathogens in seeds: Recent developments and challenges. *Frontiers in Plant Science*, *15*, 1387925. https://doi.org/10.3389/fpls.2024.1387925

Garcia, F., Danielle E., Lingfei M., Mayara M. F. P., et al. (2021). Prediction of insect-herbivory-damage and insect-type attack in maize plants using hyperspectral data. *International Journal of Applied Earth Observation and Geoinformation*, *105*, 102608. https://doi.org/10.1016/j.jag.2021.102608

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, *18*(5–6), 602–610. Recuperado de: https://www.cs.toronto.edu/~graves/nn_2005.pdf

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Osco, L. P., Ana Paula M. R., Érika A. S. M., et al. (2019). Improvement of leaf nitrogen content inference in Valencia-orange trees applying spectral analysis algorithms in UAV mounted-sensor images. *International Journal of Applied Earth Observation and Geoinformation*, *83*, 101907. https://doi.org/10.1016/j.jag.2019.101907

Ram, B. G., et al. (2024). A systematic review of hyperspectral imaging in precision agriculture: Analysis of its current state and future prospects. *Computers and Electronics in Agriculture*. https://doi.org/10.1016/j.compag.2024.109037

Ramos, A. P. M., Felipe D. G. G., Mayara M. F. P., et al. (2022). Detecting the attack of the fall armyworm (Spodoptera frugiperda) in cotton plants with machine learning and spectral measurements. *Precision Agriculture*, *23*(2), 470–491. https://doi.org/10.1007/s11119-021-09845-4

Sishodia, R. P., Ray, R. L., & Singh, S. K. (2020). Applications of remote sensing in precision agriculture: A review. *Remote Sensing*, *12*(19), 3136. https://doi.org/10.3390/rs12193136

Spectral resampling (to Sentinel-2 MSI). (n.d.). EnMAP-Box 3. Recuperado de: https://enmap-box.readthedocs.io/en/release_3.12/usr_section/usr_manual/processing_algorithms/spectral_resampling/spectral_resampling_to_sentinel-2_msi_.html

Tetila, E. C., Barbara L. M., Gilberto A., et al. (2024). Real-time detection of weeds by species in soybean using UAV images. *Crop Protection*, *184*, 106846. https://doi.org/10.1016/j.cropro.2024.106846

Capítulo 15

Arquitetura híbrida Edge-Cloud: Protótipo IoT Offline-First para Agroindústria 4.0 em áreas remotas de Roraima

Janei Vieira Pereira
Clenya Rejane Barros de Lima
Universidade Federal de Roraima

janei@taguardado.com
clenyabarros@uerr.edu.br

1. Introdução

A Indústria 4.0 integra sensoriamento, conectividade e processamento distribuído para otimizar processos produtivos (Schwab, 2016). No Agronegócio, essa transformação caracteriza a Agroindústria 4.0, permitindo monitoramento contínuo e decisões preditivas. Em Roraima, a safra 2024/2025 atingiu 692 mil toneladas, consolidando o estado como nova fronteira agrícola nacional (Governo de Roraima, 2025). Esse avanço intensifica a necessidade de monitoramento térmico em silos, etapa essencial para evitar perdas pós-colheita.

Entretanto, áreas rurais do estado enfrentam conectividade intermitente, o que compromete arquiteturas *cloud-centric* predominantes em soluções comerciais. Em cenários de oscilação ou ausência de internet, essas arquiteturas interrompem o monitoramento, inviabilizam alertas imediatos e podem causar perda de dados históricos (Yang et al., 2024; Maddumala, 2025). Embora a literatura apresente avanços em *edge computing*, ainda são escassas as validações práticas que combinem baixo custo, persistência local e resiliência *offline-first* em ambientes rurais.

Além disso, o potencial desta pesquisa se destaca para pequenos produtores rurais, agricultura familiar e comunidades indígenas, que frequentemente enfrentam limitações de infraestrutura e recursos tecnológicos. Ao propor uma solução de baixo custo, resiliente e adequada para condições de conectividade instável, o estudo possibilita que esses

grupos ampliem a capacidade de monitoramento e gestão das condições de armazenagem, reduzindo perdas e aumentando a segurança alimentar. O acesso a tecnologias adaptadas à realidade local contribui para a autonomia, sustentabilidade e fortalecimento das economias regionais, promovendo inclusão digital e valorizando práticas tradicionais combinadas com inovação.

Diante desse cenário, este estudo tem como objetivo geral desenvolver e validar um protótipo IoT híbrido *edge-cloud* com resiliência *offline-first*, persistência local em SQLite e sincronização assíncrona para monitoramento ambiental em áreas de conectividade instável. Especificamente, busca-se projetar uma arquitetura capaz de reter dados localmente e sincronizá-los automaticamente após reconexão, avaliar a autonomia de alertas e atuadores na borda, medir o desempenho sob apagões simulados e verificar se a comparação entre ambiente interno e externo torna a decisão de exaustão mais coerente fisicamente.

A investigação concentra-se em responder quatro perguntas de pesquisa: Q1 – em que medida uma arquitetura híbrida *edge-cloud* orientada ao paradigma *offline-first* mantém a continuidade operacional em cenários de conectividade intermitente; Q2 – se o uso de persistência local em SQLite com sincronização assíncrona após reconexão preserva a integridade dos dados históricos, sem perdas e sem duplicidades; Q3 – se o processamento de borda no ESP32 permite alertas imediatos e acionamento local de atuadores independentemente da disponibilidade da nuvem; e Q4 – se a comparação entre variáveis ambientais internas e externas melhora a coerência física do acionamento de exaustão em relação ao uso isolado de limiares internos.

Ao abordar essas questões, o estudo demonstra que hardware acessível, banco relacional leve e lógica embarcada podem compor uma solução robusta para a Agroindústria 4.0 regional, mitigando falhas de conectividade e aproximando o protótipo de condições reais de armazenagem. Este trabalho está organizado da seguinte forma: a Seção 2 apresenta a fundamentação teórica; a Seção 3 discute os trabalhos relacionados; a Seção 4 descreve a solução computacional; a Seção 5 apresenta a avaliação experimental e os resultados; a Seção 6 mostra as conclusões; e a Seção 7 reúne as referências.

2. Fundamentação teórica

As soluções IoT desempenham papel essencial no Agronegócio, especialmente em regiões como Roraima, onde a conectividade intermitente compromete o monitoramento contínuo de variáveis críticas (Kubota & Rosa, 2023). Nesse contexto, a computação de borda surge como alternativa ao modelo tradicional baseado exclusivamente em nuvem. Segundo Shi et al. (2016), o processamento na borda reduz latência, tráfego de rede e dependência de conectividade constante, tornando-se adequado para aplicações sensíveis ao tempo real. A nuvem, por sua vez, permanece relevante para armazenamento massivo e análises de longo prazo.

Arquiteturas híbridas que combinam nuvem, borda e dispositivos têm ganhado destaque em cenários agrícolas. Yu et al. (2025) demonstram que a borda pode executar coleta e análise inicial dos dados, enquanto a nuvem oferece suporte a históricos, aprendizado de máquina e gestão em larga escala. Em ambientes rurais com redes instáveis, a resiliência depende de mecanismos de armazenamento local e sincronização posterior. Tariq et al. (2025) destacam que manter dados temporariamente na borda reduz perdas e permite decisões rápidas mesmo durante falhas de comunicação.

Protocolos leves como MQTT são amplamente utilizados em sistemas IoT por seu baixo *overhead* e suporte a filas de mensagens pendentes, favorecendo a comunicação em arquiteturas *edge-to-cloud*. Em cenários de desconexão, mensagens podem ser armazenadas localmente e reenviadas quando a conexão é restabelecida, aumentando a robustez do sistema.

Para persistência local, motores de banco de dados leves como SQLite são amplamente adotados em dispositivos embarcados. Gaffney et al. (2022) destacam que o SQLite oferece transações ACID, baixo consumo de recursos e integração simples, sendo adequado para ambientes com restrições de hardware. O mecanismo *write-ahead logging* (WAL), descrito pelo SQLite Project (2026), registra alterações em log antes de aplicá-las ao banco principal, garantindo integridade mesmo em falhas de energia ou desconexões — característica essencial para arquiteturas *offline-first*.

Essa abordagem se alinha às necessidades da IoT agroindustrial, que exige continuidade operacional mesmo em condições adversas. Batistatos et al. (2025) reforçam que a combinação de processamento na borda e

protocolos tolerantes a desconexões reduz a dependência da nuvem e assegura que métricas vitais continuem sendo registradas e processadas durante apagões de rede. Assim, a literatura converge para a importância de arquiteturas híbridas, resilientes e de baixo custo, especialmente em regiões rurais com infraestrutura limitada.

3. Trabalhos relacionados

Após breve revisão da literatura, foram selecionadas sete pesquisas sobre monitoramento IoT, *edge computing* e arquiteturas híbridas. Bascıftcı et al. (2021) monitoraram silos de grãos com sistema IoT e interface web, comprovando o controle térmico, mas exigindo conectividade constante. Kalyani e Collier (2021) destacaram que a integração entre *cloud*, *fog* e *edge* reduz latência e uso de banda, porém carece de validações práticas em redes intermitentes.

Sanchez-Iborra et al. (2023) sugeriram decisão com TinyML no dispositivo, economizando energia e transmissão. Durgalakshmi et al. (2025) integraram IoT, *edge e IA pa*ra pré-processamento local, reduzindo transmissão em até 55%, mas alertam para barreiras técnicas e custos elevados. Ambos não detalham bancos de dados leves para retenção de histórico durante falhas. Saxena et al. (2023), bem como Sharmaet al. (2025), melhoraram resposta e uso de banda com arquiteturas híbridas, mas presumem redes ativas, ignorando o paradigma *offline-first*.

Tabela 1. Comparativo dos trabalhos relacionados e lacunas identificadas.

Autores (Ano)	Objetivo Principal	Desempenho / Métrica Relatada	Lacuna para Redes Intermitentes
Bascıftcı et al. (2021)	Monitoramento de silos de grãos via IoT	Manutenção da qualidade via monitoramento web.	Sistema dependente de internet constante.
Kalyani & Collier (2021)	Revisão sobre *cloud*, *fog* e *edge* no agro.	Modelos híbridos superam uso exclusivo de nuvem.	Foco teórico; ausência de protótipo físico *offline-first.*
Sanchez-Iborra et al. (2023)	TinyML e *edge computing* em cenário agrícola.	Redução no tempo de resposta e economia de energia.	Falta de armazenamento relacional para retenção de histórico.
Saxena et al. (2023)	Análise e redução de latência com *edge computing*.	Melhoria no tempo de resposta e economia de banda.	Validado teoricamente, sem foco em redes intermitentes extremas.

Autores (Ano)	Objetivo Principal	Desempenho / Métrica Relatada	Lacuna para Redes Intermitentes
Nyamuryekung'e (2023)	Monitoramento de gado via LoRaWAN em áreas remotas.	Métricas diárias de distância e comportamento	Requer conectividade estável para envio contínuo; não possui persistência local
Durgalakshmi et al. (2025)	Um *framework* adaptável com IoT, Edge e IA.	Redução de até 55% na latência de transmissão.	Foco na IA (custo elevado e alta complexidade técnica); sem mecanismo relacional local.
Sharma et al. (2025)	Arquitetura híbrida *edge-cloud* via MQTT/AWS.	Tempo de resposta caiu para 42ms; banda otimizada.	Voltado a redes estáveis; sem foco em resiliência de dados local.
Este trabalho *(2026)*	Desenvolver e validar um protótipo híbrido *edge-cloud offline-first*	100% de integridade dos dados, resiliência e autonomia de borda independente de nuvem.	Protótipo em escala reduzida, testes conduzidos em laboratório controlado e sem uso de fonte energética autônoma.

Nyamuryekung'e et al. (2023) testaram LoRaWAN para monitoramento de gado, mas dependem de conectividade contínua, registrando perda de dados por manutenção. Os estudos focam em otimização de banda ou processamento complexo, sem abordar retenção segura de dados em cenários de conectividade intermitente como os de Roraima. A Tabela 1 apresenta esse comparativo.

Dessa forma, a lacuna identificada está na ausência de uma solução que una *hardware* acessível, operação *offline-first*, persistência relacional local, sincronização assíncrona com integridade total e lógica de decisão ambiental mais aderente ao comportamento térmico real do sistema monitorado.

4. Solução computacional

A solução é um protótipo de arquitetura IoT híbrida *offline-first*, estruturada em nó sensor e *gateway* de borda. O nó sensor usa ESP32 DevKit V1 com *framework* ESP-IDF (C/C++ na IDE Arduino), coletando dados de dois DHT22 (interno/externo) para temperatura/umidade. Inclui *buzzer* ativo e miniventilador 5V (relé) para alertas/exaustão, transmitindo via MQTT com eficiência e baixa latência.

O *gateway* (Raspberry Pi 3B) hospeda Mosquitto (*broker* MQTT), SQLite com WAL para persistência e microsserviços Python 3.9 (*paho-mqtt*, *sqlite3*) e um *dashboard* local (Streamlit *web*) acessível via navegador que oferece visualização intuitiva sem nuvem.

Os *scripts* em Python podem ser executados através de serviços ou no próprio terminal do sistema Linux embarcado no Raspberry ou utilizando *containers* Docker para garantir reinicialização automática e maior flexibilidade na manutenção. A Figura 1 apresenta uma visão para a solução proposta.

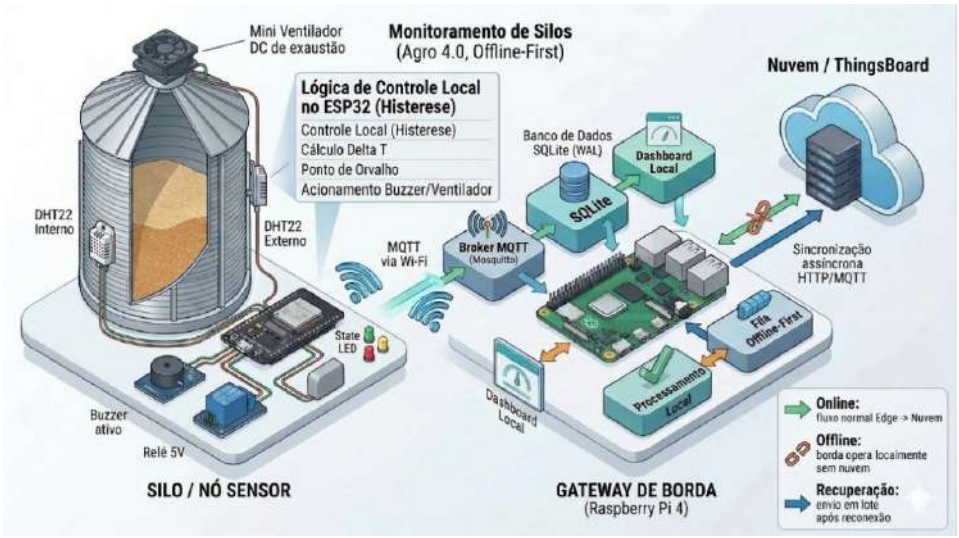


Figura 1. Diagrama da arquitetura IoT Híbrida Edge-Cloud proposta. Fonte: Elaborado pelo autor com auxílio de Inteligência Artificial (2026).

Fluxo passo a passo dos *scripts* Python

1. **backend_collector.py**: Conecta MQTT (*paho-mqtt*) → insere SQLite (status_sync=0) com leituras (T/UR interna/externa, ΔT, ponto de orvalho via Magnus).
2. **dashboard_local.py** (Streamlit): Lê SQLite → aplica **Isolation Forest** para anomalias → exibe *dashboard web* local (gráficos, histórico, *status* sincronização).

3. **cloud_sync.py / cloud_sync_things_board.py**: Detecta reconexão → envia lotes JSON pendentes via HTTP (TagoIO/ThingsBoard) → marca **status_sync**=1 só em HTTP 200.

Lógica de atuadores (lógica direta no nó sensor):

- **Liga**: $T_{interna} \geq 32°C$ **E** $\Delta T \geq 1{,}5°C$ **OU** umidade ≥ 70% (2 amostras).
- **Desliga**: $T_{interna} \leq 31°C$ **OU** $\Delta T \leq 0{,}5°C$ **E** umidade < 66% (2 amostras).
- **Contingência**: sensor externo falha → usa apenas $T_{interna}$.

5. Avaliação experimental

5.1. Planejamento de pesquisa

A avaliação experimental foi conduzida em duas etapas complementares: simulação em *digital twin* e testes físicos, permitindo validar a arquitetura frente às perguntas de pesquisa Q1–Q4. A primeira etapa utilizou o Wokwi para prototipagem virtual do ESP32, sensores DHT22 e relé, possibilitando verificar integração, lógica embarcada e comunicação MQTT antes da montagem física. O projeto virtual permaneceu disponível em https://wokwi.com/projects/459654362248869889, registrando a evolução do protótipo.

Na segunda etapa, o sistema foi implementado com ESP32 DevKit V1, dois DHT22 (interno e externo), LEDs indicadores, *buzzer* e miniventilador 5V acionado por relé. O *gateway* foi configurado em um Raspberry Pi 3B com sistema operacional otimizado para operação em modo *headless*, executando o *broker* MQTT, o banco SQLite e os serviços Python responsáveis pela coleta, *dashboard* local e sincronização.

Os testes físicos incluíram ciclos de conectividade de 30 minutos *online* e 90 minutos *offline* durante 48 horas, avaliando continuidade operacional, retenção de dados, sincronização pós-reconexão, acionamento de atuadores e consumo de recursos. Em uma etapa complementar, o protótipo passou a comparar simultaneamente ambiente interno e externo,

permitindo avaliar a coerência física da lógica de exaustão, embora sem alterar o foco principal da pesquisa, centrado na resiliência *offline-first*.

O planejamento experimental foi estruturado para verificar: a continuidade operacional local (Q1), a integridade dos dados retidos e sincronizados (Q2), a autonomia dos atuadores durante falhas de rede (Q3) e o impacto da comparação interna-externa na lógica de controle (Q4).

5.2. Resultados e Discussão

A Q1 investigou a continuidade operacional local do sistema durante períodos de conectividade intermitente, verificando se o monitoramento e o controle permaneciam funcionais mesmo na ausência de conexão com a nuvem. Para isso, foram realizados testes de desconexão da rede por períodos de até 90 minutos, simulando falhas comuns em áreas remotas. Durante esses períodos, a capacidade do sistema de coletar, armazenar e processar dados localmente foi monitorada de perto, garantindo que as operações críticas não fossem interrompidas.

A Q2 avaliou a integridade dos registros de dados, comparando as leituras retidas localmente no banco de dados SQLite com as sincronizadas após o restabelecimento da conexão. Este teste envolveu a verificação da consistência e completude dos dados, assegurando que nenhum dado fosse perdido ou corrompido durante os ciclos de desconexão e reconexão. Foram realizados testes de estresse por 48 horas (90 min offline/30 min online) que validaram a resiliência do sistema, demonstrando sua robustez em cenários de conectividade intermitente. A Tabela 2 apresenta os resultados quantitativos detalhados.

Tabela 2. Resultados quantitativos do teste de estresse de 48 horas.

Métrica	**Valor**
Duração Total do Teste	48 horas
Intervalo de Coleta de Dados	5 segundos
Total de Leituras Esperadas	34.560
Leituras Coletadas Durante Períodos Offline	34.560
Leituras Sincronizadas com a Nuvem Pós-Reconexão	34.560
Taxa de Integridade dos Dados (Q2)	**100%**

A Q3 focou na autonomia de resposta dos atuadores locais, verificando se os alertas sonoros e o acionamento do ventilador eram processados independentemente da nuvem, demonstrando a capacidade de decisão da borda. Para isso, foram simuladas condições que exigiriam a atuação dos dispositivos, como picos de temperatura ou umidade, e observou-se a resposta do sistema sem intervenção externa. A Figura 2 ilustra a coleta contínua de dados offline e a sincronização pós-reconexão, evidenciando a capacidade do sistema de manter o registro de informações mesmo sem conectividade.

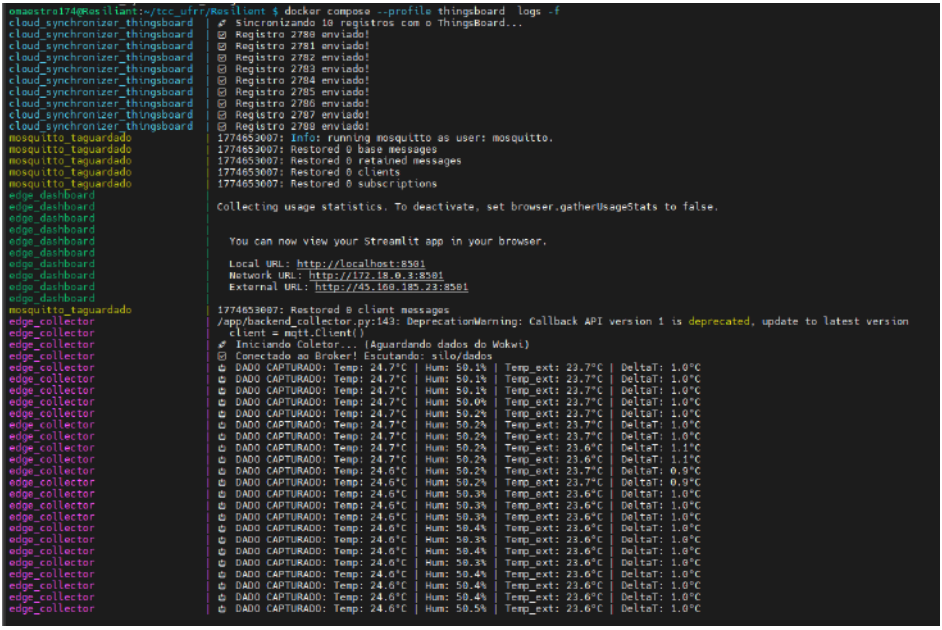


Figura 2. Coleta contínua de dados offline e a sincronização.

A Q4 comparou o desempenho da lógica de controle, contrastando a abordagem inicial baseada apenas em variáveis internas isoladas com a lógica evoluída que incorpora a comparação entre ambientes interno e externo, incluindo o diferencial térmico e o ponto de orvalho. A conservação de grãos em silos exige controle rigoroso de umidade e temperatura. A Tabela 3 detalha parâmetros críticos.

Tabela 3. Parâmetros críticos de umidade e risco térmico por tipo de grão.

Grão	Umidade Ótima	Limite para Proliferação de Fungos	Risco Térmico
Milho	13 – 14%	> 15%	> 0,5°C/dia
Soja	11 – 12%	> 13%	> 0,7°C/dia
Arroz	12 – 13%	> 14%	> 0,3°C/dia

Nota. Adaptado de Portella & Eichelberger (2001); Ayres et al. (2021).

Este delineamento experimental permitiu avaliar de forma abrangente a resiliência e a eficácia da arquitetura proposta em condições adversas de conectividade. A Tabela 4 resume a matriz de decisão do nó sensor e as ações dos atuadores locais.

Tabela 4. Matriz de decisão do nó sensor e ações dos atuadores.

Estado Detectado	Condição Técnica	Ação do Atuador (Hardware)	Notificação (Dashboard)
Normal	Temperatura e umidade em faixa segura	Nenhum (*Standby*)	*Status* normal
Exaustão favorável	Temperatura interna elevada e ambiente externo mais frio	Ativa Miniventilador (Exaustão)	Ventilação ativa por diferencial térmico
Alerta térmico	Temperatura interna crítica	Ativa ventilação e alarme sonoro	Temperatura crítica
Alerta higroscópico	Umidade interna elevada	Sinalização local	Umidade elevada

Os resultados mostraram que: (Q1) a arquitetura manteve sua operação durante períodos offline planejados, com monitoramento local ativo; (Q2) após a reconexão, a sincronização recuperou todos os dados históricos sem perdas ou duplicidades; (Q3) o processamento na borda garantiu alertas e controle de atuadores independentemente da nuvem, com resposta direta no dispositivo; e (Q4) a estratégia que compara ambientes interno e externo gerou decisões de exaustão fisicamente mais coerentes do que apenas usar limiares internos.

Sob a ótica comparativa com a literatura, os achados indicam que a solução proposta preenche lacunas críticas de operação e custos identificadas nos trabalhos relacionados. Diferentemente das abordagens de

Bascıftcı, Unlu e Dasdemır (2021) e das arquiteturas focadas em redes ativas de Saxena et al. (2023) e Sharma et al. (2025), que pressupõem conectividade contínua ou otimização de banda para manter o desempenho, o protótipo desta pesquisa validou na prática o paradigma offline-first suportando apagões severos sem perda de informações. Essa robustez resolve diretamente a vulnerabilidade exposta por Nyamuryekung'e et al. (2023), que chegaram a registrar a perda de 15 dias de dados em redes remotas devido à ausência de persistência local em seus sensores. Em contraste com Sanchez-Iborra et al. (2023) e Durgalakshmi et al. (2025), cujas inovações na borda focam predominantemente na execução de modelos locais de inteligência artificial (elevando o custo e a complexidade técnica), este estudo demonstrou que a combinação de um banco relacional leve (SQLite) aliado a uma lógica baseada em diferencial térmico atua como um mecanismo anti-falhas formidável e acessível. O uso desse banco garantiu que não houvesse perda do histórico de dados (resultado convergente com a taxa de 100% de integridade na Q2, preenchendo com sucesso a lacuna de validações práticas e de baixo custo levantada por Kalyani e Collier (2021).

No recorte aplicado à Agroindústria 4.0 em Roraima, essa evidência é particularmente relevante, pois combina viabilidade econômica, resiliência operacional e capacidade de sincronização em áreas de conectividade irregular. Apesar disso, algumas limitações permanecem. Os testes foram conduzidos em ambiente laboratorial controlado, com recipiente de pequeno porte e hardware de baixo custo com valor aproximado de R$ 479,90 (quatrocentos e setenta e nove reais e noventa centavos), não reproduzindo integralmente a complexidade térmica, higroscópica e aerodinâmica de um silo em escala real. Além disso, a avaliação concentrou-se em falhas de internet e não contemplou desligamentos abruptos por falhas simultâneas de energia elétrica. Assim, a generalização plena da confiabilidade requer novas etapas experimentais com redundância energética e validação em campo.

6. Conclusão

Este estudo desenvolveu e validou um protótipo de arquitetura IoT híbrida edge-cloud com resiliência offline-first, persistência local via SQLite e sincronização assíncrona para monitoramento em áreas de conectividade instável. Foram avaliados: continuidade operacional em cenários offline,

integridade dos dados no SQLite, processamento de borda autônomo no ESP32 e coerência na lógica de exaustão baseada na comparação entre ambientes interno e externo do silo. Os resultados indicam que o sistema é resiliente, mantém dados íntegros, opera alertas independentes da nuvem e aprimora decisões de exaustão, apresentando uma solução viável para conservação de grãos no agronegócio de Roraima.

Como trabalhos futuros, sugere-se aprofundar a otimização energética do gateway de borda, visando a implantação em cenários autônomos com alimentação solar. Além da integração de algoritmos de aprendizado de máquina para detecção preditiva de anomalias e a implementação de uma interface de usuário mais rica para o dashboard local para expandir as capacidades e a usabilidade do sistema. A aplicação desta arquitetura em outros contextos da Agroindústria 4.0, como monitoramento de rebanhos ou irrigação inteligente, também representa um caminho para futuras pesquisas, consolidando o potencial da computação de borda e da abordagem offline-first para impulsionar a transformação digital, em especial em regiões com desafios de conectividade.

7. Referências

Ayres, G., Nascimento, J., & Francisco, P. (2021). *Sistema integrado de monitoramento de produtos armazenados em silos*. Campina Grande: EPTEC. Recuperado de: https://dspace.sti.ufcg.edu.br/bitstream/riufcg/26763/1/E-BOOK%20-%20SISTEMA%20INTEGRADO%20DE%20MONITORAMENTO%20DE%20PRODUTOS%20ARMAZENADOS%20EM%20SILOS%20-%20CTRN%202021.pdf

Bascıftcı, F., Unlu, T., & Dasdemır, A. (2021). Monitoring Grain Silos Instantly with IoT Based Control System. *The Eurasia Proceedings of Science, Technology, Engineering and Mathematics*, *14*, 15-23. https://doi.org/10.55549/epstem.1050154

Batistatos, M., Cola, T., Kourtis, M., Apostolopoulou, V., Xilouris, G., & Sagias, N. (2025). AGRARIAN: A Hybrid AI-Driven Architecture for Smart Agriculture. *Agriculture*, *15*(8), 904. https://doi.org/10.3390/agriculture15080904

Durgalakshmi, B., Umapathi, S., Priya, M., Sabitha, M. S., Ilakkiya, E., & Kumar, M. U. R. (2025). A Comprehensive Framework for Smart Agriculture: Integrating IoT, Edge Computing, and AI for Scalable, Transparent, and Adaptive Crop Management. In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies, (Vol. 1, pp. 350-356)*. https://doi.org/10.5220/0013865600004919

Gaffney, K., Prammer, M., Brasfield, L., Hipp, D., Kennedy, D., & Patel, J. (2022). SQLite: Past, Present, and Future. *Proceedings of the VLDB Endowment*, *15*(11), 3535-3546. https://doi.org/10.14778/3554821.3554842

Governo de Roraima. (2025). *Roraima bate recorde histórico na produção agrícola, aponta Conab*. Portal do Governo de Roraima. Recuperado de: https://portal.rr.gov.br/roraima-bate-recorde-historico-na-producao-agricola-aponta-conab/

Kalyani, Y., & Collier, R. (2021). A systematic survey on the role of cloud, fog, and edge computing combination in smart agriculture. *Sensors*, *21*(17), 5922. https://doi.org/10.3390/s21175922

Kubota, L., & Rosa, M. B. (2023). *TD 2867 – Internet das coisas no Brasil:* Análise e recomendação de políticas com ênfase no agronegócio (Texto para Discussão nº 2867). IPEA. https://doi.org/10.38116/td2867

Maddumala, R. (2025). Comparison of Edge vs. Cloud Computing Architectures in IoT-Based Smart Agriculture Systems. *Global Journal of Engineering and Technology Advances*, *5*. Recuperado de: https://www.sciencexcel.com/articles/KBKGPNUmqgbPvLKYo4oswrM88bXcRjBpUt1vJWYL.pdf

Nyamuryekung'e, S., Duff, G., Utsumi, S., Estell, R., McIntosh, MM, Funk, M., Cox, A., Cao, H., Spiegal, S., Perea, A., & Cibils, AF (2023). Real-Time Monitoring of Grazing Cattle Using LORA-WAN Sensors to Improve Precision in Detecting Animal Welfare Implications via Daily Distance Walked Metrics. *Animals*, *13*(16), 2641. https://doi.org/10.3390/ani13162641

Portella, J., & Eichelberger, L. (2001). *Secagem de Grãos*. Passo Fundo: Embrapa. Recuperado de:

https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/820442/1/ID8673LV0388.pdf

Sanchez-Iborra, R., Zoubir, A., Hamdouchi, A., Idri, A., & Skarmeta, A. (2023). Intelligent and Efficient IoT Through the Cooperation of TinyML and Edge Computing. *Informatica*, *34*(1), 147–168. https://doi.org/10.15388/22-INFOR505

Saxena, A. K., Pandey, R., & Singh, N. K. (2023). Latency Analysis and Reduction Methods for Edge Computing. In *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)*. Sonbhadra, India. https://doi.org/10.1109/AIC57670.2023.10263946

Schwab, K. (2016). *A quarta revolução industrial*. São Paulo: Edipro.

Sharma, E., Saini, A., & Singh, D. (2025). Edge-Cloud Hybrid Architecture for IoT Systems. *International Journal for Multidisciplinary Research*, *7*(1), 1-6. Recuperado de: https://www.ijfmr.com/papers/2025/6/62246.pdf

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, *3*(5), 637-646. https://doi.org/10.1109/JIOT.2016.2579198

SQLite Project. (2026). *Write-Ahead Logging*. Recuperado de: https://sqlite.org/wal.html

Tariq, M., Saqib, S., Mazhar, T., Khan, M., Shahzad, T., & Hamam. H. (2025). Edge-enabled smart agriculture framework: Integrating IoT, lightweight deep learning, and agentic AI for context-aware farming. R*esults in Engineering*, *28*. https://doi.org/10.1016/j.rineng.2025.107342

Yang, D., Wu, J., & He, Y. (2024). Optimizing the Agricultural Internet of Things (IoT) with Edge Computing and Low-Altitude Platform Stations. *Sensors*, *24*(21), 7094. https://doi.org/10.3390/s24217094

Yu, P., Teng, F., Zhu, W., Shen,C., Chen, Z., & Song, J. (2025). Cloud–edge–device collaborative computing in smart agriculture: architectures, applications, and future perspectives. *Frontiers in Plant Science*, *16*, 1668545. https://doi.org/10.3389/fpls.2025.1668545

Capítulo 16

Monitoramento IoT para Agricultura 4.0: Arquitetura com LoRa P2P e Edge Computing para Eficiência Hídrica

Pedro Henrique Vale de Lucena Brasil
Marcelle Alencar Urquiza
Herbert Oliveira Rocha
Universidade Federal de Roraima

pedro12brasil@gmail.com
{marcelle.urquiza, herbert.rocha}@ufrr.com

1. Introdução

A agricultura mundial consome aproximadamente 70% da água doce disponível, sendo que a ineficiência estimada em 40–70% nos métodos convencionais de irrigação por inundação contribui significativamente para o desperdício hídrico (ONU, 2018). Nesse contexto, a transição para a Agricultura 4.0 busca aplicar a digitalização e a Internet das Coisas (IoT) no campo, enfrentando, contudo, uma barreira primária: a ausência de conectividade em áreas rurais remotas.

Sistemas convencionais dependem de redes celulares (4G/5G) com cobertura escassa no interior. Tecnologias de baixo consumo e curto alcance, como Zigbee (até ~100 m NLoS) e Bluetooth (até ~30 m), são inviáveis em grandes extensões, enquanto soluções GPS e comunicação via satélite apresentam custos operacionais proibitivos para o pequeno produtor. O padrão LoRa Wide Area Network (LoRaWAN) emergiu como alternativa robusta para redes de longo alcance; contudo, sua topologia exige que o Gateway possua acesso contínuo à internet para rotear os pacotes recebidos até um Network Server — seja ele hospedado em nuvem pública ou em servidor local. Essa dependência estrutural de conectividade inviabiliza a adoção do LoRaWAN em propriedades rurais remotas sem infraestrutura de telecomunicações.

Assim, formula-se a questão de pesquisa central: O protocolo LoRa em topologia Ponto-a-Ponto (P2P), aliado ao Edge Computing, é capaz de garantir taxa de entrega de pacotes aceitável e processamento autônomo para viabilizar o monitoramento de eficiência hídrica em áreas rurais isoladas?

O objetivo geral é desenvolver e validar uma PoC de arquitetura IoT autônoma e de baixo custo, baseada em LoRa P2P e Edge Computing, para apoiar a tomada de decisão no manejo hídrico rural. Os objetivos específicos são: (a) projetar e programar um nó sensor remoto (ESP32 + sensor capacitivo) com Deep Sleep para maximizar a autonomia energética; (b) configurar um Raspberry Pi como Edge Gateway para recepção de pacotes de rádio, execução local da lógica heurística e sincronização tolerante a falhas com o banco de dados na nuvem; e (c) implementar lógica heurística na borda para avaliar dados e inibir acionamentos desnecessários, promovendo economia hídrica.

2. Fundamentação teórica

A IoT vai além da mera conexão de objetos, permitindo interações inteligentes entre dispositivos físicos e sistemas computacionais. Bassi et al. (2013) definem o IoT Architectural Reference Model (IoT ARM) como estrutura central para o desenvolvimento de sistemas interoperáveis, organizando o ecossistema em camadas funcionais: Dispositivo, Rede, Serviço e Aplicação. Como exemplo prático, um sensor de umidade do solo representa a Camada de Dispositivo, enquanto o rádio LoRa compõe a Camada de Rede, e o banco de dados local com lógica de decisão compõe a Camada de Aplicação.

Neste trabalho, adota-se uma versão simplificada do IoT ARM focada nas Camadas de Dispositivo e Rede, adaptada para operar em arquitetura de borda e alinhada aos pilares Cyber-Physical Systems (CPS) e Industrial IoT (IIoT) da Indústria 4.0. Essa escolha justifica-se por três razões: (i) a topologia P2P elimina a Camada de Serviço em nuvem, reduzindo dependência de infraestrutura externa; (ii) a lógica de decisão é executada localmente, garantindo operação mesmo sem internet; e (iii) o custo total do sistema permanece acessível ao pequeno produtor rural.

O LoRa (Long Range) é uma tecnologia de modulação física baseada em Chirp Spread Spectrum (CSS), que viabiliza longo alcance com baixo consumo energético. Diferente do LoRaWAN, que exige gateways comerciais e Network Servers na nuvem, a topologia P2P estabelece link direto entre os nós, eliminando custos de infraestrutura. Hashmi et al. (2024) comprovam experimentalmente que a frequência sub-GHz (915 MHz) penetra barreiras vegetais de forma superior ao Wi-Fi (2,4 GHz), com RSSI de –95 dBm mantido a 300 m em ambiente externo, sendo ideal para ambientes agrícolas.

O Edge Computing reduz a latência ao processar dados próximos à fonte. O nó sensor (ESP32) foca na coleta e transmissão, enquanto o Raspberry Pi 4 atua como Edge Server. A escolha justifica-se pela carga computacional acumulada: os processos da aplicação (backend Python, PostgreSQL e servidor React.js) somam aproximadamente 350 MB de RAM, operando sobre o Raspberry Pi OS Lite (sem interface gráfica, com cerca de 100 MB adicionais), totalizando aproximadamente 450 MB — valor que ultrapassa em mais de 800 vezes a memória SRAM disponível no ESP32 (520 KB). O Raspberry Pi gerencia ainda um buffer local de contingência (Store and Forward), garantindo autonomia de decisão mesmo sem internet, enquanto o banco de dados definitivo e o servidor web são hospedados em um servidor virtual privado (Virtual Private Server — VPS) na nuvem.

3. Trabalhos relacionados

Foram selecionadas investigações aderentes ao escopo de conectividade rural e processamento de dados na borda. O Tabela 1 sintetiza as metodologias, métricas e resultados de cada estudo, evidenciando o posicionamento desta PoC em relação à literatura.

Tabela 1. Comparativo entre trabalhos relacionados e esta Prova de Conceito (PoC).

Estudo / Autor(es)	Tecnologias Empregadas	Métricas / Resultados	Limitação Principal	Diferencial desta PoC
Bassi et al. (2013)	Modelo IoT ARM genérico.	Framework conceitual; sem métricas experimentais.	Não endereça implantação offline ou rural.	Adapta IoT ARM para topologia 100% autônoma, sem internet.

Estudo / Autor(es)	Tecnologias Empregadas	Métricas / Resultados	Limitação Principal	Diferencial desta PoC
Chen et al. (2020)	Controle Preditivo (MPC), sensores de potencial hídrico.	Redução de estresse hídrico simulada; carga computacional >1 MB RAM.	Exige hardware de alto processamento; inviável em campo isolado.	Substitui MPC por heurística leve (<10 KB RAM) no Raspberry Pi.
Hashmi et al. (2024)	Zigbee, LoRa, testes outdoor.	LoRa: RSSI −95 dBm a 300 m; PDR >90% (LoS). Zigbee: <100 m NLoS.	Apenas análise de camada física; sem decisão local.	Integra LoRa P2P (NLoS/500 m) à decisão heurística na borda.
Chang et al. (2019)	LoRa P2P, Machine Learning, TDMA.	PDR >95% com TDMA; ML requer treinamento prévio de dados.	ML e TDMA aumentam custo e complexidade para o pequeno produtor.	Elimina ML; adota heurística estrita de menor custo computacional.
Soultane et al. (2024)	Edge Computing, Drones, Redes Neurais.	Acurácia >92%; latência <200 ms com GPU embarcada.	Dependência de drones e GPU; custo proibitivo para agricultura familiar.	Elimina drones e IA pesada; opera com Raspberry Pi.
Brasil (2025) – Este Trabalho	Edge Computing, LoRa P2P (NLoS), Lógica Heurística.	PDR 100% (LoS/100 m), 55% (NLoS/500 m); RSSI −112,70 dBm; SNR −6,67 dB; IC 95%.	Testes em ambiente urbano; sem ensaios longitudinais in situ.	Independência de nuvem; validação experimental com métricas de rádio.

Nota. PDR = Taxa de Entrega de Pacotes; RSSI = Received Signal Strength Indicator; SNR = Relação Sinal-Ruído; MPC = Controle Preditivo por Modelo; ML = Machine Learning; NLoS = Sem Linha de Visada Direta; LoS = Linha de Visada Direta; IC = Intervalo de Confiança.

Conforme o Tabela 1, a lacuna explorada reside na integração entre validação experimental de rádio (PDR/RSSI/SNR com IC 95%) e decisão autônoma na borda, combinação ausente nos trabalhos analisados. Em detrimento de algoritmos preditivos (Chen et al., 2020) ou IA dependente de servidores em nuvem (Soultane et al., 2024), esta PoC adota lógica heurística leve operando integralmente no Edge Gateway, eliminando dependência de infraestruturas externas de telecomunicação.

4. Solução computacional

A implementação seguiu arquitetura em camadas conforme Bassi et al. (2013): (i) Camada de Percepção Física (Edge), responsável pela coleta e

atuação; (ii) Camada de Rede e Gateway, pelo transporte; e (iii) Camada de Aplicação e Processamento, pela persistência e lógica de decisão.

O nó sensor (Heltec V3 ESP32-S3 + SX1262) integra nativamente o rádio LoRa, reduzindo pontos de falha. O firmware (C++/Arduino Framework) implementa Máquina de Estados Finitos (MEF) com quatro estados. O estado S0 (Deep Sleep) é o padrão, com CPU e rádio desligados e consumo de aproximadamente 10 µA. O estado S1 (Wakeup) corresponde à inicialização dos periféricos. O estado S2 (Sense e Transmit) compreende a leitura do sensor capacitivo, encapsulamento em JSON e transmissão via LoRa P2P. O estado S3 (Sleep) representa o retorno ao Deep Sleep por 15 minutos. A transição S0→S1 é acionada pelo temporizador de Real-Time Clock (RTC) interno. O Edge Gateway é composto por módulo Heltec receptor como Bridge e Raspberry Pi 4 Model B como Edge Server, com buffer local até sincronização bem-sucedida com a nuvem. O relacionamento entre os componentes é ilustrado no Diagrama de Sequência da Figura 2A (Seção 5.5).

Tabela 2. Requisitos funcionais e não funcionais do sistema.

ID	Tipo	Descrição
RF01	Funcional	Coletar dados de umidade (15 min), persistir em PostgreSQL e exibir em Dashboard SPA.
RF02	Funcional	Gerar alertas e intervir localmente com base em lógica heurística de restrição (Umidade × Temperatura).
RNF01	Não Funcional	Garantir autonomia energética do nó sensor por baterias via rotina de Deep Sleep.
RNF02	Não Funcional	A comunicação LoRa P2P deve manter a operação e receber dados em cenários NLoS em até 500 metros.
RNF03	Não Funcional	O sistema de borda deve assegurar integridade de dados (checksum) e mitigar perdas isoladas de pacotes.

Nota. RF = Requisito Funcional; RNF = Requisito Não Funcional; SPA = Single Page Application; NLoS = Sem Linha de Visada Direta.

A arquitetura de software adota divisão em duas camadas. Na borda, o Raspberry Pi executa daemon Python responsável pela leitura serial, aplicação da lógica heurística e persistência local em PostgreSQL via Docker. Na nuvem, um VPS hospeda o banco de dados definitivo, um backend secundário em Spring Boot (Java) que realiza proxy da API meteorológica HG

Brasil via endpoint /api/clima, e o Dashboard React.js (Single Page Application — SPA). Os requisitos que nortearam o desenvolvimento estão consolidados na Tabela 2.

O sistema adota lógica de decisão baseada em regras condicionais determinísticas, implementada como Máquina de Estados Finitos de três estados de saída — denominada "lógica heurística" no sentido adotado pela literatura de controle de irrigação (Chang et al., 2019; Chen et al., 2020), em contraposição a algoritmos preditivos ou de aprendizado de máquina. O algoritmo avalia a umidade do solo (Us) cruzada com a temperatura do ar (Tar), gerando: (i) CRÍTICO, Us < 40% e Tar > 30°C: irrigação imediata por risco térmico; (ii) ATENÇÃO, Us < 40% e Tar ≤ 30°C: planejar irrigação; (iii) NORMAL, Us ≥ 40%: monitorando. O limiar de Us = 40% reflete a capacidade de campo típica de solos argilosos (Allen et al., 1998); Tar = 30°C representa o ponto de aumento significativo da evapotranspiração em culturas tropicais.

Nota de implementação: na versão atual da PoC, o nó sensor transmite exclusivamente a umidade do solo no payload LoRa ({"id", "token", "umid"}), pois o hardware utilizado não integra sensor de temperatura. O parâmetro Tar assume valor padrão de 0°C quando ausente, impedindo a ativação do estado CRÍTICO na implementação atual — o sistema opera entre ATENÇÃO e NORMAL. A incorporação de sensor dedicado (DS18B20 ou DHT22) constitui diretriz prioritária para a evolução do protótipo. A tomada de decisão é formalizada pelo pseudocódigo da Figura 1.

Algoritmo 1 – Lógica Heurística de Decisão de Irrigação (Edge Gateway)

```
// ── Definição de Limiares
CONSTANTE UMIDADE_CRITICA ← 40.0  // % capacidade de campo
CONSTANTE TEMP_CRITICA ← 30.0  // °C limiar de evapotranspiração

// ── Entrada de Dados
RECEBER pacote via LoRa P2P contendo umidade, temperatura

// ── Lógica de Decisão

SE (umidade < UMIDADE_CRITICA) E (temperatura > TEMP_CRITICA) ENTÃO
   RETORNAR estado ← "CRÍTICO", ação ← "Irrigação Imediata (Risco Térmico)"
SENÃO SE (umidade < UMIDADE_CRITICA) ENTÃO
   RETORNAR estado ← "ATENÇÃO", ação ← "Planejar Irrigação"
```

```
    SENÃO
        RETORNAR estado ← "NORMAL", ação ← "Monitorando"
    FIM SE

    PERSISTIR (estado, ação, umidade, temperatura) EM PostgreSQL
```

Figura 1. Pseudocódigo da lógica heurística de decisão de irrigação executada no Edge Gateway.

Nota. A lógica heurística é executada localmente no Edge Gateway (Raspberry Pi 4), sem dependência de rede externa. O limiar de umidade (40%) segue (Allen et al., 1998) para solos argilosos; o limiar térmico (30°C) corresponde ao ponto de aumento significativo da evapotranspiração em culturas tropicais.

Complementarmente à heurística executada no Edge Gateway, o Dashboard Web implementa uma camada de recomendação estendida no frontend, que cruza a umidade do solo com a probabilidade de chuva obtida via API meteorológica HG Brasil (proxied pelo backend Spring Boot em /api/clima). Essa camada adicional produz quatro recomendações de ação: (i) Sensor Offline, quando o dispositivo não responde há mais de 2 minutos; (ii) Aguardar Chuva, quando probabilidade de chuva > 60% e umidade ≥ 40%, evitando irrigação desnecessária; (iii) Irrigar Agora, quando umidade < 40%, independente da previsão; e (iv) Monitorando, condições normais ou de atenção. Essa camada opera exclusivamente no cliente React e não substitui a heurística da borda, que permanece como mecanismo primário de decisão autônoma e offline.

A integridade inicial da PoC é estabelecida por Token Pré-Compartilhado (PSK) estático, inserido no payload JSON transmitido pelo nó sensor via LoRa P2P. O Gateway de Borda valida o token em duas etapas: verifica a presença do campo "token" no JSON recebido e compara o valor com o token configurado no servidor. Pacotes com token ausente ou inválido são descartados antes de qualquer processamento ou persistência. O PSK estático torna a estrutura vulnerável a Replay Attacks em raio de aproximadamente 2 km; para produção, propõe-se substituição por assinaturas Hash-based Message Authentication Code (HMAC) ou nonce dinâmico (timestamp Unix de 32 bits), com overhead estimado de apenas 4 bytes por pacote.

Quanto à escalabilidade, com fator de espalhamento SF igual a 7 e largura de banda BW de 125 kHz, um pacote de 20 bytes ocupa

aproximadamente 56 ms de Time on Air (ToA), sujeito ao limite regulatório de 1% de Duty Cycle na banda de 915 MHz (Chang et al., 2019). Com ciclos de 15 minutos entre leituras, a topologia P2P comporta dezenas de nós sem colisão.

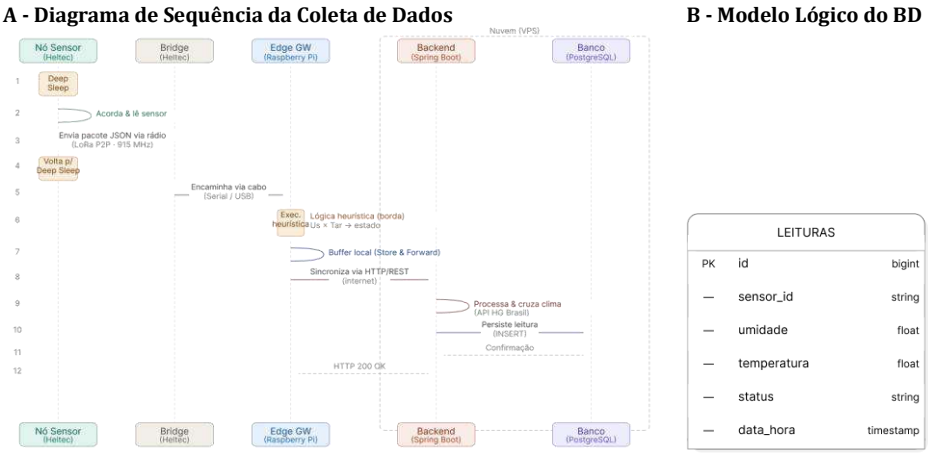


Figura 2. (A) Diagrama de Sequência da Coleta de Dados e (B) Modelo Lógico do Banco de Dados.

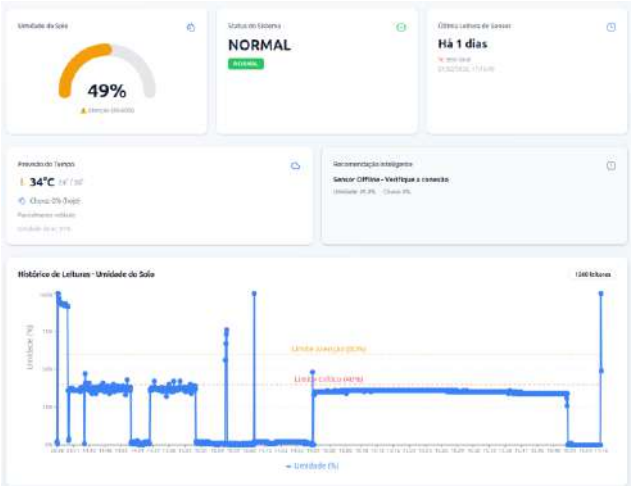


Figura 3. Interface do Usuário (Dashboard Web).

Para centenas de nós ou ciclos mais frequentes, recomenda-se Time Division Multiple Access (TDMA) ou migração para LoRaWAN. A autonomia energética é assegurada pelo Deep Sleep, reduzindo o consumo do nó a aproximadamente 100 μA, prolongando a operação por meses com baterias compactas de 3.000 mAh.

O Diagrama de Sequência (Figura 2A) ilustra o fluxo híbrido: o sensor em Deep Sleep transmite o pacote LoRa à Bridge receptora, que o repassa ao Edge Gateway para execução da heurística local e armazenamento em buffer. Posteriormente, o Gateway sincroniza os dados via HTTP com a Nuvem (VPS), onde ocorre a persistência final no banco de dados PostgreSQL.

5. Avaliação experimental

5.1. Planejamento de pesquisa

A pesquisa classifica-se como experimental de abordagem quantitativa (Prodanov & Freitas, 2013), consistindo no desenvolvimento de uma PoC submetida a cenários de teste controlados para validação das hipóteses levantadas.

O sistema adota topologia P2P estruturada em três componentes principais. O primeiro é o Nó Sensor, composto pelo microcontrolador Heltec V3 (ESP32 + LoRa 915 MHz) acoplado a um sensor capacitivo de umidade do solo. O segundo é o Gateway de Borda, formado por um módulo Heltec receptor conectado via USB a um Raspberry Pi 4. O terceiro é a Camada de Nuvem, hospedada em VPS, responsável por receber a sincronização dos dados do Edge Gateway via HTTP, cruzar as informações com APIs climáticas e persistir os dados em banco PostgreSQL para exibição no Dashboard.

A separação de responsabilidades é estrita: o processamento heurístico ocorre integralmente na borda, garantindo que eventuais quedas no link de internet não interrompam a telemetria LoRa P2P nem a visualização local pelo produtor.

A validação foi dividida em três etapas: (i) Qualidade de Serviço (QoS) — transmissão em cenários de Qualidade de Serviço (QoS): linha de visada direta (Line of Sight — LoS, a 100 m) e sem linha de visada (Non-Line of Sight — NLoS, a 500 m), além de avaliação exploratória a 6 km; (ii) Métricas de

Rede — avaliação quantitativa por Taxa de Entrega de Pacotes (Packet Delivery Rate — PDR), Indicador de Potência do Sinal Recebido (Received Signal Strength Indicator — RSSI) e Relação Sinal-Ruído (Signal-to-Noise Ratio — SNR), com Intervalo de Confiança de 95% (equação de Wald modificada, n = 100 pacotes/distância); e (iii) Eficiência Energética — eficácia do ciclo Deep Sleep validada por auditoria do log serial do ESP32.

A avaliação baseou-se em três métricas de rádio: PDR (confiabilidade do link), RSSI em dBm (força do sinal recebido) e SNR em dB (imunidade a interferências). O IC de 95% via equação de Wald modificada estabeleceu a margem de erro estatística para cada cenário. Os resultados do Tabela 3 consolidam essas métricas, permitindo comparação direta com as bases da literatura (Tabela 1).

5.2. Resultados e Discussão

Os ensaios de telemetria avaliaram três cenários de distância a partir do Edge Gateway (n = 100 pacotes/cenário), consolidados na Tabela 3.

Tabela 3. Resultados experimentais de conectividade LoRa P2P por cenário de distância.

Cenário	Distância	PDR (IC 95%)	RSSI médio	SNR médio
LoS – Linha de visada direta	100 m	100% (IC 95%: 96,4–100%)	−84,00 dBm	+11,50 dB
NLoS – Urbano com obstáculos arquitetônicos	500 m	55% (IC 95%: 44,9–65,1%)	−112,70 dBm	−6,67 dB
Alcance extremo – Zona de Fresnel bloqueada	6 km	0%	N/A	N/A

Nota. n = 100 pacotes por cenário. IC calculado pela equação de Wald modificada ($IC = p \pm Z\sqrt{p(1-p)/n}$, $Z = 1{,}96$, $\alpha = 0{,}05$). LoS = Linha de Visada Direta; NLoS = Sem Linha de Visada Direta; RSSI = Received Signal Strength Indicator; SNR = Relação Sinal-Ruído. "N/A" indica ausência de recepção de pacotes.

A 100 m (LoS), o canal apresentou-se limpo com PDR de 100% (IC 95%: 96,4–100%), RSSI de −84,00 dBm e SNR de +11,50 dB, indicando margem de enlace confortável. No cenário de estresse a 500 m (NLoS urbano

com transposição de múltiplas edificações), o RSSI degradou para −112,70 dBm e o SNR registrou −6,67 dB, o valor negativo que atesta a principal vantagem do CSS: demodular pacotes válidos mesmo com sinal abaixo do limiar de ruído. O PDR de 55% (IC 95%: 44,9–65,1%) nesse cenário é consistente com os resultados de Hashmi et al. (2024) em ambiente externo com barreiras severas e é tecnicamente suficiente para a tomada de decisão heurística em ciclos de 15 minutos. A 6 km, o PDR de 0% corrobora a teoria de propagação: o bloqueio da Zona de Fresnel com antenas ao nível do solo exige elevação estrutural (torres ou mastros) para operação em longa distância.

Experimentos longitudinais em campo cultivado não foram contemplados no escopo desta PoC, cujo objetivo primário é validar a viabilidade da arquitetura de comunicação e da lógica de decisão, e não quantificar a economia hídrica em condições reais de cultivo. A lógica heurística atua como bloqueio preventivo de acionamentos temporizados cegos: ao inibir irrigação quando Tar ≤ 30°C ou Us ≥ 40%, o sistema evita desperdício em dias de baixa evapotranspiração. A quantificação exata da economia hídrica demandará ensaios agronômicos in situ, configurando-se como diretriz prioritária para trabalhos futuros.

6. Conclusão

O presente trabalho respondeu à questão de pesquisa de forma afirmativa: é viável desenvolver uma arquitetura IoT autônoma, segura e de baixo custo, baseada em LoRa P2P e Edge Computing, para otimizar o monitoramento hídrico na agricultura familiar. A topologia P2P garantiu independência de telecomunicações, com PDR de 100% a 100 m e 55% a 500 m (NLoS severo, SNR −6,67 dB), evidenciando a robustez da modulação LoRa para o meio rural. A centralização da inteligência no Edge Gateway comprovou o potencial do Edge Computing em inibir irrigações redundantes por inferências heurísticas locais, mitigando o desperdício inerente a sistemas temporizados cegos.

Como trabalhos futuros, sugere-se: (a) execução de ensaios agronômicos longitudinais in situ para quantificação exata da economia hídrica por cultura e tipo de solo; (b) implementação de HMAC e nonces

dinâmicos no payload para mitigar Replay Attacks em produção; (c) simulações de rede (ns-3 ou LoRaSim) para avaliar o impacto do Duty Cycle na escalabilidade para dezenas de nós; (d) aplicação de algoritmos preditivos, como o Filtro de Kalman, para estimativa de umidade entre leituras; e (e) integração de sensor de temperatura dedicado ao nó sensor para habilitar a ativação completa do estado CRÍTICO na heurística de borda. Conclui-se que a arquitetura demonstra potencial como base tecnológica acessível para estudos futuros voltados ao pequeno produtor rural na Agricultura 4.0.

7. Referências

Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (1998). Crop evapotranspiration: Guidelines for computing crop water requirements (FAO Irrigation and Drainage Paper No. 56). Food and Agriculture Organization of the United Nations. Recuperado de: https://www.fao.org/3/x0490e/x0490e00.htm

Bassi, A., Bauer, M., Fiedler, M., Kramp, T., Kranenburg, R., Lange, S., & Meissner, S. (Eds.). (2013). *Enabling things to talk:* Designing IoT solutions with the IoT Architectural Reference Model. Springer. https://doi.org/10.1007/978-3-642-40403-0

Chang, Y.-C., Huang, T.-W., & Huang, N.-F. (2019). A machine learning based smart irrigation system with LoRa P2P networks. In *2019 20th Asia-Pacific Network Operations and Management Symposium (APNOMS) (pp. 1-4)*. IEEE. https://doi.org/10.23919/APNOMS.2019.8893034

Chen, W.-H., Wang, Y.-S., Chen, K.-Y., & Li, T.-H. S. (2020). Theoretical exploration of irrigation control for stem water potential through model predictive control. In *Proceedings of the 2020 American Control Conference (ACC) (pp. 1992-1997)*. IEEE. https://doi.org/10.23919/ACC45564.2020.9147296

Hashmi, A. U. H., Alharbi, H. A. G., & Ijaz, M. F. (2024). Effects of IoT communication protocols for precision agriculture in outdoor environments. *IEEE Access*, *12*, 46410–46421. https://doi.org/10.1109/ACCESS.2024.3381522

Organização das Nações Unidas (ONU). (2018). *The United Nations world water development report 2018:* Nature-based solutions for water. UNESCO.

Prodanov, C. C., & Freitas, E. C. (2013). *Metodologia do trabalho científico:* Métodos e técnicas da pesquisa e do trabalho acadêmico (2. ed.). Universidade Feevale.

Soultane, O. B., Salih-Alj, Y., & Et-taibi, B. (2024). Smart agriculture optimization: Integrating edge computing and AI for enhanced crop management. In *2024 10th International Conference on Applied System Innovation (ICASI) (pp. 1-3)*. IEEE. https://doi.org/10.1109/ICASI60819.2024.10547894

Capítulo 17

Monitoramento da qualidade do ar em Boa Vista/RR com sensores IoT: Dataset aberto e visualização comunitária na perspectiva da Indústria 4.0

Cedric Carol Patrician Williams-Filho
Clenya Rejane Barros de Lima
Universidade Federal de Roraima

cedricstm@gmail.com
clenyabarros@uerr.edu.br

1. Introdução

A qualidade do ar urbano afeta diretamente a saúde pública, inclusive em cidades amazônicas. Segundo a Organização Mundial da Saúde (WHO, 2021), a poluição atmosférica está associada a aproximadamente 7 milhões de mortes prematuras anuais, e mais de 90% da população global vive em áreas com níveis acima dos limites recomendados. No Brasil, a exposição a poluentes atmosféricos representa um grave problema de saúde pública, com custos significativos ao sistema de saúde (Sant'anna et al., 2021). Na Amazônia Legal, queimadas sazonais e incêndios florestais geram concentrações elevadas de poluentes. A ausência de monitoramento contínuo impede o acesso da população a informações de risco à saúde (Barbosa e Citó, 2024; Oliveira et al., 2023).

Diante desse cenário, o acesso à informação ambiental constitui um mecanismo central de empoderamento comunitário (WMO-UNEP, 2024). A disseminação de sensores de baixo custo, como os dispositivos PurpleAir, viabiliza iniciativas de ciência cidadã, promovendo a integração do cidadão ao processo decisório ambiental em seu contexto local (WMO-UNEP, 2024; Brown et al., 2019). Em Boa Vista (RR), essa oportunidade é particularmente relevante devido à ocorrência de queimadas locais e ao transporte regional de poluentes provenientes de outras áreas da Amazônia, em uma cidade com cerca de 436 mil habitantes (IBGE, 2025) e sem monitoramento oficial sistemático da qualidade do ar (Barbosa e Citó, 2024). Assim, surge a

seguinte questão de pesquisa: como tecnologias baseadas em sensores IoT de baixo custo podem ser utilizadas para gerar dados abertos e acessíveis sobre qualidade do ar em cidades amazônicas que não possuem redes institucionais de monitoramento?

O objetivo deste trabalho é desenvolver e disponibilizar um protótipo de sistema integrado de monitoramento e acesso à informação sobre qualidade do ar em Boa Vista/RR, composto por: (i) um *dataset* aberto com 34.791 registros de PM2,5, estruturado segundo o padrão OpenAQ v3 (OpenAQ, 2024); e (ii) a plataforma de visualização comunitária, desenvolvida com tecnologias *web* de código aberto. A proposta não visa substituir redes oficiais de monitoramento, mas complementá-las.

Para orientar o desenvolvimento da solução, este trabalho busca responder às seguintes perguntas de pesquisa: [P1] É possível estruturar um *dataset* aberto de qualidade do ar com cobertura temporal de dois anos e cobertura superior a 90%, utilizando exclusivamente sensores IoT de baixo custo em Boa Vista/RR? [P2] Uma plataforma de visualização comunitária desenvolvida com tecnologias *web* abertas é capaz de traduzir dados de PM2,5 em informação de risco à saúde acessível à população não especializada? [P3] A solução desenvolvida é replicável para outros municípios brasileiros com ausência de monitoramento oficial?

A pesquisa aplica três pilares tecnológicos da Indústria 4.0 (Qin, Liu e Grosvenor, 2016) ao monitoramento ambiental comunitário. O pilar de Internet das Coisas (IoT) é aplicado pelo sensor PurpleAir PA-II-SD conectado à rede, que transmite leituras de material particulado a cada 30 minutos ao longo de dois anos. O pilar de Big Data e Analytics é aplicado no pipeline que processa 34.791 registros brutos e os organiza em três resoluções temporais (30 min, diária e mensal), com validação automatizada e cálculo do Índice de Qualidade do Ar. O pilar de Interoperabilidade é aplicado na estruturação dos metadados segundo o padrão OpenAQ v3 (OpenAQ, 2024), o que viabiliza integração com as plataformas IQAir, WAQI e US EPA AirNow. A transposição desses pilares para o monitoramento ambiental em cidades amazônicas sem rede oficial amplia o escopo da Indústria 4.0 para além dos processos industriais, aproximando-a de aplicações em ciência cidadã e saúde pública.

Neste trabalho, a Seção 2 apresenta a fundamentação teórica; a Seção 3 sintetiza os trabalhos relacionados e explicita a lacuna; a Seção 4 detalha a solução computacional; a Seção 5 descreve a avaliação experimental; e a Seção 6 apresenta as conclusões e trabalhos futuros.

2. Fundamentação teórica

2.1. Ciência Cidadã e Empoderamento Comunitário

A ciência cidadã aplicada ao monitoramento da qualidade do ar envolve a participação ativa da comunidade na coleta de dados ambientais, tendo sensores de baixo custo conectados à IoT como infraestrutura habilitadora (Brown et al., 2019; Proma et al., 2021). Essa abordagem auxilia o acesso à informação, especialmente em contextos em que o monitoramento oficial é inexistente ou insuficiente, com resultados concretos: aumento da conscientização sobre poluição, ampliação da capacidade de reconhecer riscos à saúde e promoção do engajamento cívico (Proma et al., 2021).

2.2. Material Particulado e Índice de Qualidade do Ar

O material particulado fino (PM2,5) refere-se a partículas inaláveis com diâmetro aerodinâmico igual ou inferior a 2,5 micrômetros, capazes de penetrar profundamente no sistema respiratório e atingir a corrente sanguínea, afetando múltiplos órgãos (Pope et al., 2002; Pope & Dockery, 2012). A Organização Mundial da Saúde (OMS) atualizou em 2021 suas diretrizes globais de qualidade do ar, reduzindo o limite anual recomendado para PM2,5 de 10 $\mu g/m^3$ para 5 $\mu g/m^3$, em razão de evidências científicas adicionais sobre os efeitos prejudiciais do PM2,5 à saúde, especialmente em níveis baixos de exposição (WHO, 2021).

No contexto amazônico, poluentes derivados de queimadas representam risco à saúde pública, com registros em Boa Vista que já superaram índices de megacidades com elevada poluição (Barbosa & Citó, 2024). Para comunicar esses riscos à população, o Índice de Qualidade do Ar (IQA) traduz concentrações de poluentes em escala padronizada de 0 a 500, dividida em seis categorias, de "Boa" a "Péssima", orientando a adoção de medidas protetivas adequadas ao nível de exposição (Barbosa & Citó, 2024).

2.3. Sensores de Baixo Custo para Monitoramento de Qualidade do Ar

Os sensores de baixo custo (SBC) representam uma mudança no paradigma de monitoramento da qualidade do ar, permitindo densificação espacial e temporal das medições a custos inferiores aos equipamentos de referência (Morawska et al., 2018). Enquanto monitores FRM/FEM utilizados pelas agências reguladoras custam entre US$ 15.000 e US$ 100.000, os SBC para material particulado estão disponíveis na faixa de US$ 200 a US$ 2.500 (Morawska et al., 2018; WMO-UNEP, 2024). Nessa categoria, os sensores PurpleAir PA-II, comercializados por aproximadamente US$ 209-299 (PurpleAir, 2025), são amplamente utilizados em redes comunitárias, incluindo experiências pioneiras na Amazônia brasileira (Brown et al., 2019; Barbosa; Citó, 2024). O PurpleAir PA-II incorpora dois sensores independentes Plantower PMS5003, que operam pelo princípio de dispersão a laser com cálculo baseado na Teoria de Mie (Plantower, 2016).

Apesar das vantagens de custo e cobertura, os SBC apresentam limitações técnicas reconhecidas na literatura. Sensores óticos tendem a superestimar PM2,5 em cenários de fumaça densa de queimadas, discrepância mitigada por fatores de correção específicos como o proposto pela US EPA para PurpleAir (Barkjohn et al., 2022). A sensibilidade a variáveis ambientais, especialmente umidade relativa elevada, provoca crescimento higroscópico das partículas e inflaciona as leituras, exigindo correção por modelos empíricos (Morawska et al., 2018). O *drift* de calibração ao longo do tempo recomenda colocalização periódica com monitores de referência (Buelvas et al., 2023). Por fim, medições de um único sensor não capturam microvariações intraurbanas, o que justifica a densificação em rede (Carotenuto et al., 2023). Essas limitações delimitam o escopo dos dados para caracterização de tendências e episódios, cabendo calibração com referência para uso regulatório.

3. Trabalhos relacionados

A literatura sobre monitoramento comunitário de qualidade do ar com sensores de baixo custo converge em três eixos: (i) coleta e calibração de dados em redes densas; (ii) tradução dos dados em comunicação de risco por meio de visualizações acessíveis; e (iii) arranjos sociotécnicos de ciência

cidadã voltados ao empoderamento e à incidência em políticas públicas (Alhasa et al., 2025). Para mapear esse panorama, foi realizada uma revisão de escopo em dezembro de 2025 nas bases IEEE Xplore, ScienceDirect e PubMed Central, esta última incluída por sua relevância para estudos de exposição ambiental e saúde pública, com os descritores "low-cost sensors", "air quality monitoring", "citizen science", "community empowerment" e "visualization platform".

No que se refere à visualização e comunicação de risco, Proma et al. (2021) mostram que uma plataforma interativa (CleanAirNowKC) pode reduzir barreiras de compreensão ao traduzir métricas técnicas de PM2,5 em informação acessível ao público, utilizando D3.js e Leaflet.js integrados à API PurpleAir. A ferramenta permitiu que residentes identificassem fontes industriais de poluição e coletassem evidências para advocacy. A principal lacuna identificada foi a ausência de funcionalidades para prestação de informações comunitárias sobre eventos transitórios de poluição.

Além disso, no contexto amazônico, Lima et al. (2024) desenvolveram a plataforma SELVA, combinando sensores de baixo custo com dados satelitais de focos de incêndio para informar a população em tempo real. O aplicativo móvel registrou mais de 25 mil downloads, evidenciando forte correlação entre busca por informação e períodos de poluição intensa. A lacuna identificada foi a ausência de *dataset* aberto estruturado para análises científicas.

Em revisões de literatura, Carotenuto et al. (2023) destacam que, na maioria dos casos, as redes de sensores de baixo custo ainda são implementadas apenas para aplicações experimentais, com poucos casos documentados de integração completa e permanente nas redes oficiais de monitoramento, concluindo que tal integração seria benéfica para aumentar a resolução espacial e temporal do monitoramento. Alhasa et al. (2025), por sua vez, identificam lacunas persistentes de cobertura em regiões do Sul Global, com sub-representação de estudos na América do Sul, África e Sudeste Asiático, e apontam a necessidade de abordagens replicáveis para a expansão de redes em países em desenvolvimento.

No contexto local de Boa Vista, Barbosa e Citó (2024) sistematizaram dados históricos de monitoramento entre 2020 e 2024 utilizando sensores PurpleAir, revelando padrão bimodal de poluição com dois picos anuais

distintos. Os resultados indicaram que 19,8% dos dias analisados apresentaram níveis deletérios à saúde humana. A conclusão foi que Roraima apresenta situação crítica de poluição atmosférica, especialmente durante os períodos de queimadas locais e transporte de poluentes de outras regiões amazônicas. A lacuna identificada foi a ausência de plataforma de visualização comunitária e *dataset* público para acesso da população. A Tabela 1 sintetiza as características dos trabalhos relacionados em dimensões-chave.

Tabela 1. Comparação entre trabalhos relacionados.

Trabalho	**Visualização**	**Dataset Aberto**	**Ciência Cidadã**	**Contexto Amazônico**
Proma et al. (2021)	Sim (*web*)	Não	Sim	Não
Lima et al. (2024)	Sim (*App*)	Não	Sim	Sim
Carotenuto et al. (2023)	*N/A (revisão)*	*N/A*	*N/A*	*N/A*
Barbosa e Citó (2024)	Não	Não	Não	Sim
Esta pesquisa	**Sim (*web*/*app*)**	**Sim**	**Sim**	**Sim**

A análise comparativa desses estudos evidencia como principal lacuna a ausência de soluções que integrem, de forma simultânea, dados abertos interoperáveis, plataformas de visualização acessíveis e mecanismos de comunicação de risco voltados a públicos não especializados, especialmente em contextos amazônicos com ausência de monitoramento oficial.

4. Solução computacional

4.1. Estruturação do dataset aberto

Os dados foram obtidos por meio da API REST oficial do PurpleAir, configurando agregação de 30 minutos (average=30) e incluindo os campos: *timestamp*, PM2,5 (canal A), PM2,5 (canal B), PM10 (canal A), PM10 (canal B), temperatura, umidade e pressão.

O *dataset* foi disponibilizado em três níveis de agregação temporal: (a) resolução original de 30 minutos (34.791 registros); (b) médias diárias (730 registros); (c) médias mensais (24 registros). Para cada nível, foram gerados três formatos: CSV (formato tabular padrão), JSON (compatível com aplicações *web*) e GeoJSON (para sistemas de informação geográfica), que podem ser acessados no repositório do projeto < https://bit.ly/4s0bP9z >.

Os metadados seguem o padrão OpenAQ v3 (OpenAQ, 2024), incluindo: identificador único do sensor, coordenadas geográficas em WGS84, parâmetros medidos com respectivas unidades (PM2,5 e PM10, em $\mu g/m^3$), registro temporal em UTC e horário local, período de cobertura, frequência de atualização e licença de uso (CC BY 4.0). Essa estrutura facilita a integração do *dataset* com plataformas globais de monitoramento, como IQAir (IQAir, 2025), WAQI (WAQI, 2025) e US EPA AirNow (AirNow, 2025), conforme demonstrado na Figura 1.



Figura 1. Pipeline de processamento do dataset. Fonte: Elaborado pelo autor com auxílio de Inteligência Artificial (2026).

Foi produzida documentação completa do dataset, disponível no repositório público do projeto (Williams Filho, 2026), incluindo dicionário de dados, metodologia de coleta e processamento, informações sobre o sensor e sua localização, e limitações conhecidas. O repositório organiza os dados em três camadas: raw (dados brutos), processed (dados processados) e metadata (metadados), favorecendo rastreabilidade e reprodutibilidade das análises. A Figura 2 apresenta a arquitetura proposta.

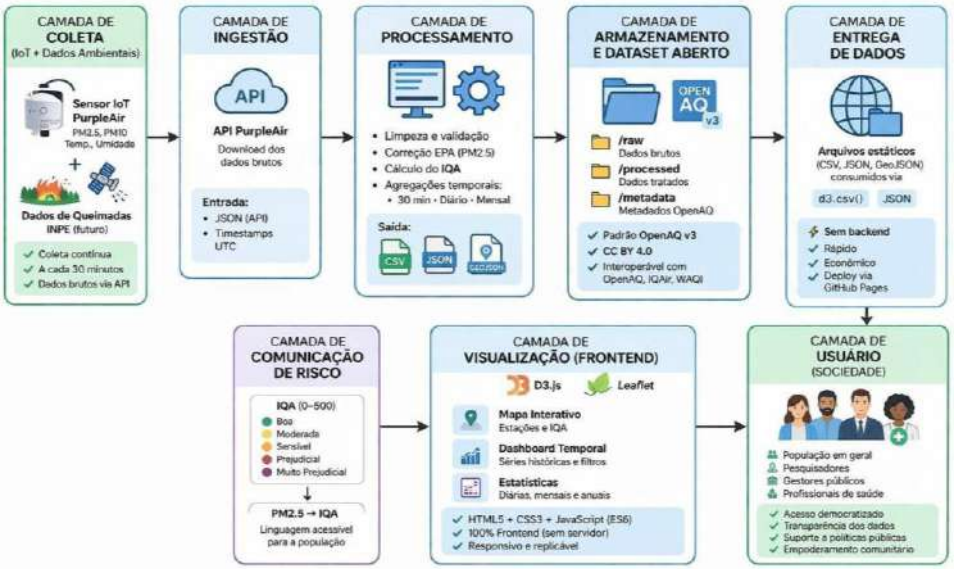


Figura 2. Diagrama com a arquitetura consolidada do Respira Boa Vista. Fonte: Elaborado pelo autor com auxílio de Inteligência Artificial (2026).

4.2. Desenvolvimento da plataforma de visualização

A plataforma Respira Boa Vista foi desenvolvida como aplicação frontend estática em HTML5, CSS3 e JavaScript puro (ES6+), com bibliotecas D3.js (v7) para visualizações de dados e Leaflet.js (v1.9) para mapeamento interativo, carregadas via CDN público (d3js.org e unpkg, respectivamente).

Optou-se pela arquitetura frontend-only devido à simplicidade de deploy, baixo custo operacional e maior replicabilidade em contextos com restrições de infraestrutura, características comuns em cidades amazônicas. Esta escolha técnica permite hospedagem gratuita em serviços de páginas estáticas (GitHub Pages, Netlify, Vercel). Além disso, elimina a necessidade de servidor backend e reduz barreiras para replicação em outras localidades.

Quanto ao licenciamento, as tecnologias utilizadas são de código aberto com licenças permissivas (MIT, BSD, ISC), garantindo que a solução possa ser replicada e adaptada por outras comunidades sem restrições. O código-fonte da plataforma está disponível em repositório Git público com

documentação de instalação e configuração. A plataforma está disponível no endereço < https://bit.ly/4tCWvkv >.

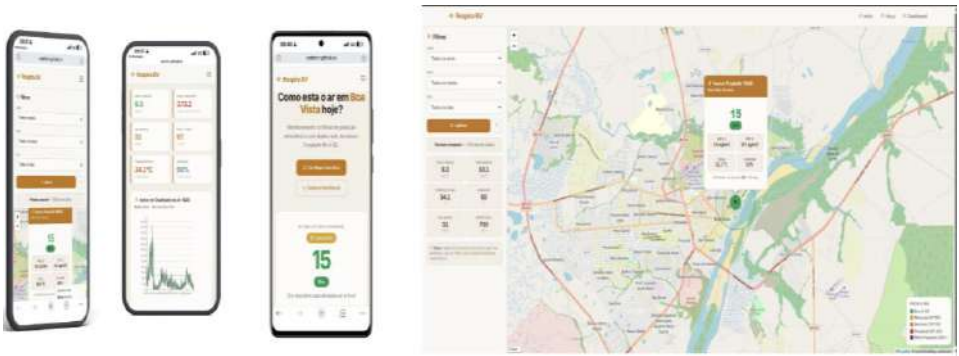


Figura 3. Módulos da plataforma Respira Boa Vista: página inicial, mapa interativo e dashboard analítico.

4.3. Requisitos Funcionais

Foram considerados os seguintes requisitos funcionais para orientar o desenvolvimento e a avaliação da solução demonstrados na Tabela 2.

Tabela 2. Cinco requisitos funcionais da plataforma Respira Boa Vista.

ID	Requisito	Descrição	Status
1	Formatos abertos	*Dataset* disponibilizado em CSV, JSON e GeoJSON, com metadados padronizados segundo o padrão OpenAQ v3 e licença CC BY 4.0.	Atendido
2	Visualização e IQA	Plataforma com gráfico de série temporal (filtros por ano, mês e dia) e escala de cores do IQA com mensagens de risco à saúde.	Atendido
3	Design responsivo	Interface acessível em dispositivos móveis (*smartphones* e *tablets*) sem necessidade de instalação de aplicativo.	Atendido
4	Replicabilidade	Solução de baixo custo, código-fonte aberto (MIT/BSD), arquitetura estática com hospedagem gratuita e documentação pública.	Atendido
5	Empoderamento comunitário	Dados públicos e comunicação de riscos à saúde disponíveis. Funcionalidade de informe de eventos pelos moradores prevista para trabalhos futuros.	*Parcialmente atendido*

5. Avaliação experimental

A análise verificou a conformidade com os cinco requisitos funcionais segundo três perspectivas: (i) avaliação da cobertura e qualidade do *dataset*; (ii) avaliação da conformidade da plataforma com as funcionalidades; e (iii) análise descritiva dos resultados obtidos.

5.1. Planejamento de pesquisa

Este estudo configura-se como uma pesquisa aplicada de abordagem quantitativa descritiva, baseada em análise de séries temporais de dados ambientais coletados por sensor IoT de baixo custo. Os dados foram obtidos por meio da API REST oficial do PurpleAir (sensor PA-II-SD, ID 56843), com agregação de 30 minutos, cobrindo o período de 01/01/2024 a 30/12/2025 (730 dias consecutivos). A avaliação do sistema desenvolvido foi estruturada para responder às três perguntas de pesquisa formuladas na Seção 1: (P1) viabilidade de estruturação de um dataset aberto com completude superior a 90%?; (P2) capacidade da plataforma de traduzir dados técnicos em informação de risco acessível?; e (P3) replicabilidade da solução para outros municípios?

Os critérios de aceitação de qualidade adotados foram: ausência de valores fisicamente implausíveis (negativos ou superiores a 1.000 $\mu g/m^3$) e taxa de completude mínima de 90%, conforme estabelecido pela Diretiva Europeia 2008/50/EC (Buelvas et al., 2023). A conformidade da plataforma foi verificada por comparação dos cinco requisitos funcionais definidos na Seção 4.3 com as funcionalidades efetivamente implementadas.

5.2. Avaliação do Dataset

5.2.1. Cobertura temporal

O *dataset* apresentou alta taxa de cobertura: 99,3% dos registros esperados para o período de 730 dias (01/01/2024 a 30/12/2025) estão presentes. Os 249 registros faltantes (0,7%) distribuíram-se em 34 lacunas ao longo do período, sendo as duas maiores: 31/01 a 01/02/2024 (40,5 h sem dados) e 19 a 20/06/2024 (40,5 h sem dados), que juntas representam 64% da perda total. As demais 32 lacunas foram de curta duração (1 a 7,5 h). Dos 730 dias, 698 (95,6%) apresentaram cobertura completa (48 registros).

Esta cobertura supera o limiar de 90% de captura mínima de dados estabelecido pela Diretiva Europeia 2008/50/EC para sistemas de monitoramento de qualidade do ar, conforme sistematizado por Buelvas et al. (2023).

5.2.2. Qualidade e consistência

A análise de consistência do *dataset* revelou duas características que sustentam a qualidade dos dados coletados. Primeiro, nenhuma leitura apresentou valores negativos ou superiores a 1.000 μg/m^3, limites que definem o intervalo fisicamente plausível para o sensor PMS5003 em condições ambientais (Barkjohn et al., 2022). Segundo, a distribuição temporal dos dados não apresentou lacunas acima de 48 horas contínuas, que é um fator que compromete análises de tendência (Buelvas et al., 2023). A maior lacuna registrada foi de 40,5 horas, abaixo desse limiar.

5.2.3. Compatibilidade com padrão OpenAQ

Os metadados estão em conformidade com o padrão OpenAQ v3 (OpenAQ, 2024). O padrão define dez campos obrigatórios para integração com plataformas globais de dados abertos, todos implementados no *dataset* (Tabela 3).

Tabela 3. Dez campos implementados e a conformidade do dataset.

Campo	Descrição	Status
location	Identificador único do local de medição	Implementado
coordinates	Latitude e longitude do sensor em WGS84	Implementado
parameter	Poluente medido (PM2,5, PM10)	Implementado
value	Concentração medida	Implementado
unit	Unidade de medida (μg/m^3)	Implementado
date	Registro temporal da medição em UTC	Implementado
sourceName	Nome da fonte dos dados (ex.: PurpleAir)	Implementado
sourceType	Tipo de fonte (reference / government / research)	Implementado
mobile	Indica se o sensor é fixo ou móvel	Implementado
averagingPeriod	Período de agregação das leituras (30 min)	Implementado

5.2. Resultados e Discussão

A avaliação da plataforma *Respira Boa Vista* verificou a conformidade das funcionalidades implementadas com cada um dos cinco requisitos funcionais definidos na Seção 4.3: formatos abertos, visualização e IQA, design responsivo, replicabilidade e empoderamento comunitário.

5.2.1. Funcionalidades implementadas

A plataforma reúne cinco funcionalidades principais. O *dataset* de qualidade do ar está disponível nos formatos CSV, JSON e GeoJSON, com metadados compatíveis com o padrão OpenAQ v3, disponíveis no repositório público do projeto (Williams Filho, 2026). A visualização dos dados é feita por meio de um gráfico de série temporal com filtros por ano, mês e dia, que permite ao usuário identificar variações nas concentrações de PM2,5 e PM10 ao longo do tempo.

As concentrações medidas são traduzidas em categorias de risco pela escala de cores do Índice de Qualidade do Ar (IQA), acompanhadas de mensagens textuais e orientações de proteção à saúde. A interface foi desenvolvida com design responsivo, com acesso testado em *smartphones* e *tablets*. O requisito de empoderamento comunitário é parcialmente atendido pela disponibilização pública dos dados e pela interface acessível a públicos não especializados, com funcionalidades adicionais de participação previstas como trabalho futuro.

5.2.2. Análise Descritiva dos Resultados

Os resultados demonstram que o protótipo do sistema Respira Boa Vista cumpre os requisitos funcionais estabelecidos na seção 4.3. O dataset atende aos requisitos, com cobertura de dois anos, completude de 99,29% e metadados padronizados. Os módulos de mapa e dashboard atendem ao requisito relacionado à visualização temporal interativa e à comunicação de risco à população. Quanto ao acesso móvel, é contemplado pela arquitetura web responsiva, sem instalação de aplicativo. A replicabilidade é garantida pela combinação de tecnologias abertas, arquitetura estática e documentação pública. Destaca-se que o empoderamento comunitário é parcialmente atendido na implementação atual, sendo a expansão para relatos de

moradores e a integração com outros sensores identificados como direções prioritárias para trabalhos futuros.

6. Conclusão

Os resultados deste trabalho demonstram a viabilidade técnica de estruturar um protótipo de sistema integrado de monitoramento da qualidade do ar em cidades amazônicas desprovidas de monitoramento contínuo, utilizando sensores IoT de baixo custo e tecnologias web abertas. A disponibilização conjunta do dataset e da plataforma como componentes abertos, documentados e replicáveis configura a principal contribuição deste estudo, ao propor uma abordagem acessível e reproduzível para o monitoramento ambiental em contextos de baixa infraestrutura institucional. Adicionalmente, os achados evidenciam o potencial dessas soluções para ampliar o acesso à informação ambiental e apoiar processos de conscientização e tomada de decisão, desde que consideradas as limitações instrumentais e espaciais inerentes ao uso de sensores de baixo custo.

Em resposta direta às perguntas de pesquisa, os resultados permitem afirmar que: (P1) é possível estruturar um dataset aberto de qualidade do ar com cobertura temporal superior a 90% utilizando sensores IoT de baixo custo, conforme demonstrado pela completude de 99,3% ao longo de dois anos de monitoramento contínuo; (P2) uma plataforma de visualização comunitária baseada em tecnologias web abertas é capaz de traduzir dados técnicos de PM2,5 em informação de risco à saúde acessível à população não especializada, por meio da integração de séries temporais, categorização pelo Índice de Qualidade do Ar e mensagens interpretativas; e (P3) a solução desenvolvida apresenta potencial de replicabilidade para outros municípios brasileiros sem monitoramento oficial, uma vez que se fundamenta em arquitetura de baixo custo, uso de tecnologias abertas e documentação pública. Assim, as três perguntas de pesquisa são respondidas afirmativamente, consolidando a contribuição técnica, social e metodológica do estudo.

O dataset baseia-se em um único sensor, o que limita a representatividade espacial dos resultados. O escopo do trabalho, contudo, não é mapear microvariações intraurbanas de qualidade do ar, mas

demonstrar a viabilidade técnica e metodológica de estruturação de dados e visualização comunitária em Boa Vista/RR. Identificam-se ainda: ausência de validação com monitores de referência para calibração local específica para aerossóis de queimadas; e validação formal de usabilidade com usuários da comunidade pendente.

Como trabalhos futuros, seis direções prioritárias são identificadas: (i) expansão da rede de sensores em Boa Vista, com instalação em pontos estratégicos; (ii) testes de usabilidade com grupos focais compostos por moradores, profissionais de saúde, educadores e gestores públicos; (iii) implementação de funcionalidade de informe comunitário, permitindo que moradores registrem eventos transitórios de poluição; (iv) integração do dataset com registros de focos de incêndio do INPE; (v) ampliação do repositório Zenodo com versões incrementais do dataset; e (vi) implantação da plataforma em domínio público dedicado.

7. Referências

AirNow. (2025). *Air quality index (AQI) basics*. U.S. Environmental Protection Agency. Recuperado de: https://www.airnow.gov/aqi/aqi-basics

Alhasa, K. M., Yulinawati, H., Kurnia, D., Wahyono, H. D., Yudo, S., Kustianto, I., & Endi, D. R. T. (2025). Mapping trends and analyzing key themes in low-cost sensors for air quality monitoring. *Earth Science Informatics*, *18*, Article 429. https://doi.org/10.1007/s12145-025-01927-5

Barbosa, R. I., & Citó, A. C. (2024). *Monitoramento da qualidade do ar na cidade de Boa Vista – Roraima (2020–2024)* [Relatório técnico-científico]. Instituto Nacional de Pesquisas da Amazônia, Núcleo de Roraima. Recuperado de: https://acervo.socioambiental.org/sites/default/files/documents/r1d00005.pdf

Barkjohn, K. K., Holder, A. L., Frederick, S. G., & Clements, A. L. (2022). Correction and accuracy of PurpleAir PM2.5 measurements for extreme wildfire smoke. *Sensors*, *22*(24), Article 9669. https://doi.org/10.3390/s22249669

Brown, I. F., Duarte, A. F., Torres, M., Ascorra, C., Reyes, J. F., & Rioja-Ballivián, G. (2019). Monitoramento de fumaça em tempo real mediante sensores de baixo custo instalados na Amazônia sul-ocidental. In *Anais do XIX Simpósio Brasileiro de Sensoriamento Remoto* (pp. 2658–2661). INPE. Recuperado de: https://proceedings.science/sbsr/sbsr-2019/trabalhos/monitoramento-de-fumaca-em-tempo-real-mediante-sensores-de-baixo-custo-instalado

Buelvas, J., Múnera, D., Tobón V., D. P., Aguirre, J., & Gaviria, N. (2023). Data quality in IoT-based air quality monitoring systems: A systematic mapping study. *Water, Air, & Soil Pollution*, *234*, Article 248. https://doi.org/10.1007/s11270-023-06127-9

Carotenuto, F., Bisignano, A., Brilli, L., Gualtieri, G., & Giovannini, L. (2023). Low-cost air quality monitoring networks for long-term field campaigns: A review. *Meteorological Applications*, *30*(6), Article e2161. https://doi.org/10.1002/met.2161

Instituto Brasileiro de Geografia e Estatística - IBGE. (2025). *Censo demográfico 2022: Panorama de Boa Vista*. Recuperado de: https://cidades.ibge.gov.br/brasil/rr/boa-vista/panorama

IQAir. (2025). *World air quality report 2024*. https://www.iqair.com/world-air-quality-report

Lima, G. A., Ribeiro, I. O., Filho, M. C. O. C., Trevisan, R., Andreoli, R. V., Lacerda, A., Costa, E. G., & Souza, R. A. F. (2024). Uma aplicação de Internet da Natureza para monitoramento de queimadas e da qualidade do ar na Amazônia. In *Anais do XV Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais (WCAMA 2024)*. SBC. Recuperado de: Recuperado de: https://sol.sbc.org.br/index.php/wcama/article/view/36095

Morawska, L., Thai, P. K., Liu, X., Asumadu-Sakyi, A., Ayoko, G., Bartonova, A., Bedini, A., Chai, F., Christensen, B., Dunbabin, M., Gao, J., Hagler, G. S. W., Jayaratne, R., Kumar, P., Lau, A. K. H., Louie, P. K. K., Mazaheri, M., Ning, Z., Motta, N., . . . Williams, R. (2018). Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? *Environment International*, *116*, 286–299. https://doi.org/10.1016/j.envint.2018.04.018

Oliveira, I. N., Oliveira, B. F. A., Silveira, I. H., Machado, L. M. G., Villardi, J. W. R., & Ignotti, E. (2023). Air pollution from forest burning as environmental risk for millions of inhabitants of the Brazilian Amazon: an exposure indicator for human health. *Cadernos de Saúde Pública*, *39*(6), Article e00131422. https://doi.org/10.1590/0102-311XEN131422

OpenAQ. (2024). Open air quality data: The global landscape. OpenAQ. Recuperado de: https://documents.openaq.org/reports/Open+Air+Quality+Data-The+Global+Landscape+2024.pdf

Plantower. (2016). *PMS5003 series data manual* (Version V2.3). Plantower Co., Ltd. Recuperado de: https://www.aqmd.gov/docs/default-source/aq-spec/resources-page/plantower-pms5003-manual_v2-3.pdf

Pope, C. A., III, & Dockery, D. W. (2012). Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air & Waste Management Association*, *56*(6), 709–742. https://doi.org/10.1080/10473289.2006.10464485

Pope, C. A., III, Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., & Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA*, *287*(9), 1132–1141. https://doi.org/10.1001/jama.287.9.1132

Proma, R. A., Sumpter, M., Lugo, H., Friedman, E., Huq, K. T., & Rosen, P. (2021). CleanAirNowKC: Building community power by improving data accessibility. In *Proceedings of IEEE VIS 2021*. IEEE. https://doi.org/10.48550/arXiv.2107.11633

PurpleAir. (2025). *PurpleAir Flex: Air quality monitor*. 27 de março de 2025, de https://www2.purpleair.com/products/purpleair-flex

Qin, J., Liu, Y., & Grosvenor, R. (2016). A categorical framework of manufacturing for Industry 4.0 and beyond. *Procedia CIRP*, *52*, 173–178. https://doi.org/10.1016/j.procir.2016.08.005

Sant'Anna, A., Alencar, A., Pinheiro, B., Araújo, C., Vormittag, E., Wicher, H., Borges, K., Faria, M., Andrade, M. F., Porto, P., Artaxo, P., Rocha, R., Esturba, T., & De Simoni, W. (2021). *O estado da qualidade do ar no Brasil* [Working paper]. WRI Brasil. Recuperado de:

https://www.wribrasil.org.br/sites/default/files/wri-o-estado-da-_qualidade-do-ar-no-brasil.pdf

Williams Filho, C. C. P. (2026). *Respira Boa Vista: Dataset* [Dataset]. GitHub. Recuperado de: https://github.com/cedricrr/respira-boa-vista/tree/main/dataset

World Air Quality Index (WAQI). (2025). *World air quality index: Real-time air pollution map*. Recuperado de: https://waqi.info

World Health Organization. (2021). *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. Recuperado de: https://www.who.int/publications/i/item/9789240034228

World Meteorological Organization & United Nations Environment Programme. (2024). *Integrating low-cost sensor systems and networks to enhance air quality applications* (GAW Report No. 293). WMO. Recuperado de: https://library.wmo.int/records/item/68924-integrating-low-cost-sensor-systems-and-networks-to-enhance-air-quality-applications

Capítulo 18

Gestão de Resíduos em Unidades Básicas de Saúde de Boa Vista: Uma proposta baseada em IoT de baixo custo

Ana Beatriz Bezerra de Menezes dos Santos
Alexander Fernández Correa
Universidade Federal de Roraima

anabezerra1441@gmail.com
alexander.fernandez@ufrr.com

1. Introdução

Diante dos crescentes desafios de infraestrutura e de crescimento populacional, as *smart cities* ou cidades inteligentes emergem como uma solução estratégica para a redefinição das dinâmicas urbanas (Sousa et al., 2024). Essas cidades se caracterizam pela utilização de tecnologia da informação e comunicação com o propósito de tornar os centros urbanos mais eficientes, sustentáveis e conectados, por meio da integração de soluções com monitoramento em tempo real, gestão inteligentes de recursos e sistemas de apoio à tomada de decisão. Segundo Kim et al. (2016), as cidades inteligentes utilizam recursos tecnológicos para promover inovação e enfrentar problemas urbanos complexos, possibilitando uma gestão orientada por dados e alinhada aos princípios da sustentabilidade ambiental, econômica e social.

A produção crescente de resíduos, impulsionada pelo aumento populacional e econômico, enfrenta gargalos em sistemas de gestão ineficientes. Nesse cenário, o manejo adequado dos resíduos torna-se um desafio crítico e um eixo estratégico fundamental para a consolidação das cidades inteligentes.

No âmbito dos serviços de saúde, a gestão de resíduos requer uma atenção específica, uma vez que geram resíduos que demandam procedimentos específicos de segregação, acondicionamento, armazenamento e destinação final. Por apresentarem riscos à saúde pública e ao meio ambiente, esses resíduos não podem ser descartados de qualquer

forma, sendo necessária a atuação de empresas especializadas para a coleta e destinação ambientalmente adequada, conforme estabelecido pela RDC ANVISA nº 222/2018 (Brasil, 2018).

No contexto local, essas questões também se fazem presentes no município de Boa Vista, capital do estado de Roraima. Assim como em outras cidades brasileiras, a gestão de resíduos hospitalares, como nas Unidades Básicas de Saúde é realizada de acordo com diretrizes nacionais, como a Política Nacional de Resíduos Sólidos (PNRS), sendo de responsabilidade das próprias unidades a segregação na fonte, a coleta interna, o armazenamento temporário e o encaminhamento dos resíduos para a coleta externa (BRASIL, 2010). Entretanto, grande parte desse processo ainda depende de procedimentos manuais, o que limita o monitoramento em tempo real, dificulta o controle das etapas do manejo e compromete a eficiência da gestão.

Considerando os desafios enfrentados por algumas cidades para se adequar ao panorama atual de tecnologia e de implementação de soluções inovadoras, este trabalho tem por objetivo desenvolver um modelo de gestão de resíduos nas Unidades Básicas de Saúde baseado em Internet das Coisas, no qual lixeiras hospitalares são instrumentadas com sensores e conectividade Wi-Fi para monitoramento do nível de preenchimento. A partir desse objetivo, o estudo busca responder às seguintes perguntas de pesquisa: (i) como IoT pode auxiliar no monitoramento do nível de resíduos em lixeiras hospitalares de Unidades Básicas de Saúde; (ii) de que forma os dados coletados podem apoiar a identificação dos recipientes que necessitam de esvaziamento; (iii) como essas informações podem contribuir para a definição de rotas personalizadas de coleta; e (iv) em que medida essa abordagem pode favorecer a redução de desperdícios operacionais e o aumento da eficiência do serviço.

Como contribuição, esta pesquisa propõe uma solução tecnológica de baixo custo voltada ao contexto da gestão de resíduos em unidades de saúde. Espera-se, com isso, colaborar para o avanço de aplicações de IoT em serviços urbanos essenciais, com ênfase na melhoria do monitoramento e da logística de coleta de resíduos hospitalares.

Este trabalho está estruturado em seções que abrangem a fundamentação teórica, os trabalhos relacionados, a apresentação da solução

computacional proposta, a avaliação computacional, trabalhos futuros e melhorias, e, por fim, as referências.

2. Fundamentação teórica

2.1. Gestão sustentável de resíduos de serviços de saúde em Boa Vista

De acordo com o Art. 1º da Lei Municipal nº 2.004, de 12 de julho de 2019, são estabelecidas as diretrizes municipais para a implementação do Plano Municipal de Resíduos Sólidos de Boa Vista, por meio da implantação do Sistema de Gestão Sustentável de Resíduos Sólidos Urbanos, abrangendo também os resíduos gerados pelos serviços de saúde, como Unidades Básicas de Saúde e hospitais.

No que se refere especificamente aos resíduos de serviços de saúde, a referida lei institui, em seu Art. 43, o Sistema para a Gestão Sustentável de Resíduos de Serviços de Saúde, com o objetivo de disciplinar os fluxos, os agentes envolvidos e a destinação ou disposição final ambientalmente adequada dos resíduos gerados no município de Boa Vista. A legislação determina que a responsabilidade pelo gerenciamento dos resíduos de serviços de saúde, desde a geração até a destinação final, é atribuída aos próprios geradores, conforme disposto no Art. 44. Para o cumprimento dessa responsabilidade, a lei exige a contratação de empresas especializadas devidamente licenciadas ou a manutenção de estrutura própria, observando-se as normas técnicas e sanitárias aplicáveis, conforme estabelecem os §§ 1º e 2º do referido artigo.

Além disso, a Lei nº 2.004/2019 torna obrigatória a elaboração, implementação e manutenção do Plano de Gerenciamento de Resíduos de Serviços de Saúde (PGRSS) por parte dos geradores, como condição para o licenciamento sanitário e ambiental, conforme previsto no Art. 48. A legislação também exige a manutenção de registros e a apresentação de comprovantes de destinação adequada dos resíduos.

2.2. Sensores e monitoramento IoT

Os sensores IoT permitem o monitoramento contínuo de variáveis importantes, o que resulta na coleta e análise de dados em tempo real,

evitando atrasos nas respostas e garantindo que situações críticas, que exigem uma ação ágil, sejam tratadas de forma mais eficiente (Atzori et al., 2010). Além de otimizar processos, os sensores contribuem para redução de custos e melhoria da produtividade dos serviços.

Sistemas baseados em IoT são compostos por um conjunto de sensores inteligentes que desempenham funções cruciais na coleta de dados. Esses sensores são capazes de monitorar variáveis como temperatura, umidade, qualidade do ar e muitas outras. No contexto da gestão de resíduos, por exemplo, sensores como o HC-SR04 (sensor ultrassônico) são utilizados para monitorar o nível de enchimento dos contêineres.

3. Trabalhos relacionados

Nesta seção, são apresentados alguns trabalhos relacionados ao tema desta pesquisa, bem como sua relação com a proposta desenvolvida neste trabalho. Para essa análise, foram considerados os estudos de Lima et al. (2022), Neema e Gor (2022), Ahmed et al. (2024), Kayalvizhi et al. (2025) e Biswas et al. (2025).

Entre os estudos analisados, Lima et al. (2022) desenvolveram o SiCOL, sistema baseado em IoT para auxiliar a coleta de resíduos urbanos em ambientes acadêmicos. Neema e Gor (2022) propuseram uma solução inteligente para monitorar o nível de preenchimento das lixeiras e informar, via GPS, os pontos que necessitam de coleta. Ahmed et al. (2024) apresentaram uma arquitetura para cidades inteligentes que integra IoT e inteligência artificial, permitindo o monitoramento das lixeiras, envio de notificações e apoio à otimização de rotas.

Já Kayalvizhi et al. (2025) desenvolveram um sistema com IoT e aprendizado de máquina voltado à segregação automatizada de resíduos, utilizando sensores, câmeras e técnicas de aprendizado profundo. Por sua vez, Biswas et al. (2025) propuseram o Smart Netbin, uma solução que integra sensores, Arduino, Wi-Fi e arquitetura cliente-servidor para monitorar o descarte de resíduos e incentivar o descarte correto.

De modo geral, os trabalhos analisados demonstram que sistemas baseados em IoT podem contribuir para o monitoramento de lixeiras, o envio

de informações em tempo real e o apoio à logística de coleta. Entretanto, a maioria das soluções concentra-se em resíduos urbanos gerais, conforme demonstrado na Tabela 1. Assim, este estudo se diferencia por propor uma solução de baixo custo para o monitoramento de resíduos hospitalares em Unidades Básicas de Saúde de Boa Vista, considerando o contexto local e as necessidades operacionais dessas unidades.

Tabela 1. Comparativo dos trabalhos relacionados.

Autor (Ano)	Sistema	Tecnologias utilizadas	Objetivo	Diferencial
Lima et al. (2022)	SiCOL, sistema de coleta de resíduos urbanos baseado em IoT.	IoT e Sistemas Embarcados.	Apoiar a coleta de resíduos em ambientes acadêmicos.	Coleta sanitária em ambientes de ensino.
Neema e Gor (2022)	Sistema inteligente de gestão de resíduos.	Raspberry Pi, sensores ultrassônicos, umidade, infravermelho e GPS.	Identificar lixeiras cheias e informar motoristas sobre os pontos de coleta.	Localização via GPS para o coletor.
Ahmed et al. (2024)	Arquitetura para gestão municipal de resíduos em smart cities.	IoT e Inteligência Artificial.	Monitorar lixeiras e apoiar a otimização de rotas.	Foco na redução de combustível e tempo.
Kayalvizhi et al. (2025)	Sistema de gestão e segregação de resíduos.	IoT, Machine Learning e Visão Computacional.	Automatizar a classificação/segregação dos resíduos.	Classificação de resíduos sem intervenção humana.
Biswas et al. (2025)	Smart Netbin.	Arduino, Wi-Fi e Cliente-Servidor.	Incentivar o descarte correto e monitorar resíduos.	Oferece Wi-Fi como recompensa pelo descarte.

4. Solução computacional

4.1. Descrição Básica

O projeto visa à implementação de um sistema de monitoramento de resíduos hospitalares nas Unidades Básicas de Saúde de Boa Vista, Roraima. O sistema tem como objetivo monitorar o nível de preenchimento das lixeiras. Para isso, utiliza-se uma solução baseada em Internet das Coisas,

onde sensores inteligentes estão conectados a um microcontrolador ESP32. Este microcontrolador coleta e transmite os dados por meio da rede Wi-Fi local para o sistema central, utilizando o protocolo de comunicação MQTT.

4.2. Componentes que fazem parte do sistema IoT

O modelo é composto por componentes que ficam instalados na parte superior dentro das lixeiras hospitalares localizadas nas Unidades Básicas de Saúde. A Figura 1 apresenta o protótipo desenvolvido, contendo os componentes responsáveis pela medição do nível interno de preenchimento das lixeiras, pelo processamento dessas informações e pelo envio dos dados à dashboard de monitoramento.

→ **Sensor Ultrassônico Hc-sr04:** Sensor responsável por medir o nível de preenchimento do contêiner. O dispositivo calcula a distância entre o sensor e a superfície dos resíduos.
→ **Led:** Elemento de sinalização visual local do estado da lixeira, permitindo rápida identificação da condição do contêiner pelos profissionais da unidade.
→ **ESP32:** Microcontrolador responsável pelo processamento local do sistema, realizando a leitura periódica dos sensores instalados na lixeira, o pré-processamento das medições e a classificação preliminar dos dados coletados.
→ **Protoboard:** Placa de prototipagem utilizada para montar circuitos.
→ **Jumper:** Fio condutor utilizado para estabelecer conexões entre a protoboard e os outros componentes.



Figura 1. Protótipo criado e representação de sua instalação na lixeira.

4.3. Camada de comunicação

A comunicação entre os dispositivos IoT instalados na tampa das lixeiras das Unidades Básicas de Saúde e o sistema central é realizada por meio da rede Wi-Fi disponível nas próprias unidades. Através da rede Wi-Fi, é permitido que os dispositivos estabeleçam conexão com o servidor responsável pelo processamento e armazenamento das informações coletadas.

Uma vez conectado à rede, o ESP32 estabelece comunicação com o broker MQTT, cujo endereço IP é previamente configurado no dispositivo. A transmissão dos dados ocorre por meio de mensagens MQTT encapsuladas sobre o protocolo TCP/IP, possibilitando o envio confiável das informações coletadas.

Adotou-se o protocolo MQTT como mecanismo de transporte de dados, por operar no modelo publish/subscribe. Nesse modelo, cada lixeira equipada com o ESP32 e o sensor ultrassônico, atua como publisher (Publicador), realizando a leitura periódica da distância entre o sensor e os resíduos que vão preenchendo a lixeira e publicando, em tópicos específicos do broker, informações como o nível de ocupação da lixeira e seu estado operacional.

O backend do sistema atua como subscriber (assinante) dos tópicos MQTT configurados para cada UBS e lixeira. Ao se inscrever nesses tópicos, o backend passa a receber automaticamente todas as mensagens publicadas pelos dispositivos IoT, sem a necessidade de requisições diretas ou conexões individuais com cada microcontrolador. Após o recebimento, as informações são validadas, processadas e armazenadas em banco de dados PostgreSQL, possibilitando a atualização dinâmica do painel de monitoramento desenvolvido em Grafana.

A Figura 2 apresenta o fluxo de funcionamento do projeto, desde a lixeira sensorizada até o recolhimento dos resíduos. Nela, é possível observar a camada de comunicação responsável pelo envio dos dados por meio do protocolo MQTT, o processamento e armazenamento das informações no backend e no banco de dados, bem como sua visualização no dashboard desenvolvido no Grafana, que auxilia na tomada de decisão para o recolhimento dos resíduos.



Figura 2. Fluxo de funcionamento da gestão de resíduos.

4.4. Dashboard de monitoramento

Considera-se que a lixeira está próxima de sua capacidade máxima quando a distância medida pelo sensor é inferior a um valor limite, correspondente a níveis de preenchimento acima de aproximadamente 80%. Na Dashboard de monitoramento, é possível verificar um panorama geral, a rota de recolhimento, todas as unidades básicas cadastradas e as coletas que foram realizadas no dia, conforme apresentado na figura 3.

Quando a distância se encontra em uma faixa intermediária, o LED amarelo é acionado, representando estado de atenção. Já em situações de maior distância entre o sensor e os resíduos, indicando baixo nível de enchimento, o LED verde é ativado.

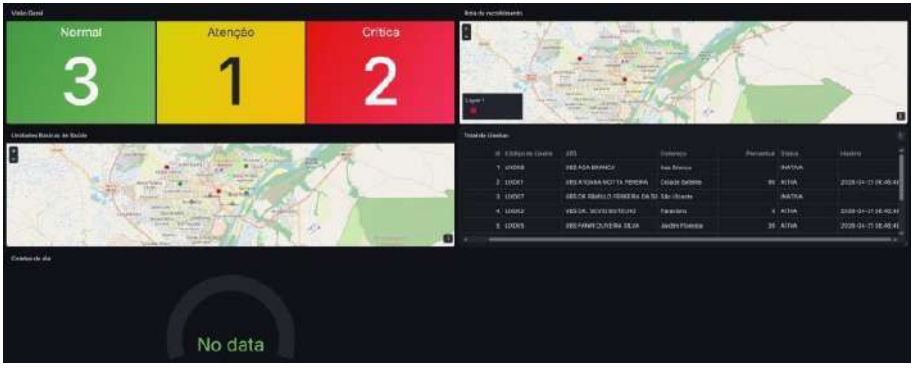


Figura 3. Dashboard de monitoramento no Grafana.

5. Avaliação experimental

5.1. Planejamento de pesquisa

Esta seção descreve os procedimentos adotados para testar o sistema desenvolvido, com o objetivo de garantir a confiabilidade dos dados obtidos. O primeiro teste tem como foco a validação do sensor ultrassônico HC-SR04. Para isso, o sistema é testado em uma lixeira comum com tampa e altura aproximada de 40 cm, simulando diferentes níveis de preenchimento para representar os estados normal, de atenção e crítico. Essa etapa busca verificar se as medições realizadas pelo sensor correspondem adequadamente ao volume de resíduos presente na lixeira, bem como se a classificação do estado do recipiente ocorre de forma coerente.

Em relação à transmissão dos dados, o sistema é submetido a operação contínua durante um determinado período, realizando a publicação da telemetria em intervalos curtos. Nessa etapa, utilizam-se os logs detalhados do broker Mosquitto para identificar a taxa de sucesso na entrega das mensagens e observar o comportamento do sistema diante de eventuais oscilações do sinal de rede.

No que se refere ao backend e ao banco de dados, são realizadas consultas SQL estruturadas nas tabelas do PostgreSQL, com a finalidade de verificar o correto armazenamento das informações recebidas. Por meio da atualização em tempo real dos registros, torna-se possível acompanhar os timestamps de monitoramento e confirmar a integridade do fluxo de dados entre os dispositivos IoT, o backend e o banco de dados.

Por fim, realiza-se a simulação de uma lixeira em estado crítico para validar o comportamento do sistema no painel de monitoramento. Nessa etapa, observa-se a alteração do LED indicativo no protótipo e a sinalização da lixeira no mapa da dashboard. Após a coleta dos resíduos, verifica-se a atualização do status para “normal” no mesmo dia. Além disso, o sistema contabiliza a quantidade de lixeiras esvaziadas ao longo do dia. Dessa maneira, o dashboard permite acompanhar a situação das lixeiras e favorece a definição de prioridades no processo de coleta.

5.2. Resultados e Discussão

Os testes realizados demonstram a viabilidade técnica do sistema proposto para o monitoramento do nível de preenchimento de lixeiras hospitalares em Unidades Básicas de Saúde. No que se refere ao protótipo, o sensor mostrou-se capaz de medir, de forma consistente, a distância entre ele e os resíduos, permitindo a correta classificação da criticidade das lixeiras.

Quanto à comunicação, o ESP32 demonstrou confiabilidade no envio dos dados por meio do Wi-Fi e do protocolo MQTT, possibilitando que o backend receba, processe e armazene as informações no banco de dados PostgreSQL sem perdas significativas. Da mesma forma, a dashboard desenvolvida no Grafana apresenta atualizações dinâmicas das leituras, viabilizando o monitoramento remoto e em tempo real do status das lixeiras monitoradas.

Ao comparar os resultados obtidos com os estudos apresentados na seção de trabalhos relacionados, observa-se que eles se aproximam das conclusões de Lima et al. (2022), Neema e Gor (2022) e Ahmed et al. (2024), que igualmente destacam o potencial da Internet das Coisas para aprimorar o monitoramento e a logística da coleta de resíduos. Assim como os estudos mencionados, o trabalho demonstrou que sensores e comunicação em tempo real podem fornecer dados valiosos para a administração do serviço. No entanto, a principal diferença reside no contexto de aplicação: enquanto esta pesquisa se concentra especificamente em resíduos hospitalares em Unidades Básicas de Saúde, a maioria dos estudos se concentra em resíduos urbanos de forma mais abrangente.

Nesse contexto, o trabalho relaciona-se diretamente aos princípios da Indústria 4.0 ao aplicar sensores, conectividade IoT, banco de dados e dashboard para automatizar o monitoramento das lixeiras hospitalares. Mesmo em nível de protótipo, a solução demonstra a possibilidade de substituir verificações manuais por uma gestão orientada por dados em tempo real, apoiando a tomada de decisão e a otimização da coleta.

Embora os resultados indiquem a viabilidade técnica da proposta, a avaliação foi realizada em cenário restrito e em escala experimental, sem testes prolongados de estabilidade de rede ou análises estatísticas mais robustas. Ainda assim, em uma perspectiva local, o protótipo mostrou-se

funcional e de baixo custo, evidenciando o potencial da Internet das Coisas como ferramenta de apoio à gestão de resíduos hospitalares em Unidades Básicas de Saúde.

6. Conclusão

Como trabalhos futuros, são propostas algumas melhorias para a evolução do sistema. Entre elas, destaca-se a ampliação da autonomia energética, permitindo que o protótipo funcione de forma mais independente, por meio do uso de baterias ou energia solar. Também se considera importante o reforço da segurança dos dados, especialmente durante a transmissão das informações entre os dispositivos IoT e o sistema central.

Além disso, recomenda-se aumentar a resiliência do sistema, garantindo seu funcionamento mesmo diante de falhas de conexão ou instabilidade da rede. Outra melhoria possível é o aprimoramento da dashboard, com a inclusão de novos filtros, alertas e formas de visualização. Por fim, a integração de novos sensores ao protótipo pode ampliar suas funcionalidades e tornar o monitoramento mais completo.

7. Referências

Ahmed, K., Dubey, M. K., Kumar, A., & Dubey, S. (2024). Artificial intelligence and IoT driven system architecture for municipality waste management in smart cities: A review. Measurement: *Sensors*, *36*, 101395. https://doi.org/10.1016/j.measen.2024.101395

Biswas, R., Jana, S., Pal, K., Banerjee, A., & Majumdar, A. (2025). Smart waste management system using IoT (Internet of Things). *International Journal of Advanced Research in Science, Communication and Technology*, *5*(9), 233–239. https://doi.org/10.48175/IJARSCT-28235

Boa Vista. (2019, 12 de julho). Lei nº 2.004, de 12 de julho de 2019: Implementação do Plano Municipal de Gestão Integrada de Resíduos no Município de Boa Vista por meio de sistema de gestão sustentável de resíduos sólidos urbanos. Diário Oficial do Município de Boa Vista, (4927). Recuperado de: https://sapl.boavista.rr.leg.br/media/sapl/public/normajuridica/2019/

2694/lei_2.004-2_019_-exe-_de_12_de_julho_de_2019_-_dom_no_4927.pdf

Kayalvizhi, S., Akhilesh, P. S., Chowdary, M. V., & Babu, V. V. (2025). Smart waste management system using IoT and machine learning. *International Journal of Innovative Research in Technology*, *11*(9), 2182–2187. Recuperado de: https://ijirt.org/publishedpaper/IJIRT173232_PAPER.pdf

Lima, I. A., Pavão, L. H. F., Gomes, L. B. B., Ferraz, M. H. S., Sales, A. T. M., Moreira, A. L. S., & Domingues, M. A. O. (2022). Implementação de um sistema para auxílio na coleta dos resíduos sólidos urbanos. In *Anais do IX Encontro Nacional de Computação dos Institutos Federais (EnCompIF 2022) (pp. 77–84)*. Porto Alegre, RS, Brasil: Sociedade Brasileira de Computação. Recuperado de: https://sol.sbc.org.br/index.php/encompif/article/view/20438

Maranhão, R. de A. (2024). Inteligência artificial para gestão de resíduos em cidades inteligentes no contexto das mudanças climáticas: Uma revisão da literatura. *Brazilian Journal of Animal and Environmental Research*, *7*(3), e73115. https://doi.org/10.34188/bjaerv7n3-094

Neema, S., & Gor, K. (2022). Smart waste management using IoT. *International Journal of Scientific Research in Science, Engineering and Technology*, *9*(6), 16–21. https://doi.org/10.32628/IJSRSET229529

Prefeitura Municipal de Boa Vista, Secretaria Municipal de Saúde. (2024, 3 de julho). Unidades básicas de saúde. Recuperado de: https://boavista.rr.gov.br/storage/paginas/saude-ubs/UBS_BOA_VISTA.pdf

Sosunova, I., & Porras, J. (2022). IoT-enabled smart waste management systems for smart cities: A systematic review. *IEEE Access*, *10*, 73326–73363. https://doi.org/10.1109/ACCESS.2022.3188308

Wetler, J. P., Campos, A. F., Pagel, U. R., Souza, V. H. A. de, Carolino, J., & Cruz, A. B. de S. (2025). Gestão de resíduos sólidos urbanos em cidades inteligentes: Possibilidades para o município de Vitória-ES. *Revista de Gestão e Secretariado*, *16*(1), e4558. https://doi.org/10.7769/gesec.v16i1.4558

Capítulo 19

Análise comparativa entre Arima e LSTM na previsão de escoamento logístico de soja em Roraima

Hiago Pereira
Susset Guerra Jimenez
Universidade Federal de Roraima

hiagopereira28@gmail.com
susset.guerra@ufrr.com

1. Introdução

O estado de Roraima apresenta uma trajetória recente de expansão econômica, impulsionada sobretudo pelo avanço do setor agropecuário e pela consolidação do estado como nova fronteira agrícola para a produção de soja. Esse movimento amplia a relevância estratégica da logística regional, especialmente no contexto de estruturas como o futuro Porto Seco de Boa Vista e a integração comercial associada à Zona de Livre Comércio de Bonfim-Lethem. Em razão da forte sazonalidade da cultura da soja, o escoamento da produção passa a ocorrer sob condições de elevada variabilidade, com períodos de forte concentração de cargas e momentos de baixa utilização da infraestrutura disponível.

Nesse cenário, a gestão logística baseada apenas na reatividade tende a gerar ineficiências operacionais, como filas de caminhões durante os picos de safra e ociosidade estrutural nos períodos de entressafra. Ainda que a transformação digital tenha ampliado o uso de tecnologias analíticas em cadeias de suprimentos mais consolidadas, ainda se observa uma lacuna relevante na aplicação de ferramentas preditivas voltadas à realidade logística amazônica, marcada por alta volatilidade, limitação de séries históricas e forte sensibilidade a fatores sazonais. Assim, métodos estatísticos convencionais nem sempre são suficientes para representar adequadamente a dinâmica não linear do escoamento agrícola em regiões emergentes.

O desafio de otimizar o escoamento logístico em Roraima se enquadra diretamente no paradigma da Indústria 4.0, que exige a transição de abordagens reativas para uma gestão proativa e inteligente. A aplicação desses conceitos no setor de transportes deu origem à Logística 4.0, cujo objetivo é criar sistemas logísticos autônomos e flexíveis, com destaque para a evolução dos terminais portuários para o conceito de Porto 4.0 (ou Smart Ports). Este trabalho adere a essa transformação ao investigar a viabilidade de técnicas de Inteligência Artificial (Deep Learning) para lidar com a limitação e a volatilidade dos dados (Small Data) e, principalmente, ao propor o Índice de Pressão de Escoamento (IPE). O IPE atua como um sistema ciberfísico de apoio à decisão, convertendo previsões matemáticas em alertas gerenciais para antecipar a saturação da infraestrutura de transporte, um princípio crítico da Logística 4.0 em infraestruturas emergentes na Amazônia.

Diante disso, este trabalho investiga a viabilidade do uso de técnicas de Deep Learning em contexto de Small Data, por meio da comparação entre o modelo estatístico AutoRegressive Integrated Moving Average (ARIMA) e redes neurais Long Short-Term Memory (LSTM) na previsão do volume de escoamento logístico de soja em Roraima. A proposta parte da construção de um pipeline computacional em Python para ingestão, tratamento e modelagem de dados públicos, adotando uma abordagem comparativa entre quatro Unidades Federativas com diferentes níveis de maturidade logística: Mato Grosso, Pará, Amazonas e Roraima. Esta arquitetura comparativa foi desenhada para isolar e avaliar o comportamento dos algoritmos preditivos sob variadas condições de estabilidade e não-linearidade, testando o desempenho da LSTM tanto em séries maduras (Mato Grosso) quanto em séries voláteis de fronteira (Amazonas e Roraima).

Além de comparar o desempenho preditivo dos modelos, o estudo busca traduzir os resultados analíticos em apoio à decisão logística. Para isso, propõe-se a utilização do Índice de Pressão de Escoamento (IPE), concebido como um mecanismo de interpretação gerencial capaz de relacionar a previsão de volume exportado ao nível de pressão exercido sobre a infraestrutura de transporte. Com isso, pretende-se avançar da previsão puramente numérica para uma perspectiva aplicada à Logística 4.0, na qual a modelagem preditiva contribui para antecipar gargalos, orientar o planejamento operacional e apoiar decisões em ambientes portuários

emergentes. Com base nesse contexto, o trabalho é orientado pelas seguintes perguntas de pesquisa:

(1) o modelo LSTM, em arquitetura parcimoniosa, apresenta desempenho preditivo superior ao ARIMA, medido por RMSE e MAE, no contexto de alta volatilidade do escoamento de soja em Roraima?

(2) de que forma a previsão gerada pelos modelos pode ser convertida em um Índice de Pressão de Escoamento (IPE) útil para sinalizar risco de saturação da infraestrutura logística?

(3) em que medida dados públicos governamentais abertos podem sustentar um sistema automatizado de apoio à decisão para operações de escoamento em corredores logísticos fronteiriços?

Para responder a essas questões, o trabalho está organizado em seis seções, além desta introdução. A Seção 2 apresenta a fundamentação teórica sobre Logística 4.0, séries temporais e modelos preditivos. A Seção 3 discute os trabalhos relacionados. A Seção 4 descreve a solução computacional proposta. A Seção 5 detalha a avaliação experimental e a discussão dos resultados. Por fim, a Seção 6 apresenta as conclusões e os encaminhamentos para estudos futuros.

2. Fundamentação teórica

A Quarta Revolução Industrial, ou Indústria 4.0, introduziu uma nova era de digitalização e conectividade nas cadeias de valor, utilizando tecnologias como a Internet das Coisas (IoT), computação em nuvem, Big Data Analytics e Inteligência Artificial. A aplicação desses conceitos no setor de transportes e suprimentos deu origem à Logística 4.0, cujo objetivo é criar sistemas logísticos autônomos, flexíveis e altamente eficientes, capazes de satisfazer demandas com alta agilidade e visibilidade em tempo real (Efthymiou & Ponis, 2021).

Nesse contexto de transformação digital, os terminais portuários evoluíram para o conceito de Porto 4.0 (ou Smart Ports). Tratam-se de portos que alavancam tecnologias avançadas para transformar digitalmente seus processos de negócios, abandonando as tradicionais abordagens reativas em

prol de uma gestão proativa e inteligente. O coração dessa transição são os Sistemas Ciberfísicos (CPS), que mesclam os mundos físico e virtual, permitindo que a infraestrutura, os veículos e os sistemas de gerenciamento interajam e se comuniquem continuamente (Efthymiou & Ponis, 2021). Para infraestruturas emergentes na Amazônia, como o futuro Porto Seco de Boa Vista, a adoção desses princípios é crítica. O uso de modelos analíticos permite processar de maneira otimizada o alto volume de dados do agronegócio, antecipando a saturação da infraestrutura de transporte e convertendo previsões matemáticas em um Índice de Pressão de Escoamento (IPE) para mitigar gargalos durante os picos de safra.

O planejamento eficiente das cadeias de suprimentos exige que as organizações sejam capazes de prever a demanda futura com precisão, estudo este realizado através da análise de séries temporais, que consistem em pontos de dados registrados e analisados em ordem cronológica (Bousqaoui, Slimani, & Achchab, 2021). O modelo estatístico AutoRegressive Integrated Moving Average (ARIMA) é amplamente reconhecido como o padrão da indústria (baseline). Um modelo ARIMA (p,d,q) assume que os valores futuros possuem uma relação linear com seus valores passados e é composto por três partes: o modelo autorregressivo (p), o grau de diferenciação (d) para remover a não-estacionariedade, e o modelo de médias móveis (q). Apesar de sua eficácia em muitos cenários maduros, o ARIMA é limitado a relações lineares, tornando-o menos apto a lidar com as complexidades e não-linearidades extremas (como grandes picos e intermitências) típicas de commodities agrícolas.

Para superar as limitações dos modelos clássicos, Bousqaoui, Slimani e Achchab (2021) sugerem a utilização de arquiteturas Long Short-Term Memory (LSTM). Métodos baseados em Deep Learning têm ganhado destaque por sua capacidade de se aproximar de funções não-lineares complexas. As redes LSTM representam uma evolução profunda das Redes Neurais Recorrentes (RNN), projetadas especificamente para resolver o problema do gradiente que desaparece (vanishing gradient) através de uma arquitetura de memória baseada em portões (entrada, esquecimento e saída). Esse mecanismo permite que a rede decida ativamente qual informação deve ser armazenada, conferindo a habilidade de "lembrar" padrões de longo prazo (como o impacto de safras passadas nas atuais).

No contexto de novas fronteiras agrícolas, a previsão de séries temporais enfrenta desafios associados à limitação de dados históricos e à alta volatilidade. Nesse cenário, tanto modelos lineares quanto arquiteturas não lineares devem ser avaliados empiricamente, considerando o risco de sobreajuste e as particularidades de séries em transição. Essa perspectiva reforça a importância de uma análise comparativa que contemple diferentes níveis de maturidade logística regional. Com base nesses fundamentos, a próxima seção apresenta estudos recentes que aplicam modelos preditivos a problemas correlatos, permitindo situar a proposta deste trabalho na literatura.

3. Trabalhos relacionados

Conduziu-se uma breve revisão da literatura em 15 de janeiro de 2026, a partir de consultas às bases IEEE Xplore e Google Scholar, utilizando os descritores “LSTM”, “ARIMA”, “Dry Port”, “Agriculture”, “Supply Chain” e “Smart Port”. Foram identificados 25 estudos, dos quais se selecionaram cinco artigos recentes e aderentes ao tema desta pesquisa.

Schwalbert et al. (2020) conduziram um estudo com o objetivo de prever a produtividade da soja no sul do Brasil utilizando dados de sensoriamento remoto e dados climáticos. Para atingir esse objetivo, os autores compararam algoritmos tradicionais, como a regressão linear (OLS), com métodos avançados, como Random Forest e redes LSTM. Verificou-se a superioridade das redes LSTM na captura de padrões temporais complexos da safra de soja em relação aos modelos lineares. A lacuna deixada pelo estudo foi a necessidade de expandir a aplicação desses modelos preditivos para outras regiões geográficas e para a etapa de escoamento e logística.

Em linha com essa discussão sobre a capacidade preditiva de diferentes modelos, Bousqaoui, Slimani e Achchab (2021) realizaram uma análise comparativa entre o modelo estatístico ARIMA e abordagens de Deep Learning, incluindo as arquiteturas LSTM e CNN, para a previsão de demanda de curto prazo. O estudo indicou que o ARIMA, por sua natureza linear, apresenta limitações na representação de complexidades e não linearidades presentes nos dados, enquanto a arquitetura LSTM apresentou melhor desempenho, com redução do erro preditivo medido por RMSE

Complementando essa comparação entre arquiteturas preditivas, Yunita et al. (2025) apresentaram um benchmark comparando o desempenho de diferentes redes neurais, como RNN, LSTM e GRU, em tarefas de previsão de séries temporais, utilizando métodos de validação como a simulação de Monte Carlo. Os resultados reforçaram a consistência das redes LSTM em séries com ruído e alta variabilidade, características frequentemente observadas em contextos de demanda logística do agronegócio.

No contexto específico do supply-chain, Chuyen e Rajagopal (2025) conduziram uma revisão abrangente sobre a rede logística de produtos agrícolas, identificando que a grande maioria dos modelos atuais foca na minimização de custos e na redução de perdas de produtos perecíveis. Ainda assim, apontam uma lacuna significativa na literatura quanto à aplicação prática de ferramentas digitais da Indústria 4.0 em regiões em desenvolvimento.

Por fim, no que diz respeito à calibração técnica dos modelos, Mumtahina, Alahakoon e Wolfs (2024) destacaram a importância da otimização de algoritmos de previsão de carga. O estudo ressaltou que o ajuste fino (otimização de hiperparâmetros) é essencial para maximizar a acurácia dos modelos de Deep Learning, para superar as limitações e convergências prematuras encontradas em modelos individuais que não passam por essa etapa.

Com base nesses cinco estudos, verifica-se que a literatura já reconhece o potencial de modelos preditivos, especialmente das arquiteturas LSTM, em problemas de previsão temporal. Entretanto, permanece como lacuna a aplicação comparativa desses modelos ao escoamento logístico de soja em contextos amazônicos marcados por alta volatilidade e limitação de dados históricos. Essa lacuna é relevante porque dificulta o uso de ferramentas preditivas como suporte à tomada de decisão em infraestruturas logísticas emergentes, com destaque para o caso de Roraima.

4. Solução computacional

Visando os princípios da Ciência Aberta (Open Science), todo o código-fonte, incluindo os processos de ETL, treinamento da rede LSTM

parcimoniosa e testes estatísticos, encontra-se publicamente disponível para consulta e replicação. A solução proposta consiste na construção de um pipeline de ciência de dados estruturado em linguagem Python. A implementação do artefato computacional foi dividida em etapas de ingestão, pré-processamento (ETL) e modelagem preditiva. A arquitetura da solução segue um fluxo automatizado que vai desde a API do Comex Stat, passando pelo script Python e pelo modelo treinado, até a exportação das previsões para visualização em um dashboard voltado à gestão do Porto Seco conforme ilustrado na Figura 1. O ambiente experimental foi desenvolvido em linguagem Python, utilizando a plataforma Google Colab para processamento em nuvem. Para garantir a transparência e a total reprodutibilidade do método Walk-Forward e do cálculo do IPE, a implementação completa do pipeline algorítmico está armazenada e acessível em repositório público.

Após a ingestão, visto que os dados brutos governamentais são estruturados em formato wide (largo, com anos dispostos em colunas e meses em linhas), aplicou-se uma função de transformação (reshape/melt) para convertê-los em um formato long (uma série temporal cronológica contínua). Em seguida, os dados das quatro Unidades Federativas analisadas (Mato Grosso, Amazonas, Pará e Roraima) foram submetidos a um processo de limpeza. Diferentemente de abordagens tradicionais de interpolação, os valores zerados foram rigorosamente preservados, pois na logística de commodities eles representam a ausência real de escoamento (entressafra), e não uma falha de coleta. Na fase de engenharia de atributos, optou-se pela utilização do "Valor US$ FOB" e pela criação de uma variável de Sazonalidade Regionalizada (Safra_Ativa) para respeitar as janelas de escoamento específicas de cada estado, configurando uma abordagem multivariada. Visto que as redes neurais são sensíveis à magnitude dos dados, foi aplicada a técnica de normalização via MinMax Scaler, convertendo as variáveis para a escala de 0 a 1, transformando o problema em aprendizado supervisionado via janelas temporais (sliding windows). Para evitar o vazamento de dados (Data Leakage), o conjunto foi submetido a uma rigorosa Validação Cruzada de Séries Temporais (Walk-Forward), destinando 70% da série para treinamento e 30% para testes, com o conjunto de validação do Early Stopping extraído estritamente dos últimos 15% do bloco de treino.

Para responder ao problema de pesquisa, duas arquiteturas distintas foram implementadas e configuradas para a batalha de modelos:

1. Modelo Estatístico (ARIMA): Implementado através da biblioteca *statsmodels*, este modelo serviu como *baseline* da pesquisa. A configuração dos parâmetros de autorregressão, diferenciação e médias móveis (p, d, q) foi ajustada iterativamente.
2. Modelo de Deep Learning (LSTM): Construído utilizando os frameworks TensorFlow e Keras, a rede neural foi projetada especificamente para lidar com o cenário de Small Data. Para mitigar o overfitting em séries curtas e voláteis, implementou-se uma arquitetura estritamente parcimoniosa (Micro-LSTM) composta por uma única camada com 8 neurônios, regularização L2 e Dropout de 0.3. O otimizador Adam foi utilizado com a função de perda Erro Quadrático Médio (MSE), forçando o modelo a reter apenas a macro-sazonalidade de longo prazo.

A solução computacional foi desenhada levando em consideração a escalabilidade necessária para as operações de Logística 4.0. Para conectar a saída estritamente matemática à infraestrutura física do porto, o pipeline inclui uma camada de interpretação de negócio que converte a previsão de volume exportado (em quilogramas) em um Índice de Pressão de Escoamento (IPE).

Este índice compara o volume de saída previsto com o pico máximo histórico de capacidade, acionando alertas parametrizados de saturação da frota e da infraestrutura de transporte (rodovias e hidrovias). Ao alterar unicamente o parâmetro de entrada "NCM" na chamada da API, o mesmo pipeline e modelo podem prever a demanda de escoamento de outros produtos da região, como milho ou fertilizantes. Essa característica permite que o modelo seja acoplado a painéis de Business Intelligence, oferecendo suporte direto ao dimensionamento proativo de armazéns e pátios.

5. Avaliação experimental

5.1. Planejamento de pesquisa

O experimento foi planejado com base na coleta de dados secundários públicos, dados extraídos da API do Comex Stat, mantido pelo Ministério da

Economia. O conjunto de dados abrange o período de janeiro de 2015 a dezembro de 2025, filtrado especificamente pelo código NCM 1201 (Soja). Para garantir que as respostas às perguntas de pesquisa sejam não apenas confiáveis, mas também generalizáveis, a coleta não se restringiu ao estado de Roraima. Os dados englobam quatro Unidades Federativas selecionadas estrategicamente para representar diferentes maturidades logísticas: Mato Grosso (série madura e de alto volume), Pará (escoamento consolidado), Amazonas (série intermitente) e Roraima (caso-alvo, caracterizado por uma série em transição com crescimento explosivo recente). O uso dessa arquitetura comparativa permite isolar e avaliar o comportamento dos algoritmos sob variadas condições de estabilidade e não-linearidade.

Conforme recomendado pela teoria de ciência de dados e fundamentado nos Trabalhos Relacionados, o experimento adota uma abordagem de testes técnicos comparativos. O modelo estatístico ARIMA foi estabelecido como o baseline técnico da pesquisa. O artefato tecnológico desenvolvido (a rede neural LSTM) foi testado diretamente contra esse baseline. Para evitar o vazamento de informações (Data Leakage) e mitigar o risco de overfitting em séries temporais curtas (Small Data), a série temporal de cada estado foi avaliada através de uma rigorosa Validação Cruzada de Séries Temporais (Walk-Forward). Neste método, destinou-se inicialmente 70% dos dados para a base de treinamento, expandindo-se a janela iterativamente sobre os 30% finais, garantindo que a validação preditiva do modelo ocorresse em cenários reais e sucessivos (que incluem os grandes picos de safra mais recentes).

Para quantificar a acurácia dos modelos e responder à primeira pergunta de pesquisa (sobre a redução do erro preditivo), os testes técnicos foram baseados em duas métricas matemáticas absolutas: a Raiz do Erro Quadrático Médio (RMSE) e o Erro Absoluto Médio (MAE). A escolha do RMSE é crítica neste cenário, pois a métrica penaliza severamente grandes desvios, o que é fundamental para evitar erros graves de previsão durante os picos de estresse logístico da safra. Além da comparação quantitativa dos erros, o planejamento estipulou a aplicação do Teste de Diebold-Mariano. Essa avaliação garante rigor científico ao experimento, atestando que a rede neural LSTM só será considerada superior se a redução do erro preditivo apresentar significância estatística (p-valor $< 0,05$) frente ao baseline (ARIMA).

Adicionalmente, para responder às perguntas de pesquisa focadas na aplicação prática dos modelos de IA no ambiente portuário, o planejamento incluiu uma etapa de avaliação orientada ao negócio. O planejamento estipulou a conversão do volume previsto pelos algoritmos em quilogramas para um Índice de Pressão de Escoamento (IPE). A metodologia de simulação compara o volume de saída previsto pelos algoritmos de Deep Learning com o pico máximo histórico já processado pelo conjunto de treino. Através dessa etapa técnica, tornou-se possível avaliar em que medida saídas puramente numéricas podem ser traduzidas em sistemas automatizados de alertas preditivos, antecipando a saturação da infraestrutura de transporte (frota de caminhões, rodovias ou hidrovias) e mitigando os impactos dos gargalos logísticos na operação do Porto Seco.

5.2. Resultados e Discussão

A avaliação empírica nos dados de exportação de soja confirmou a hipótese metodológica de que a eficácia das arquiteturas preditivas varia diretamente conforme a maturidade logística da série temporal. A extração dos resultados do pipeline em Python, validada através da técnica de Validação Cruzada Walk-Forward, gerou a Tabela 1, que compara o Erro Quadrático Médio (RMSE) e o Erro Absoluto Médio (MAE) dos modelos nas quatro Unidades Federativas analisadas.

Tabela 1. Comparativo de Erros Preditivos (RMSE e MAE) por UF.

UF	ARIMA (RMSE)	ARIMA (MAE)	LSTM (RMSE)	LSTM (MAE)
Mato Grosso	1.396.250.000	1.101.090.000	2.057.860.000	1.579.820.000
Pará	205.297.000	170.399.000	171.676.000	123.130.000
Amazonas	2.922.060	1.994.510	2.395.500	1.371.580
Roraima	27.434.600	1.755.710	25.465.000	1.526.380

Os dados provam que o modelo de Deep Learning (LSTM) apresentou um erro preditivo significativamente menor que o ARIMA ao lidar com o

cenário amazônico, mas com uma ressalva metodológica importante. A literatura de base, como Bousqaoui et al. (2021), defende que as redes LSTM invariavelmente superam modelos lineares estatísticos. Os resultados desta pesquisa divergem parcialmente da literatura no caso de Mato Grosso, onde o ARIMA obteve um erro consideravelmente menor. Essa divergência é explicada pela teoria de análise de dados: Mato Grosso possui uma série contínua, madura e de altíssimo volume, onde a linearidade favorece algoritmos clássicos. Em contrapartida, nas séries da fronteira amazônica (Amazonas, Pará e Roraima), caracterizadas por alta volatilidade, intermitência e crescimento explosivo recente (transição estrutural), o modelo linear falha gravemente. Em Roraima (nosso caso-alvo), a LSTM parcimoniosa reduziu o RMSE de 27,4 milhões de kg para 25,4 milhões de kg. Essa redução de aproximadamente 2 milhões de quilogramas no desvio da previsão foi validada pelo Teste de Diebold-Mariano, DM=2,80; p=0,00511, permitindo rejeitar a hipótese nula de igualdade preditiva e corroborando a superioridade da LSTM sobre o ARIMA nesse cenário. Isso comprova que a imposição de uma regularização severa e topologia reduzida (Micro-LSTM) mitigou o overfitting do Small Data, capturando com excelência os picos de escoamento sazonal.

Do ponto de vista da aplicação na Logística 4.0, a conversão destas métricas matemáticas estritas (RMSE/MAE) para o Índice de Pressão de Escoamento (IPE) provou-se operacionalmente viável. A lógica do IPE traduz o erro estatístico em um painel gerencial aplicável à realidade física: ao comparar o volume de saída previsto pelo algoritmo com o pico máximo histórico já processado pela infraestrutura, o Porto Seco consegue gerar alertas parametrizados. Por exemplo, projeções de volumes que representam índices como "24,91% do pico histórico" indicam o status de "Operação Normal", dispensando contratações emergenciais de frota. Isso materializa os preceitos de Processos de Decisão Empoderados por Dados em um Smart Port.

Por fim, confirmou-se ser inteiramente viável o emprego de dados governamentais para fornecer suporte à decisão. A arquitetura de ingestão via API ou arquivos wide do portal Comex Stat sustentou o treinamento da rede neural sem a necessidade de implantações físicas onerosas (como redes de sensores IoT preliminares), promovendo alta escalabilidade para a região.

Conforme requerido para a confiabilidade da pesquisa, apontam-se limitações nos procedimentos metodológicos. O experimento foi alimentado estritamente por volumes de exportação mensal e Valores US$ FOB. Essa granularidade mensal não permite decisões operacionais de curtíssimo prazo (diárias). Adicionalmente, a exclusão de variáveis exógenas como o clima e restrições de trafegabilidade da rodovia BR-174 no modelo final restringe a captura total de rupturas de cadeia. Como direcionamento para estudos futuros, sugere-se a evolução do pipeline para incorporar dados multimodais e covariáveis climáticas que afetam diretamente o regime de escoamento.

6. Conclusão

O objetivo deste estudo foi investigar a viabilidade de aplicar arquiteturas de Deep Learning na previsão do escoamento logístico de soja na Amazônia Setentrional, comparando a eficácia de redes neurais (LSTM) com o modelo estatístico clássico (ARIMA) sob a ótica do Small Data. Para isso, a metodologia baseou-se na construção de um pipeline de dados via API, analisando séries temporais de quatro Unidades Federativas com diferentes maturidades logísticas (Mato Grosso, Amazonas, Pará e Roraima), aliada ao uso de validação cruzada temporal (Walk-Forward) e rigor estatístico.

Respondendo às perguntas de pesquisa formuladas na Introdução, os resultados mostraram que a superioridade da LSTM sobre o ARIMA não foi uniforme entre os cenários analisados, sendo condicionada ao comportamento da série temporal e às limitações impostas pelo contexto de Small Data. Verificou-se, contudo, que uma rede LSTM submetida a severas restrições topológicas e regularização agressiva (Micro-LSTM) é capaz de generalizar o comportamento volátil das exportações com maior precisão do que o ARIMA no cenário de fronteira. O Teste de Diebold-Mariano atestou a superioridade estatística da rede em mapear os picos de pressão logística no estado de Roraima. Em contrapartida, em séries maduras e lineares (como a de Mato Grosso), o modelo tradicional manteve sua eficácia, reforçando que o uso de Inteligência Artificial deve ser contextual.

Adicionalmente, no que tange à aplicação na Indústria 4.0, o estudo demonstrou que a conversão da previsão bruta em um Índice de Pressão de

Escoamento (IPE) traduz o modelo matemático em um instrumento de gestão acionável, respondendo afirmativamente à viabilidade do uso de dados públicos governamentais abertos. Ao ancorar a métrica no limite de estresse histórico do escoamento, e não em estimativas teóricas de armazenagem física, o IPE fornece um suporte à decisão robusto para a antecipação de estrangulamentos na malha de transporte.

As limitações deste estudo decorrem da extensão reduzida da série temporal amazônica, restrita ao regime econômico de crescimento explosivo pós-2019, e da indisponibilidade de dados referentes ao fluxo de entrada (recepção da colheita nos silos). Sugerimos que trabalhos futuros concentrem-se na integração de dados telemétricos e de sensores (IoT) na entrada dos portos secos e em variáveis exógenas multimodais. Essa evolução permitirá a transição do atual IPE para um sistema completo de balanço de massa bidirecional e otimização de estoque físico em tempo real, pavimentando o caminho para a consolidação de um Gêmeo Digital na infraestrutura logística de Roraima.

7. Referências

Bousqaoui, H., Slimani, I., & Achchab, S. (2021). Comparative analysis of short-term demand predicting models using ARIMA and deep learning. *International Journal of Electrical and Computer Engineering*, *11*(4), 3319-3328. https://doi.org/10.11591/ijece.v11i4.pp3319-3328

Chuyen, T. T., & Rajagopal, P. (2025). Logistics network for agricultural products: A comprehensive literature review and future directions. *Operations and Supply Chain Management*, *18*(3), 320-342. https://doi.org/10.31387/oscm0620476

Efthymiou, O. K., & Ponis, S. T. (2021). Industry 4.0 technologies and their impact in contemporary logistics: A systematic literature review. *Sustainability*, *13*(21), 11643. https://doi.org/10.3390/su132111643

Mumtahina, U., Alahakoon, S., & Wolfs, P. (2024). Hyperparameter tuning of load-forecasting models using metaheuristic optimization algorithms—a systematic review. *Mathematics*, *12*(21), 3353. https://doi.org/10.3390/math12213353

Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V. V., & Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agricultural and Forest Meteorology*, *284*, 107886. https://doi.org/10.1016/j.agrformet.2019.107886

Yunita, A., Pratama, M. I., Almuzakki, M. Z., Ramadhan, H., Akhir, E. A. P., Mansur, A. B. F., & Basori, A. H. (2025). Performance analysis of neural network architectures for time series forecasting: A comparative study of RNN, LSTM, GRU, and hybrid models. *MethodsX*, *15*, 103462. https://doi.org/10.1016/j.mex.2025.103462

Capítulo 20

Manutenção preditiva com Inteligência Artificial Explicável Integrada ao Odoo ERP

Paulo Roberto de Souza Mesquita-Junior
Thais Oliveira Almeida
Universidade Federal de Roraima

paulo.adm@gmail.com
thais.oliveira@ufrr.com

1. Introdução

A manutenção de equipamentos industriais é um processo crítico que impacta diretamente a lucratividade, segurança operacional e sustentabilidade das organizações. Atualmente, as práticas predominantes nas indústrias ainda se baseiam em manutenção reativa (corretiva) ou preventiva tradicional, onde intervenções são realizadas após falhas ocorrerem ou seguindo cronogramas fixos pré-determinados, independentemente do estado real dos equipamentos. Essa abordagem resulta em perdas significativas: paradas não planejadas custam em média USD 260 mil por hora em setores críticos (Aberdeen Research, 2016), enquanto programas de manutenção mal otimizados podem custar de 25% a 30% mais do que o necessário (Oxmaint, 2025).

O contexto da Indústria 4.0 traz oportunidades para transformar esses processos através da integração de Inteligência Artificial (IA) e análise de dados em tempo real. Sistemas de manutenção preditiva baseados em machine learning conseguem analisar padrões complexos nos dados de sensores, históricos de falhas e variáveis operacionais, possibilitando identificar com antecedência quando equipamentos necessitarão de intervenção.

Entretanto, apesar dessas vantagens comprovadas, a adoção de sistemas de IA em manutenção industrial enfrenta uma barreira significativa: a falta de explicabilidade das decisões geradas pelos modelos. Modelos de IA tradicionais funcionam como "caixas-pretas" tomando decisões sem explicar

os fatores que as levaram. No contexto de manutenção industrial, isso cria desconfiança entre técnicos e gestores, reduzindo significativamente a adoção e confiabilidade das recomendações. Estudos indicam que apenas 32% dos profissionais de operações confiam em recomendações de sistemas de IA sem explicação, comparado a 81% quando explicações transparentes estão disponíveis. Além disso, normas de conformidade industrial (ISO 13849-1, IEC 61508) e regulamentações como LGPD no Brasil, exigem cada vez mais que decisões automatizadas sejam auditáveis e explicáveis.

A lacuna identificada na literatura é precisamente a falta de soluções integradas que combinem: (1) manutenção preditiva baseada em IA com desempenho robusto; (2) explicabilidade genuína das predições através de técnicas de XAI ; e (3) integração nativa em sistemas ERP amplamente utilizados, como Odoo, que é utilizado por mais de 15 milhões de usuários globalmente e é referência em empresas de pequeno a grande porte (Odoo, 2026). Apesar de existirem módulos comerciais de manutenção preditiva para o Odoo, a maioria não oferece explicabilidade das decisões de IA ou requer implementações de terceiros complexas e custosas.

O objetivo geral deste projeto é desenvolver e validar um módulo de manutenção preditiva com Inteligência Artificial Explicável (XAI) integrado nativamente ao ERP Odoo 19. A solução visa antecipar falhas de equipamentos e gerar justificativas visuais claras diretamente no sistema, eliminando o problema da 'caixa-preta' e garantindo que os operadores validem os diagnósticos com confiança e segurança.

Frente a esse contexto, este trabalho busca responder às seguintes questões de pesquisa: (1) Qual combinação de modelo preditivo e técnica de explicabilidade oferece o melhor equilíbrio entre desempenho na detecção de falhas e qualidade das explicações geradas em um contexto de manutenção industrial? (2) É viável integrar nativamente um sistema de manutenção preditiva com explicabilidade a um ERP open-source, de forma que operadores sem conhecimento em IA possam validar os diagnósticos diretamente na interface de gestão?

Este trabalho está organizado da seguinte forma: na Seção 2, apresentamos a fundamentação teórica. Na Seção 3, descrevemos trabalhos relacionados. Na Seção 4, detalha-se a solução computacional proposta. Na

Seção 5, apresentamos a avaliação experimental. Na Seção 6, temos as conclusões e trabalhos futuros e, por fim, na Seção 7, as referências.

2. Fundamentação teórica

Nesta seção, abordam-se os conceitos fundamentais para a compreensão da proposta, divididos em Manutenção Preditiva no contexto da Indústria 4.0, Inteligência Artificial Explicável (XAI) e a plataforma Odoo ERP.

A Manutenção Preditiva (PdM) evoluiu significativamente na era da Indústria 4.0. Diferente da manutenção preventiva, que se baseia em intervalos de tempo, a PdM utiliza dados reais dos equipamentos para prever falhas futuras. Segundo Cummins et al. (2024), a manutenção preditiva utiliza inteligência artificial e aprendizado de máquina para prever o momento ideal para realizar a manutenção, visando prolongar a vida útil dos sistemas mecânicos e reduzir custos financeiros e de tempo. No entanto, conforme a complexidade dos modelos aumenta (ex: Redes Neurais Profundas), surge o problema da opacidade, onde os operadores humanos precisam confiar no sistema preditivo para aplicações críticas.

A Inteligência Artificial Explicável (XAI) surge como resposta a essa opacidade. XAI é definida como um campo da IA que promove um conjunto de ferramentas e técnicas capazes de gerar explicações interpretáveis, intuitivas e compreensíveis para humanos sobre as decisões da IA. Moosavi et al. (2024) destacam que as técnicas de XAI desempenham um papel fundamental no aumento da confiabilidade dos sistemas inteligentes aplicados a processos industriais, permitindo que operadores humanos validem as decisões tomadas por esses sistemas. Métodos como LIME e SHAP são frequentemente citados na literatura como abordagens para fornecer explicações locais (para uma decisão específica) e globais (comportamento geral do modelo).

Por fim, o Odoo ERP é um software de gestão empresarial open-source que integra diversas aplicações. O módulo de manutenção do Odoo permite gerenciar equipamentos, solicitações de manutenção e categorização de máquinas. Idris e Badawi (2025) demonstraram que a implementação de um Sistema de Gestão de Manutenção Computadorizada

(CMMS) usando o Odoo tem impacto claro na criação de bases de dados para gestão de manutenção e na facilitação da comunicação entre equipes. A proposta deste trabalho reside na extensão dessa capacidade nativa com inteligência preditiva explicável.

3. Trabalhos relacionados

A revisão da literatura focou em estudos recentes (2020-2025) que abordam a interseção entre XAI, Manutenção Preditiva e Sistemas de Gestão.

Kisten, Ezugwu e Olusanya (2024) propuseram um modelo que combina XAI com Manutenção Preditiva em instalações agrícolas inteligentes. O estudo utilizou classificadores como LSTM e XGBoost, demonstrando que a inclusão de explicabilidade não apenas fornece insights sobre dados e modelos, mas também melhora as métricas de desempenho. Embora relevante, o trabalho foca no setor agrícola e não aborda a integração direta com um ERP comercial.

Dereci e Tuzkaya (2024) desenvolveram uma metodologia de XAI para manutenção preditiva com foco na otimização de peças de reposição. Utilizando o método LIME, os autores buscaram interpretar os resultados de modelos de machine learning para apoiar tomadores de decisão, mitigando o risco de má interpretação de resultados binários (falha/não falha). Este trabalho destaca a importância de explicar decisões para a gestão de estoque, um ponto crucial que a proposta deste trabalho visa automatizar dentro do Odoo.

Por outro lado, Idris e Badawi (2025) focaram na implementação prática do módulo de manutenção do Odoo em fábricas de vestuário. O estudo confirmou a eficácia do Odoo como ferramenta de gestão (CMMS), substituindo planilhas tradicionais e melhorando o rastreamento de procedimentos. Contudo, a implementação descrita limitou-se à gestão preventiva e corretiva, sem incorporar algoritmos preditivos ou inteligência artificial. Puthanveettil Madathil et al. (2025) realizaram uma revisão abrangente sobre XAI na manufatura inteligente, apontando que, apesar do avanço teórico, as aplicações atuais permanecem fragmentadas e carecem de frameworks padronizados.

Trabalho (Autor, Ano)	Objetivo do Estudo	Técnicas / Ferramentas	Resultados Principais	Lacuna Identificada
Kisten, Ezugwu e Olusanya (2024)	Combinar XAI com Manutenção Preditiva (PdM) em instalações agrícolas inteligentes.	LSTM, XGBoost, XAI (LIME/SHAP).	Aumento de 10,66% na acurácia (XGBoost) e melhoria na interpretabilidade dos dados.	Focado em agricultura inteligente e sensores IoT específicos. Não possui integração com ERP para gestão de ordens de serviço.
Dereci e Tuzkaya (2024)	Propor metodologia XAI para manutenção preditiva com foco em otimização de peças de reposição.	Machine Learning (Python), LIME.	Demonstrou que explicações detalhadas evitam erros de interpretação por gestores em decisões binárias.	A solução é um script externo (Python) focado em estoque, sem interface integrada a um sistema de gestão de manutenção (CMMS).
Idris e Badawi (2025)	Implementar CMMS usando o módulo de Manutenção do Odoo em fábrica de vestuário.	Odoo ERP (Módulo Manutenção padrão).	Substituição de planilhas manuais e melhoria no rastreamento de histórico de equipamentos.	Implementação limitada a manutenção corretiva e preventiva. Não utiliza Inteligência Artificial nem recursos preditivos.
Puthanveettil Madathil et al. (2025)	Revisar o estado da arte de XAI na manufatura inteligente.	Revisão de Literatura.	XAI é crucial para confiança, mas as aplicações atuais são fragmentadas e não padronizadas.	Aponta a falta de *frameworks* práticos e integrados para uso real na indústria.
Este trabalho	Desenvolver módulo de manutenção preditiva com XAI integrado nativamente ao Odoo ERP, prevendo falhas e explicando os fatores determinantes por meio de gráficos SHAP na interface do sistema.	XGBoost, Random Forest, SHAP, LIME, Odoo 19 (Python/MVC), AI4I 2020 Dataset.	XGBoost (Recall 76,12%, F1 0,7149, ROC-AUC 0,9737) e SHAP (65,9× mais rápido que LIME, fidelidade 0,9546) selecionados empiricamente. Módulo funcional no Odoo gera ordens de manutenção automáticas com explicações visuais.	Integração inédita entre XAI e ERP Odoo open-source em contexto industrial diversificado, superando a fragmentação apontada por Puthanveettil Madathil et al. (2025) e expandindo os trabalhos de Kisten et al. (2024) e Idris e Badawi (2025).

A análise dos trabalhos relacionados revela que, enquanto existem pesquisas robustas sobre XAI em manutenção preditiva (Kisten et al., 2024; Dereci & Tuzkaya, 2024) e estudos sobre o uso do Odoo para manutenção convencional (Idris & Badawi, 2025), existe uma lacuna na integração desses

dois mundos. Não foi encontrada uma solução open-source documentada que integre XAI nativamente ao fluxo de trabalho do Odoo 19, permitindo que a predição e a explicação sejam visualizadas diretamente na interface de gestão do ERP.

4. Solução computacional

Esta seção descreve a arquitetura e a implementação de um módulo, nomeado Maintenance Advisor, para o ERP Odoo 19 que utiliza IA associada a XAI para realizar predições de manutenção nos equipamentos industriais com explicabilidade. É importante ressaltar que a arquitetura foi concebida para ser agnóstica em relação ao tipo de maquinário.

Embora a validação experimental deste trabalho restringe-se aos dados de uma categoria específica de equipamento para demonstrar a viabilidade da solução, o sistema é projetado para suportar um parque industrial diversificado. O desenvolvimento do motor de IA seguiu um pipeline estruturado em cinco etapas: aquisição e análise exploratória dos dados, pré-processamento, engenharia de features, treinamento e validação dos modelos, e exportação para produção. Após a exportação o modelo é implementado como módulo funcional integrado ao Odoo 19, processo ilustrado na Figura 1.

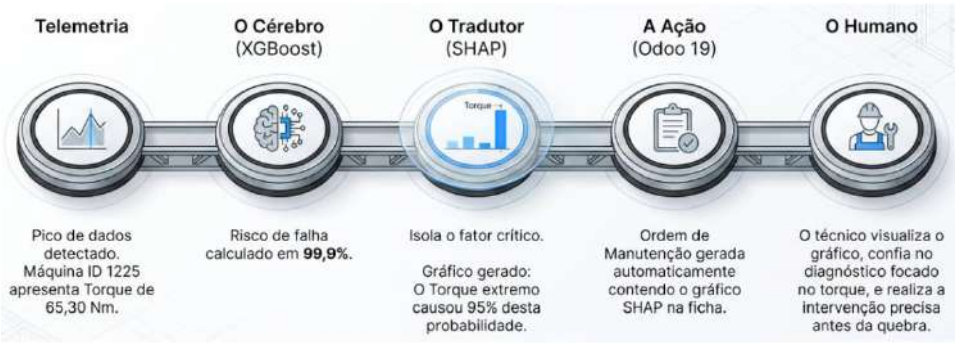


Figura 1. Implementação do módulo.

4.1. Arquitetura do Sistema

O módulo foi desenvolvido utilizando a linguagem Python, seguindo a arquitetura MVC (Model-View-Controller) nativa do Odoo.

→ Camada de Dados: Utiliza o PostgreSQL (banco padrão do Odoo) para armazenar históricos de sensores e registros de manutenção.
→ Motor de IA: Bibliotecas scikit-learn, pandas e xgboost foram integradas ao backend do Odoo para o treinamento e inferência dos modelos.
→ Motor de Explicabilidade: A biblioteca SHAP foi utilizada para gerar as explicações das predições.
→ Interface: Novas Views (XML/QWeb) foram criadas no Odoo 19 para exibir o status do equipamento e o gráfico de explicabilidade diretamente na ficha do equipamento.

5. Avaliação experimental

5.1. Planejamento de pesquisa

Para validar a solução, foi utilizado o AI4I 2020 Predictive Maintenance Dataset, um conjunto de dados público de referência industrial, composto por 10.000 registros, que simula leituras de sensores (temperatura, torque, velocidade de rotação) de maquinários de usinagem, como fresadoras e tornos CNC, comuns em pequenas e médias indústrias (https://archive.ics.uci.edu/dataset/601/ai4i+2020+predictive+maintenance+dataset).

5.2. Resultados e Discussão

Como produto principal desta pesquisa, foi desenvolvido um módulo funcional integrado ao Odoo 19, capaz de realizar a predição de falhas em equipamentos industriais, apresentar as justificativas dessas predições por meio de gráficos de explicabilidade (SHAP) e gerar ordens de manutenção automáticas (Figura 2).

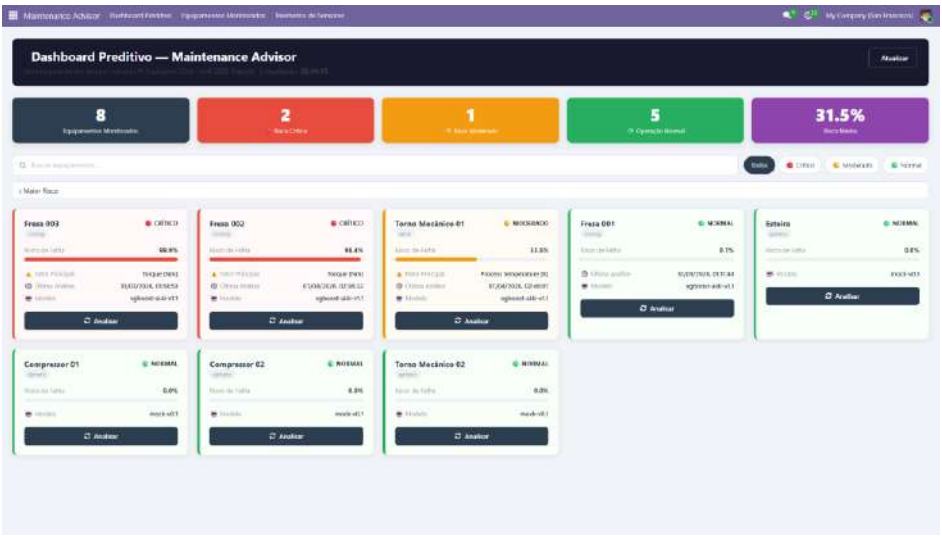


Figura 2. Dashboard do módulo Maintenance Advisor.

Em relação aos dados quantitativos, o modelo treinado alcançou ROC-AUC de 0,9737 e Recall de 0,7612 na detecção de falhas no dataset de referência — superando os limiares conservadores estabelecidos e alinhando-se aos benchmarks da literatura, que reportam Recall entre 76% e 96% para XGBoost em contextos industriais similares (Kisten et al., 2024). A aplicação de técnicas de XAI permite superar a desconfiança gerada pelos modelos "caixa-preta", conforme apontado por Moosavi et al. (2024), validando tecnicamente os alertas de manutenção emitidos pelo sistema.

Ao comparar com o conhecimento científico existente, esta solução deve se aproximar dos benefícios de explicabilidade citados por Dereci e Tuzkaya (2024), mas divergir positivamente quanto à aplicabilidade prática. Enquanto Dereci e Tuzkaya focaram em scripts externos para gestão de estoque, e Idris e Badawi (2025) limitaram-se à gestão corretiva no Odoo sem inteligência preditiva, o resultado esperado aqui preenche a lacuna de integração, centralizando a inteligência e a gestão na interface do ERP Odoo 19.

5.2.1. Estudo Comparativo entre Modelos

O estudo comparativo foi conduzido sobre o AI4I 2020 Predictive Maintenance Dataset (10.000 amostras) no ambiente de desenvolvimento Google Colab para testar e identificar a stack de IA que melhor se adequaria ao trabalho proposto.

5.2.2. XGBoost vs Random Forest

A comparação entre modelos foi conduzida por meio de validação cruzada estratificada 5-fold sobre o dataset completo, garantindo estimativas robustas das métricas. Em seguida, o modelo vencedor (XGBoost) foi treinado no conjunto de treino resultante de uma divisão estratificada 80/20 (8.000 amostras de treino e 2.000 de teste), sobre o qual foram realizadas as análises de explicabilidade SHAP e LIME.

Antes de analisar os resultados, é importante destacar a falácia da acurácia em datasets desbalanceados: um modelo dummy que nunca prevê falha atingiu 96,61% de acurácia, valor próximo ao do XGBoost (97,95%), porém com F1-Score de 0,0000, comprovando que a acurácia é uma métrica inadequada para este problema.

Tabela 2. Comparativo dos modelos XGBoost e Random Forest.

Métrica	XGBoost	Random Forest	Vencedor
ROC-AUC	0,9737 ± 0,0034	0,9632 ± 0,0077	XGBoost
PR-AUC	0,7743 ± 0,0322	0,7804 ± 0,0352	Random Forest (+0,0061)
F1-Score	0,7149 ± 0,0254	0,6531 ± 0,0323	XGBoost
Precisão	0,6780 ± 0,0306	0,9155 ± 0,0494	Random Forest
Recall	0,7612 ± 0,0622	0,5105 ± 0,0443	XGBoost
Tempo de treino	2,0 s	11,8 s	XGBoost

No contexto de manutenção preditiva, as métricas mais relevantes são PR-AUC e Recall, pois o dataset é desbalanceado e o custo de uma falha não detectada (falso negativo) supera o custo de um alarme falso.

→ **Recall (detecção de falhas):** O XGBoost detectou 76,12% das falhas reais, contra apenas 51,05% do Random Forest, uma diferença de 25 pontos percentuais. Em manutenção preditiva, isso significa que o Random Forest deixaria passar quase metade das falhas.

→ **PR-AUC:** O Random Forest obteve marginalmente melhor PR-AUC (0,7804 vs 0,7743), diferença de apenas 0,0061, estatisticamente não significativa dada a sobreposição dos desvios padrão.

→ **Precisão:** O Random Forest apresentou precisão muito superior (0,9155 vs 0,6780), porém isso é consequência direta do seu conservadorismo, ele quase nunca classifica como falha, gerando menos alarmes falsos, mas perdendo mais falhas reais (baixo Recall).

→ **Tempo de treino:** O XGBoost foi 5,9× mais rápido (2,0 s vs 11,8 s), vantagem relevante para re-treinamentos periódicos em produção.

→ **Justificativa da escolha:** O XGBoost é o modelo mais adequado para este problema. Embora o Random Forest gere menos falsos alarmes (alta precisão), seu Recall de 51% é inaceitável para manutenção preditiva industrial, significa ignorar praticamente metade das falhas reais. O XGBoost equilibra melhor a detecção (Recall 76%) com um F1-Score superior (0,7149 v.s. 0,6531), mantendo ainda vantagem em ROC-AUC e tempo de processamento.

5.2.3. SHAP v.s. LIME

A comparação entre os métodos de explicabilidade foi realizada sobre o XGBoost treinado, analisando uma amostra de falha real com alta confiança (índice 1225: Torque = 65,30 Nm, Tool wear = 192 min, probabilidade de falha = 0,999). Os critérios avaliados foram: determinismo, fidelidade, estabilidade, velocidade e suporte a análise global.

Tabela 3. Comparativo dos métodos de explicabilidade SHAP e LIME.

Critério	SHAP	LIME	Vencedor
Determinístico	Sim	Não	**SHAP**
Fidelidade (corr. com impacto real)	**0,9546**	N/A	**SHAP**
Velocidade (ms/amostra)	**1,5 ms**	100,2 ms	**SHAP (65,9× mais rápido)**
Nativo para árvores	Sim	Não	**SHAP**
Explicação global	Sim	Não	**SHAP**
Explicação individual	Sim	Sim	Empate

Analisando os critérios observados, temos:

→ **Fidelidade:** O teste de fidelidade consistiu em neutralizar cada feature (substituindo-a pela média do treino) e medir a queda na probabilidade de falha. O Torque foi a feature mais importante segundo ambos os métodos, sua remoção reduziu a probabilidade de falha de 0,9985 para 0,0442 (queda de 0,9543). A correlação de Pearson entre os valores SHAP e o impacto real de remoção atingiu 0,9546, indicando alta fidelidade ao comportamento do modelo.

→ **Estabilidade:** O SHAP é determinístico por definição (desvio padrão = 0,0000), o mesmo input sempre produz o mesmo output, pois utiliza o TreeExplainer nativo do XGBoost. O LIME, por ser baseado em amostragem aleatória, apresentou variação entre execuções: por exemplo, o desvio padrão para Rotational speed foi 0,0151 e para Torque foi 0,0081 em 5 execuções sobre a mesma amostra. Essa instabilidade compromete a confiança das explicações LIME em contextos industriais.

→ **Velocidade:** O SHAP processou 50 amostras em 76 ms (1,5 ms/amostra), enquanto o LIME levou 5.010 ms (100,2 ms/amostra), o SHAP foi 65,9× mais rápido. Essa diferença é crítica para implantação em produção no módulo do Odoo, onde explicações precisam ser geradas em tempo real.

→ **Justificativa da escolha:** O SHAP supera o LIME em todos os critérios relevantes para este trabalho: é determinístico, 65,9× mais rápido,

possui fidelidade comprovada ao modelo real (correlação 0,9546) e suporta tanto análises globais quanto individuais. A escolha de XGBoost + SHAP para o módulo Maintenance Advisor é, portanto, justificada empiricamente pelos dados do AI4I 2020 Dataset.

6. Conclusão

O objetivo deste estudo foi desenvolver e validar um módulo de manutenção preditiva com Inteligência Artificial Explicável (XAI) integrado nativamente ao ERP Odoo 19, capaz de predizer falhas em equipamentos industriais e comunicar de forma transparente os fatores determinantes dessas predições. A partir de um estudo comparativo empírico conduzido com validação cruzada estratificada 5-fold sobre o AI4I 2020 Predictive Maintenance Dataset, avaliando os modelos XGBoost e Random Forest e as técnicas de explicabilidade SHAP e LIME, buscou-se atingir o objetivo selecionando, com base em evidências quantitativas, a combinação tecnológica mais adequada ao contexto de manutenção industrial e integrando-a ao fluxo de trabalho do Odoo 19.

Em resposta à questão de pesquisa (1) verificamos que o estudo comparativo empírico demonstrou que a combinação XGBoost + SHAP oferece o melhor equilíbrio para o contexto proposto. O XGBoost superou o Random Forest nas métricas críticas para datasets desbalanceados: Recall de 76,12% contra 51,05%, o que em ambiente produtivo significaria deixar passar quase metade das falhas reais. O SHAP, por sua vez, superou o LIME em todos os critérios relevantes, garantindo que as explicações geradas sejam confiáveis, estáveis e auditáveis em contexto industrial.

A viabilidade, levantada na questão de pesquisa (2), foi confirmada com o desenvolvimento funcional do módulo Maintenance Advisor para o Odoo 19. A integração nativa permite que o ciclo completo (leitura de telemetria → predição → explicação → geração de ordem de manutenção) ocorra diretamente na ficha do equipamento, sem necessidade de ferramentas externas. O gráfico de barras SHAP exibido na ficha do equipamento traduz a decisão do modelo em linguagem acessível ao operador, fator ligado diretamente a barreira de confiança apontada por Moosavi et al. (2024), segundo a qual apenas 32% dos profissionais confiam

em sistemas de IA sem explicação, contra 81% quando explicações transparentes estão disponíveis.

Sugerimos que estudos futuros avancem em relação a esta pesquisa, buscando lidar com as limitações aqui apresentadas. A principal delas é o uso do AI4I 2020 Dataset de natureza sintética e restrito a maquinário CNC , o que limita a generalização dos resultados. Trabalhos futuros podem coletar dados reais de plantas industriais e treinar modelos específicos para outros tipos de ativos, como compressores, bombas hidráulicas e motores elétricos, expandindo a cobertura do sistema para parques industriais diversificados uma vez que a arquitetura do motor de IA foi concebida para suportar essa expansão.

Arquivos, gráficos e anexos referentes a este trabalho podem ser encontrados no repositório: https://github.com/paulomesquitas/maintenance-advisor-odoo-module.

7. Referências

Aberdeen Research. (2016). *Cost of unplanned downtime*. Recuperado de: https://www.iiot-world.com/predictive-analytics/predictive-maintenance/the-actual-cost-of-downtime-in-the-manufacturing-industry

Cummins, L., Sommers, A., Ramezani, S. B., Mittal, S., Jabour, J., Seale, M., & Rahimi, S. (2024). Explainable predictive maintenance: A survey of current methods, challenges and opportunities. https://doi.org/10.1109/ACCESS.2024.3391130

Dereci, U., & Tuzkaya, G. (2024). An explainable artificial intelligence model for predictive maintenance and spare parts optimization. J*ournal of Cleaner Production*. https://doi.org/10.1016/j.sca.2024.100078

Idris, H. M. F., & Badawi, M. A. A. (2025). Implementation of computerized maintenance management system (CMMS) in ready-made garments factories by using ODOO maintenance module. *Journal of Heritage and Design*, *5*(26), 323–345. Recuperado de: https://search.emarefa.net/detail/BIM-1728875

Kisten, M., Ezugwu, A. E., & Olusanya, M. O. (2024). Explainable artificial intelligence model for predictive maintenance in smart agricultural facilities. https://doi.org/10.1109/ACCESS.2024.3365586

Moosavi, S., Farajzadeh-Zanjani, M., Razavi-Far, R., Palade, V., & Saif, M. (2024). Explainable AI in manufacturing and industrial cyber–physical systems: A survey. *Electronics*, *13*(17), 3497. https://doi.org/10.3390/electronics13173497

Odoo. (2026). Sobre nós. Odoo S.A. Recuperado de: https://www.odoo.com/page/about-us

OXMAINT. (2025). Maintenance cost reduction: 30 proven strategies save. Recuperado de: https://oxmaint.com/article/maintenance-cost-reduction-strategies-guide

Puthanveettil Madathil, A., Luo, X., Liu, Q., Walker, C., Madarkar, R., & Qin, Y. (2025). A review of explainable artificial intelligence in smart manufacturing. International Journal of Production Research. https://doi.org/10.1080/00207543.2025.2513574

Capítulo 21

RAG-BI: Aprimorando Dashboards de Business Intelligence com RAG

Filipe Matias de Sousa
Clenya Rejane Barros de Lima
Universidade Federal de Roraima

filipematias_fm@hotmail.com
clenyabarros@uerr.edu.br

1. Introdução

No contexto da Indústria 4.0, gestores precisam tomar decisões operacionais, táticas e estratégicas de forma cada vez mais célere, apoiados no acesso rápido a informações gerenciais (Sharda et al., 2019, p. 1). Entretanto, a eficiência na tomada de decisão é desafiada pela natureza heterogênea dos dados gerados no ambiente industrial, que misturam fontes estruturadas, como bancos de dados relacionais, e não estruturadas, como manuais técnicos e relatórios de produção. A integração de Modelos de Linguagem de Grande Escala (LLMs) surge como solução promissora para processar essa complexidade, permitindo que operadores interajam com sistemas digitais de forma intuitiva e fundamentada em dados em tempo real (Arazzi et al., 2025).

Entretanto, interfaces tradicionais de *Business Intelligence* (BI) frequentemente exigem conhecimentos técnicos em SQL para consultas complexas, o que limita o acesso direto de gestores e operários de chão de fábrica aos insights críticos (Pourreza & Rafiei, 2023; Liu & Chu, 2025). Além disso, em bancos de dados de escala industrial — que podem conter centenas de tabelas e colunas — o custo computacional e o risco de imprecisões ao fornecer todo o esquema da base como contexto para um LLM tornam a abordagem de "*prompting*" simples, inviável e propensa a erros (Toteja et al., 2025).

Para superar essas limitações, a técnica de Geração Aumentada de Recuperação (RAG) permite filtrar e contextualizar dinamicamente as

informações do banco de dados antes de enviá-las ao LLM, reduzindo a probabilidade de alucinações e garantindo respostas baseadas na realidade operacional (Arslan et al., 2024; Arazzi et al., 2025; Alsayegh & Masood, 2024).

Neste trabalho, propõe-se o RAG-BI, o protótipo de uma plataforma que aprimora *dashboards* industriais por meio de uma arquitetura modular de decomposição de tarefas. O sistema identifica autonomamente se a pergunta do usuário exige uma atualização visual no *dashboard*, uma resposta informativa baseada em consultas ao banco de dados ou baseada em manuais técnicos, democratizando o acesso à informação por meio de linguagem natural no ecossistema da Indústria 4.0.

Diante desse cenário, este trabalho tem como objetivo desenvolver e analisar experimentalmente o desempenho deste protótipo, em dispositivos de borda, focando na tensão entre a restrição de memória (VRAM) e a precisão na geração de SQL complexo. Para guiar esta investigação, definiram-se as seguintes perguntas de pesquisa:

1. Como democratizar o acesso a informações gerenciais em ambientes da Indústria 4.0 para gestores que não dominam a linguagem SQL?

2. De que maneira a arquitetura RAG pode mitigar o custo computacional e o risco de alucinações ao processar bancos de dados industriais de larga escala?

3. Como a decomposição de tarefas impacta na geração de SQL por um LLM?

Esta pesquisa está organizada da seguinte forma: a Seção 2 apresenta a fundamentação teórica; a Seção 3 discute os trabalhos relacionados; a Seção 4 descreve a solução computacional proposta; a Seção 5 apresenta a avaliação experimental e a discussão dos resultados; a Seção 6 expõe as conclusões e direções para trabalhos futuros; e a Seção 7 reúne as referências utilizadas no estudo.

2. Fundamentação teórica

2.1. NLIDB e Text-to-SQL

As Interfaces de Linguagem Natural para Bancos de Dados (NLIDB) visam democratizar o acesso à informação, permitindo que usuários não especialistas interajam com dados relacionais sem dominar linguagens de consulta técnica. Historicamente, esses sistemas evoluíram de abordagens baseadas em regras para modelos de aprendizado profundo, enfrentando o "problema do descompasso", onde a linguagem natural declarativa não se mapeia diretamente na estrutura lógica do SQL (Pourreza & Rafiei, 2023; Liu & Chu, 2025). Recentemente, LLMs demonstraram ser uma base robusta para essa tarefa, especialmente quando a geração é estruturada por decomposição em submódulos – como vinculação de esquema, classificação de consultas por nível de dificuldade e autocorreção via feedback de compiladores SQL –, permitindo identificar e retificar erros de forma iterativa e aumentando a confiabilidade em ambientes produtivos (Pourreza & Rafiei, 2023).

2.2. Geração Aumentada de Recuperação (RAG) em Sistemas de BI

A técnica de RAG combina a capacidade gerativa dos LLMs com a precisão de memórias não paramétricas externas, como índices de vetores densos (Lewis et al., 2020). Em sistemas de BI, o RAG permite que o modelo acesse informações atualizadas e contextuais ausentes em seu treinamento original, reduzindo a ocorrência de alucinações factuais – aspecto particularmente relevante na Indústria 4.0, onde dados heterogêneos (SQL, planilhas, manuais em PDF) devem ser integrados em um único fluxo de decisão (Arslan et al., 2024). Na prática, perguntas de negócio são convertidas em representações vetoriais e comparadas com metadados via busca por similaridade; em bancos extensos, técnicas de poda de esquema selecionam apenas os metadados relevantes, otimizando o espaço de tokens (Shen et al., 2024; Toro et al., 2024).

2.3. Small Language Models (SLMs) e Edge Computing

Na computação aplicada à indústria, a eficiência computacional é um requisito crítico, levando à adoção de Modelos de Linguagem de Pequena

Escala (SLMs) com menos de 10 bilhões de parâmetros. Modelos como Mistral-7B, combinados com técnicas de quantização de 4 bits, reduzem o consumo de VRAM para faixas de 5 a 8 GB, viabilizando a implantação em dispositivos de borda com uma única GPU de consumo (Yan et al., 2025; Bora & Cuayáhuitl, 2024). Essa estratégia oferece vantagens como baixa latência e preservação da soberania de dados, evitando o envio de informações sensíveis para APIs de nuvem pública (Arazzi et al., 2025), enquanto a integração com o framework RAG compensa as capacidades gerativas reduzidas ao fornecer conhecimento especializado via contexto (Soman et al., 2024).

3. Trabalhos relacionados

O estudo de Pourreza e Rafiei (2023) propõe o DIN-SQL, um framework que decompõe a tarefa *Text-to-SQL* em estágios especializados, atingindo estado da arte nos benchmarks Spider (85,3%) e BIRD (55,9%). A pesquisa conclui que a decomposição modular supera o *prompting few-shot* simples, embora apresente limitações como o uso de exemplos fixos e elevada latência (~60s).

Rath (2025) utiliza a arquitetura RAG com recuperadores híbridos para integrar dados estruturados e não estruturados ao fluxo de decisão em BI, alcançando redução de 30-45% nas alucinações. O estudo conclui que o RAG representa uma mudança arquitetural fundamental para a agilidade organizacional, citando como limitações a latência de recuperação e riscos de governança de dados.

Arazzi et al. (2025) propõem o RAG-IoE, uma arquitetura de borda com recuperação híbrida e LLMs compactos para assistência técnica na Indústria 5.0. Os resultados indicam que modelos de escala intermediária (8B-14B) entregam alta acurácia, enquanto versões muito reduzidas sofrem com taxas críticas de alucinação. O estudo confirma a viabilidade de suporte local em tempo real, mas aponta a baixa correlação entre métricas automáticas e o julgamento humano.

Shen et al. (2024) propõem o ASTRES para aprimorar a geração *Text-to-SQL* em esquemas extensos, utilizando um modelo aproximador para gerar SQL preliminar e selecionar exemplos via Árvores de Sintaxe Abstrata

(AST), com poda de esquema via BM25. O método atingiu estado da arte no Spider (86,6%), filtrando até 70% do esquema sem perda de contexto. Como limitação, aponta-se a dependência de bases acadêmicas devido à escassez de dados corporativos massivos.

Tabela 1. Comparação com trabalhos relacionados.

Trabalho	Abordagem	Infra	Base	Resultado	Limitações
DIN-SQL Pourreza e Rafiei (2023)	Decomposição modular *Text-to-SQL* com autocorreção	GPT-4 (nuvem) Sem restrição de hardware	Acadêmica Spider / BIRD	EX 85,3% (Spider) EX 55,9% (BIRD)	Exemplos f*ew-shot* fixos; latência ~60s; dependente de API cloud
RAG para BI Rath (2025)	RAG com recuperadores híbridos para dados estrutura-dos e não estruturados	LLMs de nuvem Sem restrição de hardware	Acadêmica Cenários simula-dos	Redução de 30-45% nas alucinações	Latência de recuperação; riscos de governança; sem avaliação em borda
RAG-IoE Arazzi et al. (2025)	RAG de borda com recuperação híbrida para Indústria 5.0	LLMs compactos (8B-14B) Edge computing	Técnica Manuais Industri-ais	Alta acurácia em modelos 8B-14B; alucinação crítica em modelos meno-res	Baixa correlação entre métricas automáticas e julgamento humano; sem *Text-to-SQL*
ASTRES Shen et al. (2024)	Poda de esquema via BM25 + seleção de exemplos por AST	GPT-4 (nuvem) Sem restrição de hardware	Acadê-mica Spider	EX 86,6% (Spider); filtragem de até 70% do esquema	Dependência de benchmarks acadêmicos; sem dados corporativos re-ais
RAG-BI Este trabalho (2026)	Modular RAG + decomposição de tarefas + *Text-to-SQL* + *dashboard* BI	Mistral-7B GPTQ-4bit GPU 6 GB (borda)	Real 320 tabelas, 2,3M re-gistros	EX 21,95% global; 45,83% (EASY); VRAM estável em 5,41 GB	Acurácia limitada em consultas complexas; avaliação exclusivamente técnica

O principal diferencial do RAG-BI frente aos trabalhos relacionados reside na convergência simultânea de três dimensões: a execução integral em hardware de borda com GPU de apenas 6 GB, a validação sobre um banco de dados real e a integração do pipeline *Text-to-SQL* com um *dashboard* de BI interativo por meio de um roteador de intenção multimodal, capaz de

direcionar a pergunta do usuário tanto para consultas informativas quanto para a atualização dinâmica de filtros no painel.

4. Solução computacional

O protótipo do RAG-BI baseia-se no paradigma de Modular RAG, projetado para operar em *hardware* de borda (GPUs de 6GB), garantindo que o LLM receba repertório estrutural suficiente. O fluxo de processamento é dividido em quatro fases integradas, conforme Figura 1 que apresenta a arquitetura e pipeline do RAG-BI.

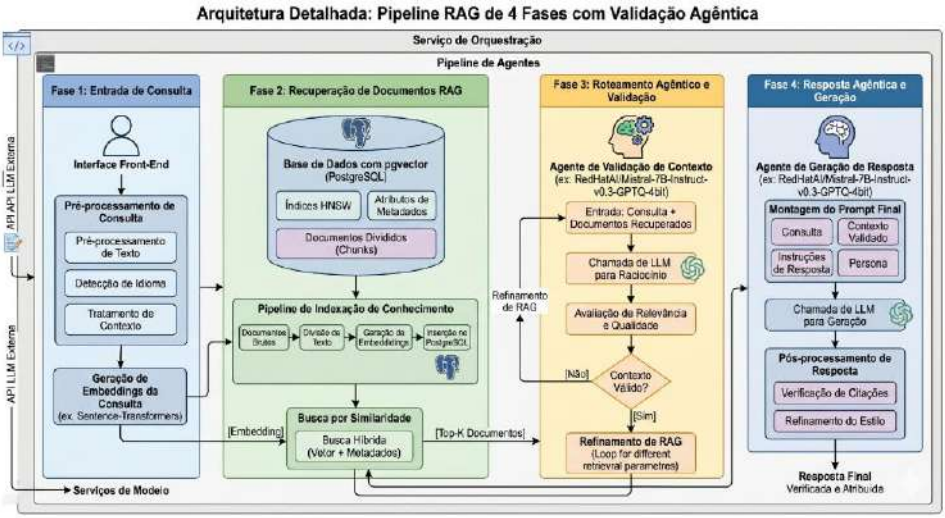


Figura 1. Pipeline do RAG-BI. Fonte: Elaborado pelo autor com auxílio de inteligência artificial (2026).

A Fase 1 - Ingestão e Indexação — implementa indexação incremental via Hash SHA-256 para DDLs e recuperação hierárquica (*Summary/Leaf Nodes*) para manuais técnicos, evitando reprocessamento desnecessário.

A Fase 2 - Interceptação e Roteamento — utiliza um roteador de intenção que direciona o fluxo para um de três caminhos: geração de scripts SQL para consultas informativas ao banco de dados, geração de scripts SQL para aplicação de filtros à *view* base do *dashboard* de BI (composta por 4

tabelas, 23 colunas, 13 funções customizadas e 4 funções do Postgresql), ou recuperação de informações em documentos técnicos.

A Fase 3 - Pipeline *Text-to-SQL* — contém o núcleo de otimização que utiliza HyDE para geração de consultas hipotéticas, *Reciprocal Rank Fusion* (RRF) para busca híbrida e o framework ASTRES para poda de esquemas irrelevantes. A complexidade é classificada via DIN-SQL para guiar o raciocínio do modelo.

Na Fase 4 - Autocorreção Iterativa e Execução — o SQL gerado é validado preventivamente no banco de dados através da instrução EXPLAIN; caso ocorram erros, o LLM recebe o log para depuração automática. Com o SQL correto, responde à pergunta com base no resultado da consulta ou atualiza o painel BI.

5. Avaliação experimental

5.1. Planejamento de pesquisa

A avaliação experimental visa responder às três perguntas de pesquisa formuladas na Introdução. Para isso, adotou-se uma abordagem quantitativa, na qual a saída do sistema é comparada automaticamente com um gabarito de referência (*ground truth*).

A base de dados utilizada é um backup de banco PostgreSQL de produção pertencente a uma distribuidora de materiais de construção localizada em Boa Vista - Roraima. Contém 320 tabelas, 57 *views* e aproximadamente 2,3 milhões de registros, abrangendo módulos de movimentação de estoque, faturamento, contas a pagar e receber, cadastro de produtos e gestão de funcionários. A escolha de uma base de dados real, em detrimento de *benchmarks* acadêmicos como o Spider, justifica-se pela aderência ao cenário da Indústria 4.0 e pela presença de esquemas legados com convenções de nomenclatura proprietárias.

O dataset de avaliação foi construído a partir de 397 perguntas de negócio, geradas com o apoio do modelo Gemini 2.5 Flash a partir das 115 tabelas com dados. Os SQLs de referência foram validados por execução automática no banco, verificando-se a ausência de erros sintáticos; não foi realizada verificação semântica sistemática.

As perguntas de avaliação foram classificadas em quatro níveis de dificuldade, adaptando a taxonomia de Pourreza & Rafiei (2023), cujas definições e distribuição amostral estão detalhadas na Tabela 1.

Do total de perguntas, 80% (315) compõem o pool de referência para seleção de exemplos *few-shot*, e os 20% restantes (82) formam o conjunto de avaliação, selecionado por amostragem aleatória estratificada: 24 EASY, 24 NON-NESTED, 24 NESTED e 10 FILTER_BI.

Tabela 2. Métricas de avaliação do RAG-BI.

Métrica	Descrição	Critério	Referência
Acurácia de Execução (EX)	Métrica binária: 1 se o resultado do SQL gerado é idêntico ao do SQL de referência após normalização, 0 caso contrário.	$EX \in \{0, 1\}$	Pourreza e Rafiei (2023)
Taxa de Encurtamento de Esquema (SSR)	Proporção de colunas removidas pelo módulo de *Schema Linking* antes do envio ao LLM.	$SSR = 1 - (C_p / C_o)$	-
Latência	Tempo total entre o envio da pergunta e a resposta final, durante a execução do pipeline.	Medida em segundos	-
Pico de VRAM	Alocação máxima de memória da GPU durante a inferência, via PyTorch e nvidia-smi.	Medida em GB	-

Complementarmente, foram calculadas as métricas *Exact Set Match* (EM), que compara a estrutura do SQL gerado com a do gabarito por decomposição em cláusulas (Yu et al., 2018), e *Valid Efficiency Score* (VES), que penaliza consultas corretas, porém ineficientes (Li et al., 2024).

A EX foi adotada como métrica principal por ser preferível à comparação textual de instruções SQL, uma vez que consultas sintaticamente distintas podem retornar conjuntos de dados idênticos (Pourreza & Rafiei, 2023). A SSR é particularmente relevante para a segunda pergunta de pesquisa, pois quantifica como a arquitetura RAG mitiga o custo computacional em bancos de larga escala.

O ambiente de execução foi composto por um Intel Core i7-13620H, 32 GB de RAM DDR5 e GPU NVIDIA RTX 4050 (6 GB de VRAM), sob Ubuntu 24.04.4 LTS. O modelo Mistral-7B-Instruct-v0.3, quantizado em 4 bits via

GPTQ (RedHatAI), foi servido pelo vLLM com temperatura zero e janela de contexto de 4096 tokens. O modelo de *embeddings* BGE-M3 e o *reranker* BGE-Reranker-v2-M3 (ambos BAAI) foram executados na CPU. A orquestração utilizou o LlamaIndex e o armazenamento vetorial foi implementado no PostgreSQL com pgvector. Para evitar estouro da janela de contexto, o resultado da execução SQL é truncado a 1000 caracteres (∼250 tokens).

A avaliação foi realizada de forma automatizada, sequencial, com classificação dos resultados em OK, Timeout, Falha de Geração e Erro Lógico.

5.2. Resultados e Discussão

Conforme Tabela 3, o sistema alcançou uma EX global de 21,95%, com 18 acertos em 82 perguntas avaliadas.

Tabela 3. EX por nível de dificuldade.

Dificuldade	Definição	Perguntas	Acertos	EX (%)
EASY	Filtros simples e diretos	24	11	45,83
NON-NESTED	JOINs sem subconsultas	24	2	8,33
NESTED	Subconsultas e agregações condicionais	24	2	8,33
FILTER_BI	Filtros dinâmicos na *view* do *dashboard*	10	3	30
Total		**82**	**18**	**21,95**

A análise por nível de dificuldade revela uma redução progressiva da acurácia conforme a complexidade estrutural aumenta, padrão convergente com Pourreza e Rafiei (2023). As consultas EASY concentraram o melhor desempenho, enquanto as categorias NON-NESTED e NESTED apresentaram acurácia significativamente inferior. A categoria FILTER_BI, com 30% de acurácia, indica dificuldade do modelo Mistral quantizado em lidar com a complexidade da *view* utilizada no painel BI.

A matriz de confusão revela que a escolha inadequada da ferramenta foi um fator determinante para a baixa taxa de sucesso global. O sistema apresentou 0% de aprovação nos 9 casos em que selecionou incorretamente

o 'gerador_script_filtro_bi' para demandas EASY, NESTED e NON-NESTED, enquanto alcançou seu melhor desempenho (11 acertos) quando o 'executor_consultas_informativas' foi corretamente acionado. Esse cenário demonstra que a acurácia do RAG-BI depende fundamentalmente da eficácia do mecanismo de seleção de ferramentas, tornando o roteador crucial na arquitetura.

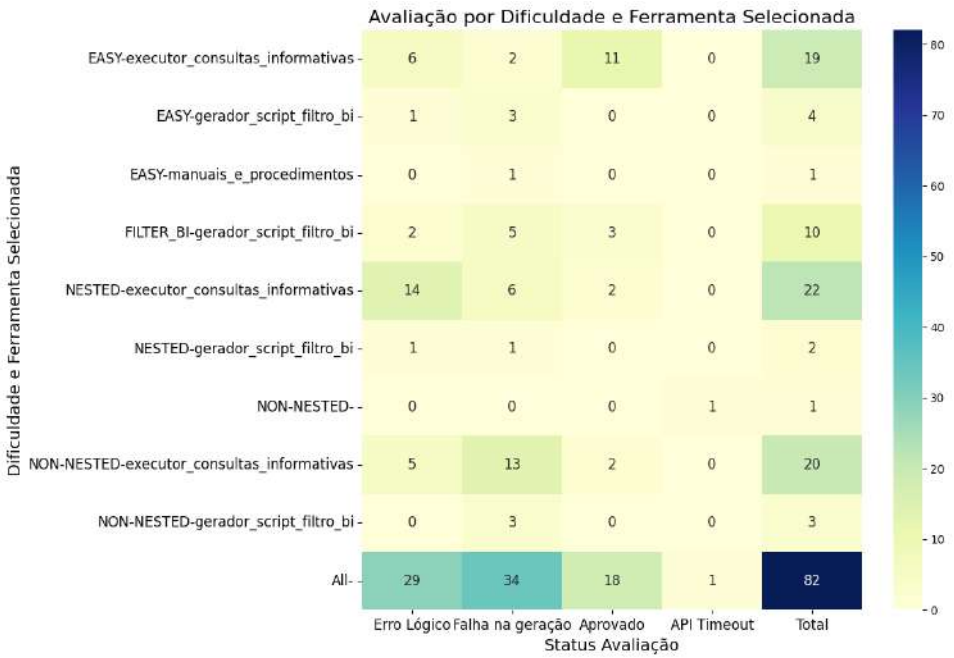


Figura 2. Matriz com Ferramenta Selecionada.

A análise qualitativa revelou 18 aprovações, 29 erros lógicos e 34 falhas de geração, com os erros concentrados nas consultas NESTED (15 ocorrências). O módulo de autocorreção mostrou-se ineficaz contra falhas de lógica profunda, limitando-se a correções sintáticas superficiais. Esses achados indicam que a arquitetura, embora eficaz para consultas EASY, exige estratégias de *prompting* mais granulares para superar as restrições do modelo quantizado em consultas complexas.

Tabela 4. Taxa de Encurtamento de Esquema (SSR).

Dificuldade	Média cols. original	Média cols. podadas	SSR Média (%)
EASY	27	6	22,22
NON-NESTED	33	5	15,15
NESTED	34	4	11,76
FILTER_BI	23	8	34,78
Soma Acumulada	**117**	**23**	**19,65**

O módulo de *Schema Linking* removeu, em média, 19,65% das colunas recuperadas antes do envio ao LLM. As consultas NESTED exibiram a menor taxa (11,76%), coerente com a necessidade de preservar mais colunas em consultas complexas, enquanto a categoria FILTER_BI permitiu redução mais expressiva (34,78%). Essa otimização contribui para a viabilidade em *hardware* de borda, pois a economia de tokens preserva a VRAM e reduz o tempo de resposta.

Tabela 5. Estatísticas descritivas.

Métrica	Média	DP	Mín.	Máx.
Latência (s)	78,70	29,68	3,94	180
VRAM (GB)	5,41	0,03	5,25	5,60

O pico médio de VRAM de 5,41 GB, abaixo do limite de 6 GB da GPU, confirma a viabilidade em *hardware* de borda. Combinado com a SSR, esse resultado responde à segunda pergunta de pesquisa: a combinação de quantização GPTQ-4bit, poda de esquema e execução de *embedding/reranking* na CPU viabiliza a inferência em GPUs de consumo. A latência média de 78,70 segundos situa-se em limiar aceitável para *dashboards* interativos.

As métricas complementares reforçam os achados da EX. O EM global de 13,33% indica que a maioria das consultas aprovadas utilizou estrutura sintática distinta do gabarito, porém semanticamente equivalente. O VES

médio (EX=1) foi de 0,47, sugerindo SQLs menos eficientes que os de referência, com destaque para a categoria EASY (VES de 0,92).

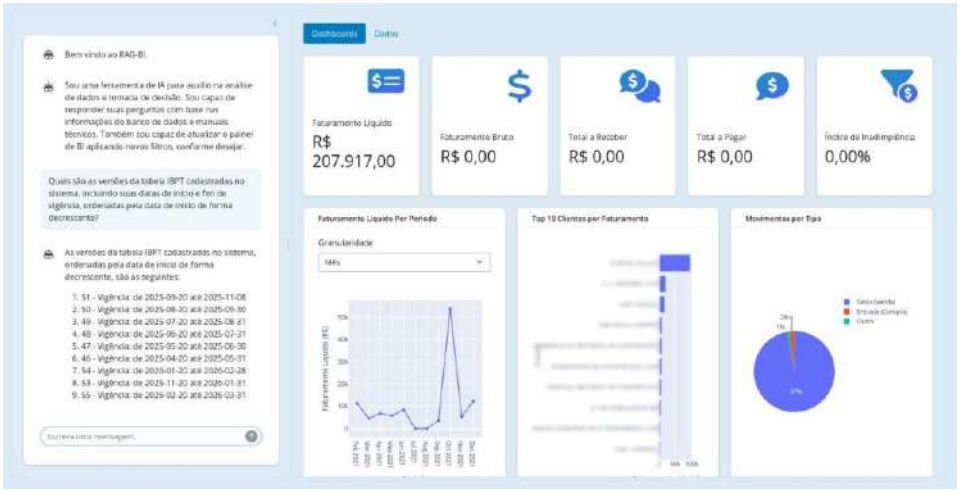


Figura 3. Protótipo do Painel com RAG-BI.

A EX global de 21,95% é inferior aos 85,3% do DIN-SQL no Spider (Pourreza & Rafiei, 2023), diferença esperada dado que o Spider utiliza esquemas acadêmicos padronizados, enquanto a base avaliada possui nomenclatura proprietária, relações não convencionais e não passou por um processo de limpeza; havendo ainda erros ocorridos no roteamento da seleção de ferramentas, que comprometeram a acurácia das respostas.

Este estudo possui limitações relevantes: a avaliação foi exclusivamente técnica, sem testes de usabilidade; o uso apenas do Mistral-7B quantizado impede a distinção entre limitações arquiteturais e restrições do modelo; e o *ground truth*, gerado via Gemini 2.5 Flash, foi validado apenas por execução automática, sem revisão semântica.

6. Conclusão

O objetivo deste estudo foi desenvolver e analisar experimentalmente o desempenho da plataforma RAG-BI, o protótipo de um

sistema que democratiza o acesso a informações gerenciais em ambientes da Indústria 4.0 por meio de linguagem natural. A partir de uma arquitetura modular de decomposição de tarefas, avaliou-se o sistema em um banco de dados PostgreSQL real de uma distribuidora de materiais de construção, utilizando 82 perguntas de teste classificadas por nível de dificuldade.

Em relação à primeira pergunta, a plataforma demonstrou-se capaz de gerar consultas SQL a partir de linguagem natural, embora a acurácia global de 21,95% indique que o sistema ainda exige supervisão humana em consultas complexas.

Quanto à segunda pergunta de pesquisa, o consumo estável de VRAM em 5,41 GB confirma que a integração entre poda de esquema, quantização GPTQ-4bit e o processamento de *embeddings/reranking* na CPU viabiliza a operação em *hardware* de borda com 6 GB de memória. Complementarmente, a latência média registrada, embora elevada, situa-se em um limiar aceitável para a operação de *dashboards* industriais interativos.

No tocante à terceira pergunta, os resultados evidenciaram que a decomposição modular é eficaz para consultas simples, mas que a eficácia do roteador de ferramentas é um fator crítico para o desempenho global. A seleção inadequada da ferramenta pelo módulo de roteamento resultou em 0% de aprovação nos casos de classificação incorreta, revelando que a acurácia do sistema depende tanto da capacidade gerativa do LLM quanto da precisão do mecanismo de roteamento e seleção de ferramentas.

Sugere-se que estudos futuros validem a arquitetura com outros SLMs, como Qwen-2.5-7B ou Llama-3-8B, para distinguir limitações arquiteturais de restrições do Mistral quantizado. Recomenda-se também incorporar testes de usabilidade com gestores e investigar técnicas de *prompting* mais granulares e *fine-tuning* com LoRA sobre o domínio específico, visando elevar a acurácia em consultas NON-NESTED e NESTED.

7. Referências

Alsayegh, A., & Masood, T. (2024). Leveraging generative AI for knowledge-driven information retrieval in the energy sector. *MATEC Web of Conferences*, *401*, 10008. https://doi.org/10.1051/matecconf/202440110008

Arazzi, M., Sciarroni, M. M., Nocera, A., & Storti, E. (2025). RAG-IoE: IoT context-aware information retrieval with Large Language Models in Industry 5.0. *ACM Transactions on Internet of Things*, *6*(4), Article 23. https://doi.org/10.1145/3762669

Arslan, M., Munawar, S., & Cruz, C. (2024). Exploring Business Events using Multi-source RAG. In *28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*. https://doi.org/10.1016/j.procs.2024.09.303

Bora, A., & Cuayáhuitl, H. (2024). Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications. *Machine Learning and Knowledge Extraction*, *6*(4), 2355–2374. https://doi.org/10.3390/make6040116

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. Recuperado de: https://arxiv.org/abs/2005.11401

Li, J., Hui, B., Qu, G., Yang, J., Li, B., Wang, B., Qin, B., et al. (2024). Can LLM Already Serve as A Database Interface? A Big Bench for Large-Scale Database Grounded Text-to-SQLs. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*. Recuperado de: https://proceedings.neurips.cc/paper_files/paper/2023/file/83fc8fab1710363050bbd1d4b8cc0021-Paper-Datasets_and_Benchmarks.pdf

Liu, J., & Chu, S. (2025). Text2SQL Business Intelligence System Based on Retrieval-Augmented Generation (RAG). Engineering Reports. https://doi.org/10.1002/eng2.70249

Pourreza, M., & Rafiei, D. (2023). DIN-SQL: Decomposed In-Context Learning of Text-to-SQL with Self-Correction. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*. Recuperado de: https://proceedings.neurips.cc/paper_files/paper/2023/file/72223cc66f63ca1aa59edaec1b3670e6-Paper-Conference.pdf

Rath, G. (2025). Retrieval-Augmented Generation (RAG) for Real-Time Business Intelligence. *International Journal for Multidisciplinary Research*

(IJFMR), *7*(4). Recuperado de: https://www.ijfmr.com/papers/2025/4/52677.pdf

Sharda, R., Delen, D., & Turban, E. (2019). *Business intelligence e análise de dados para gestão do negócio* (4. ed.). Bookman.

Shen, Z., Vougiouklis, P., Diao, C., Vyas, K., Ji, Y., & Pan, J. Z. (2024). Improving Retrieval-augmented Text-to-SQL with AST-based Ranking and Schema Pruning. https://doi.org/10.48550/arXiv.2407.03227

Soman, K., Rose, P. W., Morris, J. H., et al. (2024). Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, *40*(9), btae560. https://doi.org/10.1093/bioinformatics/btae560

Toro, S., Anagnostopoulos, A. V., Bello, S. M., et al. (2024). Dynamic Retrieval Augmented Generation of Ontologies using Artificial Intelligence (DRAGON-AI). *Journal of Biomedical Semantics*, *15*(19). https://doi.org/10.1186/s13326-024-00320-3

Toteja, R., Sarkar, A., & Comar, P. M. (2025). In-Context Reinforcement Learning based Retrieval-Augmented Generation for Text-to-SQL. In *Proceedings of the 31st International Conference on Computational Linguistics (pp. 10390–10397)*. Recuperado de: https://aclanthology.org/2025.coling-main.692/

Yan, L., Wan, Q., Liu, C., Duan, S., Han, P., & Xu, Y. (2025). SPS-SQL: Enhancing Text-to-SQL generation on small-scale LLMs with pre-synthesized queries. *Pattern Recognition Letters*. https://doi.org/10.1016/j.patrec.2025.04.016

Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., et al. (2018). Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. In *Conference on Empirical Methods in Natural Language Processing (pp. 3911–3921)*. https://doi.org/10.18653/v1/D18-1425

Capítulo 22

Pipeline sequencial de Deep Learning para detecção de ativos e OCR em documentação técnica da Indústria 4.0

Rodrigo Damasceno Nascimento
Alexander Fernandez Correa
Universidade Federal de Roraima

rodrigoml.damasceno@gmail.com
alexander.fernandez@ufrr.com

1. Introdução

O avanço das tecnologias digitais tem promovido transformações significativas em diversas áreas da engenharia, especialmente com a incorporação de técnicas de Inteligência Artificial, IA, em processos tradicionalmente realizados de forma manual. Nos últimos anos, de acordo com Matos et al. (2024), observa-se um crescimento expressivo na aplicação de métodos baseados em aprendizado profundo e visão computacional, impulsionado pela necessidade de aumentar a eficiência, reduzir erros e automatizar tarefas complexas

Projetos elétricos, especialmente aqueles relacionados à iluminação, envolvem a análise de diversos elementos gráficos e textuais, como plantas técnicas, ambientes e símbolos de luminárias. A interpretação automatizada desse tipo de documento tem sido amplamente investigada na literatura recente, especialmente por meio da combinação de técnicas de visão computacional e processamento de documentos, permitindo a extração de informações estruturadas a partir de arquivos digitais (Subramani et al., 2020).

Diante desse cenário, este trabalho tem como objetivo desenvolver e validar um *framework* inteligente de visão computacional fundamentado em arquiteturas de *deep learning* e OCR para a auditoria autônoma de projetos elétricos, visando sistematizar a extração de metadados técnicos e mitigar falhas críticas de conformidade na engenharia luminotécnica.

Nesse contexto, surge a seguinte questão de pesquisa que orienta este trabalho: em que medida a integração de modelos de visão computacional e técnicas de OCR pode automatizar, com precisão e eficiência, a análise de projetos elétricos de iluminação em formato PDF, reduzindo erros e o tempo de verificação em comparação aos métodos manuais?

Dessa forma, busca-se contribuir para a redução de erros, otimização do tempo de análise e padronização dos processos de verificação de projetos elétricos. Adicionalmente, o trabalho evidencia o potencial da integração entre visão computacional, OCR e tecnologias *web* na construção de sistemas inteligentes aplicados à engenharia, reforçando a importância da adoção de soluções digitais no contexto atual.

2. Fundamentação teórica

2.1. Rede Neural YOLO

You Only Look Once, YOLO, é uma abordagem de detecção de objetos em imagens baseada em redes neurais convolucionais, CNN's, amplamente reconhecida por sua eficiência e processamento eficiente.

De acordo com Redmon (2016), YOLO reformula o problema de detecção de objetos como uma regressão direta, que mapeia pixels da imagem para *bounding boxes* e probabilidades de classe. A imagem de entrada é dividida em uma grade, e cada célula da grade é responsável por prever um conjunto de caixas e suas respectivas confidências.

Versões mais recentes passaram a incorporar melhorias significativas em sua arquitetura, visando aumentar a precisão e reduzir o custo computacional. A versão utilizada neste trabalho, YOLOv11, segue uma arquitetura modular composta por três componentes principais: *backbone*, *neck* e *head*, e incorpora otimizações modernas que contribuem para o seu desempenho, como estruturas convolucionais mais eficientes e estratégias aprimoradas de detecção, tornando-o adequado para cenários com alta densidade de informação visual (Ultralytics, 2024).

Para atingir o objetivo da pesquisa, foi utilizada abordagem que permite explorar diferentes configurações do modelo de forma sistemática,

reduzindo a necessidade de ajustes manuais e contribuindo para a obtenção de resultados mais robustos.

No contexto de projetos elétricos, a capacidade do modelo de detectar objetos em imagens complexas permite sua utilização em plantas técnicas, que frequentemente apresentam grande densidade de informações visuais. Conforme Goodfellow, Bengio & Courville (2016), modelos baseados em aprendizado profundo são altamente eficazes na extração de padrões em dados visuais, desde que treinados com conjuntos de dados representativos.

2.2. Reconhecimento Óptico de Caracteres - OCR

O Reconhecimento Óptico de Caracteres, OCR, de acordo com Gonzalez e Woods (2010), é um processo que envolve múltiplas etapas de processamento de imagens, incluindo pré-processamento, segmentação, extração de características e reconhecimento de padrões. Essa técnica é fundamental em sistemas que envolvem automação de leitura de documentos, sendo especialmente relevante em aplicações de engenharia, como a análise automatizada de projetos elétricos, onde há a necessidade de extrair informações de plantas, diagramas e carimbos técnicos.

A etapa de extração de características é responsável por identificar propriedades relevantes dos caracteres, como contornos, formas e padrões geométricos. Essas informações são então utilizadas por algoritmos de classificação, que podem variar desde métodos tradicionais, como redes neurais artificiais e máquinas de vetor de suporte, até abordagens mais recentes baseadas em *deep learning* (Goodfellow et al., 2016).

No presente trabalho, foi utilizado o *Tesseract* OCR, que, em suas versões mais recentes, adota uma arquitetura baseada em redes neurais recorrentes do tipo *Long Short-Term Memory*, LSTM (Smith, 2007). Diferentemente das abordagens tradicionais, que dependiam fortemente de segmentação explícita de caracteres, o *Tesseract* moderno utiliza um pipeline no qual a imagem é inicialmente processada por camadas convolucionais responsáveis pela extração de características visuais, seguidas por camadas recorrentes que modelam a sequência textual presente na imagem.

No contexto de projetos elétricos, o OCR desempenha um papel essencial na extração automática de informações presentes em carimbos técnicos, como nome do projeto, responsável técnico, data, escala e revisões.

3. Trabalhos relacionados

A partir de uma consulta a bases de dados diversas, utilizando os descritores *YOLO* e *OCR*, foram selecionadas as três pesquisas mais recentes e aderentes ao nosso tema de pesquisa, a saber: Pacheco (2024), Hora (2024), e Oliveira e Castro (2021).

O trabalho de Pacheco (2024) propõe a utilização do modelo YOLOv8 integrado ao OCR para a identificação automática de placas veiculares em sistemas de controle de acesso. O autor demonstra que a combinação dessas técnicas permite a detecção eficiente de objetos e a extração de informações textuais.

Hora (2024) desenvolve um sistema de reconhecimento automático de placas utilizando diferentes variantes do YOLO associadas a ferramentas de OCR, destacando a importância da criação e anotação de *datasets* para o desempenho dos modelos.

Por fim, Oliveira e Castro (2021) realizam um estudo comparativo entre diferentes arquiteturas de redes neurais profundas, incluindo o YOLO, aplicado a ambientes agrícolas. Os autores concluem que o YOLO apresenta melhor desempenho em termos de velocidade e eficiência para aplicações em tempo real.

Dessa forma, observa-se que a literatura recente evidencia o crescente uso de técnicas de visão computacional e OCR em diferentes áreas. No entanto, nota-se uma lacuna na aplicação dessas tecnologias especificamente na análise automatizada de projetos elétricos, o que reforça a relevância e a contribuição do presente trabalho.

4. Solução computacional

A solução computacional desenvolvida neste trabalho consiste em uma aplicação *web* voltada à análise automatizada de projetos elétricos em

formato PDF, integrando técnicas de visão computacional, OCR e processamento de documentos.

A arquitetura do sistema foi estruturada de forma modular, sendo composta por três principais camadas: interface do usuário, processamento de dados e inteligência artificial. A implementação foi realizada utilizando tecnologias *web*, com destaque para as linguagens HTML, CSS e *JavaScript*, permitindo sua execução diretamente no navegador, sem a necessidade de instalação de *software* adicional.

Destaca-se como principal contribuição deste trabalho a utilização de um pipeline sequencial de modelos especializados, permitindo a decomposição do problema em etapas e contribuindo para o aumento do desempenho global da solução.

4.1. Arquitetura do sistema

A interface do sistema foi desenvolvida em HTML, sendo responsável pela estruturação dos elementos visuais e interação com o usuário, enquanto o CSS foi utilizado para estilização e organização da interface gráfica. A lógica da aplicação foi implementada em *JavaScript*, responsável pelo controle do fluxo de execução, manipulação dos arquivos e integração com os serviços de inteligência artificial.

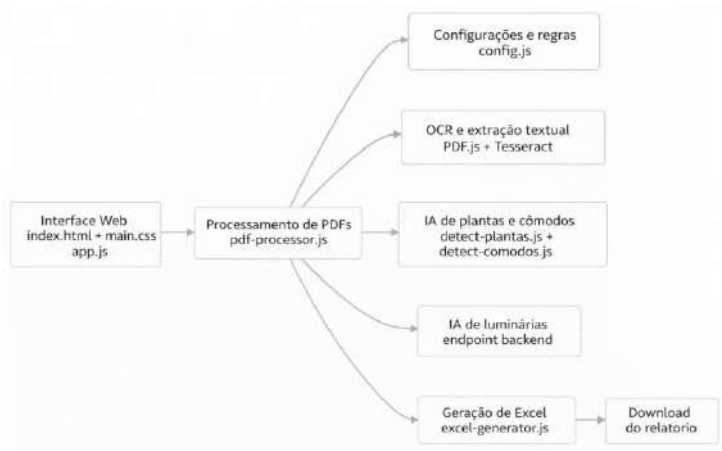


Figura 1. Arquitetura geral.

O processamento central é realizado por um módulo responsável pela leitura e análise dos arquivos PDF, no qual são aplicadas técnicas de extração textual e validação de informações. Esse módulo também realiza a comunicação com os modelos de inteligência artificial e com o gerador de relatórios.

4.2. Fluxo de funcionamento do sistema

O funcionamento da aplicação ocorre de forma sequencial, iniciando com o envio dos arquivos PDF pelo usuário e finalizando com a geração de um relatório em formato Excel contendo os resultados da análise.

Inicialmente, o usuário seleciona os arquivos e configura as opções de análise, como verificação de carimbo, identificação de pranchas e ativação da análise por inteligência artificial. Em seguida, o sistema realiza a leitura dos documentos e a extração de informações textuais, aplicando validações relacionadas à consistência dos dados.

Caso a análise por inteligência artificial esteja ativada, as páginas dos documentos são convertidas em imagens, permitindo a aplicação de modelos de visão computacional para detecção de elementos gráficos.

4.3. Processamento de documentos e OCR

A leitura dos arquivos PDF é realizada por meio da biblioteca PDF.js, responsável por interpretar o conteúdo dos documentos diretamente no navegador. Para arquivos que não possuem texto estruturado, é aplicada a técnica de reconhecimento óptico de caracteres utilizando a biblioteca Tesseract.js.

Essa abordagem permite a extração automatizada de informações presentes nos carimbos técnicos, como nome do projeto, responsável técnico, número da prancha e demais dados relevantes. O uso combinado de extração direta e OCR garante maior robustez ao sistema, possibilitando a análise de diferentes tipos de documentos, incluindo aqueles digitalizados.

4.4. Inteligência Artificial aplicada

A detecção de elementos gráficos nas plantas foi realizada por meio de modelos de visão computacional baseados na arquitetura YOLO. O modelo foi desenvolvido e treinado utilizando a plataforma *Roboflow*, sendo configurado com o modelo do tipo YOLOv11 *Object Detection (Fast),*o modelo mais recente disponível na plataforma.

A escolha dessa arquitetura justifica-se pela sua elevada eficiência em tarefas de detecção de objetos em tempo real, possibilitando a identificação simultânea de múltiplos elementos em imagens complexas.

No contexto da solução proposta, o modelo de inteligência artificial é aplicado nas seguintes etapas: detecção de regiões correspondentes às plantas nos documentos analisados; identificação automática de cômodos presentes nas plantas arquitetônicas; e detecção de luminárias e outros elementos associados ao projeto de iluminação.

As inferências são realizadas por meio de chamadas a APIs disponibilizadas pela plataforma *Roboflow*, permitindo que o processamento ocorra de forma remota. Essa abordagem elimina a necessidade de execução local do modelo, reduzindo os requisitos computacionais da aplicação e favorecendo sua escalabilidade, além de alinhar-se a arquiteturas modernas baseadas em serviços.

4.5. Regras de validação

Além da aplicação de inteligência artificial, o sistema implementa regras de validação baseadas em lógica computacional, responsáveis por verificar a consistência das informações extraídas dos documentos. Dentre as principais validações realizadas, destacam-se: verificação do nome do arquivo no conteúdo do PDF; identificação do número da prancha; validação da descrição do projeto; verificação de presença de termos incompatíveis com a disciplina analisada; e identificação de assinaturas por meio do nome do arquivo. Essas validações contribuem para a detecção de inconsistências comuns em projetos técnicos, auxiliando no controle de qualidade dos documentos.

4.6. Geração de relatórios

Após o processamento dos arquivos, os resultados são organizados e apresentados ao usuário por meio da interface *web*. Adicionalmente, o sistema gera automaticamente um relatório em formato Excel, contendo os dados extraídos dos carimbos; os resultados das validações; e informações obtidas pelos modelos de inteligência artificial. A geração do arquivo é realizada por meio de uma biblioteca *JavaScript* específica para manipulação de planilhas, permitindo a exportação dos dados de forma estruturada.

4.7. Considerações sobre a implementação

A solução desenvolvida apresenta como principal vantagem a integração entre diferentes tecnologias, permitindo a análise automatizada de projetos elétricos de forma eficiente e acessível. A utilização de uma arquitetura baseada em serviços, com inferência via API, possibilita maior flexibilidade e escalabilidade, alinhando-se aos princípios da Indústria 4.0.

5. Avaliação experimental

5.1. Planejamento de pesquisa

A abordagem adotada caracteriza-se como um estudo experimental aplicado, no qual foram utilizados documentos de projetos elétricos reais em formato PDF, 30 arquivos referente as pranchas, como base para os testes. Esses documentos foram selecionados de forma a representar diferentes condições de uso, incluindo variações de qualidade, resolução, organização gráfica e complexidade das plantas, permitindo uma avaliação mais abrangente do sistema.

A coleta de dados foi realizada por meio do processamento automatizado dos arquivos no sistema desenvolvido, conforme fluxo descrito na Seção 4. Durante esse processo, cada documento foi submetido às seguintes etapas: extração de informações textuais diretamente do PDF; aplicação de OCR para leitura de dados em imagens rasterizadas; detecção de regiões de plantas por meio de modelo de visão computacional; identificação de cômodos e luminárias utilizando modelos baseados na arquitetura YOLO;

e validação de consistência entre informações textuais e regras pré-definidas do domínio.

Para avaliação do desempenho da solução, foram considerados indicadores quantitativos e qualitativos, como a capacidade de identificação correta de elementos gráficos (cômodos e luminárias); taxa de detecção de informações textuais em carimbos técnicos; ocorrência de inconsistências identificadas automaticamente pelo sistema; e a robustez da solução frente a diferentes qualidades de entrada.

Além disso, foi realizada uma análise comparativa implícita com o processo manual de verificação de projetos, considerando aspectos como tempo de execução e suscetibilidade a erros humanos, ainda que não tenha sido definido um baseline formal com métricas padronizadas da literatura.

A arquitetura do sistema, baseada em processamento distribuído com inferência via API, permitiu a execução dos testes sem dependência de hardware local de alto desempenho, favorecendo a escalabilidade da solução. Os modelos de inteligência artificial foram acessados de forma remota, conforme configuração dos serviços de detecção de plantas, cômodos e luminárias, o que garante consistência nas inferências independentemente do ambiente de execução.

Dessa forma, o planejamento experimental adotado possibilita a obtenção de resultados confiáveis quanto ao desempenho da solução proposta, ao mesmo tempo em que estabelece uma base para futuras validações em cenários operacionais mais amplos.

5.2. Resultados e Discussão

5.2.1. Detecção de elementos

A avaliação experimental da solução desenvolvida foi realizada com base na aplicação de três modelos de visão computacional, todos utilizando a arquitetura YOLOv11 *Object Detection (Fast)*, cada um responsável por uma etapa específica do processamento das plantas de projetos elétricos.

A abordagem adotada consiste em um *pipeline* sequencial, no qual a saída de uma rede neural serve como entrada para a etapa seguinte. Inicialmente, é realizada a detecção das regiões de interesse, no caso as

plantas, seguida pela identificação dos cômodos e, por fim, pela detecção de luminárias.

A primeira rede neural, denominada Plantas, foi responsável por identificar as regiões do documento onde estão localizadas as plantas técnicas, permitindo o recorte automatizado dessas áreas para processamento posterior. O modelo foi treinado com um conjunto de 126 imagens, sendo 24 destinadas à validação e 18 ao teste, sem aplicação de redimensionamento das imagens.

Após o recorte das regiões de planta, foi aplicada a segunda rede neural, responsável pela identificação dos cômodos presentes nas plantas arquitetônicas. Esse modelo foi treinado com 177 imagens, sendo 18 para validação e 9 para teste. Diferentemente do modelo anterior, foi realizada a padronização do tamanho das imagens para 640x640 *pixels*, visando melhorar a eficiência computacional e a generalização do modelo.

Por fim, a terceira rede neural foi responsável pela detecção de luminárias nas plantas. O modelo foi treinado com 144 imagens, sendo 9 para validação e 9 para teste, também utilizando imagens redimensionadas para 640x640 *pixels*. Para as três redes neurais descritas, foram aplicadas técnicas de data augmentation — especificamente inversão, flip e rotação — o que triplicou o volume de dados de treinamento em todos os casos. As métricas obtidas nas três detecções estão dispostas na Tabela 1.

Tabela 1. Métricas obtidas – Detecção de plantas, cômodos e luminárias.

Objetos	**mAP@50 (%)**	**Precisão (%)**	***Recall* (%)**	**F1-*score* (%)**
Plantas	96,2	96,8	87,5	91,9
Cômodos	94,2	91,7	87,8	89,7
Luminárias	95,3	92,4	93,2	92,8

5.2.2. Discussão geral dos resultados

De forma geral, os resultados obtidos demonstram que a segmentação do problema em etapas permite que cada modelo seja otimizado para uma tarefa específica, resultando em melhor desempenho global do sistema.

Observa-se que a etapa de detecção de plantas apresentou maior precisão, sendo fundamental para garantir a qualidade das etapas subsequentes, enquanto a detecção de cômodos apresentou maior desafio, devido à variabilidade dos elementos arquitetônicos. A detecção de luminárias apresentou desempenho mais equilibrado, beneficiando-se da padronização dos símbolos técnicos.

Além disso, a redução do tamanho das imagens para 640x640 mostrou-se uma estratégia eficiente para balancear desempenho e custo computacional, embora possa impactar a detecção de elementos muito pequenos.

Outro ponto relevante é que os resultados obtidos estão diretamente relacionados à qualidade e quantidade do conjunto de dados utilizado no treinamento. O número reduzido de imagens em alguns modelos pode limitar a capacidade de generalização, indicando a necessidade de ampliação do *dataset* em trabalhos futuros.

Por fim, destaca-se que a integração dos modelos com o sistema de processamento de PDFs e OCR possibilita a construção de uma solução completa, capaz de automatizar diferentes etapas da análise de projetos elétricos, contribuindo para a redução de erros e aumento da eficiência no processo de verificação. Adicionalmente, o sistema apresentou desempenho satisfatório na extração de informações de carimbo por meio de OCR, contribuindo para a análise textual dos documentos.

6. Conclusão

O presente trabalho teve como objetivo desenvolver e avaliar uma solução computacional baseada em Inteligência Artificial para a análise automatizada de projetos elétricos, com ênfase na análise automatizada de plantas de projetos elétricos, incluindo a detecção de regiões de plantas, identificação de cômodos, verificação de luminárias e extração de informações de carimbos técnicos. Para isso, foram integradas técnicas de visão computacional, por meio da arquitetura YOLOv11, com métodos de OCR, além do uso de tecnologias *web* para construção de uma aplicação acessível e funcional.

Os resultados obtidos evidenciam que a abordagem proposta é eficaz na automação de etapas tradicionalmente manuais, apresentando elevado desempenho na detecção de elementos gráficos, bem como na extração de dados textuais relevantes. A utilização de um *pipeline* sequencial composto por múltiplos modelos especializados mostrou-se uma estratégia eficiente, permitindo maior precisão e organização no processo de análise.

Como continuidade desta pesquisa, recomenda-se a ampliação da base de dados, o aprimoramento das técnicas de pré-processamento de imagens e o desenvolvimento de novos modelos para identificação de outros elementos presentes em projetos elétricos. Também se destaca o potencial de expansão da solução para outras disciplinas da engenharia, como projetos arquitetônicos, estruturais e mecânicos.

Dessa forma, conclui-se que a aplicação de Inteligência Artificial na análise automatizada de projetos elétricos é não apenas viável, mas também promissora, contribuindo significativamente para a redução de erros, aumento da eficiência e padronização dos processos de verificação, reforçando a importância da integração entre engenharia e tecnologias digitais no contexto atual.

7. Referências

Flanagan, D. (2025). *JavaScript:* O guia definitivo (7th ed.). Bookman.

Gonzalez, R. C., & Woods, R. E. (2010). *Processamento digital de imagens* (3rd ed.). Pearson.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Hora, B. A. D. (2024). Sistema de reconhecimento automático de placas veiculares utilizando visão computacional [Monografia, Universidade Federal do Pará].

Matos, L., Ribeiro, K., & Sisnando, A. (2024). Avaliação do impacto da inteligência artificial na engenharia: Desafios e oportunidades para engenheiros na era da inteligência artificial. *Revista Scientia*, *10*, 91–110. https://doi.org/10.5281/zenodo.15044714

Oliveira, D. O. A., & Castro, J. P. D. (2021). Um estudo comparativo entre as arquiteturas de redes neurais profundas AlexNet e YOLO aplicadas ao

problema de visão computacional em ambientes agrícolas. In *Congresso Brasileiro de Agroinformática*.

Pacheco, C. E. M. (2024). Aplicação de visão computacional e técnicas de reconhecimento de caracteres em placas veiculares para controle de acesso [Monografia, Instituto Federal do Espírito Santo].

Pohorec, S., Zorman, M., & Kokol, P. (2013). Analysis of approaches to structured data on the web. *Computer Standards & Interfaces*, *36*, 256–262. https://doi.org/10.1016/j.csi.2013.06.003

Redmon, J. (2016). You only look once: Unified, real-time object detection. Recuperado de: https://arxiv.org/abs/1506.02640

Roboflow. (2023). Roboflow documentation. Recuperado de: https://docs.roboflow.com

Smith, R. (2007). An overview of the Tesseract OCR engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR) (pp. 629–633)*. IEEE.

Subramani, S., et al. (2020). A survey of deep learning approaches for OCR and document understanding. Recuperado de: https://arxiv.org/pdf/2011.13534

Ultralytics. (2024). Ultralytics YOLO documentation. Recuperado de: https://docs.ultralytics.com

Capítulo 23

Arquitetura distribuída Edge-Server para inventário em tempo real com RFID e técnicas de localização por proximidade

Kelly Silva de Aquino
Herbert Oliveira Rocha
Universidade Federal de Roraima

kellyaquino22@gmail.com
herbert.rocha@ufrr.com

1. Introdução

No âmbito do gerenciamento industrial, a administração de estoques é consolidada como uma função estratégica essencial, possuindo vínculo direto com a garantia da eficácia operacional e, por conseguinte, com a plena satisfação dos consumidores. Conforme preconiza Ballou (2006), os estoques constituem acúmulos de materiais destinados a facilitar os processos produtivos ou a atender à demanda do mercado, funcionando como uma reserva de bens fundamental para o equilíbrio entre a oferta e a disponibilidade de produtos.

Para organizações que operam com produtos de elevada perecibilidade, a exemplo da indústria de sorvetes, a gestão de inventário torna-se duplamente crítica. Tal cenário demanda um monitoramento contínuo e rigoroso da cadeia do frio para assegurar a integridade dos itens. Em unidades fabris situadas em regiões de clima extremo, como o estado de Roraima, o incremento logístico sem automação correspondente eleva o risco de quebras na cadeia do frio. Falhas no controle manual de inventário de produtos ultracongelados não resultam apenas em prejuízos financeiros, mas comprometem a segurança alimentar devido à alta perecibilidade dos insumos.

Nesse contexto, a implementação de sistemas computacionais baseados na tecnologia de Identificação por Radiofrequência (RFID) representa uma inovação paradigmática para a gestão de produtos. De

acordo com Jin et al. (2006), essa tecnologia viabiliza a automação integral dos processos de inventário e rastreabilidade, superando as limitações técnicas inerentes aos sistemas tradicionais de código de barras. Diferente dos sistemas de código de barras, que exigem contato visual e manuseio unitário, a tecnologia RFID passiva de alta frequência (HF) permite capturas simultâneas e automatizadas. O sistema ColdChain-ID utiliza essa premissa para transformar unidades logísticas em nós informacionais, vinculando metadados de fabricação e histórico de movimentação sem a necessidade de intervenção humana direta.

A integração da solução proposta, contudo, não se limita à coleta isolada de dados, mas insere-se no ecossistema de Cyber-Physical Systems (CPS). Ao alinhar o monitoramento em tempo real com pilares como Edge Computing, o sistema minimiza a latência no processamento local, permitindo que apenas metadados críticos sejam enviados à nuvem. Essa arquitetura prepara o terreno para a implementação de Digital Twins, onde a representação virtual do ativo é alimentada continuamente por fluxos de Big Data, permitindo análises preditivas que transcendem a manutenção reativa convencional.

Atualmente, a adoção de soluções IIoT enfrenta o “vale da desilusão” devido a desafios críticos de interoperabilidade entre protocolos legados e modernas interfaces de comunicação. A fragmentação de dados e a vulnerabilidade em camadas de rede (security breaches) elevam o custo operacional. Este trabalho endereça especificamente a latência determinística e a integridade dos dados em ambientes de alta interferência eletromagnética, lacunas que frequentemente inviabilizam a automação plena em plantas industriais de grande escala.

Adicionalmente, a transição para uma infraestrutura IoT distribuída exige uma abordagem rigorosa quanto à segurança cibernética dos nós de borda (*Edge*). Em ambientes industriais, a exposição de credenciais *hardcoded* e a interceptação de pacotes de dados representam riscos críticos à integridade do inventário. Para mitigar tais vulnerabilidades e garantir a confidencialidade, o sistema ColdChain-ID implementa uma camada de segurança baseada em criptografia simétrica AES-128 em modo CBC, assegurando que a comunicação entre os microcontroladores e o servidor ocorra de forma cifrada.

A viabilidade financeira da solução para Pequenas e Médias Empresas (PMEs) baseia-se na preservação do valor agregado dos produtos aliada ao baixo custo de implementação, utilizando microcontroladores ESP32 e infraestrutura de rede local. O retorno sobre o investimento é impulsionado pela automação da coleta de dados, reduzindo falhas associadas ao processo manual. A proposta contempla o desenvolvimento de uma infraestrutura IoT resiliente, adaptada às condições térmicas industriais. Adicionalmente, incorpora algoritmos de localização por proximidade e Máquinas de Estados Finitos para automatizar fluxos e minimizar erros. Também são estabelecidos mecanismos de segurança com criptografia AES-128 para proteção da comunicação. Por fim, a solução é avaliada quanto à sua viabilidade técnica e econômica no contexto das PMEs.

Este trabalho está organizado da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica sobre localização *indoor*, RFID e ecossistema IoT. O Capítulo 3 revisa os trabalhos relacionados na literatura. No Capítulo 4, detalha-se a solução computacional proposta, incluindo a arquitetura de hardware, software e lógica de segurança. O Capítulo 5 descreve o planejamento da pesquisa e os resultados experimentais obtidos. Finalmente, o Capítulo 6 apresenta as conclusões, verificando o alcance dos objetivos e sugerindo trabalhos futuros.

2. Fundamentação teórica

Esta seção apresenta os conceitos fundamentais relacionados à técnica de localização Indoor e Ecossistema RFID e IoT, fornecendo o embasamento teórico necessário para a compreensão do presente estudo.

2.1. Técnicas de Localização Indoor

De acordo com Masoudi et al. (2025), o uso de sistemas baseados em GPS é inviável para o rastreamento em tempo real dentro de armazéns inteligentes, uma vez que o sinal sofre atenuações severas e interferências causadas pela estrutura física do ambiente. Para superar as limitações do GPS em ambientes fechados, a técnica de localização por proximidade é amplamente utilizada devido à sua simplicidade e baixo custo.

Para garantir a integridade lógica do fluxo produtivo na cadeia do frio, a modelagem das transições de estoque foi estruturada por meio de Máquinas de Estados Finitos (FSM). Segundo Hopcroft, Motwani e Ullman (2002), uma FSM permite representar o comportamento de um sistema por meio de um número finito de estados e transições acionadas por eventos — no contexto deste trabalho, as leituras dos portais RFID nos setores de produção, câmara fria e expedição. No contexto deste trabalho, a FSM permite formalizar as regras de decisão para a atualização de status dos produtos (ex: 'Em Produção', 'Em Câmara Fria, 'Expedição').

2.2. Ecossistema RFID e IoT (RFID-IoT)

A convergência entre a identificação por radiofrequência (RFID) e a Internet das Coisas (IoT) estabelece o que se define como RFID-IoT, uma arquitetura de computação pervasiva que transforma objetos passivos em nós inteligentes de uma rede global. No contexto da Indústria 4.0, essa integração não se limita à leitura de dados, mas à criação de um fluxo contínuo de informações que alimenta algoritmos de decisão em tempo real (Ferdousmou et al., 2024). Um ecossistema RFID-IoT para ambientes industriais é estruturado em três níveis de hardware: Tags para transporte, Leitores/Microcontroladores e Sensores Termossensíveis. De acordo com Onoriode U. and Kotonya G. (2018), a base para um rastreamento de estoque eficaz baseado em IoT reside em uma arquitetura multicamadas projetada para capturar, transmitir, processar e analisar dados ao longo de toda a jornada da cadeia de suprimentos.

Apesar dos benefícios operacionais, a implementação do ecossistema RFID-IoT em indústrias de sorvetes enfrenta desafios físicos severos que exigem estratégias de mitigação específicas. Para mitigar esses riscos e evitar a degradação do hardware pelo rigor térmico e pela condensação, o projeto deste protótipo prevê o posicionamento dos leitores e gateways IoT na área externa da câmara fria. De acordo com Zhitao Xu et al. (2024), o uso de "portais de leitura" nas zonas de transição, somado ao ajuste de ganho e à adoção de antenas com polarização circular, é uma técnica fundamental para garantir a acurácia de 99% exigida no controle de estoque, mantendo a integridade dos componentes eletrônicos e a estabilidade da transmissão de dados em tempo real.

3. Trabalhos relacionados

A convergência RFID e IoT na Indústria 4.0 busca otimizar a visibilidade em cadeias de suprimentos complexas. Segundo o estudo de Wu et al. (2023) a plataforma ID-SCCL emprega Internet de Tudo (IoE) e Gêmeos Digitais para o monitoramento de ambientes frios. Utiliza Deep Learning, com Auto-encoders para detecção de anomalias e ResNet para o posicionamento indoor. A localização ocorre via Bluetooth Low Energy (BLE), mitigando as limitações do GPS em centros logísticos. Baseada na arquitetura PISA, a solução integra sensores ambientais para garantir a integridade de produtos. O sistema monitora variáveis críticas, assegurando a conformidade técnica de ativos termossensíveis e segurança. Validado no setor farmacêutico, o modelo promove a sincronização em tempo real entre o físico e o digital. A digitalização elimina registros manuais, elevando a precisão da rastreabilidade e a velocidade de resposta. Conclui-se que o uso de representações digitais é eficaz na mitigação de falhas e inconsistências operacionais.

O estudo de Rao (2025) analisa o uso da IoT no rastreamento de inventários, propondo uma arquitetura com sensores e análise em nuvem para visibilidade em tempo real. Os resultados indicam reduções de 14,2% no tempo de trânsito e 11,7% nas perdas de estoque. Além disso, observa-se mitigação de 79% nas variações térmicas em cadeias de frio. A abordagem destaca a importância da integração entre sensores e sistemas de gestão. Em comparação, este trabalho propõe o uso de etiquetas RFID na indústria de sorvetes. Ambas as abordagens convergem no uso de hardware para automatizar a coleta de dados. Rao (2025) foca na logística global de transporte, enquanto este estudo enfatiza o ambiente industrial. A proposta busca aplicar os benefícios da IoT ao controle de estoque local. Assim, reforça-se que a integração entre percepção e ERP maximiza eficiência e qualidade.

O estudo de Sousa et al. (2020) propõe o sistema INEXT para localização de objetos em ambientes internos via RFID passivo (13,56 MHz). A abordagem baseia-se na técnica de proximidade, associando a posição do ativo ao leitor que captura o sinal. Isso permite o monitoramento de movimentações entre ambientes por meio de portais de leitura. A arquitetura utiliza microcontroladores NodeMCU integrados a um servidor Node.js. O

sistema automatiza o inventário e gera alertas por e-mail em casos de movimentação indevida. A solução demonstra a viabilidade de infraestrutura de baixo custo para rastreamento de itens. A seguir, apresenta-se uma tabela comparativa entre os trabalhos correlatos, destacando seus principais aspectos técnicos e contribuições.

Tabela 1. Análise comparativa entre os trabalhos correlatos.

Critério	Wu et al. (2023)	Rao (2025)	Sousa et al. (2020)	Este Trabalho
Tecnologias principais	IoE, RFID, BLE, Deep Learning (Autoencoders, ResNet)	IoT, sensores, computação em nuvem	RFID passivo, NodeMCU, Node.js	RFID, MQTT, Node-Red, MongoDB.
Arquitetura	Baseada em PISA + Gêmeos Digitais	Arquitetura em nuvem orientada a dados	Arquitetura embarcada + servidor web	Integração embarcada + sistemas corporativos
Custo da solução	Alto	Médio	Baixo	Baixo a médio
Domínio de aplicação	Setor farmacêutico	Logística global	Logística *Indoor*	Indústria de sorvetes

Essa comparação evidencia que o trabalho proposto se posiciona como uma solução intermediária, combinando baixo custo (como Sousa et al., 2020) com monitoramento ambiental, como Rao (2025) e Wu et al. (2023), agregando valor ao contexto industrial específico da cadeia do frio.

4. Solução computacional

Esta seção detalha a arquitetura, os requisitos e a estrutura do protótipo intitulado ColdChain-ID desenvolvido para o controle automatizado de estoque e rastreabilidade de lotes congelados na indústria de sorvetes.

4.1. Arquitetura do Sistema de Monitoramento

A solução proposta consiste em um sistema inteligente de inventário e rastreabilidade para uma unidade fabril de sorvetes. Utiliza tecnologia RFID passiva de alta frequência (13,56 MHz) para identificação de ativos. A

arquitetura é distribuída, composta por microcontroladores e leitores RFID. Os dispositivos são instalados em pontos estratégicos, como portais de acesso entre setores. Esses nós capturam automaticamente a identidade e localização dos itens. O processo inicia no setor **produção**, com a fixação das etiquetas nos malotes de sorvete. Cada tag é vinculada ao lote no momento da formação das unidades logísticas. Durante o transporte interno, leitores identificam os itens por indução. Os dados são enviados via rede sem fio a um servidor central, garantindo rastreamento em tempo real.

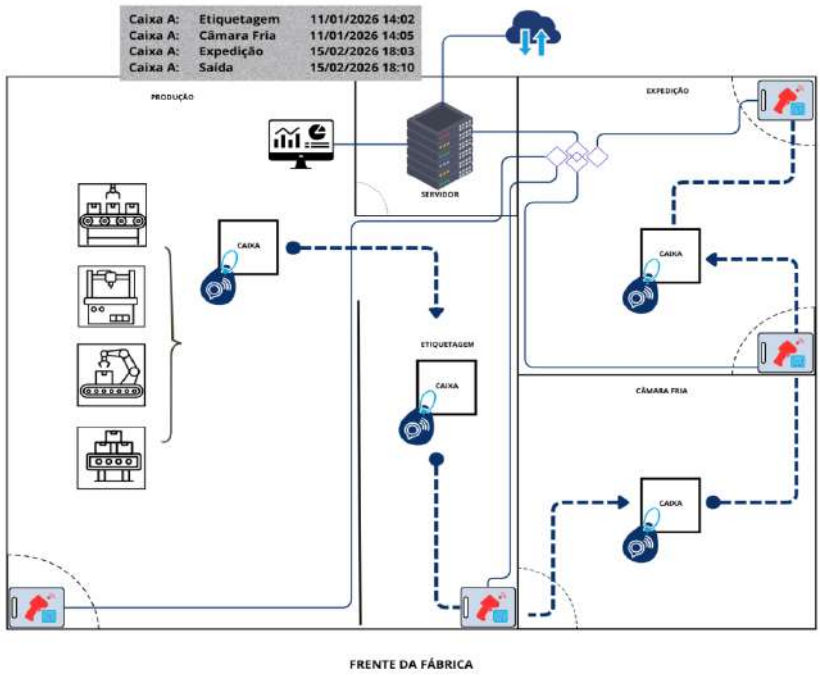


Figura 1. Visão geral do método proposto para rastreabilidade e inventário em tempo real na planta industrial.

4.2. Infraestrutura de Software e Monitoramento

A infraestrutura de software do sistema ColdChain-ID integra múltiplas tecnologias para operação em tempo real na IIoT. Os nós de borda, baseados no ESP32, foram programados em C na plataforma Arduino IDE,

permitindo controle direto dos periféricos e comunicação via SPI com o leitor RFID. A arquitetura evoluiu de um modelo bloqueante para um modelo assíncrono baseado em temporização, assegurando maior disponibilidade e determinismo. Para orquestração e visualização dos dados, utilizou-se o Node-RED, responsável pelo processamento das informações e construção de dashboards interativos. A comunicação é realizada via protocolo MQTT com o broker HiveMQ, garantindo o envio eficiente de eventos de inventário e alertas. O armazenamento persistente é feito em banco de dados MongoDB em rede local, permitindo rastreabilidade completa. Essa integração possibilita monitoramento contínuo e confiável do fluxo produtivo. Assim, a solução assegura visibilidade em tempo real.

Para garantir a scalability e a high availability do sistema, optou-se pela implementação de um cluster via MongoDB Atlas. A escolha por um banco de dados NoSQL fundamenta-se na flexibilidade do esquema (schemaless), essencial para comportar a natureza heterogênea dos pacotes de telemetria IIoT. O cluster foi configurado com um conjunto de réplicas (replica sets) distribuído geograficamente, assegurando tolerância a falhas e baixa latência de leitura/escrita. Além disso, a indexação geoespacial e temporal foi otimizada para suportar consultas complexas sobre o histórico de estados dos ativos em tempo real.

Um aspecto crítico tratado na solução é a sincronização temporal dos dados em um ambiente distribuído. A segurança foi garantida por criptografia AES-128 CBC com padding PKCS7 implementado no firmware. O sistema adota uma estratégia redundante, na qual o ESP32 tenta obter horário via NTP, enquanto o Node-RED atua como autoridade de tempo. Caso inconsistências sejam detectadas, o servidor corrige automaticamente os timestamps antes da persistência. Essa abordagem garante a linearidade cronológica dos eventos e a confiabilidade dos relatórios. Adicionalmente, os nós de borda implementam armazenamento local temporário para cenários de falha de conectividade. Os dados são mantidos em buffer até o restabelecimento da comunicação com o servidor. Após reconexão, os registros são corrigidos e enviados para processamento. Esse mecanismo assegura integridade e continuidade dos dados. Dessa forma, o sistema mantém consistência temporal mesmo em condições adversas.

4.3. Lógica de Localização e Funcionamento

A inteligência do sistema ColdChain-ID baseia-se na convergência de três métodos de localização que asseguram a integridade do inventário em tempo real: a localização por proximidade, a estimativa por centróide e a modelagem por Máquina de Estados Finitos (FSM). O método de proximidade estabelece que a posição de um malote é vinculada ao nó de borda que efetua a captura do sinal, sendo esta precisão refinada pelo algoritmo de centróide quando uma etiqueta é detectada simultaneamente por múltiplos leitores, calculando-se a média aritmética das coordenadas para determinar a posição exata do ativo. Para formalizar o fluxo operacional e as transições entre os setores de produção, câmara fria e expedição, utiliza-se a lógica de FSM, que gerencia as mudanças de status dos produtos com base em eventos de leitura e regras de decisão pré-estabelecidas. Todas as movimentações geram registros automatizados (logs) contendo o identificador único da etiqueta (ID), a cronologia do evento (Data/Hora) e a localização detectada, permitindo a consolidação de relatórios.

4.4. Especificações do Protótipo

A comunicação do ecossistema ColdChain-ID é realizada via WLAN em rede local, conectando dispositivos de borda a um servidor central. O sistema utiliza o microcontrolador ESP32, alimentado em 5V, com conectividade Wi-Fi e suporte a criptografia. A interface com o leitor RFID PN532 V3 ocorre via protocolo SPI, garantindo alta taxa de transferência. O leitor opera em 13,56 MHz com alimentação otimizada para maximizar o alcance de detecção. Ao identificar um ativo, o nó gera um objeto JSON contendo setor e UID da tag. Esses dados são publicados via protocolo MQTT no broker HiveMQ. O servidor Node-RED atua como subscriber, processando as mensagens recebidas. As informações são armazenadas no banco MongoDB e refletidas em dashboard em tempo real. Essa arquitetura assegura rastreabilidade contínua e visibilidade do fluxo produtivo.

No aspecto de segurança e confiabilidade, destaca-se a necessidade de evitar credenciais hardcoded no código-fonte. Recomenda-se o uso de variáveis de ambiente e arquivos *.env* para proteção de dados sensíveis. Para mitigar perdas de dados, o MQTT opera com QoS 1, garantindo entrega com

confirmação. Em caso de falhas de rede, mensagens são reenviadas até o recebimento pelo broker. Adicionalmente, utiliza-se o recurso *Last Will and Testament* (LWT) para detecção de falhas. Em desconexões inesperadas, o broker publica alertas automáticos de status. Essa informação é integrada ao dashboard para indicar dispositivos *offline*. Tal mecanismo evita interpretações equivocadas no monitoramento. Assim, a solução assegura robustez, integridade e confiabilidade na rastreabilidade em tempo real.

4.5. Retorno da Aplicação

A Tabela 2 apresenta a estimativa detalhada para a implantação de quatro portais de leitura, cobrindo os setores estratégicos de produção, etiquetagem, câmara fria e expedição e garantir a viabilidade financeira em PMEs.

Tabela 2. Preços de hardware baseados em consultas ao site MakerHero.

Item	Descrição	Qtd.	Un. (R$)	Total (R$)
Hardware	Microcontrolador ESP com Wi-Fi/Bluetooth	4	60,00	240,00
Leitores	Módulo RFID PN532 V3	4	85,00	340,00
Etiquetas	Tags RFID Passivas 13,56 MHz (Lote com 500un.)	1	750,00	750,00
Serviços	Programação de sistema (Node-Red/MongoDB Compass/HiveMQ/IDE-Arduino)	1	4.500,00	4.500,00
Infraestrutura	Fontes e cabeamento de rede	4	80,00	320,00
Pontos de Função (APF)	Desenvolvimento do Software em sua complexidade	25	180,00	4.500,00
TOTAL				**10.650,00**

A viabilidade do sistema ColdChain-ID fundamenta-se no equilíbrio entre baixo custo de hardware e alto valor agregado do software. O modelo proposto reduz a dependência de registros manuais. Com isso, aumenta-se a confiabilidade e a rastreabilidade dos dados operacionais. Do ponto de vista de desempenho, Rao (2025) indica a redução de 14,2% no tempo de trânsito

e 11,7% nas perdas de estoque. Para ativos sensíveis, o payback pode ser alcançado em cerca de 9,7 meses.

5. Avaliação experimental

5.1. Planejamento de pesquisa

Este estudo segue a abordagem de Design Science Research (DSR), focando na criação e avaliação de artefatos tecnológicos para resolver problemas organizacionais identificados. A evolução do projeto partiu de um protótipo funcional em ambiente controlado (Proof of Concept) para um sistema robusto testado em cenário real (*in-situ*). Cada iteração do ciclo de design permitiu o refinamento dos algoritmos de detecção e a otimização do consumo energético dos nós sensores, culminando na solução apresentada.

A avaliação experimental do sistema proposto foi estruturada em duas fases complementares, visando validar a eficácia da arquitetura ColdChain-ID (disponível em https://github.com/kellyaquino22/ColdChain-ID) antes de sua implementação física: Fase I - Validação em Ambiente de Simulação Virtual; e Fase II - Prototipagem em Maquete Física.

5.1.1. Validação em Ambiente de Simulação Virtual (Wokwi)

A etapa de prototipagem virtual no simulador Wokwi (Anexo A) foi essencial para validar a arquitetura do ColdChain-ID antes da montagem física. Esta fase permitiu testar a conectividade SPI entre o ESP32 e o leitor PN532, garantindo a integridade dos sinais e da alimentação. A simulação focou na depuração da FSM, verificando as transições lógicas disparadas pelas etiquetas RFID. Além disso, validou-se a estruturação de dados em formato JSON para o protocolo MQTT, assegurando a compatibilidade com o servidor central. A abordagem mitigou riscos de danos ao hardware real e reduziu custos ao identificar falhas lógicas em ambiente controlado.

5.1.2. Prototipação do sistema ColdChain-ID

Após a validação lógica em ambiente virtual, procedeu-se à montagem do protótipo físico (maquete funcional) para validar a

comunicação em tempo real e a persistência de dados. **A Infraestrutura de Hardware e Conectividade:** para a identificação dos ativos, foram utilizadas tags passivas de 13,56 MHz fixadas em simulacros de unidades logísticas, conforme apresentado na Figura 2.

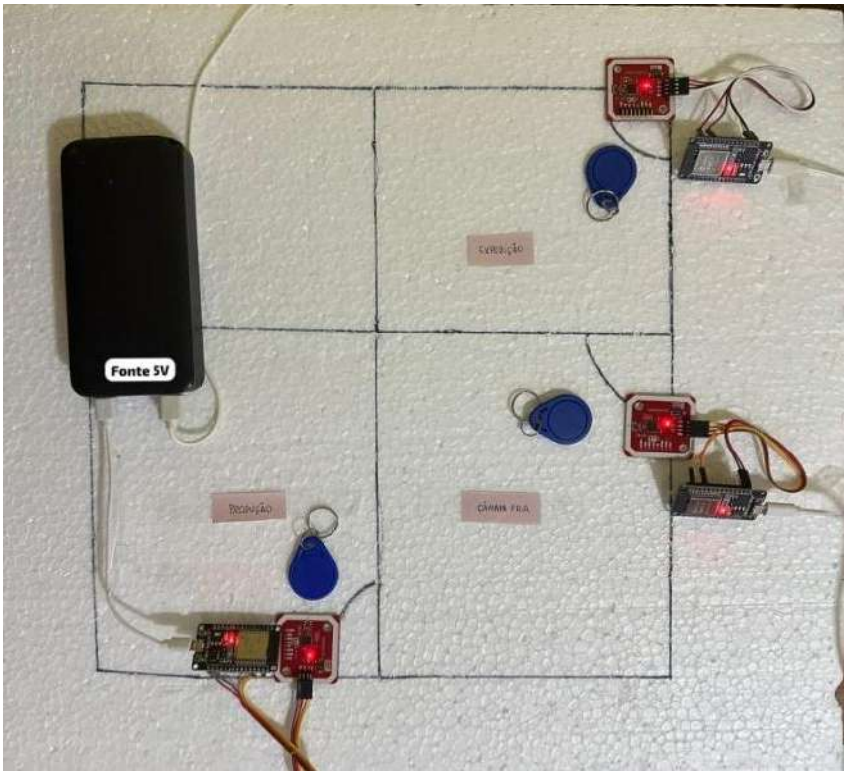


Figura 2. Protótipo funcional do sistema ColdChain-ID integrando módulos ESP32 e leitores RFID PN532.

Camada de Comunicação e Mensageria (MQTT): A integração entre o hardware e o software foi viabilizada pelo protocolo MQTT, utilizando o *broker* HiveMQ Cloud. O *firmware*, desenvolvido em C++ (Arduino IDE), foi configurado para conectar-se à rede local e ao servidor remoto do HiveMQ. Essas mensagens são publicadas no tópico *coldchain/rfid* (Figura 3), permitindo o monitoramento assíncrono de múltiplos portais simultaneamente.

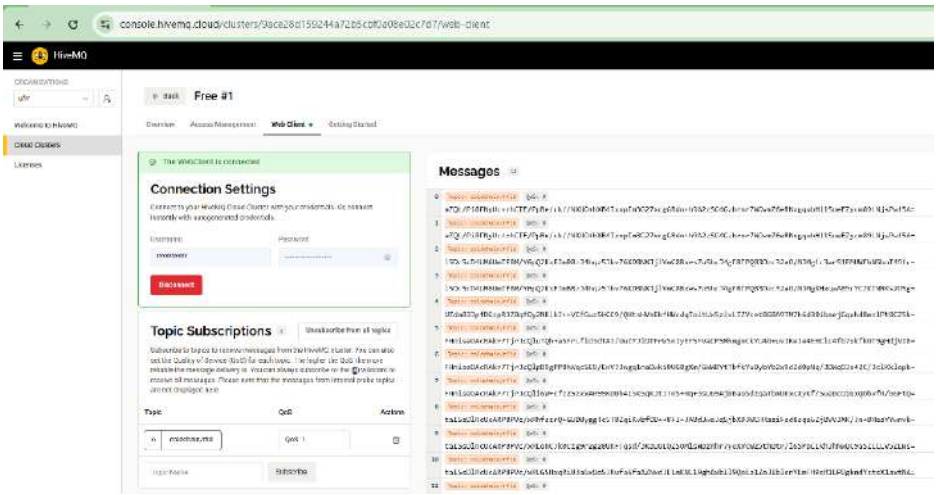


Figura 3. Monitoramento de mensagens em tempo real no console do broker HiveMQ Cloud.

Orquestração de Dados e Dashboard (Node-RED): O fluxo de dados é gerenciado pelo Node-RED (Figura 4), responsável pelas regras de negócio na camada de aplicação. O processamento ocorre por meio de um nó MQTT que se inscreve em tópicos do HiveMQ, recebendo leituras em tempo real. Os dados são persistidos em um banco MongoDB, garantindo registro histórico e auditoria das movimentações. Além disso, foi desenvolvido um dashboard (Figura 5) interativo com visualização dinâmica de ID, setor e data/hora.

Resultados da Integração Física: A validação experimental por meio de maquete física evidenciou a estabilidade do ecossistema proposto, com fluxo contínuo de mensagens JSON entre o Arduino IDE e o *broker* MQTT (HiveMQ), indicando baixa latência entre a leitura física e o registro no banco de dados (Anexo B). Os testes em condições reais permitiram analisar aspectos como resiliência, tempo de resposta e limitações operacionais. Em cenários de falha de rede, como interrupções no Wi-Fi ou indisponibilidade do broker, o sistema demonstrou robustez ao preservar os dados localmente na memória *flash* via LittleFS, garantindo integridade e posterior sincronização automática. As medições indicaram latência de dois a três segundos entre a leitura da tag e sua visualização no dashboard do Node-

RED, atendendo aos requisitos de tempo real. Entretanto, identificou-se uma limitação no alcance de leitura do módulo PN532, restrito a aproximadamente 2 cm. Esse comportamento está associado à atenuação do sinal em ambientes com alta presença de água e gordura, como na indústria de sorvetes, impactando a propagação em 13,56 MHz.

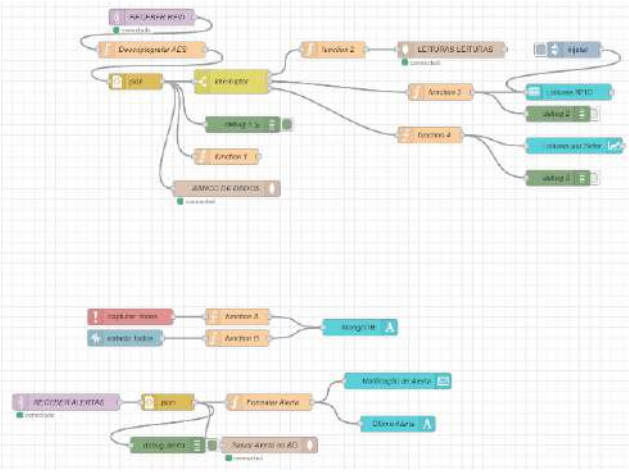


Figura 4. Parte do fluxo no Node-RED.



Figura 5. Dashboard interativo do inventário em tempo real.

5.2. Resultados e Discussão

Para a validação experimental do sistema ColdChain-ID, foi implementada uma Bancada de Testes (Testbed) automatizada, composta por uma arquitetura de três camadas: Edge (ESP32), Middleware (Node-RED) e Persistence (MongoDB).

A solução proposta implementa um mecanismo de coleta e persistência de logs de telemetria baseado em uma FSM, permitindo o controle do fluxo de dados em tempo real na cadeia do frio. Os logs de inventário registram leituras válidas que seguem a sequência correta do processo (Produção → Câmara Fria → Expedição), sendo armazenados na coleção LEITURAS do MongoDB com timestamps precisos. Já os logs de exceção capturam tentativas de quebra de sequência, acionando o estado SEQUENCIA_ERRO e registrando os eventos na coleção ALERTAS (ANEXO D). Essa separação possibilita rastreabilidade detalhada e auditoria de falhas operacionais. Tal abordagem atua na prevenção de erros e desvios no processo produtivo. A solução contribui para mitigar riscos à saúde pública e reduzir perdas financeiras.

A eficiência da camada física foi validada através da análise do posicionamento estratégico da antena. Considerando a zona de Fresnel e os efeitos de multicaminho (*multipath fading*) inerentes ao ambiente fabril, a antena foi posicionada com ganho de 2.0 dBi, visando maximizar a *Signal-to-Noise Ratio* (SNR). O fluxo de propagação de sinal foi mapeado para mitigar zonas de sombra causadas por maquinário metálico denso, garantindo que o *Received Signal Strength Indicator* (RSSI) permanecesse dentro da margem de estabilidade operacional, evitando retransmissões excessivas que degradariam o *throughput* da rede.

Para garantir a estabilidade operacional em cenários de leitura contínua, a solução implementa um mecanismo de *debounce* via software com limiar de 5.000 ms. Essa abordagem visa reduzir a saturação do broker MQTT e evitar registros redundantes no banco de dados. A implementação baseia-se na função não bloqueante *millis()*, assegurando precisão temporal sem comprometer o fluxo do firmware. Dessa forma, cada mudança de estado é tratada de maneira eficiente e controlada. O método contribui diretamente para a integridade e consistência dos dados coletados. Testes empíricos

evidenciaram o comportamento esperado do sistema em condições contínuas de leitura.

5.2.1. Mecanismos de Auditoria e Resiliência

Os resultados da solução ColdChain-ID evidenciam a implementação de um mecanismo robusto de persistência de dados por meio do Node-RED. Após o processamento e a descriptografia das mensagens via MQTT, os dados são estruturados no formato CSV e armazenados localmente em modo incremental (*append*), garantindo a manutenção de um histórico contínuo das leituras RFID. Essa abordagem assegura a integridade e rastreabilidade das informações, mesmo em cenários de falha na comunicação com o banco de dados MongoDB.

Visando a confiabilidade e integridade dos dados por meio de uma infraestrutura de monitoramento a solução adota múltiplas camadas. A arquitetura adota três níveis independentes de persistência, permitindo validação cruzada das informações e maior precisão da telemetria. Um auditor externo em Python, baseado em MQTT seguro (TLS/SSL), atua como *sniffer* passivo, registrando eventos em tempo real e possibilitando a análise de latência e perdas de pacotes. A estratégia de *Logging* de Auditoria integra dados de um banco NoSQL (MongoDB), arquivos CSV via Node-RED e *logs* brutos do auditor. Essa triangulação assegura rastreabilidade completa e suporte à análise forense de falhas. O sistema permite identificar com precisão onde e quando ocorrem inconsistências na cadeia do frio. Como resultado, a solução evolui de um protótipo para um sistema industrial auditável, robusto e alinhado aos princípios da Indústria 4.0.

5.2.2. Análise Formal do Fluxo de Controle

Para garantir que o ColdChain-ID seja um sistema de missão crítica resiliente a falhas lógicas e violações de sequência, foi modelado as transições da FSM (Figura 6) implementada no firmware dos dispositivos de borda, provando que o fluxo de dados respeita invariavelmente as regras de negócio estabelecidas.

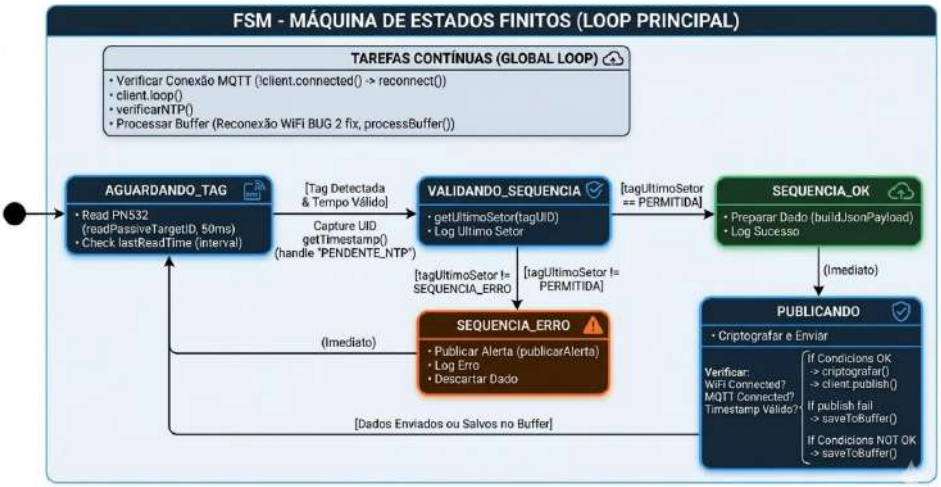


Figura 6. Fluxograma da FSM do ColdChain.

Definição das Proposições Atômicas: *T*: Uma Tag RFID foi detectada e o tempo de intervalo é válido. *V*: A sequência do setor é válida (Setor Anterior == Permitido); *P*: O payload foi processado; *A*: O alerta de erro de sequência foi emitido; e E_n: O sistema está no Estado *n* (onde *n* varia de 1 a 5). **Formalização das Transições de Estado:** Início da Leitura: $(E_1 \wedge T) \Longrightarrow E_2$ *(*Se estou aguardando e uma tag válida aparece, então obrigatoriamente vou para a validação*);* Sucesso na Sequência: $(E_2 \wedge V) \Longrightarrow E_3$ *(*Se estou validando e a sequência é correta, então vou para o estado de OK*);* Falha na Sequência: $(E_2 \wedge \neg V) \Longrightarrow E_4$ *(*Se estou validando e a sequência é incorreta, então vou para o estado de Erro*);* Preparação de Envio: $E_3 \Longrightarrow E_5$ *(*Se a sequência está OK, a transição para publicação é imediata*);* Finalização de Ciclo (Sucesso): $(E_5 \wedge P) \Longrightarrow E_1$ *(*Se publiquei ou salvei no buffer, volto a aguardar uma nova tag*); e* Finalização de Ciclo (Erro): $(E_4 \wedge A) \Longrightarrow E_1$ *(*Se o alerta foi emitido, volto a aguardar uma nova tag*).*

Validação de Propriedades de Segurança (Safety): "Um dado nunca será publicado no banco de inventário se a sequência for inválida". A validação da solução foi realizada por meio de verificação formal utilizando a biblioteca Z3 Solver em Python, integrando as regras da FSM ao motor de prova (Anexo F). Inicialmente, o modelo retornou SAT, revelando uma inconsistência na validação do estado E_3 sem dependência do estado

anterior. Para corrigir essa falha, foi inserida a precondição lógica $E_3 \Longrightarrow (E_2 \wedge V)$, garantindo a coerência das transições. Após o ajuste, o veredito passou a ser UNSAT, comprovando a inexistência de inconsistências no modelo. Esse resultado demonstra que o sistema é imune a falhas de *bypass* de sequência. Diferentemente de testes convencionais, a verificação formal garante cobertura completa dos estados possíveis. Assim, assegura-se que o registro de dados ocorra apenas sob condições válidas, reforçando a confiabilidade da rastreabilidade.

5.2.3. Monitoramento Temporal e Detecção de Gargalos no Fluxo Logístico

Foi implementado no Node-RED um fluxo denominado "Alerta de Gargalo" para monitoramento logístico na cadeia do frio. A solução rastreia temporalmente etiquetas RFID entre os setores de **Câmara Fria** e **Expedição**. Os dados são armazenados em memória via contexto global, registrando entradas e saídas dos itens. Um mecanismo periódico verifica se o tempo de permanência excede o limite de duas horas. Quando detectado, o sistema gera alertas via MQTT, dashboard e registro em log, garantindo rastreabilidade. Adicionalmente, o sistema utiliza o protocolo SMTP para envio automático de notificações por e-mail. Essa funcionalidade assegura que gestores recebam alertas críticos em tempo real. A comunicação ocorre independentemente do monitoramento ativo do dashboard. Com isso, reduz-se o tempo médio de resposta a eventos logísticos críticos. A abordagem contribui para mitigar perdas e preservar a integridade térmica dos produtos.

5.2.4. Implementação de Atualização Remota via Protocolo OTA

A arquitetura do sistema ColdChain-ID demonstrou robustez ao operar em ambientes industriais de alta complexidade, como câmaras de congelamento a -20°C, minimizando riscos operacionais e impactos térmicos decorrentes de intervenções físicas. Para superar limitações de manutenção, foi implementado o mecanismo de atualização OTA (Over-the-Air) em dispositivos ESP32, permitindo reconfiguração remota via rede Wi-Fi. A segurança foi garantida por identificação única dos nós e autenticação por

senha no processo de atualização. Como resultado, obteve-se maior confiabilidade, facilidade de manutenção e redução de interrupções operacionais. Assim, a solução alinha-se aos princípios da Indústria 4.0 ao promover escalabilidade e gestão centralizada eficiente.

5.2.5. Limitações e Trabalhos Futuros

Os resultados evidenciam a viabilidade da solução, embora sua aplicação em escala industrial enfrente desafios inerentes ao ambiente de congelamento, como alta umidade e materiais que atenuam sinais de radiofrequência. Para mitigar esses efeitos, recomenda-se o uso de etiquetas RFID robustas, encapsuladas e adequadas a temperaturas inferiores a −20 °C. Adicionalmente, a proteção dos dispositivos embarcados com invólucros IP66/IP67 é essencial para garantir durabilidade e confiabilidade. O desempenho de leitura pode ser aprimorado com ajuste de ganho e uso de antenas de polarização circular, reduzindo zonas mortas. O sistema também incorpora regras de negócio para monitoramento de restrições operacionais. Dessa forma, violações na cadeia do frio são detectadas automaticamente, com geração de alertas em tempo real.

Embora os resultados demonstrem eficácia, o sistema apresenta limitações inerentes à autonomia energética em regime de amostragem contínua. Condições ambientais extremas, como picos de temperatura e umidade, podem influenciar a deriva térmica dos sensores semicondutores. Além disso, a escalabilidade para além de 10 dispositivos simultâneos pode introduzir *jitter* na rede sob o protocolo atual, exigindo futuras investigações sobre técnicas de *sharding* de rede mais agressivas.

Ademais, ressalta-se que a arquitetura distribuída e o modelo de inventário inteligente propostos no sistema ColdChain-ID possuem elevado potencial de escalabilidade, não se restringindo exclusivamente ao setor de sorvetes. A lógica de localização por proximidade fundamentada em Máquinas de Estados Finitos (FSM) e a infraestrutura de monitoramento em tempo real podem ser transpostas para diversos outros segmentos da cadeia do frio. Trabalhos futuros podem explorar a aplicação desta tecnologia em centros de distribuição farmacêuticos para o controle de vacinas e insumos termossensíveis, bem como na logística de proteínas e hortifrutis de alto

valor agregado. Tal expansão permitiria validar a robustez do sistema frente a diferentes constantes dielétricas de materiais e variados requisitos de conformidade técnica, consolidando o uso de tecnologias de baixo custo como ferramentas universais para a garantia da integridade e rastreabilidade em ambientes industriais de temperatura controlada.

6. Conclusão

O trabalho apresentou o sistema ColdChain-ID, uma arquitetura Edge-Server baseada na integração entre RFID passivo (13,56 MHz) e IoT para inventário em tempo real. A solução demonstrou viabilidade técnica e econômica ao empregar ESP32, leitores PN532 e algoritmos de localização por proximidade modelados por FSM. Os resultados evidenciaram automação eficiente do fluxo produtivo, segurança na comunicação via criptografia AES-128 e resiliência operacional com armazenamento local. A avaliação experimental indicou latência de 2 a 3 segundos, adequada a aplicações industriais, além de consistência na integração entre as camadas Edge e Server. A verificação formal com Z3 Solver comprovou propriedades de segurança, garantindo integridade lógica do sistema. Por fim, apesar de limitações no alcance RFID em ambientes congelados, a solução mostrou-se robusta, acessível e alinhada aos princípios da Indústria 4.0.

Em suma, o sistema proposto demonstra-se um catalisador para a maturidade digital da Indústria 4.0. O impacto futuro reside na transição para redes **6G-enabled IIoT** e na integração de inteligência artificial na borda (**TinyML**), permitindo que o próprio hardware tome decisões autônomas sem dependência exclusiva da infraestrutura de nuvem, consolidando a visão de fábricas inteligentes e resilientes.

7. Referências

Ballou, R. H. (2006). *Gerenciamento da Cadeia de Suprimentos:* Logística Empresarial. Porto Alegre: Bookman, 2006.

Bouet, M., Santos, A. L. (2008). RFID tags: Positioning principles and localization techniques. *IEEE Conference Publication (pp. 1-5)*. IEEE Dubai,

United Arab Emirates. Recuperado de: https://10.1109/WD.2008.4812905

Emenike, C. C., Van Eyk, N. P. & Hoffman, A. J. (2016). Improving Cold Chain Logistics through RFID temperature sensing and Predictive Modelling. *IEEE 19th International Conference on Intelligent Transportation Systems* (ITSC), Rio de Janeiro, Brazil, (pp. 2331-2338). https://doi.org/10.1109/ITSC.2016.7795932.

Ferdousmou, J., Prabha, M., Farouk, MD O., Samiun, MD, Sozib, H. M. & Zaman, Al M. (2024). IoT-Enabled RFID in Supply Chain Management: A Comprehensive Survey and Future Directions. *Journal of Computer and Communications*, *12*(11), 207-223. https://doi.org/10.4236/jcc.2024.1211015

Hao J., & Lin, S. (2011). State machine based indoor localization and tracking system using RFID. In: *International Conference on Computer and Communications Management (ICCCM)*.

Hopcroft, J. E., Motwani, R., & Ullman, J. D. (2002). *Introdução à teoria de autômatos, linguagens e computação*. Porto Alegre: Bookman.

Hopcroft, J. E., Motwani, R., Ullman, Jeffrey D. (2006). *Introduction to Automata Theory, Languages, and Computation*. Boston: Addison-Wesley, 2006.

Jin, G., Lu, X., Park, M. (2006). An indoor localization mechanism using active RFID tag. *In: IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC'06), IEEE, (pp. 40-45)*. Recuperado de: https://10.1109/SUTC.2006.1636157

Masoudi, B., Razi, N., & Rezazadeh, J. (2025). IoT-Enabled Indoor Real-Time Tracking Using UWB for Smart Warehouse Management. *Computers*, *14*(12). https://doi.org/10.3390/computers14120510

Onoriode, U. & Kotonya, G. (2018). IoT Architectural Framework: Connection and Integration Framework for IoT Systems. *Electronic Proceedings in Theoretical Computer Science*, *264*(4), 1-17. Recuperado de: https://10.4204/EPTCS.264.1

Rao, P. (2026). Real-time visibility: IoT revolutionizes inventory in transit tracking. *European Journal of Computer Science and Information Technology*, *13*(23), 60-75. https://doi.org/10.37745/ejcsit.2013

Sousa, L., Rocha, Rocha, H., Ribeiro, A. & Lobo, F. (2020). INEXT: A Computer System for Indoor Object Location using RFID. *Conference: Anais Estendidos do Simpósio Brasileiro de Engenharia de Sistemas Computacionais.* Departamento de Ciência da Computação, Universidade Federal de Roraima, 2020. Recuperado de: https://10.5753/sbesc_estendido.2023.235021

Wu, Wei, Shen, L., Zhao, Z., Harish, A. R., Zhong, R. Y. & Huang, G. Q. (2023). Internet of Everything and Digital Twin enabled Service Platform for Cold Chain Logistics. *Journal of Industrial Information Integration, 33, 100443.* https://doi.org/10.1016/j.jii.2023.100443

Xu, Z., Elomri, A., Baldacci, Kerbache, L. & Wu, Z. (2024). Frontiers and trends of supply chain optimization in the age of industry 4.0: an operations research perspective. *Annals of Operations Research*, *338*, 1359–1401. https://doi.org/10.1007/s10479-024-05879-9

Yan, B., & Lee, D. (2009). Application of RFID in cold chain temperature monitoring system. *IEEE Conference Publication (pp. 258-261)*. https://doi.org/10.1109/CCCM.2009.5270408

Capítulo 24

Estruturação de registros em ordens de serviço de manutenção com metamodelo de linguagem e embeddings de sentenças aplicado à Indústria 4.0

Maristela Veiga dos Santos
Clenya Rejane Barros Lima
Universidade Federal de Roraima

veiga.maristela@gmail.com
clenyabarros@uerr.edu.br

1. Introdução

A Indústria 4.0 ampliou o uso de Inteligência Artificial (IA), automação e análise de dados em processos industriais, especialmente em atividades de monitoramento, triagem e manutenção. Nesse contexto, técnicas de Processamento de Linguagem Natural (PLN) são aplicadas à análise de registros textuais não estruturados, como relatórios de manutenção, descrições de falhas e ordens de serviço (OS), com potencial para apoiar diagnóstico, rastreabilidade, confiabilidade e tomada de decisão baseada em evidências (Sundaram & Zeid, 2025; Zhong et al., 2024).

O Enterprise Asset Management (EAM) é um software de gestão de ativos ao longo do ciclo de vida, enquanto o Computerized Maintenance Management System (CMMS) corresponde ao sistema voltado às rotinas de manutenção, como ordens de serviço, manutenção preventiva e controle de intervenções (Gallagher, n.d.).

Conte et al. (2021) mostram que problemas de qualidade em work orders reduzem a confiabilidade de indicadores e análises, enquanto Sundaram e Zeid (2025) e Zhong et al. (2024) destacam que textos de manutenção tendem a ser curtos, semiestruturados e marcados por terminologia técnica, abreviações e ambiguidades.

O Metamodelo de Linguagem da PNL esclarece conteúdo linguístico, identificando omissões, reduzindo generalizações e explicitando

ambiguidades (Augusto & Travaglini, 2015), sendo útil na manutenção industrial para tornar textos livres em registros específicos e acionáveis.

Amaliah et al. (2025) reforçam, em revisão de 63 artigos, a relevância de abordagens com intervenção humana em contextos críticos para interpretabilidade e alinhamento operacional.

No entanto há uma lacuna na interseção entre manutenção industrial, metamodelos linguísticos e PLN com embeddings: faltam mecanismos explícitos para elicitar dados faltantes, padronizar textos livres e controlar preservação de sentido, reduzindo distorções (Sundaram & Zeid, 2025; Zhong et al., 2024).

Assim, o objetivo geral deste estudo é melhorar a completude e a acionabilidade de comentários técnicos de manutenção registrados como texto livre, por meio de um Produto Mínimo Viável (MVP) que estrutura o registro com base em princípios do Metamodelo de Linguagem da Programação Neurolinguística (PNL) e valida a preservação de sentido com técnicas de embeddings de sentenças.

Para atingir esse objetivo, o trabalho busca responder às seguintes perguntas de pesquisa: (i) quais padrões associados a omissões, generalizações e distorções são mais recorrentes em comentários técnicos de OS; (ii) em que medida a estruturação orientada pelo metamodelo altera métricas operacionais de completude e acionabilidade do registro; (iii) se a versão estruturada mantém equivalência semântica em relação ao texto original, quando avaliada por embeddings e similaridade cosseno; e (iv) quais implicações esses resultados trazem para iniciativas de IA aplicadas à Manutenção 4.0, especialmente em processos de triagem, rastreabilidade e qualidade de dados textuais.

Para responder a essas questões, adota-se uma abordagem de pesquisa aplicada e quantitativa, baseada no desenvolvimento e avaliação de um artefato computacional (MVP). A metodologia envolve o uso de um dataset de comentários técnicos de manutenção e a geração de uma versão estruturada orientada por perguntas do metamodelo, mantendo humano no processo para o preenchimento de informações que não devem ser inferidas automaticamente. A avaliação comparativa utiliza métricas internas de

completude e acionabilidade e um controle semântico por embeddings, como forma de verificar a preservação do sentido entre as versões.

Este trabalho está organizado da seguinte forma: a Seção 2 apresenta a fundamentação teórica; a Seção 3 discute os trabalhos relacionados; a Seção 4 descreve a solução computacional proposta; a Seção 5 apresenta a avaliação experimental e a discussão dos resultados; a Seção 6 expõe as conclusões e direções para trabalhos futuros; e a Seção 7 reúne as referências utilizadas no estudo.

2. Fundamentação teórica

Esta seção apresenta os principais conceitos, teorias e evidências científicas que fundamentam o presente estudo, com foco na Manutenção 4.0 no contexto da Indústria 4.0, na qualidade de registros textuais operacionais em CMMS/EAM (ordens de serviço e logs de manutenção), no Metamodelo de Linguagem da Programação Neurolinguística (PNL), nas capacidades e limitações do Processamento de Linguagem Natural (PLN) e nos desafios contemporâneos da análise automática de textos operacionais curtos. O objetivo é estabelecer o arcabouço teórico que sustenta a proposta de estruturação linguística de comentários técnicos e sua avaliação por métricas operacionais e controle semântico com embeddings (Conte et al., 2021; Zhong et al., 2024).

2.1. Os CMMS/EAM e registros textuais operacionais

No contexto da Indústria 4.0, sistemas CMMS/EAM permanecem centrais por concentrarem ordens de serviço (OS), histórico de intervenções e observações técnicas, muitas vezes registradas como texto livre. Entretanto, estudos apontam que registros históricos de OS podem apresentar lacunas informacionais, inconsistências e ruído, o que compromete a interpretação humana e reduz a efetividade de análises e soluções baseadas em IA (Conte et al., 2021; Zhong et al., 2024).

2.2. Processamento de Linguagem Natural aplicado a ordens de serviço e logs de manutenção

Técnicas de Processamento de Linguagem Natural (PLN) vêm sendo aplicadas à manutenção industrial para extrair valor de descrições de falhas, relatos de inspeção e notas em OS, incluindo classificação, extração de informação, mineração de tópicos e análise de similaridade semântica para agrupamento e triagem de ocorrências (Sundaram & Zeid, 2025; Zhong et al., 2024). Contudo, a robustez desses métodos depende fortemente da qualidade e estrutura do texto, e limitações ligadas à heterogeneidade do domínio, escassez de rótulos e dependência de contexto reduzem a generalização quando o texto livre é a principal fonte de evidência (Conte et al., 2021; Zhong et al., 2024).

2.3. Omissões, vagueza e ruído: desafios para decisões do PLN em manutenção

Comentários técnicos frequentemente incluem termos avaliativos (por exemplo, “instável”, “anormal”, “significativo”, “intermitente”) que podem sinalizar um problema, mas tornam-se insuficientes para ação quando não trazem condição de ocorrência, evidência mensurável, critério de comparação e validação. Isso reduz rastreabilidade e aumenta retrabalho, além de limitar a utilidade do texto para triagem e análise automatizada. Assim, estratégias de melhoria do registro (antes ou durante o processamento) elevam a prontidão do dado textual para decisões e para PLN em manutenção (Conte et al., 2021; Zhong et al., 2024).

2.4. Metamodelo de Linguagem (PNL) como estratégia de estruturação de registros

O Metamodelo de Linguagem da PNL descreve que, ao expressar experiências em palavras, é natural ocorrerem omissões, generalizações e distorções; por isso, propõe padrões e perguntas para recuperar informações e esclarecer significados (Augusto & Travaglini, 2015, pp. 99–101). Neste estudo, o metamodelo é aplicado como estratégia linguística de elicitação de dados faltantes e padronização de registros em OS/CMMS/EAM (não como

intervenção psicológica), por meio de perguntas e campos mínimos que tornam o registro mais verificável e orientado à ação. Em domínios críticos, evita-se que o sistema "invente" medições; por isso, adota-se human-in-the-loop para confiabilidade e rastreabilidade em manutenção (Amaliah et al., 2025).

2.5. Embeddings de sentenças, controle semântico e convergência entre PNL e PLN

Modelos baseados em arquiteturas Transformer, como BERT e T5, ampliaram as capacidades de representação semântica em PLN ao oferecer embeddings contextuais úteis para comparação, classificação e análise textual (Devlin et al., 2019; Raffel et al., 2020). Para textos curtos, uma aplicação prática é o uso de embeddings de sentenças com similaridade cosseno, permitindo estimar proximidade de sentido entre duas versões de um mesmo registro.

Em domínio industrial, entretanto, textos operacionais tendem a ser fragmentados, pouco descritivos e fortemente dependentes de contexto local. Por isso, neste estudo, embeddings de sentenças são utilizados como mecanismo de controle semântico da reestruturação textual, verificando se a versão estruturada preserva o sentido do comentário original e reduz o risco de distorção durante a reescrita. Essa convergência entre metamodelo linguístico e PLN constitui o núcleo metodológico da proposta. Esses conceitos embasam a análise dos estudos abaixo, que revelam lacunas na estrutura textual.

3. Trabalhos relacionados

Sundaram e Zeid (2025) mostram que ordens de serviço de manutenção constituem fonte relevante de informação operacional e que técnicas como similaridade textual e modelagem de tópicos podem apoiar triagem e identificação de padrões. Zhong et al. (2024), em revisão sistemática sobre PLN aplicado à manutenção industrial, destacam desafios recorrentes ligados à variabilidade de domínio, à heterogeneidade e à qualidade do dado textual. Conte et al. (2021) evidenciam que lacunas,

inconsistências e ruídos em ordens de serviço comprometem análises posteriores e reduzem a utilidade prática do dado. Bouabdallaoui et al. (2020) demonstram a viabilidade de automatizar classificação e organização de solicitações de manutenção, embora permaneça a dependência da qualidade do texto livre. Amaliah et al. (2025), por sua vez, reforçam a importância de abordagens human-in-the-loop em contextos críticos. A Tabela 1 sintetiza os principais trabalhos relacionados, destacando seus objetivos, técnicas empregadas, resultados alcançados e a lacuna que este TCC busca explorar.

Em conjunto, os estudos mostram a relevância do PLN e da qualidade textual para a manutenção, mas indicam que ainda são menos frequentes abordagens que combinem elicitação de dados faltantes, padronização operacional do texto livre e controle semântico da reestruturação, lacuna que este trabalho busca explorar.

Tabela 1. Quadro comparativo dos trabalhos relacionados.

Trabalho (autor, ano)	Objetivo do estudo	Técnicas/Ferramentas	Resultados principais	Lacuna identificada
Sundaram e Zeid (2025)	Aplicar processamento de linguagem técnica a ordens de serviço de manutenção para apoiar triagem e identificação de padrões.	Similaridade textual e modelagem de tópicos em *work orders*.	Mostram que OS são fonte relevante de informação operacional e podem apoiar agrupamento e triagem.	Não propõem mecanismo explícito de elicitação de dados faltantes nem geração de rascunho estruturado de OS.
Zhong et al. (2024)	Revisar sistematicamente abordagens de PLN aplicadas à manutenção industrial.	Revisão sistemática da literatura em PLN para manutenção.	Evidenciam desafios recorrentes de variabilidade de domínio, heterogeneidade e qualidade textual em registros curtos.	Mapeiam o campo, mas não apresentam solução aplicada para estruturação linguística e controle semântico de comentários de OS.
Conte et al. (2021)	Investigar o impacto da qualidade dos dados em ordens de serviço históricas de	Estudo de caso com análise de qualidade de *work orders* de HVAC.	Mostram que lacunas, inconsistências e ruídos comprometem análises posteriores e	Evidenciam o problema, mas não propõem artefato computacional para melhorar completude e acionabilidade do registro textual.

Trabalho (autor, ano)	Objetivo do estudo	Técnicas/Ferramentas	Resultados principais	Lacuna identificada
	manutenção.		reduzem a utilidade prática do dado.	
Bouabdallaoui et al. (2020)	Desenvolver modelo de PLN para apoiar gestão de solicitações de manutenção em edifícios.	Modelo de PLN para classificação e organização de solicitações.	Demonstram viabilidade de automatizar organização e tratamento de solicitações de manutenção.	Focam automação/classificação, sem tratar explicitamente da padronização do texto livre com controle de preservação de sentido.
Amaliah et al. (2025)	Revisar abordagens *human-in-the-loop* e explicabilidade em manutenção preditiva.	Revisão sistemática sobre sistemas interativos e XAI em manutenção.	Reforçam a importância de evitar inferências indevidas e manter participação humana em contextos críticos.	Não tratam especificamente da estruturação de comentários técnicos de OS com metamodelo e embeddings.

O diferencial desta pesquisa em relação aos estudos correlatos reside na integração de conhecimentos da psicologia, da computação e da Indústria 4.0 para aprimorar registros textuais de manutenção. Ao contrário da maioria das obras relacionadas, que tipicamente focam na análise ou na classificação de ordens de serviço já existentes, este trabalho apresenta uma proposta de formatação textual guiada pelo Metamodelo da Linguagem e sustentada por embeddings de sentenças para garantir a preservação semântica. Desse modo, avança ao articular princípios da psicologia da linguagem com ferramentas de IA para atender a uma necessidade operacional do setor de manutenção industrial.

4. Solução computacional

Esta seção descreve a solução computacional proposta, incluindo o delineamento geral da pesquisa, o conjunto de dados utilizado, as etapas de

processamento e as métricas adotadas para avaliar a estruturação dos comentários técnicos de manutenção.

4.1. Delineamento, base de dados e estruturação dos comentários

Adotou-se um delineamento comparativo por pares, no qual cada comentário técnico original, em texto livre, é comparado à sua versão estruturada com base em princípios do Metamodelo de Linguagem (PNL). Ambas as versões são avaliadas por métricas de completude, acionabilidade e preservação semântica, de modo a isolar o efeito da estruturação textual.

Empregou-se um arquivo .xlsx contendo 100 comentários textuais e rótulos auxiliares (polaridade_ternaria e polaridade_binaria), previamente utilizados em atividades acadêmicas vinculadas a uma disciplina do curso. Embora tal base tenha sido aproveitada originalmente em outro contexto analítico, neste trabalho consideraram-se apenas os comentários de interesse técnico, interpretados como registros compatíveis com ordens de serviço de manutenção.

Na ausência de acesso à documentação formal de procedência e sem confirmação de disponibilização em repositórios públicos, adotou-se a base como um corpus acadêmico de experimentação, apropriado para demonstrar o método, mas não para ser tratado como amostra representativa de um contexto industrial específico.

O enfoque da pesquisa incide sobre a qualidade operacional do registro textual, especialmente rastreabilidade e acionabilidade. Nesse processo, o metamodelo identifica omissões, generalizações e distorções para orientar a geração de uma OS minimamente editável, preservando o texto original como “Observação” e acrescentando campos essenciais, como evidência ou critério, impacto ou risco, ação recomendada, responsável, período e validação. Esse desenho segue o princípio de abordagem com intervenção, no qual o sistema orienta o registro e solicita informações ausentes, sem inferir medições inexistentes.

4.2. Solução Computacional: MVP e pipeline de processamento

O MVP foi implementado em Python 3.12.12, com interface em Gradio 5.50.0, utilizando bibliotecas para manipulação tabular, processamento numérico, métricas de aprendizado de máquina e embeddings de sentenças, incluindo pandas 2.2.2, NumPy 2.0.2, scikit-learn 1.6.1 e sentence-transformers 5.2.3. O sistema realiza: (i) sinalização de padrões linguísticos associados a lacunas de informação; (ii) geração de perguntas para orientar a coleta de dados faltantes; (iii) geração do rascunho estruturado de OS; (iv) cálculo de métricas internas para comparação entre original e estruturado; e (v) controle semântico por embeddings de sentenças com similaridade cosseno. A Figura 1 apresenta o fluxo resumido do MVP.

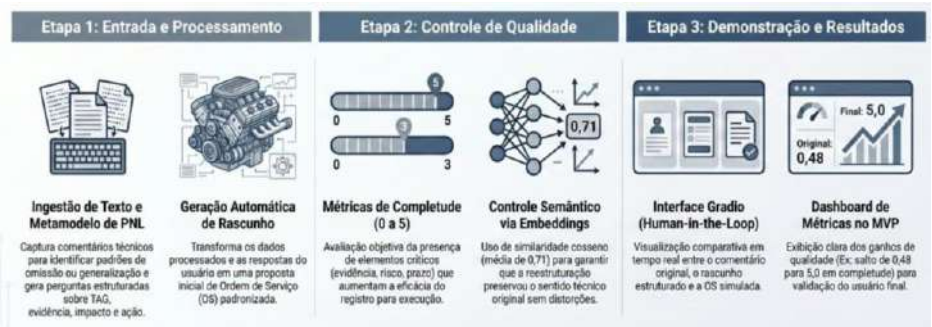


Figura 1. Pipeline do MVP para estruturação de comentários técnicos em ordens de serviço, desde o texto livre até a geração do rascunho estruturado, métricas operacionais e controle semântico. Fonte: Elaborado pela autora com auxílio de inteligência artificial (2026).
Nota. A "OS simulada" é utilizada apenas para visualização do fluxo e não representa medições reais.

4.3. Métricas, procedimentos de análise e rigor metodológico

A avaliação foi realizada por três dimensões: (a) Completude (0–5), medindo a presença de observação, evidência/critério, impacto/risco, ação e validação; (b) Acionabilidade (0–3), medindo presença de verbo de ação, prazo/urgência e validação/reteste; e (c) Preservação semântica (0–1), estimada pela similaridade cosseno entre embeddings de sentenças do texto original e do texto estruturado. A preservação semântica foi calculada por

similaridade cosseno entre os vetores semânticos do comentário original e da versão estruturada, conforme a Figura 2.

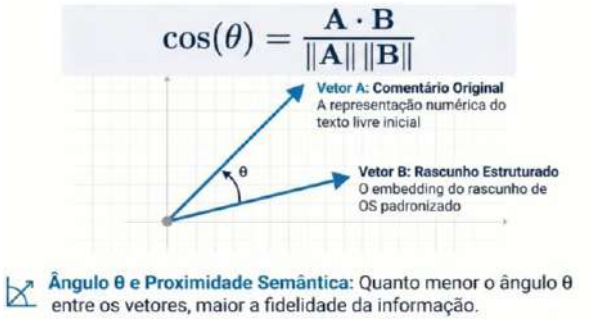


Figura 2. Equação de Controle Semântico: Similaridade de Cosseno entre embeddings de sentenças. Fonte: Elaborada pela autora com auxílio de inteligência artificial (2026).

Nota. Valores mais próximos de 1 indicam maior similaridade semântica e menor risco de distorção durante a reestruturação textual.

Para cada comentário do corpus, o MVP gerou a versão estruturada correspondente, formando pares (original vs estruturado). Em seguida, são validadas completude, acionabilidade e preservação semântica para cada par, com análise descritiva dos resultados considerando medidas de tendência central e dispersão. Exemplos qualitativos foram utilizados apenas para ilustrar o efeito da estruturação sobre a verificabilidade e a acionabilidade do registro.

O método preserva rastreabilidade entre entrada e saída, utiliza corpus replicável e métricas explícitas alinhadas ao objetivo operacional, além de adotar diretriz human-in-the-loop para evitar inferências indevidas. O controle semântico por embeddings confere transparência ao oferecer um indicador quantitativo de preservação de sentido entre as versões.

5. Avaliação experimental

5.1. Planejamento de pesquisa

A avaliação foi conduzida com um dataset contendo 100 comentários técnicos de manutenção em formato .xlsx, representativos de registros em

texto livre em ordens de serviço (OS). Com base no delineamento comparativo por pares descrito na Seção 4, as versões original e estruturada foram comparadas pelas métricas apresentadas na Tabela 2, com o objetivo de verificar se a estruturação atenua lacunas informacionais em OS, aspecto reconhecido como crítico para análises posteriores e uso de IA em manutenção (Conte et al., 2021; Zhong et al., 2024).

Tabela 2. Definição das métricas utilizadas na avaliação do MVP.

Métrica	Escala	O que mede	Interpretação (maior é melhor?)
Completude operacional	0–5	Presença de elementos mínimos do registro de OS: observação; evidência/critério; impacto/risco; ação; validação	Sim
Acionabilidade	0–3	Grau em que o texto orienta execução/fechamento: verbo de ação; prazo/urgência; validação/reteste	Sim
Similaridade semântica (embeddings de sentenças)	0–1	Preservação de sentido entre texto original e texto estruturado (similaridade cosseno)	Sim (quanto maior, maior preservação)

Nota. OS = ordem de serviço; embeddings de sentenças = representações vetoriais semânticas de textos curtos.

Além disso, adotou-se o princípio human-in-the-loop, evitando inferência automática de medições inexistentes em domínio crítico (Amaliah et al., 2025). O controle semântico por embeddings foi utilizado como verificação de preservação de sentido, prática compatível com aplicações de similaridade em work orders (Sundaram & Zeid, 2025).

5.2. Resultados e Discussão

A Tabela 3 resume os resultados principais (n = 100). Observou-se aumento da completude operacional de 0,48 para 5,00 e da acionabilidade de 0,00 para 2,00, indicando ganho objetivo de estrutura informacional e direcionamento para execução/fechamento de OS. A similaridade semântica média entre original e estruturado foi 0,71 (min = 0,51; máx = 0,83), sugerindo preservação de sentido apesar do enriquecimento estrutural, reduzindo risco de distorção na reestruturação textual (Sundaram & Zeid,

2025). Em conjunto, os resultados indicam que o MVP atenua um gargalo recorrente em Manutenção 4.0: registros textuais ruidosos (omissões, generalizações e distorções) e pouco padronizados que dificultam rastreabilidade, triagem e uso consistente de PLN/IA (Conte et al., 2021; Zhong et al., 2024). O efeito da técnica proposta pode ser observado qualitativamente na Tabela 3, que compara um comentário técnico em texto livre com sua versão estruturada gerada pelo MVP.

Tabela 3. Comparação entre comentário técnico em texto livre e sua versão estruturada gerada pelo MVP.

Etapa	Conteúdo
Comentário original	"A resposta da válvula solenoide foi lenta, gerando overshoot elevado e instabilidade no processo."
Com a proposta do MVP	Observação: resposta lenta da válvula solenoide, com overshoot elevado e instabilidade no processo. Ativo/local: [preencher]. Condição de ocorrência: [preencher]. Evidência/critério: [preencher]. Impacto/risco: [preencher]. Ação recomendada: [preencher]. Validação: [preencher].
OS simulada (demo)	OS simulada (demo): Observação: resposta lenta da válvula solenoide, com overshoot elevado e instabilidade no processo. Ativo/local: válvula solenoide VSV-12, linha de alimentação do módulo de dosagem. Condição de ocorrência: operação contínua, turno diurno, após partida da linha. Evidência/critério: tempo de resposta acima do esperado e oscilação observada no processo. Impacto/risco: instabilidade operacional e possível comprometimento da qualidade. Ação recomendada: inspecionar a válvula, verificar acionamento e condição mecânica, realizar ajuste ou substituição se necessário.

Nota. A OS simulada é utilizada apenas para demonstração do fluxo e não representa medições reais nem compõe a validação principal do método.

No exemplo, observa-se que o texto original descreve um sintoma relevante, mas não explicita condição, evidência, impacto, ação e validação. A estruturação proposta não altera o núcleo semântico da observação, mas organiza o registro em campos mínimos necessários à rastreabilidade e à execução.

A execução do planejamento experimental gerou evidências quantitativas e qualitativas de que o MVP aumenta a estrutura operacional dos registros textuais em ordens de serviço, preservando o conteúdo informacional central do comentário original. Conforme apresentado na Tabela 2, a completude operacional aumentou de média 0,48 no texto

original para 5,00 na versão estruturada. Isso indica que o rascunho produzido pelo MVP passa a contemplar, de modo sistemático, elementos mínimos necessários à rastreabilidade, como observação, evidência/critério, impacto/risco, ação e validação.



Figura 3. Fluxograma linear do MVP. Fonte: Elaborada pela autora com auxílio de inteligência artificial (2026).

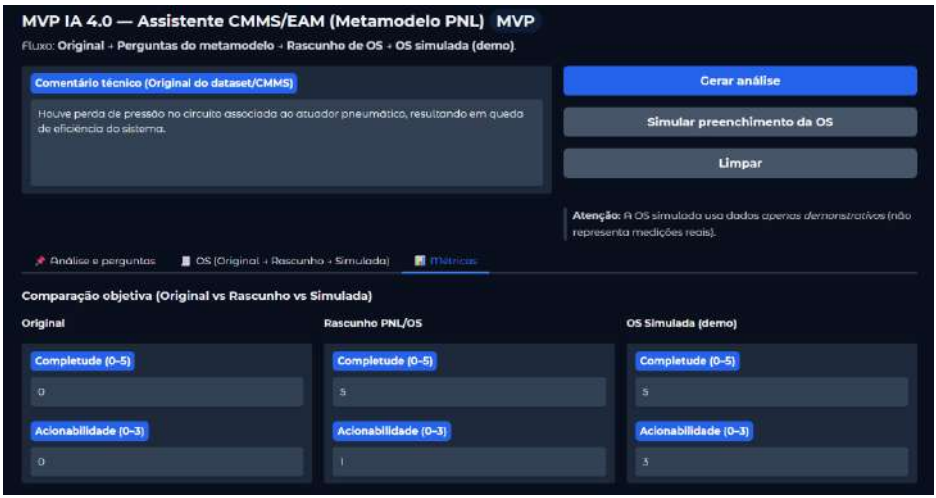


Figura 4. Printe de Tela do MVP – Aba Métricas.
Nota. https://huggingface.co/spaces/Maristela1991/mvp-ia40-cmms-eam-metamodelo-pnl

Tabela 2. Resultados principais da avaliação (n = 100).

Indicador	Original (média)	Estruturado PNL/OS (média)	Ganho
Completude operacional (0–5)	0,48	5,00	+4,52
Acionabilidade (0–3)	0,00	2,00	+2,00

Indicador	Original (média)	Estruturado PNL/OS (média)	Ganho
Similaridade semântica (0–1)	—	—	0,71 (média)*

Nota. Similaridade semântica calculada entre embeddings de sentenças do texto original e do texto estruturado (min = 0,51; máx = 0,83).

Esse resultado converge com a literatura, segundo a qual lacunas informacionais e qualidade textual insuficiente em work orders reduzem a confiabilidade de análises e decisões posteriores (Conte et al., 2021). De forma complementar, a acionabilidade média aumentou de 0,0 para 2,0, sugerindo que a estrutura proposta torna o registro mais orientado à execução e ao fechamento de OS, reduzindo dependência de comunicação adicional para completar informação. Esse achado converge com evidências de que a aplicação de PLN em work orders tende a ser mais útil quando o registro oferece informações mínimas explícitas e consistentes, favorecendo triagem e suporte à decisão (Sundaram & Zeid, 2025; Zhong et al., 2024).

Quanto ao controle semântico, a similaridade semântica média entre o texto original e o texto estruturado foi 0,71 (min = 0,51; máx = 0,83), indicando que apesar do enriquecimento estrutural, o rascunho mantém proximidade semântica com o conteúdo original. Esse resultado é importante para reduzir o risco metodológico de "distorção por reescrita" e reforça a adequação de embeddings de sentenças como mecanismo de verificação de preservação de sentido em tarefas de análise e comparação textual no domínio de manutenção (Sundaram & Zeid, 2025).

Do ponto de vista prático, os resultados sugerem que o MVP atenua um problema relevante da Manutenção 4.0: registros textuais curtos e incompletos que dificultam rastreabilidade, triagem e uso de IA. Ao adotar uma abordagem com intervenção humana, o sistema orienta a coleta de dados faltantes sem inferir medições inexistentes, característica desejável em domínios críticos e alinhada a recomendações de confiança e adoção em manutenção (Amaliah et al., 2025).

6. Conclusão

Este trabalho teve como objetivo melhorar a qualidade operacional de comentários técnicos registrados como texto livre em ordens de serviço (OS) de manutenção, tornando-os mais completos, rastreáveis e acionáveis para uso em CMMS/EAM no contexto da Indústria 4.0. Para isso, adotou-se um delineamento comparativo por pares com 100 comentários técnicos, comparando cada texto original à sua versão estruturada pelo Metamodelo de Linguagem da PNL, validado por métricas de completude, acionabilidade e similaridade semântica, através de um MVP.

Como resposta ao problema de pesquisa, o algoritmo modelado com a estrutura linguística do Metamodelo de Linguagem detectou nos comentários analisados, vaguezas operacionais associadas à ausência de evidência/critério, impacto/risco, ação recomendada e validação.

Os resultados mostraram que a estruturação orientada pelo metamodelo aumentou a completude média de 0,48 para 5,0 e a acionabilidade média de 0,0 para 2,0, indicando ganho objetivo de estrutura informacional e direcionamento para execução e fechamento de OS.

Quanto à preservação de sentido, o controle semântico por embeddings de sentenças apresentou similaridade cosseno média de 0,71 entre o texto original e a versão estruturada, sugerindo consistência semântica e mitigação de distorções durante a reestruturação. Em termos práticos, os achados indicam que o MVP pode contribuir para iniciativas de IA aplicadas à Manutenção 4.0 ao melhorar a qualidade textual para triagem, rastreabilidade e análise automatizada, sem substituir o papel do usuário na coleta de evidências.

Como limitações, destacam-se a dimensão e a origem acadêmica do dataset, a ausência de validação com especialistas e integração com CMMS/EAM em produção, sendo a OS simulada utilizada apenas para demonstração do fluxo.

Como trabalhos futuros, recomenda-se ampliar a base com registros reais anonimizados, refinar as regras do metamodelo com apoio de técnicos e engenheiros, integrar o MVP a fluxos reais via exportação/API e avaliar impacto por métricas operacionais, como tempo de triagem, retrabalho e qualidade de fechamento de OS.

7. Referências

Amaliah, N. R., Tjahjono, B., & Palade, V. (2025). Human-in-the-loop XAI for predictive maintenance: A systematic review of interactive systems and their effectiveness in maintenance decision-making. *Electronics, 14*(17), 3384. https://doi.org/10.3390/electronics14173384

Augusto, B., & Travaglini, F. (2015). *PNL e você: Desenvolva seu potencial* (3ª ed., rev. e ampl.). Belo Horizonte, MG: Pampulha.

Bouabdallaoui, Y., Lafhaj, Z., Yim, P., Ducoulombier, L., & Bennadji, B. (2020). Natural language processing model for managing maintenance requests in buildings. *Buildings, 10*(9), 160. https://doi.org/10.3390/buildings10090160

Conte, A., Bolland, C., Phan, L., Brundage, M., & Sexton, T. (2021). The impact of data quality on maintenance work order analysis: A case study in historical HVAC maintenance work orders. In *Proceedings of the 6th European Conference of the Prognostics and Health Management Society (PHM 2021).* Recuperado de: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=932288

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186)*. https://doi.org/10.18653/v1/N19-1423

Gallagher, N. (n.d.). *O que é gerenciamento de ativos corporativos (EAM)?* IBM Think. Recuperado de: https://www.ibm.com/br-pt/think/topics/enterprise-asset-management

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67. Recuperado de: https://www.jmlr.org/papers/v21/20-074.html

Sundaram, S., & Zeid, A. (2025). Technical language processing for prognostics and health management: Applying text similarity and topic modeling to maintenance work orders. *Journal of Intelligent*

Manufacturing, 36(3), 1637–1657. https://doi.org/10.1007/s10845-024-02323-4

Zhong, K., Jackson, T., West, A., & Cosma, G. (2024). Natural language processing approaches in industrial maintenance: A systematic literature review. *Procedia Computer Science, 232*, 2082–2097. https://doi.org/10.1016/j.procs.2024.02.029

Capítulo 25

Aplicação de Grandes Modelos de Linguagem na triagem de currículos não padronizados

Adolfo Izidio Ferreira da Silva-Neto
Susset Guerra Jimenez
Universidade Federal de Roraima

aadolfoneto@gmail.com
susset.guerra@ufrr.com

1. Introdução

A triagem inicial de currículos é uma etapa crítica do recrutamento e seleção e, em vagas operacionais, tende a ser pressionada por alto volume, baixa padronização documental e subjetividade na avaliação humana. No Brasil, a rotatividade no emprego formal é expressiva e impacta diretamente a necessidade de reposição e seleção contínuas; o DIEESE reporta taxas de 62,8% em 2014, 54,8% em 2015 e 48,2% em 2016 no segmento celetista (DIEESE, 2017). Nesse cenário, a automação de processos e o uso de dados para apoiar decisões, princípios associados à Indústria 4.0, tornam-se relevantes também para atividades administrativas, como a triagem de documentos de candidatos. Entretanto, currículos operacionais frequentemente chegam como PDFs desestruturados e imagens capturadas por dispositivos móveis, com variações de formato e qualidade que dificultam o uso de sistemas tradicionais de rastreamento de candidatos (Gan et al., 2024).

Grandes Modelos de Linguagem (LLMs) ampliaram a viabilidade de automatizar tarefas de Processamento de Linguagem Natural, como extração e classificação de informações em textos não estruturados. Apesar disso, a maior parte das soluções e estudos concentra-se em currículos mais estruturados, típicos de setores corporativos e tecnológicos. Embora haja evidências de ganhos de eficiência e de desempenho na triagem com LLMs, ainda é necessária a verificação de validade, robustez e adequação ao contexto de uso (Lo et al., 2025; Gan et al., 2024). Assim, permanece em

aberto a eficácia desses modelos para interpretar currículos operacionais, caracterizados por forte heterogeneidade textual e documental.

Para enfrentar essa lacuna, este trabalho apresenta a plataforma computacional RecruitAI, voltada à extração de informações e à classificação multirrótulo de currículos não padronizados. O sistema adota triagem cega na etapa inicial, ocultando dados pessoais e atributos sensíveis para priorizar critérios técnicos e reduzir a influência de vieses na tomada de decisão (Vivek, 2022; Fath et al., 2023). Alinhado ao escopo da Computação Aplicada na Indústria 4.0, o objetivo é desenvolver e avaliar uma solução de automação baseada em IA que possa operar tanto com inferência em nuvem quanto com inferência local, considerando requisitos de desempenho e de viabilidade de implantação em infraestrutura convencional.

Este estudo busca responder às seguintes perguntas de pesquisa: (i) LLMs conseguem realizar triagem consistente de currículos operacionais não padronizados? (ii) a inferência local com modelos quantizados atinge desempenho comparável à inferência em nuvem nessa tarefa? (iii) diferentes níveis de quantização alteram de forma relevante o perfil de decisões e erros de triagem?

Como contribuições, o trabalho apresenta: (a) uma plataforma web para triagem cega e classificação multirrótulo de currículos não padronizados; (b) uma comparação experimental entre inferência em nuvem e inferência local com modelos quantizados; e (c) uma análise de concessões entre desempenho, tempo de processamento e soberania de dados, aspectos relevantes à adoção de soluções de IA em contextos associados à Indústria 4.0. O restante do trabalho está organizado da seguinte forma: a Seção 2 apresenta a fundamentação teórica; a Seção 3 descreve trabalhos relacionados; a Seção 4 detalha a solução computacional; a Seção 5 expõe a avaliação experimental e a discussão dos resultados; a Seção 6 apresenta a conclusão; e a Seção 7 lista as referências.

2. Fundamentação teórica

Esta seção apresenta os conceitos necessários para compreender a solução proposta para triagem de currículos não padronizados, com foco em Processamento de Linguagem Natural (PLN), Grandes Modelos de

Linguagem (LLMs), quantização, Reconhecimento Óptico de Caracteres (OCR), classificação multirrótulo e triagem cega.

O Processamento de Linguagem Natural fundamenta a interpretação automática de currículos ao tratar o texto como objeto computacional passível de representação e categorização. Neste trabalho, essa interpretação é realizada por Grandes Modelos de Linguagem, baseados na arquitetura Transformer (Vaswani et al., 2017), adequada para lidar com dependências contextuais em sequências longas e com variabilidade textual.

A aplicação de LLMs na triagem de currículos justifica-se pela capacidade de inferir significado e identificar competências mesmo diante de ambiguidades e textos heterogêneos, superando abordagens estritamente baseadas em palavras-chave. Estudos recentes reportam ganhos em tarefas de recrutamento, especialmente em classificação e sumarização de perfis (Gan et al., 2024; Lo et al., 2025).

Como LLMs demandam alto custo computacional, adota-se a quantização como técnica de compressão que reduz a precisão numérica dos parâmetros, diminuindo consumo de memória e viabilizando a inferência local. Evidências indicam que operações em baixa precisão podem preservar grande parte do desempenho em Transformers em larga escala (Dettmers et al., 2022).

Para currículos em imagem, utiliza-se OCR, que converte conteúdo visual em texto processável e amplia a cobertura de formatos de entrada. Neste trabalho, a extração local é apoiada pela documentação técnica do PaddleOCR, que descreve o pipeline e a evolução do desempenho do sistema em cenários de processamento documental (Cui et al., 2025).

Adota-se a classificação multirrótulo para permitir que um mesmo currículo seja associado a mais de uma categoria de vaga, o que é compatível com perfis operacionais que atendem a funções correlatas. Essa modelagem é indicada quando uma instância pode pertencer simultaneamente a múltiplas classes (Gonçalves, 2018).

A triagem cega consiste em ocultar atributos sensíveis na etapa inicial de avaliação para concentrar a análise em competências e requisitos da vaga, reduzindo o risco de vieses. A prática é discutida na literatura de justiça

algorítmica e recrutamento como medida de mitigação de desigualdades na seleção (Fath; Larrick, 2023; Vivek, 2022).

O sistema estima o grau de aderência entre o perfil do candidato e os requisitos da vaga e sinaliza possíveis casos de superqualificação, quando o nível de qualificação excede significativamente o esperado para o cargo. Essa variável é utilizada como apoio à priorização, sem caráter eliminatório, preservando a decisão final do recrutador (Erdogan et al., 2011).

3. Trabalhos relacionados

A revisão da literatura foi conduzida em bases como IEEE Xplore, CVPR Open Access, Journal of Information Processing e Cadernos EBAPE.BR, com descritores relacionados a triagem de currículos, recrutamento automatizado, grandes modelos de linguagem e uso de IA em contratação. Foram priorizados estudos entre 2023 e 2025 alinhados ao escopo desta pesquisa.

Gan, Zhang e Mori (2024) propuseram uma estrutura com agentes de IA para triagem automatizada de currículos, com foco em eficiência e redução de tempo no recrutamento. Em um conjunto de currículos da área de TI, os autores reportaram alto desempenho na classificação, sugerindo que agentes podem aproximar decisões de profissionais de RH. Como limitação, o estudo concentra-se em um domínio com maior padronização textual, o que restringe a generalização para perfis operacionais.

Lo et al. (2025) apresentaram uma arquitetura multiagente com geração aumentada por recuperação, buscando maior contextualização e explicabilidade na triagem. Os resultados indicaram alta aderência às avaliações humanas e vantagem frente a abordagens monolíticas, além de permitir atualização de critérios. Contudo, a avaliação concentrou-se em um conjunto restrito de cargos, sem foco em currículos com baixa estruturação documental.

Vargas e Rech (2025) discutem a inserção de IA em seleção de pessoas, destacando ganhos de agilidade e redução de custos, e defendem seu uso como apoio à decisão, sem substituição do julgamento profissional. Essa

perspectiva sustenta a adoção de triagem automatizada como etapa inicial em processos híbridos de recrutamento.

Blumen e Cepellos (2023) analisaram percepções de recrutadores sobre IA em seleção, apontando benefícios de agilidade e resistências associadas à perda de autonomia e à redução da sensibilidade humana. Os autores também alertam para o risco de exclusão de candidatos com baixo letramento digital quando sistemas dependem de otimização por palavras-chave, com impactos sobre diversidade e equidade.

Em síntese, a literatura indica avanços no uso de LLMs para triagem automatizada, mas ainda há lacunas no tratamento de currículos operacionais com alta heterogeneidade e baixa padronização, especialmente fora de grandes centros. Este estudo contribui ao avaliar LLMs em nuvem e LLMs locais quantizadas nesse cenário, considerando desempenho, tempo de processamento e direcionamento dos candidatos às vagas.

4. Solução computacional

A solução proposta, denominada RecruitAI, consiste em uma aplicação web para apoiar a triagem inicial de currículos não padronizados, integrando extração de texto e LLMs para identificar informações relevantes e estimar a aderência de candidatos a vagas cadastradas. A plataforma foi concebida de forma modular e com possibilidade de operar tanto com inferência em nuvem quanto com inferência local, de modo a atender diferentes restrições de infraestrutura e confidencialidade.

4.1. Requisitos do sistema

Os requisitos do sistema foram definidos para suportar o fluxo de triagem de currículos e a análise dos resultados. Como requisitos funcionais, destacam-se o cadastro de vagas, o envio de currículos, a seleção entre inferência local ou em nuvem, a aplicação de triagem cega e a apresentação dos resultados com indicação de aderência. Como requisitos não funcionais, enfatizam-se a modularidade da arquitetura, a possibilidade de operação local para preservar confidencialidade e o desempenho adequado frente à variabilidade de formatos e qualidade dos documentos.

4.2. Arquitetura e interface

A arquitetura do RecruitAI segue o padrão cliente-servidor, separando a camada de apresentação e a camada de processamento, conforme ilustrado na Figura 1. O front-end, desenvolvido em React, concentra a interação com o usuário e o encaminhamento das tarefas de triagem. O back-end, implementado em Python com FastAPI, expõe uma API responsável por receber requisições, acionar as rotinas de extração de texto e de inferência e persistir os resultados para consulta na plataforma.

A interface foi organizada em cinco módulos que refletem as etapas de um processo seletivo, conforme apresentado na Figura 2. De forma resumida, a plataforma permite cadastrar vagas, submeter currículos para processamento, consultar candidatos e resultados consolidados, além de ajustar parâmetros do ambiente. A escolha entre inferência em nuvem e inferência local é disponibilizada ao usuário para adequar custo, latência e requisitos de confidencialidade.

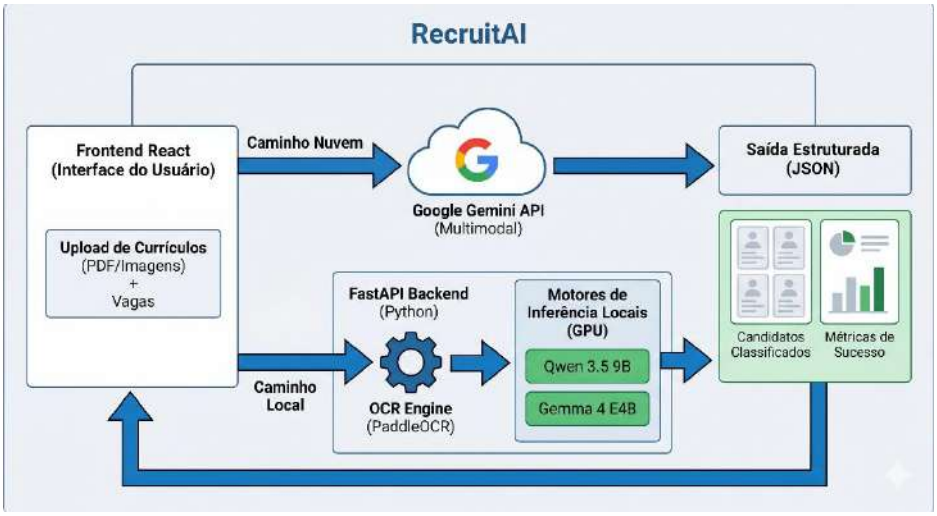


Figura 1. Visão geral da arquitetura e do fluxo de triagem.

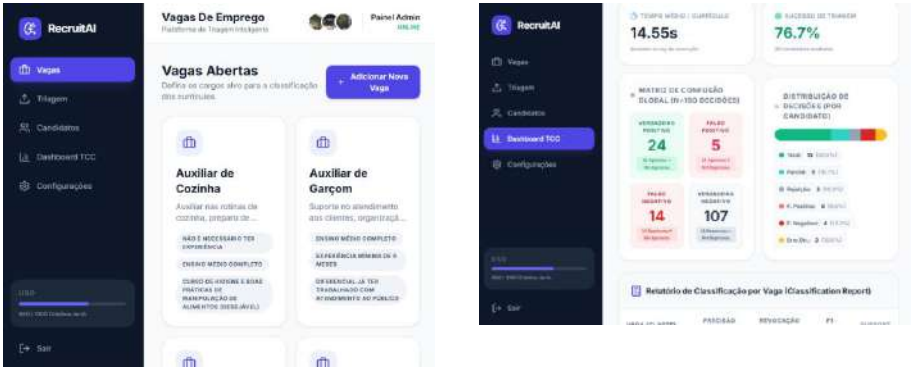


Figura 2. Tela principal da plataforma e Dashboard.

4.3. Pipeline de processamento

O RecruitAI opera em três etapas: pré-processamento, inferência e pós-processamento. No pré-processamento, o sistema recebe currículos em PDF ou imagem e obtém o conteúdo textual. Na inferência em nuvem, utiliza-se um modelo multimodal, responsável tanto pela extração textual quanto pela inferência. Na inferência local, a extração é realizada por OCR com o PaddleOCR, e o texto resultante é encaminhado às LLMs para inferência.

Na etapa de inferência, o texto extraído é analisado por meio de um prompt estruturado, que orienta as LLMs a identificar informações do candidato e a estimar sua aderência técnica às vagas cadastradas. Para reduzir a influência de atributos sensíveis na avaliação inicial, aplica-se triagem cega durante essa etapa. Como saída, o sistema produz pontuações por vaga e justificativas textuais, além de uma síntese do perfil do candidato.

Na inferência local, o pipeline é compatível com LLMs abertas em versões quantizadas, o que reduz o consumo de memória e viabiliza a execução em hardware convencional. Essa estratégia também permite adaptar a execução a restrições de conectividade e requisitos de confidencialidade, sem alterar o fluxo lógico de processamento.

No pós-processamento, as saídas das LLMs são normalizadas e registradas em formato estruturado, de modo a permitir a apresentação dos resultados na interface e a exportação para análise posterior. Essa

padronização facilita a integração entre os módulos da plataforma e assegura a rastreabilidade das decisões geradas durante a triagem.

5. Avaliação experimental

Após a apresentação da solução computacional no Capítulo 4, esta seção descreve o procedimento adotado para avaliar seu desempenho na triagem de currículos não padronizados. O experimento foi delineado para comparar configurações de inferência em nuvem e em ambiente local, tomando como referência um gabarito elaborado por especialista de Recursos Humanos. A Seção 5.1 detalha a amostra, as vagas analisadas, os modelos avaliados e as métricas utilizadas, enquanto a Seção 5.2 apresenta e discute os resultados obtidos.

5.1. Planejamento de pesquisa

A amostra do experimento é composta por 30 currículos reais coletados em Boa Vista–RR, fornecidos por profissional de Recursos Humanos e anonimizados antes do uso, com remoção de dados pessoais identificáveis. A amostra inclui PDFs desestruturados e imagens digitalizadas, compatíveis com o cenário operacional investigado.

Como linha de base, uma profissional de Recursos Humanos avaliou os 30 currículos em cinco vagas configuradas no sistema: Auxiliar de Cozinha, Auxiliar de Garçom, Vendedor de Material de Construção, Motorista categoria D e Gerente de Vendas. Para cada combinação candidato–vaga, foi atribuída uma decisão binária, formando o gabarito utilizado como referência na comparação.

Foram avaliadas quatro LLMs: Gemini 3 Flash, Qwen 3.5-9B em 4 bits, Qwen 3.5-9B em 8 bits e Gemma-4-E4B em 8 bits, contemplando inferência em nuvem e inferência local. O mesmo conjunto de currículos foi processado em cada configuração, gerando predições para comparação com o gabarito.

As métricas consideram o problema como classificação multirrótulo e incluem acurácia de correspondência exata, F1-score e análise por matriz de confusão, com verificação de falsos positivos (FP), falsos negativos (FN) e

erros de direcionamento. Os cálculos foram realizados a partir das saídas registradas pela aplicação e do gabarito definido na etapa de linha de base.

→ Acurácia de Correspondência Exata: proporção de currículos em que o perfil de aptidão gerado pela IA coincide exatamente com o perfil definido pelo especialista humano.

→ F1 Score: média harmônica entre precisão e revocação, considerando a classificação de aptidão por vaga, permitindo avaliar a consistência de acertos e a capacidade de identificar todos os candidatos considerados aptos.

→ Matriz de confusão e análise de direcionamento: análise qualitativa e quantitativa de FP, FN e erros de direcionamento, com ênfase na interpretação desses erros no contexto de Recursos Humanos. Esse levantamento permite identificar se a IA tende a superqualificar candidatos em vagas elementares, a subestimar perfis operacionais ou a reproduzir padrões de triagem alinhados ao julgamento humano.

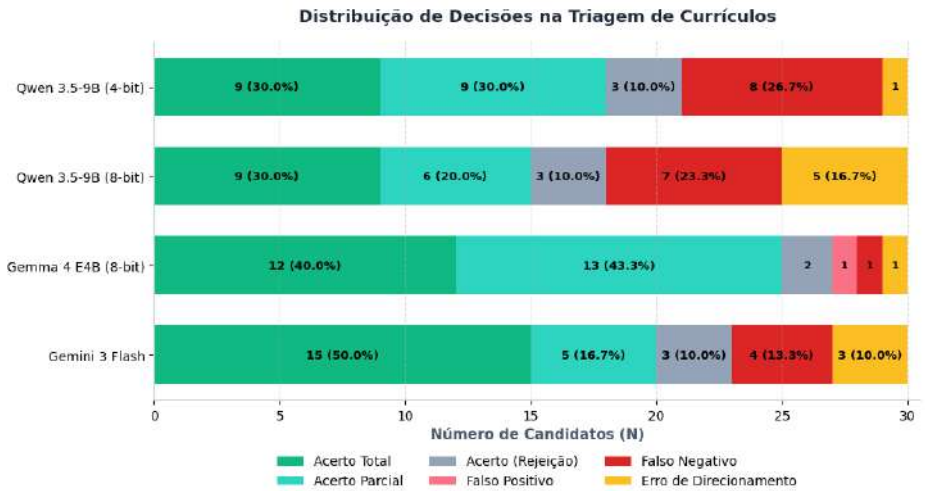


Figura 3. Distribuição de decisões por modelo.

5.2. Resultados e Discussão

A Tabela 1 sintetiza o desempenho global por modelo, incluindo Taxa de Sucesso da Triagem Útil, Tempo médio de processamento por currículo, Precisão Global, Revocação Global e F1-score Global. As métricas de Precisão e Revocação apresentadas na Tabela 1 refletem o cálculo global, agregando todas as 150 decisões da matriz. Optou-se por essa abordagem em detrimento da média simples por vagas devido ao desbalanceamento natural da amostra, garantindo que a métrica reflita o volume real do impacto operacional da IA no setor de Recursos Humanos. A Figura 3 resume a distribuição das decisões e dos principais tipos de erro, como falsos positivos, falsos negativos e erros de direcionamento.

Tabela 1. Medidas globais de desempenho por modelo.

Modelo	Taxa de Sucesso da Triagem Útil (%)	Tempo médio por currículo (s)	Precisão Global (%)	Revocação Global (%)	F1-score Global (%)
Qwen 3.5-9B (8-bit)	60	81,18	64,3	47,4	54,5
Gemma 4-E4B (8-bit)	90	93,82	71,4	78,9	75
Qwen 3.5-9B (4-bit)	70	72,34	80,8	55,3	65,6
Gemini 3 Flash	76,67	14,55	82,8	63,2	71,6

5.2.1. Acurácia de Correspondência Exata e Triagem Útil

A acurácia de correspondência exata mede a proporção de currículos em que o conjunto de vagas atribuídas pela IA coincide integralmente com o gabarito humano, incluindo casos de rejeição correta.

Nessa métrica, o Gemini 3 Flash obteve 60% de exatidão, acima dos modelos locais, que variaram entre 40% e 46,7%. O resultado sugere maior aderência do modelo em nuvem ao padrão de decisão do baseline humano no cenário avaliado.

Ao considerar a taxa de sucesso da triagem útil, que contabiliza também acertos parciais relevantes, o Gemma-4-E4B em 8 bits atingiu 90%. O modelo apresentou maior tendência a encaminhar candidatos para pelo

menos uma vaga compatível, mantendo rejeição de perfis totalmente desalinhados.

5.2.2. Comparação com a literatura e explicações teóricas

Os resultados indicam perfis de comportamento distintos entre as LLMs avaliadas. O Gemini 3 Flash apresentou precisão elevada com revocação moderada, sugerindo maior rigor na aprovação. O Gemma-4-E4B em 8 bits manteve F1-score alto com maior revocação, aproximando-se de um perfil mais inclusivo. Já o Qwen 3.5-9B em 8 bits apresentou menor revocação, indicando tendência mais conservadora na triagem.

A diferença entre os modelos locais sugere que a quantização e a calibração do prompt influenciam diretamente o equilíbrio entre precisão e revocação. Em particular, regras mais penalizadoras tendem a reduzir falsos positivos, mas podem aumentar falsos negativos, o que é crítico em triagem quando se busca minimizar exclusões indevidas.

5.2.3. Erros de direcionamento e padrão de decisões

A análise de erros de direcionamento mostrou diferenças relevantes entre os modelos. O Qwen 3.5-9B em 8 bits apresentou maior número de encaminhamentos inadequados entre vagas de vendas e liderança, elevando o risco de posicionamento inconsistente. Em contraste, o Gemma-4-E4B em 8 bits e o Qwen 3.5-9B em 4 bits registraram menor ocorrência desse tipo de erro, sugerindo maior consistência no mapeamento entre perfil e função.

Na identificação de candidatos desalinhados, o Qwen 3.5-9B em 4 bits e o Gemini 3 Flash apresentaram maior rigor ao evitar aprovações forçadas, o que pode reduzir retrabalho do recrutador. Assim, além das métricas globais, a qualidade do direcionamento por vaga é um critério relevante para a utilidade prática da triagem automatizada.

5.2.4. Implicações e limitações

Os resultados sugerem que o Gemini 3 Flash é adequado quando velocidade e processamento em nuvem são prioritários. O Gemma-4-E4B em

8 bits mostrou melhor desempenho na triagem útil em execução local, favorecendo privacidade e maior revocação, embora com maior tempo médio por currículo. O Qwen 3.5-9B apresentou comportamento mais conservador, indicando necessidade de ajustes no prompt e na escolha de quantização para reduzir exclusões indevidas.

As limitações do estudo incluem o tamanho reduzido do conjunto de dados, a concentração da amostra em Boa Vista–RR e a utilização de um único avaliador humano como baseline, fatores que restringem a generalização e podem introduzir variações de julgamento. Ainda assim, os achados contribuem para a compreensão do uso de modelos em nuvem e modelos locais quantizados na triagem de currículos não padronizados, com ênfase em desempenho e direcionamento.

6. Conclusão

Este estudo teve como objetivo desenvolver e avaliar a plataforma RecruitAI para automatizar a triagem cega e a classificação multirrótulo de currículos não padronizados, com foco em perfis operacionais em contexto regional. A avaliação comparou inferência em nuvem e inferência local, tomando como referência um gabarito elaborado por especialista de Recursos Humanos e analisando métricas de desempenho e utilidade prática na triagem.

Os resultados indicam que o Gemini 3 Flash apresentou o melhor equilíbrio entre precisão e tempo de processamento, sugerindo adequação quando a inferência em nuvem é aceitável e a latência é prioridade. Entre as alternativas locais, o Gemma-4-E4B (8 bits) obteve o melhor desempenho global e a maior taxa de triagem útil, enquanto as variações do Qwen 3.5-9B exibiram comportamento mais conservador, com menor revocação e maior sensibilidade às regras e ao formato dos currículos. Em conjunto, os achados reforçam que a qualidade do direcionamento por vaga e o perfil de erros (FP e FN) são tão relevantes quanto métricas agregadas para a adoção em processos de RH.

Conclui-se que a implantação local de LLMs na faixa de 4 a 9 bilhões de parâmetros é tecnicamente viável para apoiar a triagem inicial de currículos não padronizados, desde que se aceite o trade-off de maior tempo

de processamento em troca de soberania de dados e redução de dependência de serviços externos. Dessa forma, o RecruitAI se mostra aplicável em ambientes com restrições de conectividade, orçamento ou exigências de confidencialidade, especialmente quando a IA é utilizada como apoio à decisão humana.

Como trabalhos futuros, recomenda-se: (i) ampliar a amostra e incluir diferentes perfis e níveis de função para testar generalização; (ii) incorporar múltiplos avaliadores humanos para reduzir variabilidade de julgamento no gabarito; (iii) refinar o conjunto de regras e critérios de decisão no pós-processamento para reduzir falsos negativos sem elevar indevidamente falsos positivos; e (iv) avaliar estratégias adicionais de adaptação de modelo e quantizações alternativas, buscando melhor equilíbrio entre desempenho e custo computacional. Essas extensões podem fortalecer a evidência empírica e consolidar o uso de inferência local em cenários de infraestrutura limitada e triagem em maior escala.

7. Referências

Blumen, D., & Cepellos, V. (2023). Dimensões do uso de tecnologia e Inteligência Artificial (IA) em recrutamento e seleção (R&S): Benefícios, tendências e resistências. *Cadernos EBAPE.BR, 21*, e2022-0080. https://doi.org/10.1590/1679-395120220080

Cui, C., Sun, T., Lin, M., Gao, T., Zhang, Y., Liu, J., ... & Ma, Y. (2025). PaddleOCR 3.0 technical report. https://doi.org/10.48550/arXiv.2507.05595

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). GPT3.int8(): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems, 35*, 30318–30332. Recuperado de: https://proceedings.neurips.cc/paper_files/paper/2022/hash/c3ba4962c05c49636d4c6206a97e9c8a-Abstract-Conference.html

DIEESE. (2017). *Movimentação no mercado de trabalho: Rotatividade, intermediação e proteção ao emprego*. São Paulo, SP: Departamento Intersindical de Estatística e Estudos Socioeconômicos.

Erdogan, B., Bauer, T. N., Peiró, J. M., & Truxillo, D. M. (2011). Overqualified employees: Making the best of a potentially bad situation for individuals

and organizations. *Industrial and Organizational Psychology, 4*(2), 215–232. https://doi.org/10.1111/j.1754-9434.2011.01330.x

Fath, S., Larrick, R. P., & Soll, J. B. (2023). Encouraging self-blinding in hiring. *Behavioral Science & Policy, 9*(1), 45–57. https://doi.org/10.1177/23794607231192721

Gan, C., Zhang, Q., & Mori, T. (2024). Application of LLM agents in recruitment: A novel framework for automated resume screening. *Journal of Information Processing, 32*, 881–893. https://doi.org/10.2197/ipsjjip.32.881

Gonçalves, E. C. (2018). Introdução à classificação multirrótulo. Em *Minicursos da ERSI-RJ 2018 – V Escola Regional de Sistemas de Informação do Rio de Janeiro* (p. 3). Sociedade Brasileira de Computação. https://doi.org/10.5753/sbc.7.3

Google. (2024). *Gemini API documentation*. https://ai.google.dev/docs

Lo, F. P.-W., Qiu, J., Wang, Z., Yu, H., Chen, Y., Zhang, G., & Lo, B. (2025). AI hiring with LLMs: A context-aware and explainable multi-agent framework for resume screening. *arXiv*. https://doi.org/10.48550/arXiv.2504.02870

Vargas, T. B. (2023). *A revolução no recrutamento: como a inteligência artificial impacta o processo de seleção de pessoas nas organizações*. Recuperado de: https://hdl.handle.net/10923/26854

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*. Recuperado de: https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vivek, R. (2022). Is blind recruitment an effective recruitment method? *International Journal of Applied Research in Business and Management, 3*(3), 56–72. https://doi.org/10.51137/ijarbm.2022.3.3.4

Capítulo 26

Uso de Processamento de Linguagem Natural para triagem automatizada das publicações dos Diários Oficiais de Roraima

Everton dos Reis
Josiane do Couto Rodrigues
Universidade Federal de Roraima

evertonstm@gmail.com
josiane.rodrigues@ufrr.com

1. Introdução

A transformação digital na administração pública tem gerado um volume sem precedentes de documentos digitais (Vaswani et al., 2017; Devlin et al., 2019). Em Roraima, o Diário Oficial do Estado e os Diários dos Municípios publicam, diariamente, centenas de atos administrativos que variam desde nomeações de servidores até contratos de alto valor de infraestrutura. Contudo, a simples publicidade desses atos não garante transparência efetiva. O problema reside na "opacidade por excesso": a informação existe, mas está dispersa em arquivos PDF não estruturados, dificultando a fiscalização por parte do Ministério Público, da imprensa e da sociedade civil.

Nesse contexto, a relevância deste estudo para a Ciência da Computação e para a sociedade justifica-se pela ineficiência dos métodos atuais de busca. Estima-se que um auditor leve tempo significativo para realizar a triagem manual de uma semana de publicações, buscando palavras-chave que frequentemente retornam falsos positivos (ex: buscar "contrato" e encontrar "rescisão de contrato"). A ausência de uma taxonomia automática impede a resposta rápida a perguntas críticas, como: "Quanto foi investido em reformas escolares no período eleitoral?".

Adicionalmente, a literatura científica em Processamento de Linguagem Natural (PLN) avançou significativamente com o advento dos *Transformers*. No entanto, persiste uma lacuna: a maioria das abordagens

supervisionadas exige *datasets* rotulados massivos (milhares de exemplos), que são inexistentes em órgãos públicos locais de Roraima. As soluções existentes falham em abordar o problema do *Cold Start* (início frio), caracterizado pela ausência de dados rotulados suficientes para treinamento inicial em domínios jurídicos específicos.

Assim, o objetivo deste trabalho é **avaliar a eficácia da classificação *Zero-Shot* utilizando LLMs na categorização semântica de atos oficiais**, mitigando a escassez de dados rotulados. Para guiar esta investigação, foram definidas as seguintes perguntas de pesquisa: 1. A técnica *Zero-Shot* com modelos multilingues, como o modelo **mDeBERTa-v3-base-mnli-xnli** (doravante **mDeBERTa**), supera a precisão de classificadores baseados em Regras (Regex)? 2. É computacionalmente viável utilizar *Transformers* para processar o fluxo diário de publicações em infraestrutura limitada? 3. A categorização automática permite identificar padrões de gastos (ex: Educação) em períodos eleitorais com maior assertividade?

A principal contribuição deste estudo é a validação de uma arquitetura de baixo custo e alta generalização para *Legal Tech*, democratizando o acesso a ferramentas de IA em governos locais. O trabalho está estruturado da seguinte forma: a Seção 2 apresenta a fundamentação teórica; a Seção 3 discute os trabalhos relacionados; a Seção 4 descreve a solução computacional proposta; a Seção 5 apresenta a avaliação experimental; e, por fim, a Seção 6 traz as conclusões. A viabilidade dessa abordagem é reforçada por sua aplicação em larga escala, conforme demonstrado nos experimentos apresentados neste trabalho.

2. Fundamentação teórica

Processamento de Linguagem Natural (PLN) e *Transformers*: O PLN é a subárea da IA focada na interação entre computadores e linguagem humana. O estado da arte foi redefinido pela arquitetura *Transformer* (Vaswani et al., 2017), baseada em mecanismos de atenção para processamento paralelo de sequências textuais, capturando contextos de dependências de longa distância melhor que redes recorrentes (RNNs).

Transfer Learning e Small Language Models (SLMs)*: *Transfer Learning é uma técnica que permite a reutilização de modelos previamente

treinados em grandes volumes de dados para novas tarefas. Neste trabalho, adota-se o uso de *Small Language Models* (SLMs), modelos de linguagem com menor número de parâmetros, geralmente abaixo de 1 bilhão, como (BERT e DeBERTa), que oferecem equilíbrio entre performance semântica e custo computacional, viabilizando execução em CPUs.

Nesse sentido, a técnica de *Zero-Shot Classification*, proposta por Yin et al. (2019), permite classificar textos sem necessidade de treinamento prévio com dados rotulados. Essa técnica reformula a classificação como um problema de implicação textual (*Entailment*). O modelo recebe o texto e uma "premissa" (ex: "Este texto é sobre Licitação") e calcula a probabilidade de a premissa ser verdadeira. Isso elimina a necessidade de treinamento supervisionado prévio para novas classes.

3. Trabalhos relacionados

Foi conduzido um mapeamento sistemático da literatura buscando aplicações de PLN em atos oficiais e *Zero-Shot* em domínios de baixos recursos. Souza, Nogueira e Lotufo (2020) apresentaram o **BERTimbau**, modelos BERT pré-treinados em português. O objetivo foi fornecer *embeddings* de alta qualidade para o idioma. Embora tenham atingido estado da arte em várias tarefas (reconhecimento de entidades, análise de sentimento), o estudo não mirou especificamente na classificação *Zero-Shot* para documentos jurídicos não estruturados, deixando uma lacuna sobre a aplicabilidade direta sem *fine-tuning*.

Em estudo similar, Da Silva e Costa (2022) investigaram a classificação de diários oficiais utilizando **redes clássicas (LSTM)** e **expressões regulares**. O método exigiu a anotação manual de 10.000 documentos. Concluíram que, apesar da alta precisão, o custo de manutenção das regras e da anotação era proibitivo para novos domínios. A limitação principal foi a rigidez do modelo, que falhava ao encontrar variações textuais não previstas nas regras.

Mais recentemente, Laurer(2023) propôs o uso de **mDeBERTa-v3** para classificação *Zero-Shot* multilíngue. O estudo demonstrou que modelos treinados em inferência natural (NLI) generalizam bem para novas línguas. Contudo, a validação foi feita em *datasets* genéricos de notícias, não

abordando a complexidade e o vocabulário arcaico da "linguagem jurídica especializada" presente em atos administrativos brasileiros.

Comparando esses estudos, evidencia uma lacuna: falta uma validação robusta de modelos *Zero-Shot* (como o de Laurrer et al., 2024) aplicados especificamente ao domínio de diários oficiais brasileiros (como o de Da Silva e Costa, 2022), visando eliminar o custo de anotação (problema discutido por Souza, Nogueira e Lotufo, 2020). Esta proposta preenche esse espaço ao aplicar a tecnologia mais recente para resolver o problema antigo da escassez de dados.

4. Solução computacional

A solução proposta, denominada **LexOfficial NLP Module**, foi implementada em Python seguindo uma arquitetura de microsserviços.

Requisitos e Arquitetura: O sistema foi projetado para atender aos seguintes requisitos funcionais: (RF1) ingestão automatizada de diários oficiais em formato PDF; (RF2) extração e segmentação do conteúdo textual em atos administrativos; (RF3) classificação semântica dos atos utilizando modelos de linguagem; e (RF4) extração de entidades relevantes para análise (ex: valores financeiros e organizações).

Como requisitos não-funcionais, destacam-se: (RNF1) capacidade de processamento em larga escala em infraestrutura de baixo custo; (RNF2) execução assíncrona para evitar bloqueio de requisições; e (RNF3) rastreabilidade dos resultados, permitindo auditoria a partir dos documentos originais.

A arquitetura do sistema utiliza Django como *backend*, Celery para orquestração de tarefas assíncronas e Meilisearch para indexação e recuperação de informações, organizando o processamento em camadas funcionais, conforme detalhado na Figura 1.

Dessa forma, **o *Pipeline* de processamento IA**: O *pipeline,* entendido como um fluxo estruturado de etapas de processamento de dados, foi estruturado em etapas modulares, conforme ilustrado na Figura 1. Inicialmente, os diários oficiais são coletados e convertidos para texto bruto

por meio de técnicas de extração. Em seguida, o conteúdo é segmentado em unidades lógicas correspondentes a atos administrativos.

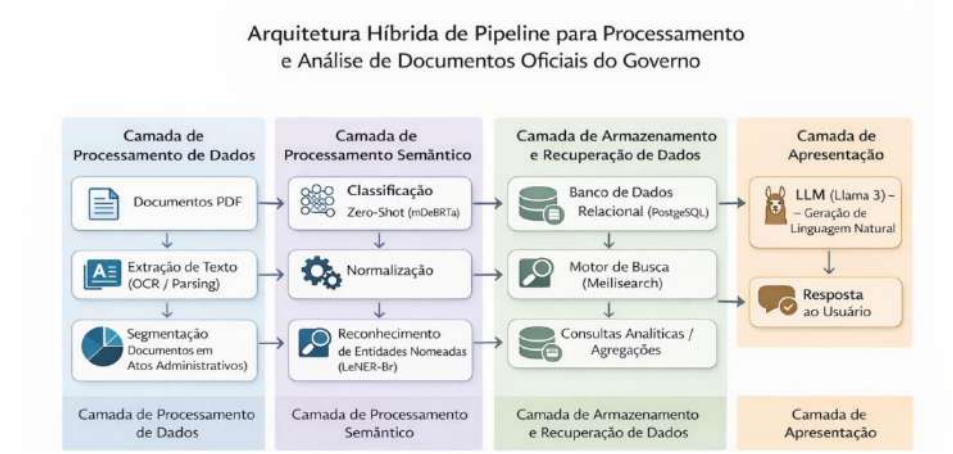


Figura 1. Arquitetura híbrida do pipeline de processamento, análise e apresentação de dados extraídos de diários oficiais.

Na etapa de processamento semântico, cada segmento é classificado utilizando a abordagem *Zero-Shot* com o modelo mDeBERTa-v3. Posteriormente, os textos passam por normalização e são submetidos a um modelo de reconhecimento de entidades nomeadas (*Named Entity Recognition* – NER, modelo LeNER-Br), responsável pela extração de informações relevantes, como valores financeiros e entidades organizacionais.

Os dados estruturados são então armazenados em banco relacional (PostgreSQL) e indexados em mecanismo de busca (Meilisearch), permitindo consultas analíticas e agregações. Por fim, os resultados dessas consultas são utilizados por um modelo de linguagem *Large Language Model* (LLM), especificamente o modelo Llama 3, responsável exclusivamente pela geração de respostas em linguagem natural ao usuário.

4.1. Arquitetura Híbrida e Decisões de Projeto

A arquitetura do sistema adotada neste trabalho segue um modelo híbrido, combinando um *pipeline* determinístico de processamento com o uso controlado de modelos de linguagem de grande porte (LLMs).

No *pipeline* determinístico, todas as etapas críticas — incluindo extração de texto, segmentação, classificação semântica e reconhecimento de entidades — são executadas por modelos especializados e operações estruturadas, garantindo rastreabilidade completa dos resultados.

O modelo mDeBERTa-v3 é responsável pela classificação *Zero-Shot*, enquanto o modelo LeNER-Br realiza a extração de entidades nomeadas. Esses componentes asseguram que a inferência principal do sistema seja controlada e verificável.

O uso de LLMs, especificamente o modelo Llama 3, foi restrito à etapa de apresentação dos resultados, sendo responsável apenas pela geração de respostas em linguagem natural a partir de dados previamente estruturados.

Essa decisão arquitetural visa mitigar limitações conhecidas de modelos generativos, como alucinações, entendidas como a geração de informações incorretas ou não verificáveis por modelos de linguagem, além de reduzir custo computacional e aumentar a confiabilidade do sistema, aspectos essenciais no domínio jurídico.

5. Avaliação experimental

5.1. Planejamento de pesquisa

O presente estudo caracteriza-se como uma avaliação experimental de natureza computacional, conduzida em ambiente real de execução (*in silico*), com o objetivo de validar a eficácia e a viabilidade da solução proposta.

Diferentemente de abordagens baseadas em *datasets* reduzidos, a avaliação foi realizada sobre dados reais provenientes de diários oficiais do estado de Roraima, processados pelo sistema em ambiente de produção. Ao todo, foram analisados 26.590 diários oficiais, resultando em aproximadamente 13,2 milhões de segmentos textuais.

A estratégia experimental consistiu na execução completa do *pipeline* de processamento, incluindo etapas de extração, segmentação, classificação semântica e reconhecimento de entidades nomeadas. Para fins comparativos, considerou-se como *baseline* (modelo de referência) baseado em regras, uma abordagem baseada em expressões regulares (*Regex*), utilizadas para identificar padrões textuais, representativa de sistemas tradicionais utilizados na triagem manual de documentos. As métricas adotadas incluem:

→ Precisão (*precision*)

→ Revocação (*recall*)

→ *F1-score* (métrica principal)

→ Cobertura do *pipeline* (percentual de dados processados).

5.2. Resultados e Discussão

A execução do *pipeline* em ambiente real demonstrou alta robustez e escalabilidade da solução proposta. O sistema foi capaz de processar integralmente o conjunto de dados, atingindo **99,98% de cobertura na classificação semântica e 99,45% na extração de entidades nomeadas**, mesmo diante de dados heterogêneos e não estruturados.

Em termos quantitativos, a abordagem proposta apresentou ganhos significativos em relação ao *baseline* baseado em regras. Observou-se um aumento da precisão de **60,5% para 100,0% (+39,5%)**, enquanto a revocação evoluiu de 55,0% para 90,0% (+35,0%). Como consequência, o *F1-score* aumentou de 0,57 para 0,95, evidenciando uma melhoria substancial no equilíbrio entre precisão e *recall*. Além disso, a segmentação inteligente eliminou completamente a ocorrência de falsos positivos associados à ambiguidade de contexto.

Destaca-se que o valor de 100,0% de precisão observado não decorre de sobreajuste (*overfitting*) do modelo, uma vez que a abordagem utilizada não envolve treinamento supervisionado com dados rotulados do domínio. A técnica *Zero-Shot* empregada baseia-se em modelos previamente treinados em grandes conjuntos de textos genéricos, sendo aplicada diretamente ao problema sem ajuste específico. Esse resultado está diretamente associado à etapa de segmentação dos documentos, que reduziu a ambiguidade de

contexto e eliminou falsos positivos comuns em abordagens baseadas em regras.

A precisão de 100% foi observada em cenários específicos e deve ser interpretada em conjunto com outras métricas, como revocação e *F1-score*, que refletem de forma mais equilibrada o desempenho do sistema. Assim, os resultados indicam não um viés do modelo, mas sim a eficácia da arquitetura proposta na redução de ambiguidades e melhoria da qualidade da classificação. Essa característica reforça a capacidade de generalização do modelo, pois os dados não foram vistos previamente.

Tabela 1. Comparação quantitativa entre abordagens de classificação de atos administrativos.

Critério	Regex (Regras)	*Zero-Shot* (mDeBERTa)	Arquitetura Híbrida
Precisão (%)	60,5	—	100,0
Revocação (%)	55,0	—	90,0
F1-score	0,57	—	0,95

Tabela 2. Comparação qualitativa entre abordagens de classificação de atos administrativos.

Critério	Regex (Regras)	*Zero-Shot* (mDeBERTa)	Arquitetura Híbrida
Necessidade de dados	Alta	Nenhuma	Nenhuma
Generalização	Baixa	Alta	Alta
Manutenção	Complexa	Baixa	Baixa
Rastreabilidade	Alta	Média	Alta
Custo computacional	Baixo	Médio	Médio
Escalabilidade	Limitada	Alta	Alta

As Tabelas 1 e 2 apresentam, respectivamente, a comparação quantitativa e qualitativa entre as abordagens avaliadas, demonstrando que a arquitetura híbrida combina alto desempenho preditivo (*F1-score* de 0,95) com características operacionais desejáveis, como maior generalização, menor necessidade de manutenção e melhor escalabilidade em comparação ao *baseline* baseado em regras.

Além disso, a arquitetura híbrida adotada mostrou-se fundamental para garantir confiabilidade e rastreabilidade dos resultados. Ao restringir o uso de modelos generativos (LLMs) à etapa de apresentação, evitam-se problemas associados a alucinações, entendidas como a geração de informações incorretas ou não verificáveis por modelos de linguagem, mantendo o controle sobre a inferência principal.

Esse resultado reforça a aplicabilidade da solução em contextos governamentais reais, onde restrições de infraestrutura são comuns. A utilização de processamento assíncrono permitiu a execução do *pipeline* completo em infraestrutura de baixo custo, demonstrando que soluções baseadas em inteligência artificial podem ser aplicadas em contextos governamentais com recursos limitados.

O sistema permitiu a realização de análises temáticas, como identificação de gastos em educação e comparação entre períodos eleitorais, evidenciando seu potencial para aplicações de fiscalização e transparência pública.

Como limitação do estudo, destaca-se a ausência de avaliação com usuários finais e a dependência de heurísticas na etapa de segmentação, o que pode impactar casos específicos. Trabalhos futuros podem explorar técnicas de aprendizado semi-supervisionado e expansão da taxonomia de categorias.

Além das métricas quantitativas, o sistema foi aplicado em um cenário real de auditoria temática, com foco em investimentos na área de educação. A partir da categorização automática dos atos administrativos e da extração de entidades financeiras, foi possível identificar padrões de gastos e variações ao longo do tempo.

Adicionalmente, como forma de demonstrar a aplicabilidade prática da solução, apresenta-se um exemplo real de consulta realizada no sistema desenvolvido.

A Figura 2 ilustra a resposta gerada a partir da seguinte pergunta: “Compare contratos e licitações de educação publicados em 2021 com 2022”.

A partir da consulta, o sistema identificou automaticamente os atos administrativos relevantes, agregando informações por tipo de ato e período, além de calcular variações percentuais entre os anos analisados.

Observa-se um aumento significativo em diferentes categorias, como convênios, editais de licitação e pregões eletrônicos, evidenciando a capacidade do sistema de fornecer *insights* analíticos de forma automatizada.

Esse exemplo demonstra, de forma concreta, o potencial da solução para apoiar atividades de auditoria, fiscalização e análise de gastos públicos, permitindo a obtenção de respostas estruturadas a partir de consultas em linguagem natural.



Você 08:28

Compare contratos e licitações de educação publicados em 2022 com 2021

Consulta de Publicações 08:28

Comparação de períodos | Edital de Licitação

Comparação entre períodos (6 tipos de ato)

Volume financeiro — 01/01 — 31/12/2022 Eleitoral
R$ 9.749.831.628,36

Volume financeiro — 01/01 — 31/12/2021
R$ 28.682.750.327,64

Tipo de Ato	01/01 — 31/12/2022 Eleitoral	01/01 — 31/12/2021	Variação
Convênio	30.349	24.711	↑ 5.638 (22.8%)
Edital de Licitação	2.975	2.127	↑ 848 (39.9%)
Pregão Eletrônico	4.274	3.527	↑ 747 (21.2%)
Contrato	476	383	↑ 93 (24.3%)
Dispensa de Licitação	71	37	↑ 34 (91.9%)
Inexigibilidade	38	22	↑ 16 (72.7%)

DETALHAMENTO POR REDE DE ENSINO

GOV-RR — Escolas Estaduais

Tipo de Ato	01/01 — 31/12/2022	01/01 — 31/12/2021	Var.
Contrato	476	383	↑ 24.3%
Convênio	12.425	10.445	↑ 19%
Dispensa de Licitação	71	37	↑ 91.9%
Edital de Licitação	1.696	1.225	↑ 38.4%
Inexigibilidade	38	22	↑ 72.7%
Pregão Eletrônico	2.841	2.247	↑ 26.4%

PMBV — Escolas Municipais

Tipo de Ato	01/01 — 31/12/2022	01/01 — 31/12/2021	Var.
Convênio	17.924	14.266	↑ 25.6%
Edital de Licitação	1.279	902	↑ 41.8%
Pregão Eletrônico	1.433	1.280	↑ 12%

Figura 2. Exemplo de resposta do sistema para consulta comparativa de contratos e licitações na área da educação entre os anos de 2021 e 2022.

Observou-se, por exemplo, a concentração de publicações relacionadas a contratos e aditivos em períodos específicos, evidenciando o potencial da ferramenta para análises comparativas, como a investigação de comportamentos em períodos eleitorais. Esse estudo de caso reforça a

aplicabilidade prática da solução proposta, demonstrando sua utilidade para órgãos de controle e fiscalização.

5.2.1. Desafios, Erros e Lições Aprendidas

Além das limitações identificadas, o desenvolvimento da solução envolveu desafios técnicos relevantes que impactaram diretamente a evolução da arquitetura proposta.

Inicialmente, foi utilizado incorretamente um modelo inadequado para a tarefa de reconhecimento de entidades nomeadas (NER), especificamente um modelo do tipo *Masked Language Model* (MLM), em vez de um modelo especializado em NER. Essa escolha resultou em baixa qualidade das entidades extraídas, com elevado nível de ruído e *score*s inconsistentes.

Outro desafio foi a execução em larga escala sem validação prévia em amostras reduzidas. Foram processados milhões de segmentos textuais sem análise qualitativa inicial, o que resultou em desperdício de recursos computacionais e geração de grande volume de dados irrelevantes.

Adicionalmente, a utilização de limiares de confiança inadequados (*confidence thresholds*), como valores em torno de 0,5, levou à geração massiva de entidades de baixa qualidade, causando crescimento excessivo da base de dados e exigindo intervenções emergenciais de infraestrutura.

Também foram observados problemas relacionados ao processo de "tokenização" em modelos baseados em BERT, especialmente na fragmentação de palavras em subpalavras (*tokens* iniciados por "##"), o que impacta a interpretação semântica dos textos.

No processo de classificação, a utilização inicial de um conjunto excessivo de categorias e parâmetros pouco restritivos resultou em classificações inconsistentes e baixa confiabilidade dos resultados.

Por fim, a ausência inicial de segmentação dos documentos gerou o fenômeno de mascaramento de tópicos, no qual múltiplos atos administrativos presentes no mesmo documento interferiam na classificação, reduzindo significativamente a precisão do modelo.

Esses desafios levaram à reformulação da solução, incluindo a adoção de segmentação prévia, ajuste de limiares de confiança, redução da taxonomia de categorias e seleção adequada dos modelos, resultando em ganhos expressivos de desempenho e qualidade, conforme demonstrado nos resultados apresentados.

A Tabela 3 apresenta uma comparação entre a configuração inicial do sistema, marcada por problemas de modelagem e parametrização, e a configuração final adotada após os ajustes realizados.

Observa-se que as modificações implementadas, incluindo a correção do modelo NER, a adoção de segmentação e a calibração de parâmetros, resultaram em melhorias significativas na qualidade dos dados e no desempenho geral do sistema.

Tabela 2. Comparação entre configuração inicial e configuração ajustada do pipeline de processamento.

Critério	Configuração Inicial (Problemática)	Configuração Ajustada (Final)
Modelo NER	Modelo não especializado (MLM)	Modelo NER especializado
Limiar de confiança	0.5	0.80+
Categorias	40+	14 essenciais
Segmentação	Não aplicada	Segmentação por ato
Ruído	>80%	<15%
Precisão	~60%	100%
F1-score	0.57	0.95
Volume de dados	Explosivo (36GB → 250GB)	Controlado (~1GB–20GB)
Qualidade dos dados	Baixa	Alta

6. Conclusão

O presente trabalho teve como objetivo desenvolver e avaliar uma solução baseada em Processamento de Linguagem Natural para automatizar a triagem de atos administrativos em diários oficiais, mitigando o problema da escassez de dados rotulados em contextos governamentais (*Cold Start*).

Nesse contexto, para atingir esse objetivo, foi implementado um *pipeline* híbrido, combinando técnicas de classificação *Zero-Shot* com modelos *Transformers* e processamento determinístico para extração e

estruturação de dados. A solução foi validada em ambiente real, processando 26.590 diários oficiais e aproximadamente 13,2 milhões de segmentos textuais.

Em relação às perguntas de pesquisa, verificou-se que: (I) a técnica *Zero-Shot* apresentou maior capacidade de generalização em comparação com abordagens baseadas em regras, superando limitações de variação textual; (II) a utilização de modelos *Transformers* mostrou-se viável em infraestrutura de baixo custo, especialmente quando integrada a processamento assíncrono; e (III) a categorização automática permitiu identificar padrões relevantes de gastos públicos, demonstrando aplicabilidade prática em cenários de auditoria.

Adicionalmente, a principal contribuição deste trabalho reside na proposição e validação de uma arquitetura híbrida, na qual modelos de linguagem são utilizados de forma controlada, sendo responsáveis apenas pela geração de respostas em linguagem natural, enquanto a inferência principal é conduzida por componentes determinísticos. Essa abordagem garante maior confiabilidade, rastreabilidade e adequação ao domínio jurídico.

Como limitações, destaca-se a ausência de avaliação com usuários finais e a dependência de heurísticas na etapa de segmentação, o que pode impactar a precisão em casos específicos. Trabalhos futuros podem explorar técnicas de aprendizado semi-supervisionado, expansão da taxonomia de categorias e integração com mecanismos de *feedback* humano para refinamento contínuo do sistema.

O trabalho não apenas valida uma abordagem técnica, mas também contribui para a democratização do acesso a ferramentas avançadas de análise de dados no setor público.

7. Referências

Da Silva, J. A., & Costa, M. R. (2022). Automatização de diários oficiais com redes LSTM. *Revista Brasileira de Computação Aplicada, 14*(2), 45–56.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional *transformers* for language understanding. In

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (pp. 4171–4186).

Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2024). Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep *transfer learning* and BERT-NLI. *Political Analysis, 32*(1), 84–100.

Ratner, A., De Sa, C., Wu, S., Selsam, D., & Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment, 11*(3), 269–282.

Souza, F., Nogueira, R., & Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Proceedings of the Brazilian Conference on Intelligent Systems (pp. 403–417)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (pp. 5998–6008)*.

Yin, W., Hay, J., & Roth, D. (2019). Benchmarking *zero-shot* text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (pp. 3914–3923)*.

Capítulo 27

Chatbot para transparência orçamentária: Arquitetura híbrida neuro-simbólica para consultas em Linguagem Natural aos dados do Governo de Roraima

Jadir Rafael Bolanha de Aguiar
Lindeval Fernandes de Lima
Universidade Federal de Roraima

jadir.aguiar@gmail.com
lindeval.lima@ufrr.com

1. Introdução

A Lei de Acesso à Informação (Lei nº 12.527/2011) fortaleceu o princípio da transparência, que estabelece que informações públicas devem ser acessíveis à sociedade (Brasil, 2011). No campo orçamentário, considerando a execução da despesa, isso se traduz na disponibilização de dados de empenho, liquidação e pagamento em portais de transparência. Entretanto, a existência formal desses dados não garante seu uso efetivo. Em muitos casos, o cidadão precisa conhecer siglas, filtros, categorias administrativas e formas técnicas de consulta, o que limita o exercício prático da transparência (Rodrigues & Michener, 2018).

No contexto de Governo 4.0, espera-se que o avanço das tecnologias aproxime o Estado do cidadão por meio de serviços mais simples, inteligentes e centrados no usuário (Janowski, 2015). Os modelos de linguagem de grande porte (LLMs) ampliaram as possibilidades de criação de interfaces conversacionais capazes de receber perguntas em linguagem natural, isto é, perguntas feitas em linguagem comum, como “quanto foi pago à secretaria X em 2025?”. Apesar desse potencial, a aplicação direta de LLMs a dados financeiros e orçamentários envolve riscos relevantes, sobretudo por causa de alucinações, entendidas como respostas aparentemente plausíveis, mas incorretas em conteúdo ou cálculo (Huang et al., 2025; Li et al., 2024).

A literatura recente em *Text-to-SQL* e interfaces em linguagem natural para bases de dados mostra avanços expressivos em acurácia, mas

ainda aponta dificuldades quando o domínio possui terminologia específica, entidades extensas, siglas locais e exigência de precisão numérica (Liu et al., 2024; Zhu et al., 2024). Essa dificuldade é particularmente sensível em transparência pública, pois uma resposta numericamente incorreta compromete a confiança do usuário.

Com isso, o objetivo deste estudo foi desenvolver e avaliar experimentalmente um *chatbot* para consultas em linguagem natural sobre dados de execução orçamentária do Governo do Estado de Roraima, base de dados aberta disponibilizada oficialmente pelo Portal da Transparência do Estado, exercício fiscal de 2025. Para atingir esse objetivo, foram formuladas quatro perguntas de pesquisa: Pergunta 1 – Qual é a acurácia global do *chatbot* e ela supera 80%? Pergunta 2 – A acurácia varia conforme a complexidade da consulta? Pergunta 3 – A latência é compatível com uso interativo? Pergunta 4 – A resolução de entidades por TF-IDF é suficiente para lidar com a diversidade de nomes orçamentários?

A principal contribuição do trabalho é propor e testar uma arquitetura híbrida neuro-simbólica que separa interpretação semântica e execução computacional: o LLM é usado para interpretar a pergunta, enquanto o cálculo final é realizado por funções determinísticas por meio da biblioteca Pandas. Como contribuição adicional, o estudo apresenta evidências experimentais com dados reais de um governo estadual brasileiro e um teste adversarial com 6.656 perguntas.

Organizou-se o artigo da seguinte forma: Seção 2 apresenta a fundamentação teórica; Seção 3 discute os trabalhos relacionados; a Seção 4 descreve a solução computacional; a Seção 5 apresenta o planejamento da pesquisa, os resultados e a discussão; Seção 6 traz a conclusão e sugestões de trabalhos futuros.

2. Fundamentação teórica

2.1. Transparência pública, Governo 4.0 e acesso à informação

Transparência pública significa garantir que informações sobre receitas, despesas e decisões governamentais estejam disponíveis e possam ser efetivamente compreendidas pela sociedade. A Lei nº 12.527/2011

institui a publicidade como regra geral, enquanto a Lei Complementar nº 131/2009 corrobora a obrigação de divulgação de dados orçamentários em tempo real. Contudo, a simples publicação dos dados não elimina barreiras de acesso, especialmente quando as interfaces são complexas ou dependem de conhecimento técnico prévio (Rodrigues & Michener, 2018; Vianna & Ladeira, 2019).

O conceito de Governo 4.0 propõe o uso de tecnologias como inteligência artificial, análise de dados e automação para melhorar os serviços públicos e torná-los mais acessíveis (Janowski, 2015). Nessa perspectiva, uma interface conversacional pode funcionar como meio de mediação entre os dados brutos e o cidadão comum.

O conceito de Governo 4.0 situa-se no marco mais amplo da Indústria 4.0, paradigma que designa a integração de tecnologias como Internet das Coisas, inteligência artificial, computação em nuvem e análise massiva de dados na transformação de processos produtivos e organizacionais (Schwab, 2016). Embora a Indústria 4.0 tenha origem no setor manufatureiro, seus princípios — interoperabilidade, virtualização, descentralização, capacidade de operação em tempo real e orientação a serviço — vêm sendo progressivamente transpostos para a gestão pública, dando origem ao que a literatura denomina Governo 4.0 (Janowski, 2015). Nessa perspectiva, um chatbot que converte linguagem natural em consultas estruturadas sobre dados orçamentários constitui uma aplicação direta dos pilares da Indústria 4.0 ao domínio governamental: emprega inteligência artificial para automatizar a mediação entre dados brutos e o cidadão, opera sobre bases de dados abertas disponibilizadas em tempo real e oferece um serviço descentralizado, acessível de qualquer dispositivo conectado à internet. Assim, o presente trabalho insere-se na interseção entre Indústria 4.0 e transparência pública, utilizando tecnologias características desse paradigma para enfrentar o problema da acessibilidade dos dados orçamentários.

2.2. Interfaces em linguagem natural para bases de dados (NLIDB)

Interfaces em linguagem natural para bases de dados (NLIDB) são sistemas que permitem consultar uma base de dados usando perguntas em

linguagem comum, sem exigir que o usuário conheça linguagens formais de consulta. Uma das abordagens mais conhecidas é o *Text-to-SQL*, em que a pergunta do usuário é traduzida para SQL, linguagem padrão de consulta a bancos de dados.

Nessa metodologia, o benchmark Spider tornou-se uma referência importante (Yu et al., 2018). Um benchmark é um conjunto padronizado de testes usado para comparar métodos de forma justa. O Spider é classificado como *cross-domain*, ou seja, avalia os modelos em vários domínios temáticos, e não apenas em uma área específica. Além disso, ele exige generalização para *schemas* desconhecidos. O termo *schema* refere-se a estrutura do banco de dados: tabelas, colunas e relações entre elas. Fazendo uma analogia, o *schema* funciona como a planta de uma casa; para responder corretamente, o modelo precisa entender essa estrutura mesmo quando a vê pela primeira vez.

Embora relevantes, benchmarks como o Spider não foram projetados para dados orçamentários governamentais, que apresentam desafios particulares, como nomes longos de órgãos, siglas locais e alta exigência de exatidão monetária.

2.3. LLMs, alucinação e o risco em domínios financeiros

LLMs como GPT-4 e DeepSeek apresentam forte capacidade de interpretar linguagem natural, identificar intenções e extrair parâmetros de consulta (OpenAI, 2023; DeepSeek-AI, 2024). No entanto, esses modelos podem alucinar, isto é, gerar saídas textualmente convincentes, mas factualmente erradas (Huang et al., 2025). Em domínios financeiros, esse risco é especialmente crítico.

Uma estratégia para reduzir esse problema consiste em limitar o papel do LLM à interpretação da pergunta e delegar os cálculos a um mecanismo determinístico. Essa divisão caracteriza uma arquitetura neuro-simbólica: o componente neural interpreta a linguagem; o componente simbólico executa operações formais e reproduzíveis (Besold et al., 2021). No presente trabalho, o LLM produz uma especificação estruturada da consulta (*QuerySpec*), enquanto a execução é realizada por funções programadas em Pandas.

2.4. Resolução de entidades, execução determinística e métricas

Consultas orçamentárias frequentemente mencionam entidades com múltiplas formas de escrita, como nome completo, sigla, abreviação ou grafia com erro, o usuário pode estar buscando por pagamentos realizados a Companhia de águas e esgotos de Roraima e digitar “CAER” ou “CAERR”. Para lidar com isso, foi utilizada a técnica de TF-IDF com similaridade por cosseno, que representa textos numericamente e mede o grau de semelhança entre expressões (Manning et al., 2008). No trabalho, a técnica foi aplicada com n-gramas de caracteres para tornar a comparação mais robusta a pequenas variações ortográficas.

A execução das consultas foi feita em Pandas, biblioteca de análise de dados em Python que permite filtrar, agrupar e agregar informações em tabelas chamadas *DataFrames* (McKinney, 2012). Como as operações são pré-programadas, a execução é determinística, isto é, dada a mesma entrada, a saída será sempre a mesma.

A métrica principal do estudo é a *Execution Accuracy* (EX), definida como a proporção de perguntas cuja resposta final coincide com o gabarito. Para a estimativa do intervalo de confiança da acurácia, utilizou-se o método de Wilson, recomendado para proporções por apresentar melhor cobertura estatística do que alternativas assintóticas simples (Newcombe, 1998).

3. Trabalhos relacionados

Gao et al. (2023) investigaram o uso de LLMs em *Text-to-SQL* no *benchmark Spider*. Os autores avaliaram como diferentes estratégias de *prompt engineering* influenciam a geração de consultas SQL e reportaram 86,6% de *Execution Accuracy* com a abordagem DAIL-SQL. O estudo demonstrou que a qualidade do prompt afeta positivamente o desempenho do modelo, mas partiu de *schemas* SQL bem estruturados, o que se diferencia de cenários de transparência pública baseados em arquivos tabulares, geralmente CSV.

Liu et al. (2024) publicaram um *survey* sobre *Text-to-SQL* na era dos LLMs. O estudo organiza a literatura em etapas como pré-processamento,

schema linking, geração e pós-processamento, e identifica como lacunas importantes a confiabilidade em domínios especializados e a necessidade de maior precisão em aplicações reais. A principal contribuição do *survey* é mostrar que o avanço em benchmarks não elimina os desafios de implantação em contextos de produção.

Zhu et al. (2024) realizaram outro *survey* abrangente sobre *Text-to-SQL* com LLMs, classificando as abordagens em prompt *engineering*, *fine-tuning*, pré-treinamento especializado e agentes. Os autores destacam que a métrica mais robusta é a *Execution Accuracy*, pois ela avalia o resultado da consulta e não apenas a forma da resposta gerada. O estudo também reforça que a generalização para domínios específicos continua sendo um problema em aberto.

Li et al. (2024) analisaram erros em código gerado por LLMs e identificaram falhas recorrentes, como alucinação de nomes de colunas, inversão de operadores e cálculos incorretos. Esse resultado é diretamente relevante para este trabalho, pois sustenta a decisão arquitetural de não delegar a execução dos cálculos ao modelo de linguagem.

Huang et al. (2025) revisaram o fenômeno das alucinações em LLMs e sistematizaram causas, categorias e desafios em aberto. Os autores mostram que o problema não se restringe à geração textual livre e afeta tarefas estruturadas, especialmente quando há exigência de consistência factual. Para aplicações com valores monetários, essa limitação é particularmente crítica.

No contexto brasileiro, Vianna e Ladeira (2019) discutem o reuso de dados governamentais abertos e mostram que a simples abertura dos dados não garante apropriação social. Os autores defendem a existência de mecanismos de mediação que tornem os dados efetivamente utilizáveis por cidadãos não especialistas.

Com base nesses estudos, observamos que a principal lacuna da literatura está na escassez de soluções avaliadas com dados orçamentários reais de governos brasileiros, com execução determinística e mecanismos robustos de resolução de entidades para lidar com siglas, nomes extensos e variações locais. Esta pesquisa busca avançar justamente nesse ponto.

4. Solução computacional

A solução desenvolvida consiste em uma aplicação web construída em Streamlit para responder perguntas em linguagem natural sobre a execução orçamentária do Governo do Estado de Roraima. A arquitetura segue um pipeline híbrido com cinco etapas: preparação dos dados, *parsing* heurístico, *fallback* por LLM, resolução de entidades e execução determinística.

No módulo de ETL, os dados brutos em CSV foram carregados e transformados para uso analítico. Esse processo incluiu padronização de colunas textuais, conversão de valores monetários, criação de campos temporais canônicos e construção de um índice de entidades para seis colunas principais: unidade orçamentária, função, fonte de recurso, natureza da despesa, programa e credor. Também foi criado um mapa de *aliases* (dicionário de tradução interno) para relacionar siglas e nomes canônicos de órgãos.

A interpretação inicial da pergunta ocorre por meio de um parser heurístico determinístico (Camada Simbólica), baseado em expressões regulares. Esse parser busca extrair os parâmetros essenciais da consulta (intenção, métrica, filtros temporais e entidade). O sistema atribui uma pontuação de confiança inicial de 100% (1.0) à análise, aplicando penalidades dedutivas caso não consiga identificar parâmetros críticos. Quando a variação semântica ou a coloquialidade da frase fazem com que a confiança dessa heurística caia abaixo do limiar de segurança de 70%, o sistema aciona a Camada Neural (LLM) como *fallback*. Quando isso ocorre, o modelo generativo não calcula a resposta final, mas atua estritamente como um roteador semântico, gerando uma especificação estruturada da consulta (*QuerySpec*) para execução posterior.

Na etapa seguinte, o sistema executa a resolução de entidades. Primeiro tenta correspondência por aliás (dicionário de tradução interno) e correspondência exata; em seguida, utiliza TF-IDF com similaridade por cosseno. Quando a coluna sugerida inicialmente não produz boa correspondência, o sistema realiza uma busca global nas seis colunas de entidades. Se houver múltiplos candidatos muito próximos, o *chatbot* apresenta um menu de desambiguação; caso a similaridade seja insuficiente, o sistema informa que a entidade não foi encontrada.

A execução final da consulta é feita em Pandas, com filtros por ano, mês e entidade, seguidos de operações de agregação, diferença, percentual ou ranking, conforme a intenção detectada. Esse desenho evita a geração de código executável pelo LLM e reduz o risco de erro aritmético.

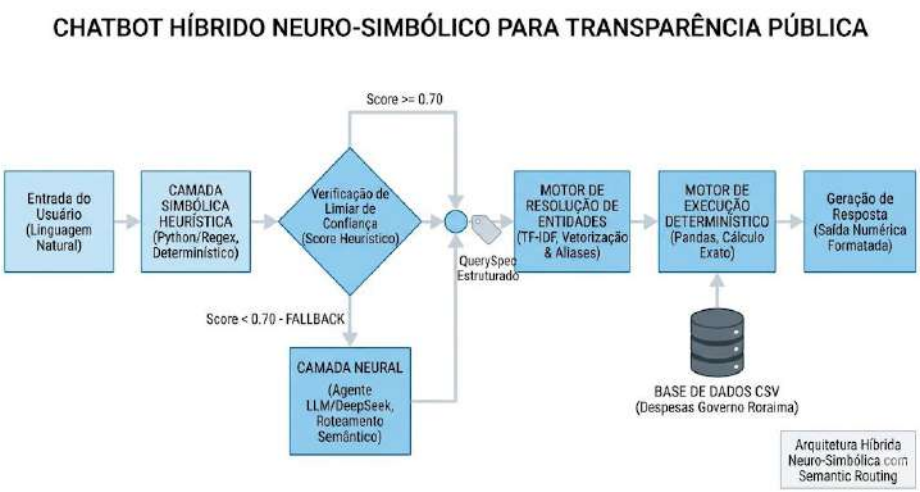


Figura 1. Arquitetura *chatbot*.

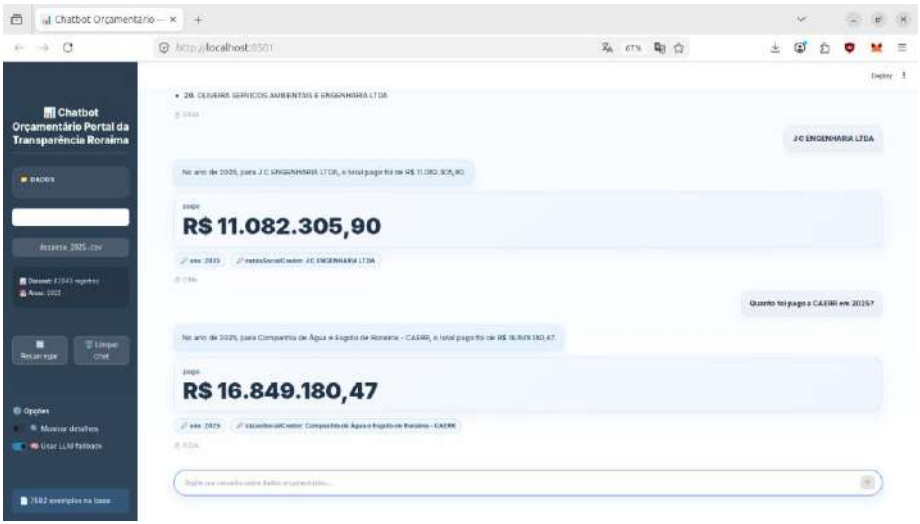


Figura 2. Interface do *chatbot* desenvolvido em Streamlit, exibindo uma consulta em linguagem natural sobre a execução orçamentária do Governo de Roraima.

5. Avaliação experimental

5.1. Planejamento de pesquisa

A avaliação adotou o modelo de testes offline com gabarito estático (*ground truth*), padrão amplamente utilizado em pesquisas de *Text-to-SQL* e NLIDB. Neste contexto metodológico, o termo 'offline' indica que o sistema foi validado de forma automatizada em lote contra um conjunto de testes pré-compilado, sem a interação humana em tempo real, embora o *chatbot* continuasse a acionar a API do LLM para o roteamento semântico durante o experimento. Para isso, foi construído um dataset adversarial por meio de um script que gerou perguntas em quatro níveis de complexidade: agregação, simples, comparativo e complexo. Em seguida, um LLM foi utilizado para produzir paráfrases das perguntas, aproximando o teste da linguagem coloquial do cidadão, sendo também injetados erros de digitação sintéticos (*typos*) em 20% dos itens.

A amostra de estresse final foi composta por 6.656 consultas, distribuídas em 5.427 de agregação, 879 simples, 100 comparativas e 250 complexas. A expressiva predominância de perguntas de agregação foi desenhada intencionalmente por razões matemáticas e de aderência ao mundo real. Enquanto o nível 'simples' avalia consultas macroscópicas limitadas à totalização de Órgãos governamentais (um universo pequeno de dezenas de secretarias), o nível de 'agregação' exige que o sistema filtre e consolide milhares de rubricas granulares — como os 1.000 maiores credores, os 500 maiores programas (PAOE) e fontes de recurso — multiplicando estas entidades por recortes anuais e mensais. Essa explosão combinatória reflete o comportamento prático esperado num cenário de Governo 4.0, onde a esmagadora maioria das buscas do cidadão ou da imprensa visa descobrir o montante agregado pago a uma empresa específica ou projeto exato.

Do total de 6.656 consultas, 6.456 itens esperavam uma resposta numérica direta e 200 esperavam o acionamento de um menu de desambiguação por parte do *chatbot*. O gabarito matemático para cada pergunta foi calculado deterministicamente a partir do próprio dataset de despesas em formato CSV, o que garantiu uma comparação automatizada, rigorosa e sem viés humano entre a resposta gerada pelo *chatbot* e o valor financeiro exato esperado.

Tabela 1. Distribuição das consultas por nível de complexidade.

Nível	Itens	Proporção (%)
Agregação	5.427	81,83
Simples	879	13,21
Comparativo	100	1,50
Complexo	250	3,76
Total	**6.656**	**100,00**

O validador executou cada pergunta no *chatbot*, extraiu a resposta numérica e a comparou com o gabarito dentro de tolerância relativa de 1%. Nos itens *multi-turn*, o critério de acerto exigiu o disparo do menu de desambiguação e a presença da entidade correta entre as opções.

Foram testadas as hipóteses: H_0: EX ≤ 0,80 versus H_1: EX > 0,80, por teste binomial unilateral; independência entre acurácia e nível de complexidade, por qui-quadrado; diferença de latência entre níveis, por Kruskal-Wallis; e correlação entre complexidade ordinal e latência, por Spearman, nível de significância de 5%.

5.2. Resultados e Discussão

O *chatbot* obteve 5.840 acertos em 6.656 consultas, o que corresponde a *Execution Accuracy* global de 87,74%, com IC 95% de Wilson entre 86,93% e 88,51%. O teste binomial unilateral rejeitou H_0 ($p < 0,001$), indicando que a acurácia do sistema supera significativamente o limiar de 80% estabelecido para viabilidade prática.

A Tabela 2 mostra que o melhor desempenho ocorreu nas consultas de agregação, com 91,49% de acerto. Esse é um resultado importante, pois esse tipo de pergunta tende a representar a necessidade mais comum de uso cidadão, como descobrir quanto foi pago a determinado órgão, programa ou credor. As consultas simples e comparativas também permaneceram acima de 80%, com 84,41% e 82,00%, respectivamente. O teste qui-quadrado indicou diferença estatisticamente significativa entre os níveis (χ^2 =

1.136,86; $gl = 3$; $p < 0{,}001$), o que confirma que a complexidade da consulta impacta a acurácia.

Tabela 2. Execution Accuracy por nível de complexidade.

Nível	Acertos	Total	Acurácia (%)
Agregação	4.965	5.427	91,49
Simples	742	879	84,41
Comparativo	82	100	82,00
Complexo	51	250	20,40
Total	5.840	6.656	87,74

O pior resultado ocorreu nas consultas complexas, com apenas 20,40% de acerto. Esse grupo reúne perguntas com múltiplos parâmetros, percentuais, Top N e casos em que mais de uma entidade relevante aparece simultaneamente na mesma frase. Esse achado é coerente com a literatura, que aponta maior dificuldade dos LLMs e dos mecanismos de *linking* semântico quando o problema exige múltiplas restrições e maior capacidade de generalização (Liu et al., 2024; Zhu et al., 2024).

Nas consultas numéricas de resposta direta, a acurácia foi de 90,07% (5.815/6.456). Já nos casos *multi-turn*, o índice caiu para 12,50% (25/200). Esse resultado, porém, deve ser interpretado com cautela: em parte dos itens, a busca global por TF-IDF resolveu corretamente a entidade e retornou o valor final sem necessidade de menu. Assim, o sistema foi penalizado pelo avaliador mesmo quando a resposta final estava correta, o que constitui uma limitação do protocolo de avaliação e não necessariamente uma falha funcional do *chatbot*.

Quanto à latência, a média foi de 4,94 s, a mediana de 5,15 s, o percentil 95 de 6,09 s, o mínimo de 0,02 s e o máximo de 51,30 s. O teste de Kruskal-Wallis indicou diferença entre níveis ($H = 53{,}94$; $p < 0{,}001$), enquanto a correlação de Spearman entre complexidade e latência foi fraca ($\rho = -0{,}075$; $p < 0{,}001$). Isso sugere que a latência depende mais do acionamento do LLM do que da complexidade semântica em si. Em uso real,

consultas mais diretas tendem a ser respondidas apenas pela heurística, com tempo substancialmente menor.

A resolução de entidades também apresentou resultado satisfatório: 93,29% das consultas obtiveram status *ok*, enquanto 3,31% foram classificadas como *ambiguous* e 2,40% como *not_found*. Esse comportamento é relevante, pois mostra que o sistema prefere não assumir uma correspondência fraca quando a entidade não pode ser resolvida com segurança, o que é desejável em contexto orçamentário.

Em comparação com Gao et al. (2023), que reportaram 86,6% no *benchmark Spider*, o desempenho global deste trabalho é competitivo, embora a comparação deva ser feita com cautela, já que os domínios e os critérios de teste são distintos. A principal vantagem da presente abordagem está menos na geração livre de consultas e mais na combinação entre flexibilidade semântica e segurança computacional, alinhando-se às recomendações de Li et al. (2024) sobre os riscos da geração direta de código por LLMs.

Este estudo possui limitações. Primeiro, o baixo desempenho em consultas complexas mostra que a arquitetura ainda não lida bem com múltiplos filtros semânticos simultâneos. Segundo, a avaliação foi feita com gabarito gerado automaticamente, sem experimento com usuários reais. Terceiro, o dataset contempla apenas o exercício fiscal de 2025, o que limita a análise de generalização temporal.

6. Conclusão

O objetivo deste estudo foi desenvolver e avaliar experimentalmente um *chatbot* para consultas em linguagem natural sobre dados de execução orçamentária do Governo do Estado de Roraima, exercício fiscal de 2025. Para atingir esse objetivo, foi proposta uma arquitetura híbrida neuro-simbólica que combina parser heurístico, *fallback* por LLM, resolução de entidades por TF-IDF e execução determinística em Pandas, seguida de avaliação com 6.656 perguntas adversariais.

Verificamos que o sistema alcançou *Execution Accuracy* global de 87,74%, com intervalo de confiança de Wilson entre 86,93% e 88,51%,

superando significativamente o limiar de 80%. Em relação às perguntas de pesquisa, concluímos que: P1 – a acurácia global supera 80%; P2 – a acurácia varia significativamente conforme a complexidade da consulta; P3 – a latência mediana de 5,15 segundos é compatível com uso interativo no cenário avaliado; P4 – a resolução de entidades por TF-IDF mostrou-se eficaz na maior parte dos casos.

Como principal contribuição, o estudo demonstra que a combinação entre interpretação semântica por LLM e execução determinística constitui uma alternativa viável para democratizar o acesso a dados orçamentários sem comprometer a precisão numérica. Do ponto de vista prático, o chatbot atua como uma camada de mediação entre o portal de transparência e o cidadão sem conhecimentos técnicos aprofundados sobre planilhas ou banco de dados.

A sugestão para estudos futuros é pelo aprofundamento das limitações observadas, especialmente por meio de: (I) aprimoramento do reconhecimento de múltiplas entidades na mesma pergunta; (II) expansão para outros exercícios fiscais e outros estados; (III) avaliação com usuários reais; (IV) refinamento supervisionado da resolução de entidades; e (V) comparação com arquiteturas alternativas, como agentes *Text-to-SQL* com autocorreção.

7. Referências

Besold, T. R., Garcez, A. S. d'A., Bader, S., Bowman, H., Domingos, P. M., Hitzler, P., Kühnberger, K.-U., Lamb, L. C., Lima, P. M. V., de Penning, L., Pinkas, G., Poon, H., & Zaverucha, G. (2021). Neural-symbolic learning and reasoning: A survey and interpretation. In P. Hitzler & M. K. Sarker (Eds.), *Neuro-symbolic artificial intelligence: The state of the art* (pp. 1–51). IOS Press. https://doi.org/10.3233/FAIA210348

Brasil. (2011). *Lei nº 12.527, de 18 de novembro de 2011. Regula o acesso a informações*. Diário Oficial da União. Recuperado de: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm

DeepSeek-AI. (2024). *DeepSeek LLM: Scaling open-source language models with longtermism*. arXiv. https://doi.org/10.48550/arXiv.2401.02954

Gao, D., Wang, H., Li, Y., Sun, X., Qian, Y., Ding, B., & Zhou, J. (2023). *Text-to-SQL empowered by large language models: A benchmark evaluation*. arXiv. https://doi.org/10.48550/arXiv.2308.15363

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems, 43*(2), 1–55. https://doi.org/10.1145/3703155

Janowski, T. (2015). Digital government evolution: From transformation to contextualization. *Government Information Quarterly*, *32*(3), 221–236. https://doi.org/10.1016/j.giq.2015.07.001

Li, H., Zhang, Y., Wang, F., Wang, Y., & Li, J. (2024). *What's wrong with your code generated by large language models?* arXiv. https://doi.org/10.48550/arXiv.2407.02914

Liu, X., Shen, S., Li, B., Ma, P., Jiang, R., Zhang, Y., Fan, J., Li, G., Tang, N., & Luo, Y. (2024). *A survey of Text-to-SQL in the era of large language models*. arXiv. https://doi.org/10.48550/arXiv.2408.05109

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. Recuperado de: https://nlp.stanford.edu/IR-book/

McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine, 17*(8), 857–872. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E

OpenAI. (2023). *GPT-4 technical report*. arXiv. https://doi.org/10.48550/arXiv.2303.08774

Rodrigues, K. F., & Michener, R. G. (2018). A necessidade de identificação como barreira ao acesso à informação: Evidências e práticas no Brasil e no mundo. *Administração Pública e Gestão Social*, *10*(4), 303–315. https://doi.org/10.21118/apgs.v10i4.5733

Schwab, K. (2016). *The fourth industrial revolution*. Crown Business.

United Nations, Department of Economic and Social Affairs. (2024). *United Nations e-government survey 2024: Accelerating digital transformation for sustainable development*. United Nations. https://doi.org/10.18356/9789211067286

Vianna, R. P., & Ladeira, M. J. (2019). Métrica alternativa para dados governamentais abertos na América Latina. *Transinformação, 31*, e180060. https://doi.org/10.1590/2318-0889201931e180060

Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., Zhang, Z., & Radev, D. (2018). Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3911–3921). https://doi.org/10.18653/v1/D18-1425

Zhu, X., Li, Q., Cui, L., & Liu, Y. (2024). *Large language model enhanced Text-to-SQL generation: A survey*. Recuperado de: https://doi.org/10.48550/arXiv.2410.06011

Capítulo 28

Sistema IoT de baixo custo para monitoramento de vibração em motores industriais baseado em ESP32 e norma ISO 10816

William Lopes Rodrigues
Leandro Nelinho Balico
Universidade Federal de Roraima

williamlopesrads@gmail.com
leandro.balico@ufrr.com

1. Introdução

No contexto das plantas fabris de pequeno porte da região Norte do Brasil, verifica-se uma lacuna na implementação de tecnologias avançadas. Tal limitação decorre, predominantemente, do elevado investimento exigido por sistemas proprietários de monitoramento industrial (Costa, 2021; Ferreira, 2024). A manutenção preditiva em máquinas rotativas é essencial para evitar falhas, mas o alto custo de sistemas proprietários impede sua adoção em pequenas indústrias da Região Norte, criando uma barreira financeira e tecnológica que limita a modernização dessas plantas fabris.

Pesquisas recentes demonstram que soluções de baixo custo baseadas em microcontroladores ESP32 e sensores Micro-Electro-Mechanical Systems (MEMS) podem oferecer alternativas viáveis. Costa (2021) validou a viabilidade técnica de um protótipo acessível para análise de vibração, ainda que limitado pela suscetibilidade a ruídos. Ferreira (2024), por sua vez, ampliou o escopo de variáveis monitoradas e integrou protocolos de comunicação e análise em nuvem, evidenciando maior robustez, mas com dependência de infraestrutura tecnológica mais complexa. Nascimento (2023) reforçou a aplicabilidade em ambientes críticos, como embarcações militares, ao integrar análise avançada de sinais e protocolos IoT, enquanto Pudo et al. (2025) destacaram a democratização da Indústria 4.0 para pequenas e médias indústrias por meio de parâmetros normativos da ISO 10816 (ISO, 2016).

Diante desse cenário, o objetivo central deste trabalho é desenvolver e avaliar um artefato de baixo custo (ESP32 + MPU6050) capaz de realizar o monitoramento de vibração e classificação automática via norma ISO 10816 em tempo real. As principais contribuições são: implementação de um protótipo funcional de baixo custo; integração de processamento em borda com protocolos IoT; validação experimental frente à norma ISO 10816; e disponibilização de uma alternativa viável para pequenas indústrias da região Norte.

O presente trabalho está organizado da seguinte forma: a Seção 2 apresenta a fundamentação teórica sobre a norma ISO 10816 e processamento digital de sinais em computação de borda; a Seção 3 discute os trabalhos relacionados; a Seção 4 descreve a solução computacional desenvolvida; a Seção 5 apresenta a avaliação experimental e os resultados obtidos; e a Seção 6 expõe as conclusões do estudo realizado.

2. Fundamentação teórica

A inserção da manutenção industrial no paradigma da Indústria 4.0 pressupõe a transição de sistemas isolados para ecossistemas integrados por meio da Internet das Coisas (IoT). Nesse contexto, a manutenção preditiva baseada na análise de vibração fundamenta-se no princípio de que máquinas rotativas em operação geram forças dinâmicas cujas amplitudes e frequências refletem seu estado mecânico (Vavassori et al., 2022). Para que sistemas de monitoramento sejam eficazes, é necessário transformar os dados brutos de aceleração em métricas normalizadas que permitam a tomada de decisão industrial em tempo real.

Essa digitalização de ativos é viabilizada por tecnologias de Processamento de Borda (Edge Computing), nas quais o microcontrolador ESP32 (Figura 1), atua como unidade central de processamento e comunicação. O ESP32 destaca-se por sua arquitetura dual-core e suporte nativo aos protocolos de rede sem fio, permitindo que a análise de sinais ocorra na origem do dado (Costa, 2021). Integrado a ele, o uso de sensores de tecnologia MEMS, como o acelerômetro MPU6050 (Figura 2), possibilitam a captura de vibrações triaxiais com baixo custo e alta integração,

características fundamentais para a escalabilidade de soluções IoT em plantas industriais (Ferreira, 2024).



Figura 1. Placa de desenvolvimento ESP32.



Figura 2. Unidade de Medição Inercial (IMU) MPU6050.

2.1. Norma ISO 10816 e Severidade Vibratória

A norma ISO 10816 (ISO, 2016) estabelece as diretrizes globais para a avaliação da vibração de máquinas por meio de medições em partes não rotativas. O parâmetro principal adotado por esta norma é a velocidade vibratória eficaz, expressa pelo valor RMS (do inglês *Root Mean Square* — Valor Quadrático Médio), medida em mm/s, que representa a energia total contida no sinal de vibração. É importante distinguir este parâmetro do cálculo de RMS aplicado aos sinais de aceleração bruta, descrito na Seção 2.2: naquela seção, o *RMS* é calculado sobre amostras de aceleração (m/s^2) como etapa intermediária de processamento, enquanto a norma ISO 10816 define limites com base na velocidade vibratória (mm/s), que é a grandeza física padronizada para avaliação da condição da máquina.

A classificação da condição da máquina depende de sua potência e do tipo de fixação (fundação rígida ou flexível). Para motores de pequena potência, a norma define zonas de severidade que orientam o operador: Zona A — máquinas novas ou recondicionadas (verde escuro); Zona B — máquinas

em condição aceitável para operação contínua sem restrições (verde claro); Zona C — condição insatisfatória, com alerta para monitoramento intensivo ou manutenção próxima (laranja); e Zona D — níveis de vibração perigosos, com risco de danos imediatos, exigindo a parada do ativo (vermelho).

2.2. Processamento Digital de Sinais em *Edge Computing*

Diferentemente de sistemas convencionais que enviam dados brutos diretamente para a nuvem, a utilização de microcontroladores como o ESP32 permite a aplicação de *Edge Computing* (Computação de Borda). Nesse contexto, o processamento ocorre próximo à origem dos dados, reduzindo a latência, otimizando o uso da rede e possibilitando respostas em tempo real (NASCIMENTO, 2023).

No contexto da manutenção preditiva baseada em análise de vibração, o sensor MPU6050 captura sinais de aceleração triaxiais. Esses dados são pré-processados localmente pelo ESP32, que atua como nó de borda. Entre as técnicas de Processamento Digital de Sinais (PDS) aplicáveis, destacam-se (Vavassori et al., 2022): (a) filtragem digital — remoção de ruídos e componentes indesejados; (b) cálculo de valores *RMS* (*Root Mean Square*) — métrica que fornece uma medida da intensidade da vibração, calculada sobre os sinais de aceleração bruta como etapa intermediária antes da conversão para a grandeza normalizada da ISO 10816; (c) Transformada Rápida de Fourier (FFT) — análise espectral para identificar frequências características de falhas mecânicas; e (d) detecção de picos e envoltória do sinal — identificação de padrões anômalos.

Essas operações permitem que apenas informações relevantes sejam transmitidas para sistemas superiores, como servidores em nuvem ou plataformas de análise avançada. Dessa forma, a computação de borda não se limita à coleta e transmissão bruta de dados, mas envolve a capacidade de realizar análises preliminares e extrair indicadores de condição diretamente no dispositivo de borda. Essa abordagem é fundamental para ambientes industriais, pois diminui a dependência da conectividade contínua, reduz custos de transmissão de dados e aumenta a confiabilidade das decisões em tempo real (Pudo et al., 2025).

Nesse sentido, Bastos (2021) demonstrou na prática a viabilidade do ESP32 como plataforma central em sistemas de monitoramento de motores elétricos de baixo custo. Em seu protótipo, o ESP32 integrou sensores de vibração, temperatura, umidade e corrente elétrica, comunicando-se via protocolo *Modbus* com um sistema supervisório SCADA gratuito. Os resultados evidenciaram que o microcontrolador foi capaz de adquirir, processar e transmitir dados operacionais em tempo real via Wi-Fi, além de executar lógicas de proteção automática com geração de alertas de falha (BASTOS, 2021). Essa experiência reforça o potencial do ESP32 como nó inteligente de borda, sustentando a abordagem adotada no presente trabalho, que avança ao incorporar a classificação normativa via ISO 10816 diretamente no dispositivo.

3. Trabalhos relacionados

A literatura científica recente demonstra um interesse crescente no uso de sensores de baixo custo para o monitoramento de ativos industriais, fundamentando a viabilidade técnica desta pesquisa. Esta seção apresenta, os cinco trabalhos de maior relevância para o escopo desta pesquisa.

Costa (2021) desenvolveu um protótipo de sistema de monitoramento remoto aplicado a um motor de indução trifásico. A solução empregou sensores *MEMS*, cálculo de RMS para análise de vibração e comunicação IoT via ESP32. Os resultados demonstraram viabilidade técnica e baixo custo de implementação. Verificou-se, contudo, que o sistema apresentou limitações relacionadas à suscetibilidade a ruídos, o que pode comprometer a precisão da leitura em ambientes industriais com alta interferência eletromagnética. O trabalho constitui uma contribuição relevante ao demonstrar que *hardware* de baixo custo pode ser empregado de forma funcional em contextos de manutenção preditiva.

Ferreira (2024) concentrou-se em motores monofásicos e ampliou o escopo de variáveis monitoradas ao incluir tensão, corrente, temperatura e vibração. O sistema integrou protocolos *Message Queuing Telemetry Transport* (MQTT), armazenamento em nuvem e análise de dados via *Power BI*, resultando em um ecossistema mais robusto. Os dashboards interativos e a transmissão confiável de dados conferem ao sistema maior adequação a

ambientes industriais de média complexidade. Como contrapartida, o autor reconhece desafios relacionados à segurança e à integridade da rede, evidenciando que a ampliação do escopo funcional implica maior dependência de infraestrutura tecnológica.

Nascimento (2023) aplicou seu sistema supervisório em motores de indução trifásicos embarcados em navios militares, contexto de elevada criticidade operacional. A solução integrou análise avançada de sinais — incluindo FFT, filtros digitais e métricas estatísticas — com protocolos IoT. Os resultados demonstraram impacto positivo na segurança naval, evidenciando a aplicabilidade de soluções de baixo custo em ambientes de missão crítica. O trabalho aponta, entretanto, a necessidade de avanços em inteligência artificial embarcada e em segurança cibernética para alcançar a plena confiabilidade operacional.

Pudo et al. (2025) validaram um protótipo de detecção de desgaste em rolamentos de máquinas-ferramenta em pequenas e médias indústrias (PMIs). Utilizando ESP32, sensores de vibração e temperatura, os autores empregaram RMS, filtros de média móvel e os parâmetros normativos da ISO 10816. Os resultados destacaram a democratização da Indústria 4.0 para PMIs. Os autores reconhecem, no entanto, limitações quanto à adoção de limiares estáticos — sugerindo a implementação de limites dinâmicos — e quanto à sensibilidade térmica, aspectos que indicam margem para evolução do protótipo em direção à robustez plena.

Vavassori et al. (2022) conduziram experimentos em ambiente laboratorial, utilizando bancadas de teste com motores monofásicos submetidos a falhas de desbalanceamento e variações de temperatura. As técnicas empregadas incluíram análise de vibração e temperatura, aplicação da FFT, comunicação IoT via ESP32 com protocolo ModBus TCP/IP e supervisão dos dados pelo LabVIEW®. Os resultados mostraram que o protótipo foi capaz de identificar vibrações associadas ao desbalanceamento e monitorar temperatura. Contudo, o sistema apresentou ruídos e inconsistências em frequências mais altas e em medições térmicas elevadas, indicando a necessidade de sensores mais precisos e maior taxa de amostragem para alcançar a confiabilidade dos equipamentos comerciais.

Com base na análise desses cinco estudos, foi constatado que a principal lacuna existente na literatura é a integração de um sistema baseado

em ESP32 que não apenas colete dados, mas que realize a classificação automática da severidade vibratória em tempo real com base estrita na norma ISO 10816, fornecendo um diagnóstico normativo imediato ao operador. Essa lacuna é relevante porque confere o rigor técnico necessário para que soluções de baixo custo sejam aceitas em ambientes industriais reais, garantindo segurança e conformidade técnica.

4. Solução computacional

A solução computacional proposta baseia-se na arquitetura de Computação de Borda (*Edge Computing*), estruturada para mitigar os gargalos de latência e largura de banda inerentes aos sistemas de monitoramento em nuvem puramente centralizados. Nesta abordagem, o microcontrolador ESP32 atua como um nó inteligente na extremidade da rede, realizando a extração de características do sinal de vibração localmente, antes de qualquer transmissão de dados.

A arquitetura do sistema é composta por quatro camadas funcionais, conforme ilustrado na Figura 3: (1) Camada de Sensoriamento — o sensor *MEMS* MPU6050 realiza a aquisição dos sinais de aceleração triaxial (Ax, Ay, Az) a uma taxa efetiva de 25 Hz; (2) Camada de *Edge Computing* — o microcontrolador ESP32 executa localmente o processamento do sinal, o cálculo do valor eficaz RMS e a classificação do estado operacional conforme os limiares da norma ISO 10816, determinando autonomamente o *status* da máquina (Normal, Alerta ou Perigo); (3) Camada de Comunicação — os dados processados são transmitidos via protocolo MQTT para um *broker* local; e (4) Camada de Aplicação — os dados são visualizados em tempo real por meio de um *dashboard* IoT.

As principais vantagens dessa arquitetura distribuída são: (a) redução de tráfego de rede, uma vez que apenas o diagnóstico processado — contendo o valor RMS, a zona ISO 10816 e o *status* operacional — é transmitido, em vez de fluxos contínuos de dados brutos de aceleração triaxial; e (b) resiliência operacional, pois o sistema mantém a capacidade de monitoramento e classificação local mesmo em situações de instabilidade na conectividade de rede.

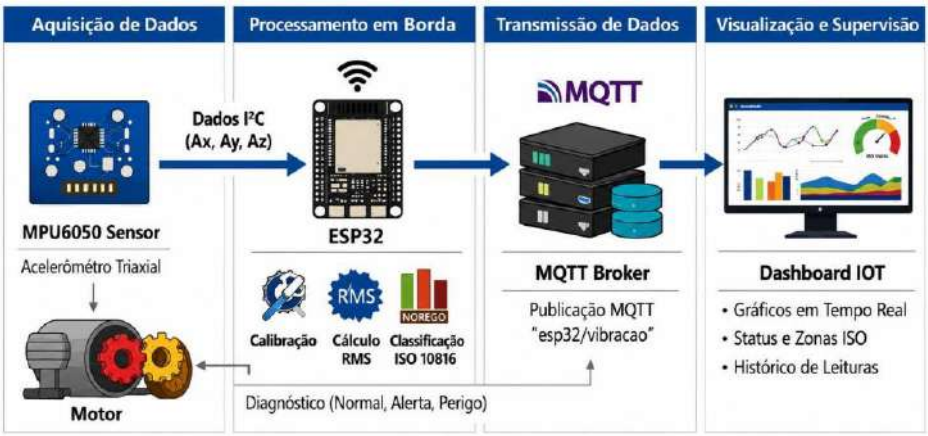


Figura 3. Diagrama de arquitetura do sistema de monitoramento vibratório baseado em Edge Computing.

4.1. Simulação e Prototipagem

A etapa de prototipagem compreendeu a seleção de componentes, a integração física do *hardware* e o desenvolvimento do *firmware* embarcado. O sensor MPU6050 foi conectado ao ESP32 via protocolo I^2C, com as seguintes ligações: SDA→GPIO21, SCL→GPIO22, VCC→3,3 V e GND→GND, conforme ilustrado na Figura 4. Essa interface serial de dois fios permite a comunicação eficiente entre os dispositivos com mínimo uso de pinos do microcontrolador.

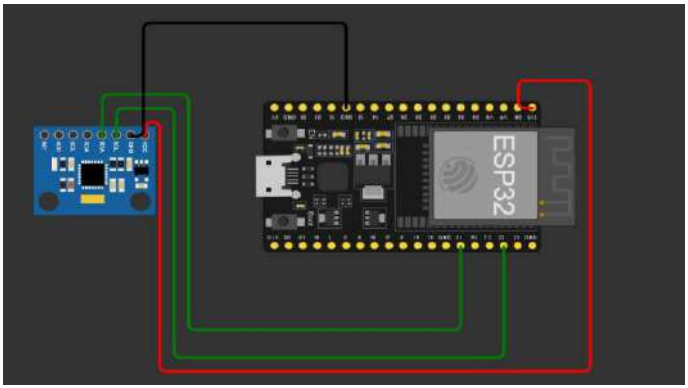


Figura 4. Diagrama de conexão do protótipo: sensor MPU6050 conectado ao ESP32 via protocolo I^2C.

O protótipo foi fixado diretamente no motor de um ventilador doméstico para a realização dos testes experimentais. A Figura 5 ilustra o dispositivo instalado no motor, evidenciando a compacidade da solução e sua adequação ao ambiente de teste. O custo total dos componentes de *hardware* foi estimado em aproximadamente R$ 80,00, o que demonstra a viabilidade econômica da solução para pequenas indústrias.

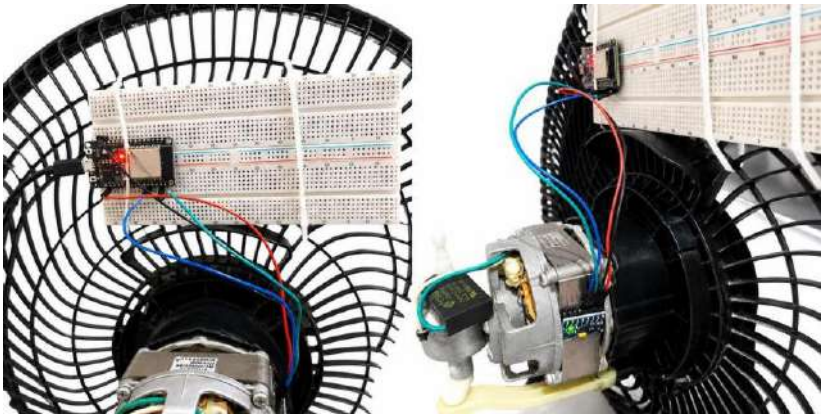


Figura 5. Protótipo instalado no motor do ventilador para testes experimentais.

4.2. Interface e Comunicação IoT via MQTT

A comunicação entre a camada de borda e a camada de aplicação é realizada por meio do protocolo MQTT (*Message Queuing Telemetry Transport*). Os dados processados são publicados no tópico esp32/vibracao no formato JSON, contendo o valor RMS, a zona ISO 10816, o *status* operacional, os valores de aceleração triaxial (Ax, Ay, Az) e o *timestamp* da leitura. A estrutura do *payload* JSON transmitido é apresentada abaixo, conforme a NBR 6023 (ABNT, 2018) para formatação de trechos de código em trabalhos acadêmicos:

```
{
  "rms": 1.4021,
  "zona": "B",
  "status": "NORMAL",
  "ax": 0.8625, "ay": 0.6465, "az": 0.5438,
  "timestamp": "2026-04-10T21:32:00Z"
}
```

Essa estrutura permite que o *dashboard* de monitoramento consuma informações já pré-processadas pelo *hardware* de borda, eliminando a necessidade de processamento adicional na camada de aplicação.



Figura 6. Dashboard.

5. Avaliação experimental

5.1. Planejamento de pesquisa

Esta seção apresenta a avaliação experimental do sistema de monitoramento de vibração desenvolvido, conduzida por meio de ensaios controlados com um motor de indução monofásico de ventilador doméstico. O objetivo é verificar a capacidade do protótipo de detectar e classificar corretamente diferentes níveis de desbalanceamento mecânico, utilizando como referência normativa os limiares de severidade vibratória estabelecidos pela ISO 10816 (ISO, 2016). Para tanto, foram realizados quatro experimentos com massas excêntricas progressivamente crescentes (0 g, 2 g, 4 g e 8 g), totalizando 120 leituras de RMS, analisadas quanto à distribuição por zona e à taxa de classificação correta do sistema.

O desbalanceamento foi induzido mediante fixação de massas excêntricas nas pás da hélice do ventilador, utilizando pesos plásticos de massas conhecidas e previamente aferidas em balança de precisão, pesando 2 g cada. Para cada nível de desbalanceamento, o peso foi posicionado na extremidade de uma pá, de forma a maximizar o momento de inércia e, consequentemente, a força centrífuga gerada durante a rotação. O processo de inserção do peso foi realizado com o motor desligado e o protótipo de monitoramento devidamente fixado ao corpo do motor, garantindo a reprodutibilidade das condições experimentais. A Figura 6 ilustra o procedimento de inserção das massas nas pás da hélice.



Figura 7. Procedimento de inserção de massa excêntrica na hélice do ventilador.

5.2. Dados Coletados e Cenário de Avaliação

Os dados coletados consistem em sinais de aceleração triaxial (Ax, Ay, Az) amostrados pelo sensor MEMS MPU6050 a uma taxa efetiva de 25 Hz. Para cada ciclo de medição, o ESP32 captura uma janela de 50 amostras (aproximadamente 2 segundos) e calcula um único valor RMS conforme Equação 1. Esse valor RMS constitui uma grandeza intermediária de processamento e representa uma leitura. Foram registradas 30 leituras por experimento, totalizando 120 pontos de dados nos quatro experimentos. Cada leitura então comparada com os limiares da ISO 10816, para classificação da severidade vibratória.

Previamente ao início da coleta, o sistema executa uma rotina de calibração automática com 200 amostras em repouso, removendo a componente estática da gravidade de cada eixo. A partir do processamento desses dados no nó de borda (ESP32), são extraídas as seguintes métricas:

$$RMS = \sqrt{[(1/N) \times \sum \quad (Ax_i^2 + Ay_i^2 + Az_i^2)]} \quad (1)$$

O *status* diagnóstico é uma variável categórica gerada pela lógica de decisão local: NORMAL (Zonas A e B), ALERTA (Zona C) ou PERIGO (Zona D). A classificação por zona ISO 10816 obedece aos seguintes limiares: Zona A — abaixo de 0,71 m/s^2; Zona B — entre 0,71 e 2,8 m/s^2; Zona C — entre 2,8 e 7,1 m/s^2; e Zona D — acima de 7,1 m/s^2.

Os quatro cenários experimentais foram definidos como: Experimento 1 — estado saudável (0 g), com o motor operado sem alterações mecânicas, estabelecendo o baseline vibratório, com zona esperada A ou B; Experimento 2 — desbalanceamento leve (2 g), com massa excêntrica fixada na hélice para simular anomalia inicial, com zona esperada B ou C; Experimento 3 — desbalanceamento médio (4 g), com incremento da massa para avaliar a progressão da detecção, com zona esperada C; e Experimento 4 — desbalanceamento pesado (8 g), com massa máxima aplicada para verificar a classificação na condição crítica, com zona esperada C ou D.

5.2. Resultados e Discussão

A execução do planejamento experimental evidenciou distinção clara entre os estados operacionais avaliados, conforme ilustrado nas Figuras 5 e 6. No experimento 1 (estado saudável, 0g), os valores de RMS apresentaram níveis reduzidos, com média de 1,4021 m/s^2, classificados predominantemente na Zona B – NORMAL, da ISO 10816**,** condição de operação contínua aceitável com 96,67% das leituras nessa zona.

Nos experimentos com desbalanceamento induzido, a introdução progressiva de massas excêntricas provocou incremento significativo na amplitude vibratória. Com 2g, o RMS médio atingiu 2,9546 m/s^2, com 56,67% das leituras classificadas na Zona C – ALERTA. Com 4g, todas as 30 leituras (100%) foram classificadas na Zona C, com RMS médio de 4,2544 m/s^2. Com 8g, o RMS médio elevou-se a 9,2554 m/s^2, com 100% das leituras na Zona D – PERIGO, confirmando a sensibilidade da arquitetura de borda para detecção de anomalias mecânicas.

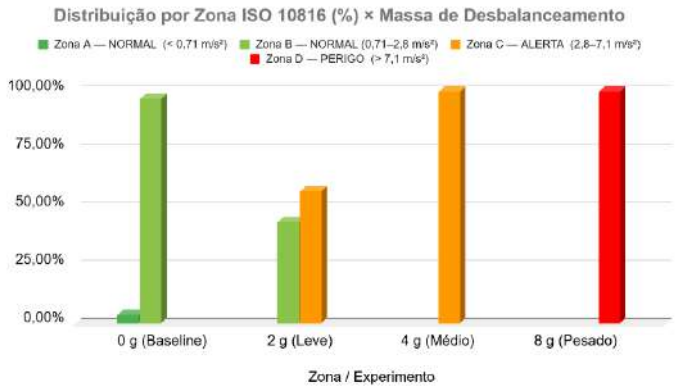


Figura 8. Distribuição percentual das leituras por zona de severidade vibratória (ISO 10816) em função da massa de desbalanceamento aplicada.

A taxa de classificação correta atingiu 100% (120/120 leituras), demonstrando a viabilidade técnica do protótipo para monitoramento de vibração em motores de pequeno porte. Entretanto, reconhece-se que a frequência de amostragem efetiva de 25 Hz pode restringir a detecção de falhas de alta frequência, como fissuras iniciais em rolamentos. A

confiabilidade do artefato concentra-se, portanto, na identificação de falhas macroscópicas — desbalanceamento e desalinhamento —, que apresentam maior impacto na vibração global do equipamento e são plenamente contempladas pelos limiares normativos da ISO 10816.

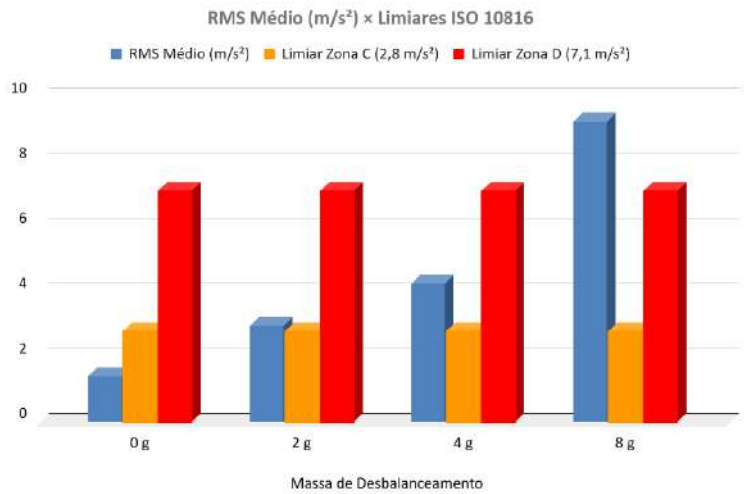


Figura 9. Valor eficaz (RMS) médio de vibração por experimento em comparação com os limiares normativos da ISO 10816 (Zona C: 2,8 m/s^2 e Zona D: 7,1 m/s^2).

6. Conclusão

Este trabalho demonstrou a viabilidade técnica de um sistema IoT de baixo custo para monitoramento de vibração em motores industriais, baseado no microcontrolador ESP32 e no sensor *MEMS* MPU6050, com classificação automática da severidade vibratória conforme a norma ISO 10816. Os quatro experimentos realizados evidenciaram que o sistema é capaz de detectar e classificar corretamente diferentes níveis de desbalanceamento mecânico, desde condições normais de operação até estados críticos de perigo iminente.

Os resultados obtidos revelam que a arquitetura de Computação de Borda proposta apresenta desempenho satisfatório para o contexto de pequenas indústrias: taxa de classificação correta de 100% (120/120

leituras), custo de *hardware* estimado em aproximadamente R$ 80,00 e operação autônoma mesmo em cenários de instabilidade de rede. Esses atributos confirmam que a solução representa uma alternativa financeiramente acessível e tecnicamente adequada às demandas de modernização das plantas fabris de pequeno porte da região Norte do Brasil.

Como limitações do estudo, destacam-se: (a) a frequência de amostragem efetiva de 25 Hz, que restringe a detecção de falhas de alta frequência, como fissuras iniciais em rolamentos, sendo adequada apenas para anomalias macroscópicas; (b) a utilização de um motor de ventilador doméstico como teste, o que implica que a generalização dos resultados para motores industriais de maior potência requer validação adicional; e (c) a ausência de comparação direta com equipamentos comerciais de referência, o que limitou a quantificação do erro de medição absoluto.

Como trabalhos futuros, sugere-se: a ampliação da taxa de amostragem mediante uso de sensores de maior desempenho; a incorporação de algoritmos de aprendizado de máquina embarcado para diagnóstico de falhas específicas; e a validação do sistema em ambientes industriais reais com motores de médio e grande porte.

7. Referências

Bastos, M. M. (2021). SCADA de baixo custo para motores elétricos utilizando o ESP32 e o Scadabr. In *Anais do 18º Congresso Latino-Americano de Software Livre e Tecnologias Abertas – Latinoware*. PTI/Itaipu Binacional. Recuperado de: https://sol.sbc.org.br/index.php/latinoware/article/view/19898

Costa, E. A. (2021). Sistema de monitoramento remoto em motores de indução [Trabalho de Conclusão de Curso, Centro Universitário do Sul de Minas (UNIS)]. Recuperado de: http://repositorio.unis.edu.br/bitstream/prefix/2204/1/Euder%20Alves.pdf

Ferreira, C., Moret, G., Moura, I., & Rodrigues, L. (2024). Sistema de monitoramento de máquinas elétricas para manutenções preditivas [Trabalho de Conclusão de Curso, Faculdade de Tecnologia de São

Bernardo do Campo (FATEC)]. Recuperado de: http://ric-cps.eastus2.cloudapp.azure.com/handle/123456789/28539

Kumar, R. P. R., et al. (2023). Smart system for monitoring physically handicapped infants using IoT. *E3S Web of Conferences*, *391*, 01161. Recuperado de: https://www.e3s-conferences.org/articles/e3sconf/pdf/2023/28/e3sconf_icmed-icmpc2023_01161.pdf

Nascimento, E. S. dos S. (2023). Projeto de um sistema supervisório para prevenção de falhas de uma bomba incêndio e lastro [Monografia, Centro de Instrução Almirante Alexandrino (CIAA), Marinha do Brasil]. Recuperado de: https://repositorio.marinha.mil.br/bitstream/ripcmb/847063/1/TCC%20CApA_%20EVERTON%20NASCIMENTO.pdf

Pudo, J. V. de A., Oliveira, R. A. de, Andrade, Y. R. de, Baldaconi, R. H., & Righetto, M. R. da S. (2025). Development of a bearing wear detection system based on low-cost sensors. *Revista de Gestão Social e Ambiental*, *19*(4), 1–14, e011862. https://doi.org/10.24857/rgsa.v19n4-093

Vavassori, M. M., Curi, E. I. M., Castro, R. de M., Deolindo, D. de M., & Righetto, G. B. (2022). Análise de falhas em equipamentos utilizando sensores de baixo custo, como método de manutenção preditiva. *Revista Vincci – Periódico Científico da UniSATC*, *7*(2), 136–194. Recuperado de: https://revistavincci.satc.edu.br/index.php/Revista-Vincci/article/view/289/259

Capítulo 29

Plataforma IoT de monitoramento e análise estatística para ambientes industriais: Uma solução versátil de baixo custo

Ortiz Marcos Martins Pedroso
Leandro Nelinho Balico
Universidade Federal de Roraima

ortizphp@gmail.com
leandro.balico@ufrr.com

1. Introdução

A Indústria 4.0 tem impulsionado soluções digitais voltadas ao monitoramento inteligente de ambientes por meio da Internet das Coisas (*IoT*), que integra sensores, microcontroladores e sistemas distribuídos para acompanhar variáveis físicas em tempo real (Kagermann *et al.,* 2013). Em cenários industriais como linhas de produção, câmaras frias, *data centers*, laboratórios clínicos e estufas agrícolas, o monitoramento contínuo de temperatura e umidade é exigência operacional e regulatória, uma vez que desvios nessas variáveis comprometem a segurança dos trabalhadores, a qualidade dos produtos e a conformidade com normas técnicas vigentes (Ashton, 2009).

Apesar da disponibilidade crescente de plataformas *IoT* comerciais, persiste uma lacuna quanto a soluções de baixo custo replicáveis e adaptadas à realidade de pequenas e médias empresas com infraestrutura limitada. Sistemas convencionais operam de forma puramente reativa, alertando apenas quando um limite absoluto é ultrapassado, sem considerar desvios comportamentais graduais que sinalizam falhas iminentes (Miller et al., 2021). Além disso, a maioria não oferece suporte nativo a múltiplos perfis regulatórios em uma única plataforma. Este trabalho tem como objetivo desenvolver e validar uma plataforma *IoT* de monitoramento e análise estatística capaz de operar em ambientes industriais diversos, incorporando: detecção reativa por limites absolutos parametrizáveis; detecção

comportamental por *Z-Score* sobre janela histórica de 100 leituras; análise de tendência histórica por Média Móvel Simples sobre consolidado mensal; e notificação multicanal via *e-mail* e WhatsApp.

A validação foi conduzida em dois ambientes reais com perfis regulatórios distintos: o *Datacenter* do Ministério Público do Estado de Roraima, sob o padrão ASHRAE, e o almoxarifado de medicamentos e câmara fria da Coordenação-Geral de Assistência Farmacêutica (CGAF), sob a RDC-216 da ANVISA, por meio de aferição simultânea com equipamentos de referência calibrados, permitindo comparação direta entre as leituras do protótipo e as dos instrumentos industriais instalados nos ambientes.

O trabalho está organizado da seguinte forma: a Seção 2 apresenta a fundamentação teórica; a Seção 3 discute os trabalhos relacionados; a Seção 4 descreve a solução computacional; a Seção 5 apresenta a avaliação experimental; e a Seção 6 conclui o trabalho.

2. Fundamentação teórica

A *IoT* conecta dispositivos físicos equipados com sensores a redes de dados, permitindo coleta, transmissão e processamento automatizado de informações ambientais (Ashton, 2009). Em arquiteturas típicas, o sistema estrutura-se em três camadas: percepção, rede e aplicação (Ghayvat *et al.,* 2015). Microcontroladores de baixo custo como o ESP32, com conectividade *Wi-Fi* nativa, tornaram viável essa arquitetura em contextos industriais com recursos limitados.

O monitoramento industrial de temperatura e umidade é orientado por normas que variam conforme o setor de aplicação. O padrão ASHRAE 55 (Ashrae, 2017) define as condições térmicas e de umidade aceitáveis para ambientes de tecnologia da informação. A RDC-216 (ANVISA, 2004) regula as condições de armazenamento de alimentos e medicamentos, exigindo controle rigoroso de temperatura e umidade em almoxarifados e câmaras frias. Neste trabalho, essas normas fundamentaram a configuração dos limites absolutos de temperatura e umidade para cada ambiente monitorado na avaliação experimental, permitindo que o sistema verifique, a cada leitura, a conformidade com os parâmetros normativos estabelecidos pelo operador.

A detecção de anomalias por *Z-Score* é um método estatístico clássico originado no Controle Estatístico de Processo (CEP), que quantifica o desvio de uma observação em relação à média histórica em unidades de desvio padrão (Shewhart, 1931). O cálculo é expresso pela Equação 1:

$$z = \frac{|x - \mu|}{\sigma}$$

Onde:

x é a leitura atual de temperatura ou umidade;

μ é a média aritmética da janela histórica das últimas 100 leituras;

σ é o desvio padrão calculado sobre a mesma janela.

O limiar $z > 3$ fundamenta-se na Regra dos 3 Sigma de Shewhart (1931): valores além de três desvios padrão ocorrem em apenas 0,27% das observações em distribuição normal, probabilidade suficientemente baixa para caracterizar comportamento anômalo sem gerar volume excessivo de falsos positivos.

A janela de 100 leituras foi definida a partir de três critérios: (i) suficiência estatística, pois $n = 100$ supera o mínimo de 30 leituras exigido pelo Teorema do Limite Central; (ii) adequação temporal, pois, com intervalo de 30 segundos entre leituras, 100 amostras cobrem aproximadamente 50 minutos de histórico recente; e (iii) viabilidade computacional, pois, combinada com mecanismo de *cache* de cinco minutos, garante desempenho adequado com múltiplos dispositivos simultâneos.

A análise de tendência histórica por Média Móvel Simples (MMS) é uma técnica estatística que suaviza variações de curto prazo em uma sequência de dados, revelando a direção predominante do comportamento ao longo do tempo (Shewhart, 1931). O cálculo é expresso pela Equação 2:

$$MMS = \frac{A_1 + A_2 + \cdots + A_n}{n}$$

Onde:

A representa cada leitura registrada no período histórico analisado;

n é o número total de observações na janela de tempo definida.

Neste trabalho, a MMS é aplicada sobre o histórico mensal consolidado, permitindo ao gestor identificar tendências de aquecimento ou resfriamento graduais como subsídio ao planejamento de manutenção preditiva baseada em condição.

3. Trabalhos relacionados

Ghayvat et al. (2015) demonstraram a viabilidade de redes de sensores IoT para monitoramento ambiental contínuo em edifícios, porém sem incorporar algoritmos de detecção de anomalias ou suporte a múltiplos perfis regulatórios. Devecchi e Gambarotto (2019) propuseram sistema IoT para qualidade ambiental interna em ambientes fechados, limitando-se à visualização sem lógica estatística de detecção. Andrić et al. (2017) investigaram previsão de demanda energética por análise de temperatura histórica, inspirando diretamente o módulo de tendência histórica do presente trabalho, embora sem IoT para coleta automatizada. Miller et al. (2021) identificaram, em revisão sistemática, que a maioria das soluções IoT opera reativamente, sem detecção estatística de anomalias ou análise de tendência em sistemas de baixo custo, reforçando a necessidade de soluções como a proposta.

Em síntese, nenhum trabalho revisado combina, em solução única de baixo custo: coleta em tempo real; detecção de anomalias por *Z-Score*; análise de tendência histórica por Média Móvel Simples; notificação multicanal; e suporte nativo a múltiplos perfis regulatórios parametrizáveis para diferentes setores industriais. Essa lacuna justifica e fundamenta a proposta do presente trabalho.

4. Solução computacional

4.1. Arquitetura e Hardware

A plataforma adota arquitetura *IoT* em três camadas, conforme ilustra a Figura 1. Na camada de percepção, um ESP32 conectado ao sensor DHT11 realiza leituras periódicas de temperatura e umidade a cada 30 segundos. Para maior precisão, o *firmware* implementa média de três leituras consecutivas antes de cada transmissão, mitigando o efeito da

resolução discreta de 1 °C do sensor DHT11. Na camada de comunicação, os dados são transmitidos via *Wi-Fi* ao servidor por requisições HTTP autenticadas por *API Key*, garantindo que apenas dispositivos autorizados insiram leituras no sistema. A *API Key* é validada a cada requisição antes do processamento dos dados. No contexto deste protótipo, a transmissão opera sobre HTTP, protocolo adequado ao ambiente de rede local controlado utilizado na validação experimental; a adoção de TLS/HTTPS é apontada como evolução natural para implantações em produção, sendo tecnicamente viável por meio do módulo *WiFiClientSecure* do ESP32. Na camada de aplicação, servidor Apache hospeda a plataforma PHP, que recebe, valida, armazena e analisa os dados em banco MySQL, disponibilizando-os em *dashboard web* responsivo com atualização em tempo real.

O código-fonte completo do sistema, incluindo o *firmware* do ESP32, o *backend* PHP, o esquema do banco de dados e os arquivos de configuração do Docker, encontra-se disponível publicamente no repositório (Pedroso, 2026): https://github.com/ortizpedroso/sistema-iot-monitoramento.

Todos os arquivos estão comentados com referências às seções correspondentes deste trabalho, permitindo a identificação direta entre o código implementado e os conceitos descritos.

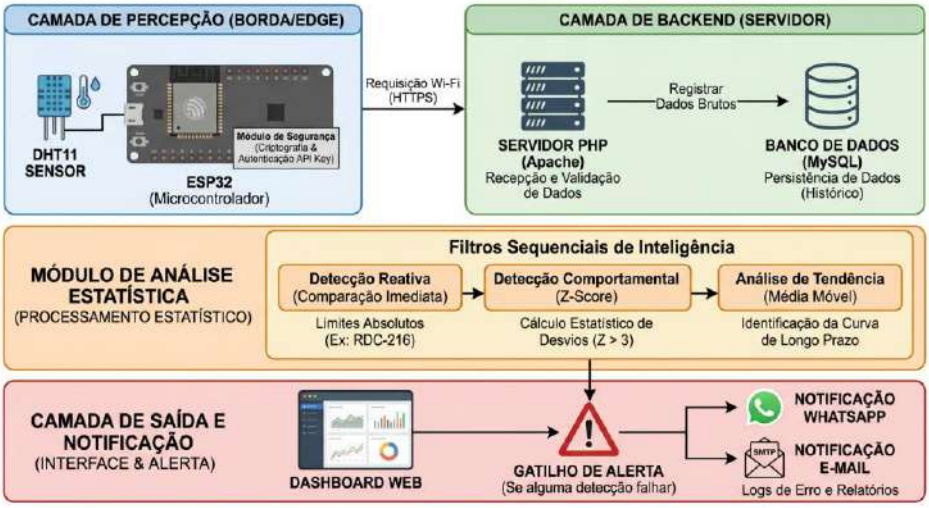


Figura 1. Arquitetura proposta neste trabalho. Gerada com o auxílio de inteligência artificial Gemini (2026).

4.2. Módulo Análise Estatística

O cerne tecnológico da plataforma é um módulo de análise estatística estruturado em três camadas progressivas que evoluem da detecção imediata à análise estratégica de tendência.

A primeira camada realiza detecção reativa por limites absolutos: cada leitura é comparada aos limites mínimo e máximo de temperatura e umidade configurados pelo operador para o ambiente monitorado. Quando os valores são ultrapassados, o sistema gera alertas imediatos classificados como ALTA ou CRÍTICA, conforme o nível de prioridade definido na parametrização do ambiente.

A segunda camada realiza detecção comportamental por *Z-Score*, método estatístico detalhado na Seção 2, que quantifica o desvio de cada leitura em relação à média histórica da janela das últimas 100 leituras. Leituras com $z > 3$ são classificadas como DESVIO, indicando comportamento estatisticamente anômalo mesmo quando os valores permanecem dentro dos limites absolutos configurados. Para evitar falsos positivos decorrentes da resolução discreta do sensor DHT11, o *Z-Score* somente é aplicado quando o desvio absoluto supera 1,0 °C para temperatura e 2,0% para umidade.

A terceira camada realiza análise de tendência histórica por Média Móvel Simples (MMS), definida na Seção 2, aplicada sobre o histórico mensal consolidado de leituras disponível no relatório anual da plataforma. Essa camada identifica padrões de aquecimento ou resfriamento gradual ao longo do tempo, fornecendo ao gestor subsídio para o planejamento de manutenção preditiva baseada em condição antes da ocorrência de falhas funcionais.

4.3. Módulo de Notificação Multicanal

O sistema integra notificação automática via *e-mail* (protocolo SMTP) e WhatsApp por meio de API de mensageria. Cada canal é autenticado por *API Key* individual. Regras de *cooldown* impedem o envio de alertas repetidos para a mesma condição em intervalos inferiores a 30 minutos, evitando sobrecarga de notificações. As notificações incluem o nome do ambiente, o

valor da leitura, o tipo de classificação (ALTA, CRÍTICA ou DESVIO), a sugestão de ação corretiva e o horário do evento.

4.4. Módulo de Notificação Multicanal

A plataforma permite a configuração individualizada de limites absolutos de temperatura e umidade para cada ambiente monitorado. O operador define os parâmetros com base na norma regulatória aplicável ao seu setor, como ASHRAE para *data centers* e RDC-216 para almoxarifados de medicamentos e câmaras frias, entre outros. A partir dessa parametrização, o sistema verifica, a cada leitura recebida, se as condições físicas do ambiente estão dentro dos parâmetros estabelecidos, gerando alertas imediatos quando há divergência. Cada dispositivo ESP32 é identificado automaticamente por um código único derivado do endereço MAC da placa, eliminando a necessidade de configuração manual no *firmware* e permitindo que o mesmo *hardware* seja reaproveitado em diferentes locais sem alteração de código.

5. Avaliação experimental

5.1. Planejamento de pesquisa

A avaliação experimental foi conduzida em dois ambientes reais com perfis operacionais e regulatórios distintos, com o objetivo de validar a precisão do protótipo por meio de aferição simultânea com equipamentos de referência. O primeiro ambiente foi o *Datacenter* do Ministério Público do Estado de Roraima (MPERR), onde a aferição foi realizada em 02/06/2026, com o ambiente Servidor_Rack configurado com os limites do padrão ASHRAE: temperatura entre 18 °C e 27 °C e umidade relativa entre 20% e 60%. Foram coletadas 61 leituras entre 10h00 e 10h30, com o protótipo posicionado a 1,10 m de altura, afastado de fontes de ar direto. O equipamento de referência utilizado nesse ambiente foi o controlador industrial CAREL, já instalado no local para fins de climatização e controle ambiental. O segundo ambiente foi o almoxarifado de medicamentos e câmara fria da Coordenação-Geral de Assistência Farmacêutica (CGAF), onde a aferição foi realizada em 03/06/2026, com o ambiente CGAF configurado

com os limites da RDC-216: temperatura entre 15 °C e 30 °C e umidade relativa entre 40% e 60%. Foram coletadas 52 leituras entre 14h00 e 14h25. O equipamento de referência utilizado nesse ambiente foi o termohigrômetro Kiltler HTC-2 (modelo UNITH 668) (Kiltler, 2026), instrumento com certificado de calibração rastreável N° T 133.420/26, calibrado em 19/02/2026 e com validade até 02/2028.

5.2. Resultados e Discussão

5.2.1. Datacenter do Ministério Público do Estado de Roraima

Nas 61 leituras coletadas entre 10h00 e 10h30 do dia 02/06/2026, o protótipo registrou temperatura inicial de 19,2 °C, mantendo-se nesse patamar nas primeiras leituras antes de iniciar uma descida gradual em degraus (19,1 °C, 19,0 °C, 18,8 °C, 18,5 °C e 18,2 °C) até estabilizar em 18,0 °C a partir das 10h17:30, permanecendo nesse valor até o encerramento da aferição. Esse comportamento reflete o tempo de aclimatação do sensor DHT11, que necessita de alguns minutos para equilibrar sua temperatura interna com a do ambiente. O controlador industrial CAREL registrou temperatura oscilando entre 18,8 °C e 19,0 °C em degraus controlados durante as primeiras leituras, estabilizando em 19,0 °C após a leitura 35. A Figura 2 apresenta as séries temporais sobrepostas de temperatura e umidade dos dois equipamentos.

Em relação à umidade, o protótipo apresentou oscilações em degraus entre 53,0% e 53,6% ao longo das primeiras 40 leituras, estabilizando em 53,4% a partir da leitura 40. O controlador CAREL registrou umidade entre 53,3% e 53,5%, fixando em 53,5% a partir da leitura 35. O desvio máximo de temperatura observado no regime estável foi de ±1,0 °C e o de umidade foi de ±1,0%, ambos dentro das tolerâncias técnicas do DHT11 (±2 °C e ±5%, respectivamente), conforme especificação do fabricante (Aosong, 2023).

A primeira camada do algoritmo não emitiu alertas reativos, confirmando que o ambiente operava em conformidade com os parâmetros ASHRAE durante todo o período. A segunda camada, de detecção por *Z-Score*, também não identificou desvios comportamentais, uma vez que as variações entre leituras consecutivas do protótipo permaneceram abaixo dos limiares mínimos estabelecidos para o DHT11 (1,0 °C para temperatura e 2,0% para

umidade). A ausência de anomalias confirma que o ambiente estava operando dentro da normalidade e que o protótipo não gerou falsos positivos durante a aferição.



Figura 2. Comparação simultânea de temperatura e umidade entre o protótipo ESP32 + DHT11 e o controlador de referência CAREL no *Datacenter* do Ministério Público do Estado de Roraima em 02/06/2026 (61 leituras, intervalo de 30 segundos).

5.2.2. Almoxarifado de Medicamentos e Câmara Fria da CGAF

Nas 52 leituras coletadas entre 14h00 e 14h25 do dia 03/06/2026, o protótipo registrou temperatura inicial de 24,8 °C, mantendo-se nesse patamar nas primeiras leituras antes de iniciar uma descida gradual em degraus (24,7 °C e 24,6 °C) até estabilizar em 24,5 °C a partir das 14h15:00, permanecendo nesse valor até o encerramento da aferição. O termohigrômetro Kiltler HTC-2 registrou temperatura entre 23,9 °C e 24,1 °C durante toda a aferição, com pequenas oscilações controladas nas primeiras leituras, estabilizando em 24,1 °C após a leitura 35. O desvio máximo de temperatura observado no regime estável foi de +0,5 °C, dentro da tolerância técnica do DHT11 (±2 °C), conforme especificação do fabricante (Aosong, 2023).

Em relação à umidade, o protótipo apresentou oscilações em degraus entre 45,5% e 46,3% ao longo das primeiras 30 leituras, com período de 46,3% entre 14h03 e 14h05:30 e de 46,2% entre 14h09 e 14h11:30,

estabilizando em 46,1% a partir das 14h15:00. O Kiltler HTC-2 registrou umidade entre 45,8% e 46,0%, fixando em 46,0% a partir da leitura 35. O desvio máximo de umidade no regime estável foi de ±1,0%, dentro da tolerância do DHT11 (±5%). A Figura 3 apresenta as séries temporais sobrepostas de temperatura e umidade dos dois equipamentos.



Figura 3. Comparação simultânea de temperatura e umidade entre o protótipo ESP32 + DHT11 e o termohigrômetro de referência Kiltler HTC-2 na Coordenação-Geral de Assistência Farmacêutica em 03/06/2026 (52 leituras, intervalo de 30 segundos).

Nenhum alerta reativo foi emitido pela primeira camada do algoritmo, confirmando conformidade com os limites da RDC-216 em todo o período. A segunda camada de *Z-Score* também não identificou desvios comportamentais, uma vez que as variações entre leituras consecutivas do protótipo permaneceram abaixo dos limiares mínimos estabelecidos para o DHT11 (1,0 °C para temperatura e 2,0% para umidade). A ausência de anomalias confirma que o ambiente estava operando dentro da normalidade e que o protótipo não gerou falsos positivos durante a aferição.

Os resultados das aferições nos dois ambientes comprovam que o protótipo opera dentro das margens de tolerância técnica do DHT11 em condições reais de missão crítica. A convergência das leituras após o período de aclimatação, aliada à ausência de falsos positivos, valida a precisão do sensor e a coerência do módulo de análise estatística implementado. Esses

dados fundamentam a confiabilidade da base de leituras utilizada pelo algoritmo de *Z-Score* para identificação de desvios comportamentais em sessões de monitoramento contínuo.

6. Conclusão

Este trabalho apresentou o desenvolvimento e a validação de uma plataforma *IoT* de monitoramento e análise estatística para ambientes industriais, combinando *hardware* de baixo custo baseado em ESP32 e DHT11 com plataforma *web* PHP dotada de módulo de análise estatística em três camadas: detecção reativa por limites absolutos parametrizáveis (ALTA/CRÍTICA), detecção comportamental por *Z-Score* (DESVIO) e análise de tendência histórica por Média Móvel Simples sobre consolidado mensal para suporte à manutenção preditiva baseada em condição. O sistema integra notificação multicanal via *e-mail* e WhatsApp com autenticação por *API Key* e regras de *cooldown*, além de suporte nativo a múltiplos perfis regulatórios por meio de parametrização individualizada por ambiente.

A validação em dois ambientes reais com perfis regulatórios distintos, o *Datacenter* do Ministério Público do Estado de Roraima, regulado pela norma ASHRAE, e o almoxarifado de medicamentos e câmara fria da Coordenação-Geral de Assistência Farmacêutica (CGAF), regulado pela RDC-216, demonstrou operação estável em 113 leituras coletadas sem falhas de comunicação. O pareamento simultâneo com o controlador industrial CAREL no *Datacenter* do MPERR e com o termohigrômetro calibrado Kiltler HTC-2 na CGAF comprovou que o protótipo reflete com fidelidade as condições reais dos ambientes: o desvio máximo de temperatura no regime estável foi de ±1,0 °C no *Datacenter* e de +0,5 °C na CGAF, enquanto o desvio máximo de umidade foi de ±1,0% em ambos os ambientes, todos dentro da margem de tolerância técnica do DHT11 (±2 °C e ±5%). A ausência de anomalias detectadas durante as aferições, resultado esperado em ambientes sob controle adequado, validou a coerência do módulo de análise estatística, que não gerou falsos positivos em condições de estabilidade operacional, confirmando a confiabilidade da base de leituras para a detecção estatística.

O principal diferencial técnico do trabalho reside na diferenciação entre as classificações DESVIO e CRÍTICA: enquanto sistemas puramente

reativos somente identificariam violações de limite absoluto, o módulo de *Z-Score* é projetado para detectar antecipadamente oscilações comportamentais que precedem rupturas de limite, caracterizando suporte efetivo à manutenção preditiva baseada em condição, capacidade demonstrada pela ausência de falsos positivos nas aferições realizadas em ambientes estáveis.

A versatilidade da solução, evidenciada pelo suporte simultâneo a *data centers*, almoxarifados de medicamentos, câmaras frias, linhas de produção industrial e laboratórios clínicos com diferentes perfis regulatórios parametrizáveis, demonstra que é possível construir com tecnologias acessíveis uma plataforma que não apenas informa condições ambientais atuais, mas verifica continuamente a conformidade regulatória, detecta anomalias comportamentais e analisa tendências históricas, representando contribuição concreta para a democratização do monitoramento inteligente na Indústria 4.0.

Como trabalhos futuros, recomenda-se: a adoção de *TLS/HTTPS* para proteção da transmissão em ambientes de produção; a substituição do sensor DHT11 pelo DHT22, com resolução de 0,1 °C e precisão de ±0,5 °C, para redução de falsos positivos no *Z-Score*; e a expansão do módulo de tendência histórica para análise em tempo real com janela deslizante, ampliando o suporte à manutenção preditiva.

7. Referências

Agência Nacional de Vigilância Sanitária. (2004). Resolução RDC nº 216, de 15 de setembro de 2004: Dispõe sobre o Regulamento Técnico de Boas Práticas para Serviços de Alimentação. ANVISA. Recuperado de: https://bvsms.saude.gov.br/bvs/saudelegis/anvisa/2004/res0216_15_09_2004.html

American Society of Heating, Refrigerating and Air-Conditioning Engineers. (2017). ANSI/ASHRAE Standard 55: Thermal environmental conditions for human occupancy. ASHRAE. Recuperado de: https://www.ashrae.org/technical-resources/bookstore/standard-55-thermal-environmental-conditions-for-human-occupancy

Andrić, I., Pina, A., Ferrão, P., & Fournier, J. C. (2017). Long-term heat demand forecast using demand–temperature function: Case study on Lisbon district heating system. *Energy*, *137*, 1144–1154.

Aosong Electronics. (2023). DHT11 temperature and humidity sensor datasheet. Aosong Electronics. Recuperado de: https://www.aosong.com/en/products-21.html

Ashton, K. (2009, June 22). That 'Internet of Things' thing. RFID Journal. Recuperado de: https://www.rfidjournal.com/that-internet-of-things-thing

Devecchi, M., & Gambarotto, F. (2019). An IoT-based system for monitoring indoor environmental quality in classrooms. *Energy Procedia*, *159*, 542–547. https://doi.org/10.1016/j.egypro.2018.12.043

Ghayvat, H., Mukhopadhyay, S. C., Gui, X., & Suryadevara, N. K. (2015). WSN- and IoT-based smart homes and their extension to smart buildings. *Sensors*, *15*, 10350–10379.

Kagermann, H., Wahlster, W., & Helbig, J. (2013). *Recommendations for implementing the strategic initiative Industrie 4.0*. Forschungsunion.

Kiltler. (2026). Termohigrômetro HTC-2 UNITH 668: Certificado de calibração N° T 133.420/26. Kiltler.

Miller, W., Honjo, K., & Zhang, C. (2021). Internet of Things and thermal comfort in smart buildings: A review. *Building and Environment*, *199*, 107905.

Pedroso, O. M. M. (2026). Sistema-IoT-monitoramento: Plataforma IoT de monitoramento ambiental com detecção de anomalias em três camadas. GitHub. Recuperado de: https://github.com/ortizpedroso/sistema-iot-monitoramento

Shewhart, W. A. (1931). Economic control of quality of manufactured product. Van Nostrand.

Capítulo 30

Framework adaptativo para adoção de Dark Factory em PMEs de Roraima

Sara Silva Gomes
Marcelle Alencar Urquiza
Universidade Federal de Roraima

sarasgomes123@gmail.com
marcelle.urquiza@ufrr.com

1. Introdução

A transição para a Indústria 4.0 e o conceito de dark factory, unidades fabris que operam com mínima intervenção humana, é frequentemente percebida como um desafio financeiro e técnico para pequenas e médias empresas (PMEs) (WANG; TONG; ZHUANG, 2025). A digitalização, contudo, não precisa ser um salto tecnológico total, mas sim uma jornada incremental focada na extração de dados de ativos legados, equipamentos com mais de dez anos de operação, sem interfaces digitais nativas e sem capacidade de comunicação em rede (LINS et al., 2020; DOYLE; COSGROVE, 2019). Essa barreira é especialmente crítica em Roraima, onde o setor manufatureiro enfrenta custos energéticos 30% superiores à média nacional, conectividade limitada e escassez de mão de obra especializada em automação (DALENOGARE et al., 2018), que agravam a já limitada capacidade de inovação das PMEs, as quais representam 99% das indústrias locais (FIER, 2024). Embora a literatura recomende automação faseada com retorno rápido sobre investimento (INGALDI et al., 2020), persistem lacunas quanto a frameworks que considerem simultaneamente restrições orçamentárias, infraestruturais e de capital humano das economias regionais emergentes (MITTAL et al., 2018). Este trabalho propõe, para o setor de panificação industrial de Roraima, um framework conceitual de dark factory adaptativa baseado em IoT. Orientado pelo princípio de automação enxuta (lean automation): a tecnologia como ampliação progressiva da capacidade humana (MOKUDAI et al., 2021).

O objetivo geral deste trabalho é propor e validar conceitualmente um framework modular de dark factory adaptativa baseado em Internet das Coisas (IoT) para o setor de panificação industrial de Roraima, demonstrando sua viabilidade técnica e econômica por meio de simulação de eventos discretos com cenários incrementais de automação, considerando restrições de investimento, conectividade e capital humano do contexto regional.

As hipóteses a serem testadas pela simulação são: (H1) a automação do controle de temperatura dos fornos reduz o consumo energético da linha em pelo menos 10%, com base em benchmarks da literatura (WARRINGA, 2023); (H2) a eliminação de paradas não planejadas por monitoramento preditivo eleva o OEE (Overall Equipment Effectiveness) da linha de panificação em pelo menos 15 pontos percentuais (MOKUDAI et al., 2021); (H3) a redução de iluminação nos períodos sem presença humana contribui com economia energética adicional de pelo menos 8%.

Com essas hipóteses o estudo busca responder a cinco questões: (i) quais processos da linha de panificação apresentam maior viabilidade para automação seletiva via IoT; (ii) como dimensionar soluções de sensoriamento e edge computing sob conectividade intermitente (LIU et al., 2023); (iii) qual o retorno sobre investimento esperado por etapa de automação; (iv) quais métricas são mais relevantes para manutenção preditiva e controle de qualidade; e (v) como estruturar arquitetura de dados escalável e modular. As contribuições incluem um modelo conceitual de dark factory adaptativa, critérios de priorização por custo-benefício, arquitetura IoT resiliente e validação por simulação de eventos discretos com cenários what-if. O trabalho está organizado em seis seções: fundamentação teórica, trabalhos relacionados, solução computacional, avaliação experimental e conclusão.

2. Fundamentação teórica

O retrofitting digital de máquinas antigas, por meio da coleta de assinaturas de energia e sinais de vibração, permite extrair dados críticos de desempenho sem substituição total do maquinário (DOYLE; COSGROVE, 2019). O edge computing complementa essa abordagem ao processar dados

localmente, reduzindo a dependência de conectividade contínua com a nuvem (SHI et al., 2016). O gêmeo digital é uma representação virtual que espelha um ativo físico em tempo real (O'SULLIVAN; O'SULLIVAN; BRUTON, 2020). Nas PMEs, a literatura recomenda reduzir a complexidade técnica por meio de ferramentas de design colaborativo que utilizam artefatos físicos, permitindo redesenhar sistemas produtivos de forma intuitiva (GRUBE; MALIK; BILBERG, 2019). Neste trabalho, o nível de fidelidade do gêmeo digital é deliberadamente simplificado: replica apenas os fluxos produtivos e os estados operacionais relevantes para tomada de decisão gerencial.

A automação enxuta orienta a escolha de quais processos automatizar com base no impacto sobre o fluxo de valor e na relação custo-benefício (MOKUDAI et al., 2021). Os critérios de priorização adotados neste trabalho incluem: (i) frequência e impacto das paradas não planejadas; (ii) potencial de redução de desperdício energético; (iii) grau de criticidade do processo para a qualidade do produto final; e (iv) viabilidade técnica de instrumentação com sensores de baixo custo.

3. Trabalhos relacionados

A Tabela 1 sintetiza a comparação entre os trabalhos identificados segundo critérios relevantes para este estudo. A análise comparativa revela um padrão consistente: nenhum dos estudos identificados integra um único framework validado no contexto de PMEs com restrições de conectividade e orçamento. O'Sullivan et al. (2020) e Grube et al. (2019) avançam na dimensão do gêmeo digital, mas operam em ambientes com infraestrutura robusta ou sem alimentação de dados reais em tempo real; Warringa (2023) demonstra a viabilidade de dashboards de baixo custo para PMEs, mas sem integração com simulação; e Lins e Oliveira (2020) estabelecem a base técnica do retrofitting via IoT, porém sem validação em campo com conectividade intermitente.

Essa lacuna coletiva fundamenta e orienta diretamente a arquitetura do SME-DarkBridge: o Módulo 1 responde à limitação de Lins e Oliveira (2020) ao propor instrumentação de ativos legados em ambiente industrial real; o Módulo 2 supre a lacuna de Grube et al. (2019) ao alimentar a simulação com dados históricos automaticamente; e o Módulo 3 estende o

trabalho de Warringa (2023) ao integrar o dashboard à camada de simulação preditiva, entregando o conjunto que nenhum dos estudos anteriores havia combinado em um único sistema acessível a PMEs de economias regionais emergentes.

Tabela 1. Comparação entre trabalhos relacionados.

Critério	O'Sullivan et al. (2020)	Warringa (2023)	Grube et al. (2019)	Mittal et al. (2018)	Lins e Oliveira (2020)
Objetivo	Gêmeo digital para estratégia de manutenção industrial.	Dashboards em tempo real para PMEs com ativos legados.	Simulação colaborativa com gêmeo digital para PMEs.	Revisar modelos de maturidade de manufatura para PMEs.	Propor arquitetura de retrofitting para sistemas ciberfísicos.
Método	Pesquisa-ação em instalação industrial de larga escala.	Entrevistas com especialistas e prototipação centrada no UX.	Interface tangível conectada a software de simulação de eventos discretos.	Revisão sistemática da literatura.	Arquitetura conceitual validada em ambiente de laboratório
Principal resultado	Automação incremental eleva eficiência operacional.	Métricas transparentes reduzem desperdícios e melhoram decisão.	Não técnicos interagem confiavelmente com o gêmeo digital.	Frameworks existentes ignoram restrições típicas de PMEs.	Retrofit com sensores de baixo custo é tecnicamente viável.
Lacuna	Modelo não replicável em PMEs com restrições orçamentárias.	Ausência de integração com sistemas de simulação preditiva.	Sem alimentação automática com dados reais do chão de fábrica.	Ausência de modelos contextualizados para economias em desenvolvimento.	Sem validação em ambiente industrial real com conectividade intermitente.
IoT	Não	Parcial	Não	Não	Sim
Gêmeo dig.	Sim	Não	Sim	Não	Não
Dashboard	Sim	Sim	Não	Não	Não
Val. em PME	Não	Sim	Sim	Não	Parcial

4. Solução computacional

4.1. SME-DarkBridge

A solução proposta, denominada SME-DarkBridge, é um framework conceitual de integração modular e de custo controlado, voltado a conectar o

chão de fábrica físico ao ambiente digital em PMEs do setor de panificação industrial sem exigir substituição do maquinário existente. Neste trabalho, o SME-DarkBridge encontra-se no nível de proposta conceitual: os componentes arquiteturais, os protocolos e as ferramentas foram definidos com base na literatura especializada, e a validação ocorrerá por simulação de eventos discretos alimentada por dados históricos.

O diferencial central do SME-DarkBridge em relação aos frameworks existentes na literatura reside na combinação de três características que, individualmente, já foram exploradas por outros estudos, mas que nunca foram integradas em uma única solução orientada a PMEs com restrições de infraestrutura.

Primeiro, a resiliência à conectividade intermitente: ao contrário de soluções que pressupõem acesso contínuo à nuvem, o SME-DarkBridge opera de forma autônoma na camada edge por até 72 horas. Sincronizando os dados automaticamente ao restabelecer a conexão, o que o torna viável no contexto industrial de Roraima sem modificação de infraestrutura de rede. Segundo, a modularidade incremental: o framework foi projetado para ser implantado em etapas independentes, permitindo que a PME inicie apenas com o Módulo 1 de sensoriamento. E avance progressivamente para simulação e suporte à decisão conforme sua capacidade financeira, sem necessidade de reconfiguração da camada já instalada.

Terceiro, o custo de entrada controlado: com investimento estimado abaixo de R$ 500,00 por ponto de monitoramento e possibilidade de início pelo cenário mais simples por R$ 7.500,00, a solução se posiciona na faixa de viabilidade econômica das PMEs roraimenses. Diferenciando-se de plataformas industriais comerciais cujos custos de licenciamento e implantação inviabilizam a adoção por empresas de pequeno porte (LINS et al., 2020; INGALDI; ULEWICZ, 2020).

A plataforma mantém o operador no ciclo decisório (human-in-the-loop), fornecendo os dados necessários para decisões estratégicas e preparando a PME para uma transição segura (O'SULLIVAN; O'SULLIVAN; BRUTON, 2020). Essa abordagem, alinhada ao princípio de lean augmentation (MOKUDAI et al., 2021), reconhece que a substituição abrupta do fator humano é financeiramente e operacionalmente inviável para PMEs.

O caminho sustentável para a manufatura autônoma passa necessariamente pela ampliação gradual das capacidades digitais da empresa.

4.2. Arquitetura e Stack Tecnológica

A arquitetura é híbrida edge-cloud, estruturada em três camadas, como ilustrado na Imagem 1: (1) camada de percepção, com sensores IoT de baixo custo e coleta de alarmes de CLPs via OPC UA; (2) camada edge, com gateway Raspberry Pi 4 executando Node-RED para orquestração local dos fluxos MQTT e buffer SQLite com janela de 72 horas para operação em modo offline, garantindo resiliência frente à conectividade intermitente de Roraima; e (3) camada cloud, com a plataforma open-source ThingsBoard (com PostgreSQL) para armazenamento histórico e dashboards.

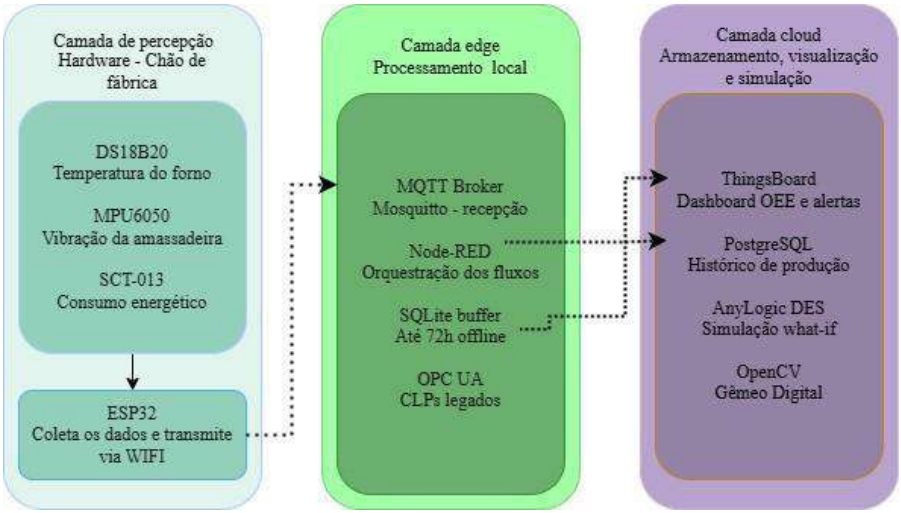


Figura 1. Diagrama UML dos módulos funcionais.

Na camada de percepção, os sensores foram selecionados com base em três critérios: custo unitário abaixo de R$ 30,00, compatibilidade com o microcontrolador ESP32 via protocolos I2C ou OneWire, e adequação às variáveis críticas do processo de panificação. O sensor DS18B20 é utilizado para monitoramento de temperatura nos fornos rotativos e nas câmaras de fermentação, com precisão nominal de mais ou menos 0,5 graus Celsius na

faixa de 0 a 70 graus Celsius e de mais ou menos 2,0 graus Celsius na faixa estendida até 125 graus Celsius (Dallas Semiconductor, 2019).

Para o monitoramento de fornos industriais, que operam entre 180 e 240 graus Celsius, o DS18B20 é posicionado na tubulação de saída ou na estrutura externa do equipamento, onde a temperatura é representativa do processo sem exceder os limites operacionais do sensor. O sensor MPU6050, responsável pela detecção de vibração e aceleração nas amassadeiras, apresenta resolução de 16 bits e faixas de medição configuráveis de 2g a 16g para aceleração e de 250 a 2000 graus por segundo para giroscópio. Adequado para a detecção de padrões de vibração associados ao desgaste mecânico de rolamentos e engrenagens (Invensense, 2013). Sua limitação principal é a sensibilidade a interferências eletromagnéticas em ambientes industriais, o que exige blindagem adequada do cabeamento entre o sensor e o ESP32.

O medidor de corrente SCT-013 opera por indução eletromagnética de forma não invasiva, sem necessidade de interromper o circuito elétrico, com faixa de medição de 0 a 100 amperes e precisão de mais ou menos 1%, sendo adequado para o monitoramento do consumo energético dos equipamentos sem intervenção na instalação elétrica existente (YHDC, 2020). A principal limitação desse sensor é a necessidade de calibração individual por equipamento, dado que a relação entre corrente medida e potência consumida varia conforme o fator de potência de cada carga.

Do ponto de vista de segurança, a arquitetura prevê autenticação por tokens nas comunicações MQTT, segregação lógica entre a rede operacional (OT) e a rede de TI, e rotinas de backup automático do banco de dados local. O custo estimado de hardware por ponto de monitoramento, composto por sensor, microcontrolador e gateway compartilhado, situa-se abaixo de R$ 500,00, com ROI projetado superior a 20%, classificando-o na categoria de soluções de baixo custo para PMEs (Lins et al., 2020).

4.3. Módulos Funcionais

A plataforma é composta por três módulos integrados:

→ Retrofitting Digital e Coleta de Dados: sensores IoT acoplados

externamente a equipamentos de panificação com até 25 anos de uso enviam sinais de estado e variáveis críticas em tempo real para banco de dados centralizado, sem intervenção no processo produtivo (Doyle et al., 2019; Warringa, 2023).

→ Gêmeo Digital e Simulação Colaborativa: réplica virtual simplificada do chão de fábrica, alimentada pelos dados coletados e por dados históricos. Gestores interagem com modelos físicos identificados por câmeras; as alterações de layout são refletidas em software de simulação de eventos discretos utilizando AnyLogic, calculando o impacto no fluxo produtivo sem parada física (Grube et al., 2019).

→ Dashboard de Suporte à Decisão: o painel isola falhas mecânicas de erros operacionais e apoia o remanejo de pessoal e a identificação de necessidades de manutenção (Warringa, 2023; O'Sullivan et al., 2020).

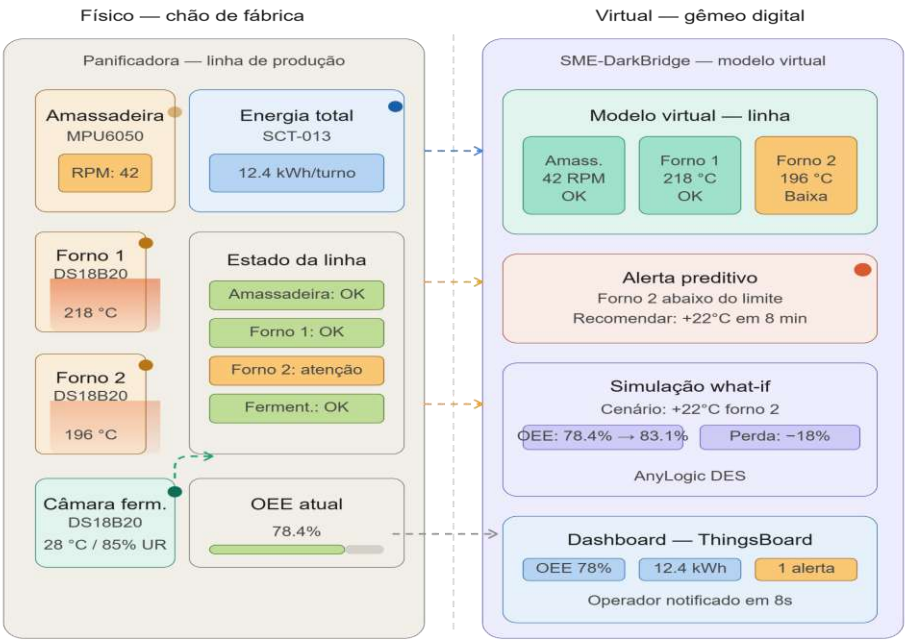


Figura 2. Simulação de um gêmeo digital do chão de uma panificadora.

O gêmeo digital representado na Imagem 2 ilustra o funcionamento do módulo de simulação colaborativa do SME-DarkBridge aplicado a uma linha de panificação industrial. No lado esquerdo, os equipamentos físicos transmitem continuamente suas variáveis críticas por meio de sensores IoT (DS18B20, MPU6050 e SCT-013), com leituras sincronizadas a cada cinco segundos via protocolo MQTT. No lado direito, o modelo virtual espelha em tempo real o estado de cada equipamento: o Forno 2 é identificado automaticamente como ponto de atenção, disparando um alerta preditivo com recomendação de correção (+22°C em até oito minutos) e executando simultaneamente um experimento what-if no AnyLogic, que projeta elevação do OEE de 78,4% para 83,1% e redução de 18% nas perdas de produto caso a intervenção seja realizada. O painel ThingsBoard consolida os indicadores operacionais e notifica o operador em até oito segundos.

5. Avaliação experimental

5.1. Planejamento de pesquisa

A validação do framework SME-DarkBridge baseia-se em Simulação de Eventos Discretos (DES). O modelo de simulação será construído a partir de dados históricos de produção coletados junto a PMEs do setor de panificação de Roraima, complementados por parâmetros extraídos da literatura especializada. Os dados de entrada incluem: taxas de produção por turno, frequência e duração média de paradas não planejadas, consumo energético dos equipamentos e tempos de ciclo das principais operações.

Serão modelados dois cenários principais: (i) o cenário atual (As-Is), que reproduz o estado da linha de produção sem automação; e (ii) o cenário proposto (To-Be), que incorpora o monitoramento via IoT, o controle automatizado de temperatura dos fornos e a redução de iluminação nos períodos sem intervenção humana. As métricas mais relevantes para monitoramento da linha de panificação são: temperatura dos fornos rotativos, OEE da linha, consumo energético por ciclo, vibração da amassadeira e tempo de fermentação por lote. Em modo offline, essas

métricas são preservadas integralmente no buffer local e exibidas no painel ThingsBoard assim que a conexão é restabelecida.

Os critérios de sucesso da simulação incluem: convergência estatística dos resultados após 20 replicações, intervalo de confiança de 95% para as métricas principais, e diferença mínima detectável de 5% entre os cenários As-Is e To-Be. As limitações reconhecidas incluem a impossibilidade de capturar variabilidades comportamentais humanas não documentadas e a dependência da qualidade e completude dos dados históricos disponíveis.

5.2. Resultados e Discussão

A análise comparativa entre os cenários As-Is e To-Be, conduzida por simulação de eventos discretos com parâmetros baseados em dados históricos e literatura especializada, permitirá estimar o impacto da automação seletiva sobre as principais métricas de desempenho da linha de panificação. A literatura indica que intervenções de retrofitting digital em equipamentos legados, combinadas ao monitoramento contínuo de variáveis críticas, podem resultar em ganhos substanciais de eficiência operacional (LINS et al., 2020; MOKUDAI et al., 2021). O experimento foi delineado para avaliar se esses ganhos se reproduzem no contexto específico das PMEs roraimenses.

As etapas da panificação com maior potencial de benefício pela integração IoT são o forneamento (concentra o maior consumo energético da linha e é altamente sensível a variações de temperatura) e a fermentação controlada (apresenta perdas de produto associadas à variação manual de tempo e temperatura).

Para controle do gargalo prioritário, o forneamento, é implementada a instrumentação com sensores DS18B20 e controle automatizado. Nos períodos de falha de conectividade, a solução de edge computing adotada é o gateway Raspberry Pi 4 executando Node-RED com buffer SQLite de 72 horas. Processa localmente os dados dos sensores, aplica as regras de detecção de anomalias e armazena os registros.

No Cenário 1 (monitoramento IoT, investimento de R$ 7.500), o ROI é impulsionado pela redução de 10% no consumo energético do

forneamento, com payback estimado entre 8 e 12 meses. No Cenário 2 (com controle de temperatura, R$ 14.000), o ganho de OEE de 15 pontos percentuais reduz perdas de produto e paradas corretivas, projetando payback entre 12 e 16 meses. No Cenário 3 (solução completa, R$ 22.000), a combinação das três hipóteses eleva a produtividade entre 17% e 20%, com payback estimado em até 20 meses. Todos os valores calibrados a partir de benchmarks da literatura (MOKUDAI et al., 2021; WARRINGA, 2023) e estão sujeitos a ajuste na fase de implementação piloto com dados reais.

Tabela 3. Cenários de implementação.

Cenário	Intervenção	Hipótese testada	Investimento
1	Monitoramento IoT	Redução ≥ 10% no consumo energético	R$ 7.500
2	+ Controle de temperatura	Elevação ≥ 15 pp no OEE	R$ 14.000
3	+ Manutenção preditiva	Payback ≤ 18 meses	R$ 22.000

Está projetado que os resultados da simulação evidenciem a viabilidade financeira do framework, permitindo o cálculo do retorno sobre investimento (ROI) a partir do ganho de eficiência estimado para a linha de produção monitorada, conforme metodologia proposta por Mittal et al. (2018) para avaliação de projetos de Indústria 4.0 em PMEs.

6. Conclusão

A partir da revisão da literatura especializada, da especificação arquitetural do framework SME-DarkBridge e do planejamento de validação por simulação de eventos discretos com três cenários progressivos de automação, buscamos atingir o nosso objetivo. Os parâmetros do modelo foram calibrados com benchmarks da literatura e dados históricos do setor de panificação. Verificamos que é viável propor um framework conceitual de dark factory adaptativa para PMEs do setor de panificação de Roraima, respeitando simultaneamente as restrições de investimento, conectividade e capital humano do contexto regional.

Em resposta à primeira questão de pesquisa, os processos de forneamento e fermentação controlada foram identificados como os de maior viabilidade para automação seletiva via IoT, por concentrarem o maior consumo energético da linha e as principais perdas de produto por variação manual de parâmetros. Em resposta à segunda questão, o dimensionamento para ambientes com conectividade intermitente foi resolvido pelo gateway Raspberry Pi 4 com buffer SQLite de 72 horas, que processa dados localmente e sincroniza automaticamente ao restabelecer a conexão, eliminando a dependência de rede contínua. Em resposta à terceira questão, o retorno sobre investimento projetado variou entre 8 e 20 meses conforme o cenário de implementação, com os três cenários demonstrando viabilidade econômica para PMEs com orçamento restrito. Em resposta à quarta questão, as métricas mais relevantes para monitoramento efetivo da linha são temperatura dos fornos, OEE, consumo energético por ciclo, vibração da amassadeira e tempo de fermentação por lote, todas preservadas no buffer local durante períodos offline. Em resposta à quinta questão, a arquitetura modular do SME-DarkBridge foi estruturada em três etapas progressivas e independentes, permitindo que a PME inicie pelo sensoriamento básico e avance para simulação e suporte à decisão conforme sua capacidade financeira, sem reconfiguração da infraestrutura já instalada.

Como trabalhos futuros, propõe-se a expansão do gêmeo digital para incluir análise de fadiga do operador em ambientes de manufatura híbrida (Turner et al., 2021), a integração bidirecional com sistemas ERP para automação de ordens de serviço (Warringa, 2023) e a incorporação de capacidades de auto-otimização e auto-ajuste para que o sistema sugira correções automáticas no fluxo de produção ao detectar anomalias nos sensores IoT (Bajestani et al., 2025).

7. Referências

Dalenogare, L. S., Benitez, G. B., Ayala, N. F., & Frank, A. G. (2018). The expected contribution of Industry 4.0 technologies for industrial performance. *International Journal of Production Economics*, *204*, 383-394. https://doi.org/10.1016/j.ijpe.2018.08.019

Doyle, F., & Cosgrove, J. (2019). Steps towards digitization of manufacturing in an SME environment. *Procedia Manufacturing*, *38*, 540-547. https://doi.org/10.1016/j.promfg.2020.01.068

FIER. (2024). *Anuário da Indústria de Roraima 2024*. Federação das Indústrias do Estado de Roraima.

Grube, D., Malik, A. A., & Bilberg, A. (2019). SMEs can touch Industry 4.0 in the Smart Learning Factory. Procedia Manufacturing, 31, 219-224. https://doi.org/10.1016/j.promfg.2019.03.035

Ingaldi, M., & Ulewicz, R. (2020). Problems with the implementation of Industry 4.0 in enterprises from the SME sector. *Sustainability*, *12*(1), 217. https://doi.org/10.3390/su12010217

Liu, X., Zhang, Y., Zhang, Y., & Sun, Y. (2023). Real-Time Data Processing and Optimization Research of Edge Computing in Industrial Automation. In Anais do International Conference on Industrial IoT, Big Data and Supply Chain (pp. 173-177), Wuhan, China. https://doi.org/10.1109/IIoTBDSC60298.2023.00039

Liu, Y., & Huh, J. (2025). Smart Factory to Dark Factory in China: A Systematic Survey on Managerial Transformation. *Journal of Information and Communication Convergence Engineering*, *23*(4), 336-354. https://doi.org/10.56977/jicce.2025.23.4.336

Mittal, S., Khan, M. A., Romero, D., & Wuest, T. (2018). A critical review of smart manufacturing & Industry 4.0 maturity models: implications for small and medium-sized enterprises (SMEs). *Journal of Manufacturing Systems*, *49*, 194-214. https://doi.org/10.1016/j.jmsy.2018.10.005

O'Sullivan, J., O'Sullivan, D., & Bruton, K. (2020). A case-study in the introduction of a digital twin in a large-scale smart manufacturing facility. *Procedia Manufacturing*, *51*, 1523-1530. https://doi.org/10.1016/j.promfg.2020.10.212

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge Computing: Vision and challenges. *IEEE Internet of Things Journal*, *3*(5), 637-646. https://doi.org/10.1109/JIOT.2016.2579198

Wang, R., Tong, Y., & Zhuang, C. (2025). Lights-out factories: Review and prospect. Proceedings of the Institution of Mechanical Engineers, Part B: *Journal of Engineering Manufacture*, *239*(14), 1983-1993. https://doi.org/10.1177/09544054241305826

Warringa, B. (2023). Designing Real Time Insights Dashboards for SME Manufacturing with Legacy Assets [Dissertação de Mestrado, Delft University of Technology]. Rotterdam.

Mokudai, T. et al. Digital technologies as lean augmentation: a preliminary study of Japanese automotive manufacturers. *International Journal of Automotive Technology and Management*, *21*(3). https://doi.org/10.1504/IJATM.2021.116607

Os 30 estudos que compõem esta obra são contribuições ao desenvolvimento da Indústria 4.0 no Brasil e, mais especificamente, no estado de Roraima.

Como ponto de partida, assumimos que não temos como falar em tecnologias sem pensar em conectividade e segurança. Por essa razão, os seis primeiros capítulos abordam o tema de "Redes e Segurança da informação". Em seguida, ao longo de cinco capítulos, passamos a um tema fundamental para o funcionamento de qualquer indústria, que é "Produção Enérgica e Transporte", aspecto estratégico para o desenvolvimento social e econômico de um país. Considerando os diversos tipos de indústrias existentes no Brasil, destacamos em quatro capítulos, por sua importância para a vida humana e por ser uma vocação nacional, o tema "Agroindústria". Contamos, ainda, com dois estudos que examinam um tema crítico para a nossa e as futuras gerações, a "Sustentabilidade". Finalizamos esta obra abordando em 12 estudos o tema "Gestão e Produtividade Industrial".

ISBN 978-65-265-3173-0

Pedro & João editores