tecnia

revista de educação, ciência e tecnologia do IFG

v. 10, Edição Especial 1 | 2025
ISSN: 2526–2130

DOSSIÊ TEMÁTICO

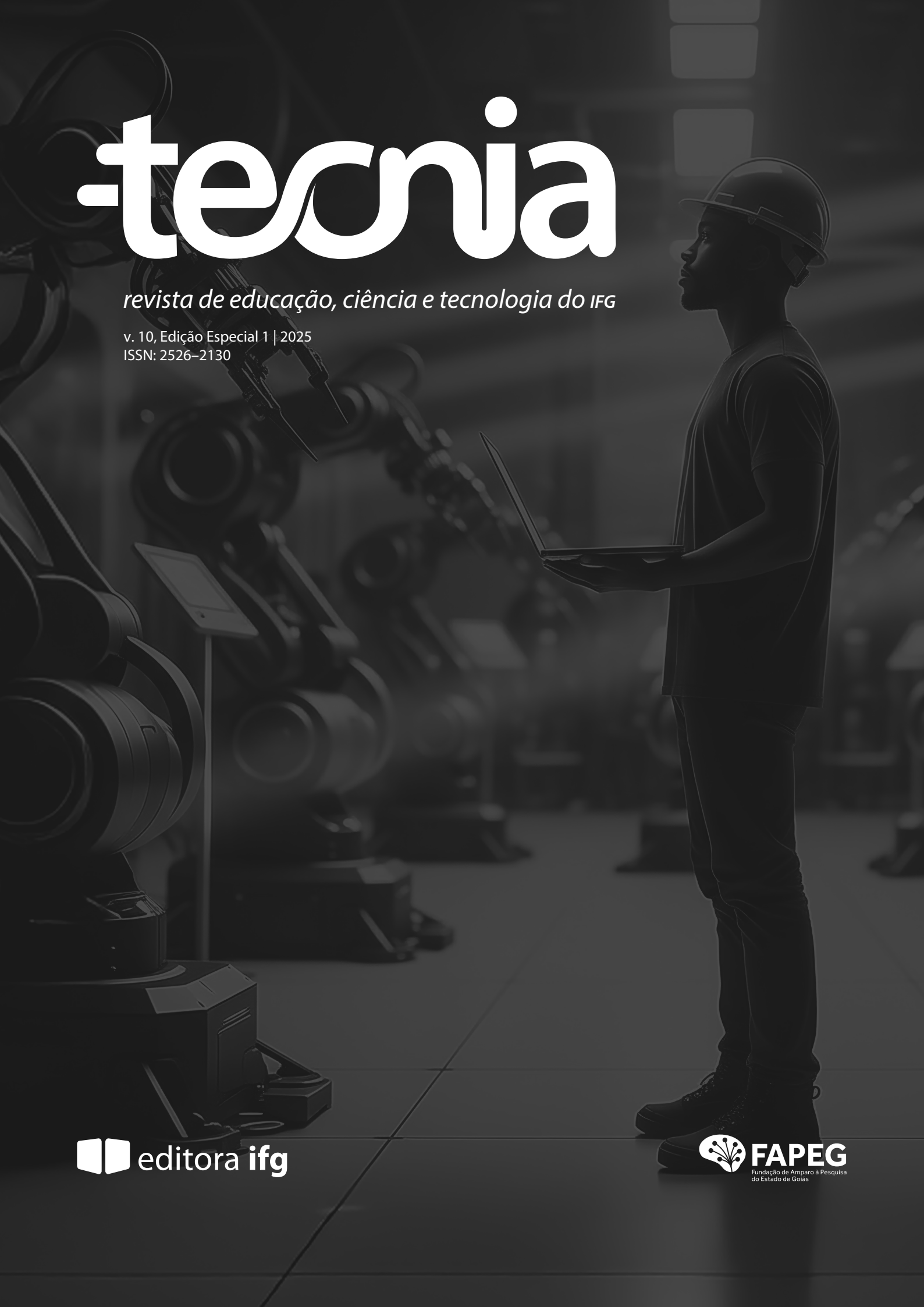
Tecnologias Habilitadoras para a Industria 4.0

editora ifg

tecnia

revista de educação, ciência e tecnologia do IFG

v. 10, Edição Especial 1 | 2025
ISSN: 2526–2130



editora ifg

FAPEG
Fundação de Amparo à Pesquisa do Estado de Goiás

Expediente

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE GOIÁS

Apoio
A revista Tecnia conta com o apoio da Fundação de Amparo à Pesquisa do Estado de Goiás (Fapeg)

Imagem da Capa
Imagem gerada pela ferramenta de inteligência artificial Adobe Firefly e manipulada digitalmente pelo programador visual Pedro Henrique Pereira de Carvalho.

Sumário

Editorial

A Indústria 4.0, também chamada de Quarta Revolução Industrial, apoia-se em tecnologias como sistemas ciber físicos, internet das coisas (IoT), inteligência artificial, big data, manufatura aditiva, processamento de linguagem natural, visão computacional, blockchain e cibersegurança. Essas tecnologias, atuando de forma integrada e descentralizada, habilitam operações em tempo real, modularidade e otimização contínua dos processos industriais.

Na prática, essas tecnologias permitem a automação inteligente, a tomada de decisão rápida e a manutenção preditiva, reduzindo custos e aumentando a eficiência produtiva. No entanto, obstáculos ainda existem, tais como a interoperabilidade entre sistemas heterogêneos, a segurança cibernética e a escassez de profissionais capacitados para operar esse novo ecossistema industrial.

A XI Escola Regional de Informática de Goiás (ERI GO), movimento itinerante que foi realizado de forma cooperada pelo Senai-Fatesg, o Instituto de Informática (UFG) e a Sociedade Brasileira de Computação (SBC), é um importante evento técnico científico e um fórum de discussão para pesquisadores, professores, estudantes e profissionais de computação do estado de Goiás e outros estados do Brasil. Em dezembro de 2023, na sua 11ª edição, a ERI GO abordou diretamente o tema "Tecnologias Habilitadoras para a Indústria 4.0", promovendo palestras, sessões técnicas e oficinas voltadas à interação entre comunidade acadêmica e industrial. A ERI GO tem sido vital para incentivar a pesquisa científica em Goiás, especialmente no âmbito da computação aplicada à indústria.

Ao incluir trilhas específicas para "Sistemas Computacionais e de Informação", o evento fortaleceu a disseminação de conhecimento e aproximou a academia das demandas do setor produtivo regional. Essa colaboração impulsionou a formação de redes entre instituições, fomentou parcerias e ampliou a visibilidade dos trabalhos produzidos por pesquisadores de todo o Brasil. Neste dossiê, destacam-se diversos trabalhos inovadores, como os apresentados a seguir: o "Aplicação da meta-heurística algoritmo genético na solução do Problema da Próxima Versão com modelagem, implementação e análise comparativa" explora meta heurísticas em engenharia de software, aplicando algoritmos genéticos para priorizar funcionalidades e demonstrando ganhos em eficiência de gestão de requisitos. Em "Previsão de Vazão de Rios usando Rede Perceptron Multi Camada Otimizada por Neural Architecture Search", observamos a aplicação de inteligência artificial e de otimização automatizada de redes neurais para modelagem hidrológica de bacias. O artigo "Modelos de Aprendizado Profundo aplicados à Detecção de Pólipo Colorretal" reforça a atuação da visão computacional na área da saúde, especificamente em um estudo de caso utilizando deep learning aplicado na detecção de objetos na medicina. No artigo "Análise Quantitativa e Qualitativa Preliminar dos Efeitos dos Algoritmos de Reamostragem no Registro de Imagens Utilizando a Detecção de Cantos", abordam-se técnicas de processamento de imagem, focando em algoritmos de reamostragem e pontos de interesse para melhor alinhamento das imagens. Em "Expansão automática de léxico para Análise de Sentimentos de textos no domínio do Mercado Financeiro Brasileiro", explora-se o processamento de linguagem natural e a lexicografia adaptada à análise de sentimentos aplicada para finanças.

Em "Fake news: a rapid tertiary study through health, deep learning, and emerging perspectives", discutem-se os desafios atuais da desinformação em saúde, evidenciando

aplicações de deep learning e a revisão de literatura para resposta emergente. O estudo “Redes Perceptron Multicamadas para Modelar Efeitos de Distorção em Sinais de Guitarra Elétrica” aplica inteligência artificial ao processamento de sinais musicais, mostrando versatilidade dos modelos em domínios criativos.

Por fim, o artigo “Adoção da Inteligência Artificial no Schema Matching: Um Levantamento Sistemático do Estado da Arte” oferece uma revisão estruturada da aplicação de inteligência artificial em integração de dados, importante para a interoperabilidade de sistemas. Esses trabalhos ilustram um ecossistema de pesquisa multidisciplinar, no qual as tecnologias habilitadoras da Indústria 4.0 se conectam com problemas práticos reais.

Diante disso, ressalta-se que este editorial reforça que a Indústria 4.0 representa um novo patamar de integração entre tecnologias avançadas que, juntas, habilitam operações automatizadas, escaláveis e modulares. A XI ERI GO, realizada em 7 e 8 de dezembro de 2023, consolida-se no Estado de Goiás como palco central de diálogo técnico e científico entre a academia e a indústria. Os artigos demonstram a diversidade de aplicações da Computação em áreas como Engenharia de Software, Hidrologia, Saúde, Finanças, Música e Integração de Dados, evidenciando a relevância das tecnologias habilitadoras no contexto real de pesquisa e desenvolvimento. Ao fomentar as várias trilhas de pesquisa, a ERI GO não só amplia a visibilidade dos trabalhos goianos, mas também fortalece redes colaborativas entre instituições regionais e o setor produtivo. Assim, o evento atua como catalisador para a formação de profissionais qualificados, o estabelecimento de parcerias duradouras e a transferência de soluções tecnológicas ao mercado, contribuindo de forma efetiva para o fortalecimento do ecossistema científico e industrial de Goiás.

Nesta edição especial, expressamos novamente um agradecimento à Fundação de Amparo à Pesquisa do Estado de Goiás (Fapeg) pelo apoio financeiro destinado aos serviços de revisão e editoração, por meio de política de fomento à ciência e à difusão do conhecimento.

Ricardo Augusto Pereira Franco
Editor do Dossiê

tecnia revista de educação, ciência e tecnologia do ifg

v. 10, Edição Especial 1 | 2025

Submetido 11/06/2024. Aprovado 10/03/2025
Avaliação: revisão duplo-anônimo

Aplicação da meta-heurística algoritmo genético na solução do Problema da Próxima Versão com modelagem, implementação e análise comparativa

THE GENETIC ALGORITHM META-HEURISTIC IS APPLIED TO SOLVE THE NEXT VERSION PROBLEM, AND THE PROCESS IS MODELED, IMPLEMENTED, AND COMPARED

APLICACIÓN DE LA METAHEURÍSTICA DEL ALGORITMO GENÉTICO EN LA SOLUCIÓN DEL PROBLEMA DE LA PRÓXIMA VERSIÓN CON MODELADO, IMPLEMENTACIÓN Y ANÁLISIS COMPARATIVE

Ana Clara A. G. da Silva
Universidade Estadual de Goiás (UEG)
anaclara.araujo@ueg.br

Celso G. C. Junior
Universidade Federal de Goiás (UFG)
celso@inf.ufg.br

Gilmar T. Junior
Universidade Estadual de Goiás (UEG)
gilmar.junior@ueg.br

Ricardo Manuel G. Martins
Pontifícia Universidade Católica de Goiás (PUC Goiás)
ricardomartins@pucgoias.edu.br

Thiago D. de C. Q. Gama
Faculdade de Tecnologia Senai de Desenvolvimento Gerencial (Fatesg)
thiagoddcqg@gmail.com

Resumo

Este estudo apresenta uma investigação sobre a aplicação da meta-heurística algoritmo genético para resolver o complexo Problema da Próxima Versão na engenharia de software. O algoritmo genético foi adaptado para tratar dessa questão, demonstrando sua eficácia em comparação com outras configurações, por meio de experimentos em conjuntos de dados reais. Os resultados indicam que essa abordagem gera soluções eficientes e balanceadas para os objetivos do projeto, oferecendo insights valiosos para a gestão de requisitos em projetos de desenvolvimento de software.

Palavras-chave: algoritmo genético; problema da próxima versão; gerenciamento de requisitos; meta-heurísticas; engenharia de software baseada em busca.

Abstract

This study explores the application of the Genetic Algorithm metaheuristic to address the complex Next Release Problem (NRP) in software engineering. The proposed approach adapts the Genetic Algorithm

to the specific characteristics of this problem and evaluates its performance through experiments conducted on real datasets. The results demonstrate that the method produces efficient and well-balanced solutions aligned with project objectives, providing valuable contributions to requirements management in software development.

Keywords: genetic algorithm; next release problem; requirements management; metaheuristics; search-based software engineering.

Resumen
Este estudio presenta una investigación sobre la aplicación de la metaheurística del algoritmo genético para resolver el complejo problema de la próxima versión en la ingeniería de software. El algoritmo genético fue adaptado para abordar esta cuestión, demostrando su eficacia en comparación con otras configuraciones mediante experimentos en conjuntos de datos reales. Los resultados indican que este enfoque genera soluciones eficientes y equilibradas para los objetivos del proyecto, ofreciendo valiosos conocimientos para la gestión de requisitos en proyectos de desarrollo de software.

Palabras clave: algoritmo genético; problema de la próxima versión; gestión de requisitos; metaheurísticas; ingeniería de software basada en búsqueda.

Introdução

Este trabalho acadêmico apresenta uma investigação sobre a aplicação da meta-heurística algoritmo genético (AG) como uma abordagem eficaz para a resolução do desafiador Problema da Próxima Versão (Next Release Problem – NRP) no contexto da engenharia de software. Esse problema consiste na seleção e priorização de requisitos ou funcionalidades a serem incluídos em uma futura versão de um sistema, considerando restrições de tempo, recursos humanos e demais limitações operacionais.

Primeiramente, foi feita uma revisão abrangente da literatura relacionada à seleção de requisitos, ao algoritmo genético e a suas aplicações em engenharia de software. Em seguida, descreveu-se a modelagem do NRP como um problema de otimização multiobjetivo, cujos objetivos incluem a maximização do valor do software, a minimização dos custos e o cumprimento de restrições técnicas e de prazo.

Na sequência, detalhou-se a implementação do algoritmo genético adaptado para abordar o NRP, destacando as representações de solução, operadores genéticos, função de aptidão e critérios de parada específicos. Na investigação foram realizados experimentos empíricos, usando, para isso, conjuntos de dados reais e comparativos associados a técnicas de seleção de requisitos, como também a outras configurações de parâmetros de entrada da heurística AG e algoritmos de busca exaustiva ou de força bruta (FB).

Os resultados demonstraram que o algoritmo genético proposto apresenta desempenho considerável em termos de eficácia na seleção de requisitos para a próxima versão do software, gerando soluções eficientes e balanceadas em relação aos objetivos concorrentes. Com base nesses resultados, discute-se as vantagens, desafios e limitações da abordagem, bem como possíveis direções futuras de pesquisa. Desse modo, acredita-se que o estudo contribui para a aplicação prática do algoritmo genético na engenharia de software, fornecendo insights valiosos para profissionais e pesquisadores que lidam com a gestão de requisitos em projetos de desenvolvimento de software. Além disso, demonstra a promissora aplicação desta heurística na resolução do NRP.

A estrutura deste trabalho está organizada da seguinte forma: Introdução, seção onde são apresentados o contexto e o tema da pesquisa; Motivação, subseção cujo objetivo é apresentar os fundamentos que impulsionaram a realização da pesquisa, destacando as razões e objetivos que justificam sua condução; Bases teóricas, seção em que são apresentados estudos prévios e contribuições relevantes que contextualizam o tema e sustentam teoricamente a abordagem adotada; Decisões metodológicas, seção que descreve a metodologia empregada, detalhando os procedimentos adotados para implementação da solução; Modelagem do com algoritmo genético, seção onde é explorada a aplicação do algoritmo genético ao problema, com ênfase na subseção Operadores genéticos, que apresenta os detalhes operacionais da técnica utilizada; Experimentação e resultados, seção que traz os experimentos realizados e os resultados obtidos, fornecendo evidências empíricas sobre o desempenho da abordagem; Discussão, seção em que se analisa os resultados à luz dos objetivos propostos, interpretando os achados e suas implicações; Proposta de nova modelagem matemática para o Problema da Próxima Versão, seção onde é apresentada uma abordagem alternativa baseada em formulação matemática; Análise da nova modelagem para o Problema da Próxima Versão, seção em que se avalia a eficácia da proposta com base em experimentos adicionais; Considerações finais e trabalhos futuros, seção que sintetiza as principais contribuições da pesquisa e propõe direções para investigações futuras; Referências, seção que reúne as fontes bibliográficas utilizadas ao longo do trabalho.

Motivação

O NRP é um desafio comum na indústria de desenvolvimento de software. Com o avanço da tecnologia, as organizações estão constantemente buscando maneiras de selecionar e priorizar os requisitos de software de forma eficaz, para atender às demandas do mercado com recursos disponíveis.

A seleção de requisitos para uma próxima versão de software envolve a consideração de múltiplos objetivos e restrições, como custo, prazo, recursos e requisitos técnicos. Isso torna o problema altamente complexo e digno de investigação aprofundada. A aplicação de heurísticas, como o algoritmo genético, oferece a oportunidade de encontrar soluções eficientes e balanceadas para o problema, otimizando a alocação de recursos e aumentando o valor das versões do software resultante.

A pesquisa sobre a aplicação do algoritmo genético na solução desse problema específico pode preencher lacunas no conhecimento existente em engenharia de software, oferecendo uma nova perspectiva para a seleção de funcionalidades. A aplicação de algoritmos de inteligência artificial, como esse, na resolução de problemas complexos de engenharia de software demonstra uma abordagem inovadora, que pode impulsionar avanços consideráveis na área. A eficácia na seleção de requisitos pode levar a economias de custos, redução de desperdícios e melhoria na qualidade do software, fatores críticos para o sucesso de projetos de desenvolvimento de software.

Por fim, o estudo do NRP oferece oportunidades para pesquisadores acadêmicos explorarem conceitos teóricos, adaptando e aplicando o algoritmo genético em um contexto prático e relevante. Portanto, a motivação para este trabalho acadêmico está enraizada na necessidade de abordar um problema real e complexo na engenharia de software, bem como na busca por soluções inovadoras e eficazes que possam beneficiar a indústria e a pesquisa acadêmica. Logo, o objetivo desta pesquisa é avaliar a eficácia do algoritmo genético na seleção de funcionalidades para a próxima versão do software.

Bases teóricas

Esta síntese de bases teóricas destaca a relevância de algoritmos genéticos na solução do NRP em engenharia de software, demonstrando sua eficácia na otimização da seleção de funcionalidades em comparação com abordagens tradicionais. Os estudos citados a seguir servem como um alicerce sólido para a pesquisa e aplicação prática da heurística algoritmo genético nesse contexto.

A natureza adaptativa e exploratória dos algoritmos genéticos os torna adequados para lidar com as complexas interações entre diferentes requisitos e restrições. Além disso, esses algoritmos podem ser configurados para incorporar preferências do usuário, prioridades de desenvolvimento e outros fatores relevantes para a tomada de decisões, conforme exposto em Elvassore (2016).

O NRP é um desafio complexo enfrentado no desenvolvimento de software, quando os desenvolvedores precisam decidir quais funcionalidades, correções e melhorias devem ser incluídas na próxima versão de um produto. Essa decisão envolve equilibrar requisitos técnicos, restrições de tempo e recursos, prioridades do cliente e objetivos estratégicos da organização.

Os algoritmos genéticos são uma classe de algoritmos de otimização inspirados no processo de seleção natural. Eles têm sido amplamente utilizados em problemas de otimização complexos, oferecendo uma abordagem eficaz em situações que envolvem múltiplos objetivos e restrições. Segundo Harman, Mansouri e Zhang (2012), a engenharia de software baseada em busca (Search-Based Software Engineering – SBSE) aplica algoritmos como os genéticos em diversas fases do ciclo de vida do software, incluindo seleção de requisitos, planejamento de projeto, manutenção e reengenharia. Essa abordagem é particularmente atrativa em cenários caracterizados por grandes espaços de solução e múltiplos objetivos conflitantes, como é o caso do NRP, reforçando a escolha dos algoritmos genéticos como estratégia de otimização adequada para esse contexto.

Algoritmos genéticos têm sido aplicados com sucesso em diversos domínios da engenharia de software. No contexto da seleção de requisitos, Niu, Huang e Jin (2008) descrevem sua utilização para otimizar a escolha de funcionalidades em sistemas de software, destacando a busca por soluções eficientes e balanceadas. Ampliando essa perspectiva, Souza e Rebouças Filho (2020) discutem aplicações mais recentes de algoritmos genéticos na engenharia de software, incluindo sua utilização em ambientes de computação distribuída e aprendizado de máquina, o que evidencia a flexibilidade e a robustez dessa meta-heurística em diferentes contextos computacionais.

O NRP é um desafio crítico em engenharia de software, pois exige a seleção e priorização dos requisitos que serão incorporados em uma próxima versão de software. Nesse sentido, Sommerville (2011) destaca a importância da seleção de requisitos na gestão de projetos de software e a necessidade de considerar restrições de recursos e objetivos de negócios.

Além dos algoritmos genéticos, diversas abordagens tradicionais têm sido utilizadas na seleção de requisitos, como métodos baseados em heurísticas e técnicas de análise multicritério. Gorschek, Wohin e Östberg (2006) fornecem uma visão geral dessas estratégias e destacam suas limitações quanto à escalabilidade e à capacidade de tratar múltiplos critérios de forma integrada. Complementando essa visão, Sarrab e Al Shibli (2019) realizaram uma revisão sistemática sobre técnicas de otimização no planejamento de lançamentos de software, evidenciando a eficácia dos algoritmos genéticos na priorização de requisitos em cenários complexos. Essa combinação de

estudos reforça a relevância dos algoritmos genéticos como ferramenta de apoio à tomada de decisão em ambientes com restrições multifatoriais.

Decisões metodológicas

A seguir, são relatadas, de forma ordenada, as decisões metodológicas aplicadas neste estudo:

Descrição detalhada do NRP em engenharia de software, incluindo suas características, desafios e objetivos.

Especificação clara do objetivo da pesquisa.

Identificação das fontes de dados necessárias para conduzir o estudo, a saber: registros de requisitos, dados de projetos anteriores e informações sobre restrições de recursos e prazos.

Descrição de como os requisitos seriam representados como indivíduos em uma modelagem de algoritmo genético, levando em consideração as características específicas do problema.

Detalhamento dos operadores genéticos que seriam utilizados, incluindo cruzamento (em inglês, *crossover*), mutação e seleção, e explicação de como eles seriam aplicados ao contexto do Problema da Próxima Versão.

Definição de uma função de aptidão que quantificasse o desempenho das soluções em relação aos objetivos do projeto, nesse caso, maximização do valor da release a partir das *features* disponíveis.

Especificação dos parâmetros do algoritmo genético, como tamanho da população, taxa de cruzamento, taxa de mutação e critérios de parada. As escolhas foram justificadas com base na bibliografia especializada abordada neste estudo.

Planejamento dos experimentos que seriam realizados para avaliar o desempenho do algoritmo genético. Isso incluiu a execução de simulações em conjuntos de dados reais – que não puderam ser abordados com profundidade devido às questões de confidencialidade –, e a comparação com outras execuções aplicando parâmetros diferentes e com a execução de um algoritmo de força bruta.

Especificação das métricas que seriam usadas para avaliar o desempenho do AG, como a qualidade das soluções encontradas, o tempo de execução e a análise da convergência dos resultados.

Descrição de como os resultados seriam analisados, incluindo a interpretação das métricas de avaliação e a discussão das implicações dos resultados.

Modelagem do Problema da Próxima Versão com algoritmo genético

A modelagem do NRP com algoritmo genético envolve requisitos como a representação dos indivíduos, a definição de operadores genéticos e a formulação da função de aptidão.

Acerca da representação do indivíduo, cada um (ou solução) apresenta uma lista de genes inteiros e contém uma variável aleatória. Cada solução candidata, que representa uma seleção de requisitos para a próxima versão do software, é codificada como um indivíduo composto por genes. Cada gene pode representar uma funcionalidade específica e seu estado (incluso ou não incluso na próxima versão), supondo-se,

por exemplo, que se tem as seguintes funcionalidades para um software: *feature* A, *feature* B, *feature* C e *feature* D.

Com relação à representação da população, ela contém uma lista de indivíduos e um método que calcula o total de aptidão da população. Pode-se representar uma solução candidata como um indivíduo contendo um vetor binário de genes, sendo cada um desses genes do vetor indica se o requisito correspondente está ou não incluso na próxima versão. Por exemplo, "1010" indicaria que as *features* A e C estão incluídas, enquanto B e D não estão.

No tocante à representação dos desenvolvedores, têm-se os seguintes atributos listados: nome, disponibilidade, nível do desenvolvedor e lista de funcionalidades. A representação dos níveis dos desenvolvedores é necessária para o cálculo da provisão da equipe de desenvolvedores, a fim de aferir qual é a quantidade de funcionalidades que esse time consegue desenvolver no decorrer da *sprint*. A Equação 1 utilizada para calcular a provisão de cada *sprint* é:

$$fp(x) = \sum_{i=0}^{n} p(d).t(d) \quad (1)$$

A quantidade de desenvolvedores da equipe é n. A capacidade de provisão de cada desenvolvedor d é p,enquanto t é a disponibilidade de tempo em horas semanais que cada desenvolvedor dispõe na *sprint*. O critério de parada usado é a quantidade de gerações.

A seguir, são identificados os atributos empregados na representação das funcionalidades, quais sejam: id; sistema/módulo; projeto/módulo; número da hierarquia do projeto; hierarquia do projeto; tipo; situação; título; desenvolvedor para o qual a *feature* foi atribuída; catálogo; HET[1]; quantidade de serviços; início; tarefa pai; quantidade de anexos; pontos de função; nível de prioridade (seguem os níveis de prioridade das *features* e suas respectivas pontuações aplicadas neste trabalho. Tais pontuações são: urgente (270), alta (90), média (30) e baixa (10), atribuídas e usadas. Esses dois últimos atributos citados indicam estados que variam entre os valores "verdadeiro" e "falso".

Nesta listagem é apresentada a descrição dos métodos do indivíduo utilizados:

- **inicializar:** insere no vetor de genes os valores 0 e 1 de forma aleatória.
- **Calcular restrição:** verifica se o atributo "valido" é verdadeiro, caso negativo chama o método "reparar" em *loop* até que o indivíduo seja válido.
- **reparar:** muda um gene 1 para 0, em uma posição aleatória.
- **função objetivo:** quantifica a aptidão do indivíduo.
- **getAptidao:** faz um somatório de todas as funcionalidades utilizadas pelo indivíduo, levando em consideração a função de aptidão.
- **getTotalPontoFuncao:** retorna ao somatório dos pontos de função.

1 Horas Efetivamente Trabalhadas: quantidade real de horas registradas como trabalho dedicado à implementação de uma funcionalidade, podendo divergir da estimativa inicial prevista (Departamento Estadual de Trânsito de Goiás, 2020).

- **mutar:** alterna o gene de 0 para 1, ou vice-versa, conforme a taxa de mutação passada via parâmetro.
- **isValido:** se o somatório de pontos de função for menor ou igual ao provisionamento retorna válido.

Algumas informações específicas da implementação da aplicação tratada neste trabalho são:

- **Parâmetros de entrada para a aplicação proposta, com seus respectivos exemplos de valores/formatos de entrada:** um arquivo com extensão .csv contendo a listagem de *features* e desenvolvedores disponíveis (colunas: #, sistema-módulo, projeto-módulo, # projeto hierarquia, projeto hierarquia, tipo, situação, título, atribuído para, catálogo, HET (real), quantidade de serviços, início, tarefa pai, quantidade de anexos); taxa de mutação (por exemplo, 0,07); tamanho da população (por exemplo, 100); quantidade de gerações (por exemplo, 1000); chance de cruzamento (por exemplo, 85%); tipo de cruzamento (opções: ponto de cruzamento ou máscara); tipo de seleção de indivíduo (opções: roleta ou torneio); e tamanho do torneio (por exemplo, 2).
- **Informações geradas e exibidas ao término da execução da aplicação:** população; máximo de gerações; taxa de mutação; tamanho do torneio; tempo; gasto; provisão; aptidão do melhor indivíduo; somatório dos pontos de função; quantidade de *features* por prioridade; chance de cruzamento; quantidade de *sprints*; quantidade de acessos à função objetivo; somatório dos pontos de aptidão do *product backlog*; *features*; benefícios; implementadores; *features* selecionadas para a próxima *sprint* e *features* usadas por *sprint*.

Os passos a seguir descrevem, de forma estruturada e em linguagem natural, o fluxo principal executado pelo método iniciar, localizado na classe NextReleaseProblemService (Figura 1). Esse método conduziu toda a lógica da aplicação proposta neste estudo, desde o carregamento dos dados de entrada até o processo evolutivo, que ocorreu ao longo de múltiplas *sprints* e gerações, utilizando operadores genéticos para encontrar soluções otimizadas para o Problema da Próxima Versão.

Carrega-se a lista de *features* a serem alocadas.
Carrega-se a lista de desenvolvedores disponíveis.
Calcula-se a quantidade total de horas disponíveis dos desenvolvedores para a *sprint*.
Cria-se a população inicial de soluções (indivíduos).
Define-se a quantidade de *sprints* necessárias com base nas restrições e no escopo.
Avalia-se a aptidão de cada solução da população inicial.
Inicia-se o processo evolutivo:
Para cada *sprint*:
a. Para cada geração:
i. Inicia-se uma nova população vazia.
ii. Para cada indivíduo da população atual:
– Selecionam-se dois pais (usando roleta ou torneio).
– Realiza-se o cruzamento (com ponto de corte ou máscara).
– Aplica-se a mutação (utilizando a técnica de *flip*).
– O novo indivíduo gerado é adicionado à nova população.
iii. A população atual é atualizada com os novos indivíduos.

Figura 1 – Algoritmo genético para solução do Problema da Próxima Versão
Fonte: Elaborado pelos(as) autores(as).

Operadores genéticos

As técnicas de seleção de indivíduos usadas na aplicação proposta foram:

- **Roleta:** a ideia é que indivíduos com maior aptidão tenham uma probabilidade maior de ser selecionados, semelhantemente a uma roleta, que gira para determinar um vencedor.
- **Torneio:** é uma técnica de otimização inspirada pelo processo de seleção natural, cujo objetivo é favorecer os indivíduos mais aptos, para que suas características positivas sejam transmitidas às gerações subsequentes. O tamanho do torneio aplicado nos experimentos apresentados neste estudo foi igual a 2.

As técnicas de cruzamento (em inglês, *crossover*) aplicadas na solução apresentada neste estudo foram:

- **Um ponto de corte:** nesta técnica, que se refere ao tamanho do vetor de genes, os genes do filho são criados com os genes do primeiro pai até o ponto de cruzamento, enquanto o restante dos genes é coletado, daquele ponto de cruzamento em diante, do segundo pai.
- **Máscara:** faz referência a um padrão binário, que determina quais genes de um indivíduo serão selecionados durante o processo de cruzamento com outro indivíduo para gerar descendentes.

A mutação é a técnica usada para introduzir pequenas alterações aleatórias em um indivíduo. Ela pode representar a adição ou remoção de requisitos da solução. Neste estudo, a técnica de mutação empregada foi a *flip*, que se baseia em trocar o valor do gen de 1 para 0 ou de 0 para 1, atendendo à taxa de mutação.

A função de aptidão é a técnica que avalia a qualidade de uma solução candidata em relação aos objetivos do projeto. Nesse estudo, a função de aptidão considerou múltiplos critérios, entre eles:

- **Prioridade:** a soma dos valores (importâncias) das *features* selecionadas. Cada requisito pode ter um peso associado, que reflete sua importância para os *stakeholders*.
- **Complexidade:** o custo estimado da implementação dos requisitos selecionados, levando em consideração o esforço de desenvolvimento, recursos necessários e custos associados.
- **Atendimento a prazos e recursos humanos disponíveis:** verificação se a seleção de *features* se adequa aos prazos estabelecidos e à força de trabalho disponível para o projeto.

A função de aptidão pode ser formulada como uma combinação ponderada desses critérios, com o objetivo de maximizar o valor do software, enquanto se mantém dentro das restrições de custo, técnica e prazo. A função de aptidão, Equação 2, usada na aplicação proposta neste trabalho está descrita abaixo.

$$fa(x) = PF(f).P(f) \quad (2)$$

A função de aptidão consiste na aptidão do indivíduo e é feita pelo somatório de todas as *features* associadas ao indivíduo, referente aos valores seguintes. Nessa

função *PF(f)* representa o ponto de função da *feature P(f)* e significa a pontuação de prioridade da *feature*. Essa é uma modelagem do NRP usando algoritmos genéticos para essa aplicação específica. A complexidade real pode variar, dependendo das nuances do problema e dos requisitos específicos do projeto. É importante ajustar essa modelagem com base em detalhes adicionais e realizar experimentos para determinar os parâmetros e configurações ideais do AG.

Em seguida gráficos produzidos e suas respectivas descrições acrescentadas ao final da execução, conforme descrito em Sommerville (2011).

- **Gráfico de Gantt:** é gerado a partir da listagem das *features* distribuídas por *sprint*. Neste modelo, cada *sprint* possui uma duração fixa de 30 dias. As *sprints* são sequencialmente distribuídas dentro deste intervalo de tempo, respeitando a data de início da primeira *sprint*. Este gráfico é essencial para visualizar o progresso do projeto, permitindo identificar o tempo previsto para a conclusão de cada *feature* e assegurar que as dependências e restrições de precedência sejam respeitadas ao longo do ciclo de desenvolvimento.
- **Gráfico de *burndown*:** representa o total de *features* ainda não concluídas ao longo das *sprints*. Essa visualização é essencial para acompanhar o progresso do projeto em relação ao tempo, permitindo verificar se a equipe está no ritmo adequado para concluir todas as *features* dentro do prazo estipulado. O gráfico evidencia a quantidade de trabalho restante em cada *sprint*, facilitando o controle do andamento das atividades e a adoção de estratégias para manter o projeto dentro do cronograma.

Experimentação e resultados

Um estudo de caso real foi conduzido por meio de um projeto de desenvolvimento de software em que o algoritmo genético foi aplicado na seleção de requisitos para a próxima versão. Os resultados práticos confirmaram a eficácia da abordagem e seguem expostos na Tabela 1, a seguir, cujas siglas exibidas (FU, PA, PF e TAP) significam, respectivamente: somatório das *features* utilizadas; somatório dos pontos de aptidão da *sprint;* somatório dos pontos de função da *sprint;* e taxa de aproveitamento da provisão na *sprint*.

As configurações do computador e das tecnologias utilizadas no experimento incluíram um processador Intel(R) Core(TM) i7-5500U, com CPU de 2.40 GHz, 16 GB de memória RAM e o sistema operacional Windows 10. No *back-end*, foi utilizado o *framework* Spring Boot, com as linguagens de programação Java 17 e Python 3.11.4. No *front-end*, foram empregados os *frameworks* Thymeleaf e Bootstrap. O Maven foi adotado como gerenciador de build e dependências, e a IDE utilizada foi o IntelliJ IDEA Community Edition.

Sprint	Seleção	População	Geração	Tempo	∑FU	∑PA	∑PF	TAP
1	Roleta	5	5	3s	156	33880	800	100,00%
	Torneio	5	5	5s	166	34080	798	99,75%
	Roleta	100	1000	1508s	157	39280	800	100,00%
	Torneio	100	1000	185s	168	42810	799	99,88%
2	Roleta	5	5	3s	140	30420	796	99,50%
	Torneio	5	5	5s	149	31190	797	99,62%
	Roleta	100	1000	1508s	153	31360	796	99,50%
	Torneio	100	1000	185s	142	28830	799	99,88%
3	Roleta	5	5	3s	156	30240	792	99,00%
	Torneio	5	5	5s	135	28180	800	100,00%
	Roleta	100	1000	1508s	147	24520	800	100,00%
	Torneio	100	1000	185s	156	24000	800	100,00%
4	Roleta	5	5	3s	156	23910	799	99,88%
	Torneio	5	5	5s	159	25380	800	100,00%
	Roleta	100	1000	1508s	137	24000	800	100,00%
	Torneio	100	1000	185s	149	24000	800	100,00%
5	Roleta	5	5	3s	138	23290	797	99,62%
	Torneio	5	5	5s	143	23110	797	99,62%
	Roleta	100	1000	1508s	159	23940	800	100,00%
	Torneio	100	1000	185s	139	23480	800	100,00%
6	Roleta	5	5	3s	18	2070	83	10,38%
	Torneio	5	5	5s	12	1870	75	9,38%
	Roleta	100	1000	1508s	11	710	71	8,88%
	Torneio	100	1000	185s	10	690	69	8,62%

Tabela 1 – Dados obtidos no experimento realizado por método de seleção
Fonte: Elaborado pelos(as) autores(as).

Na Figura 2 são mostrados os diagramas de engenharia de software gerados com emprego do *framework* JFreeChart ao término da execução do algoritmo genético:

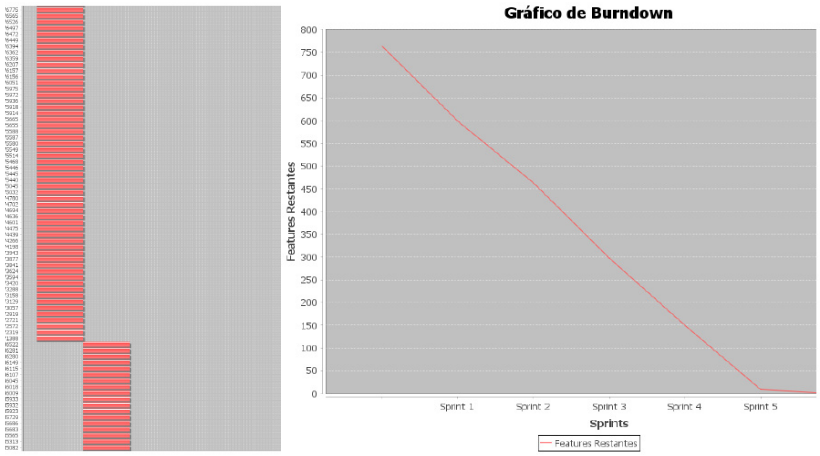


Figura 2 - Diagramas de engenharia de software gerados pelo sistema referido: (a) Diagrama de Gantt, (b) Diagrama de burndown.
Fonte: Elaborado pelos(as) autores(as).

A Tabela 2 apresenta a comparação das execuções do AG com os seguintes parâmetros: população de 100 e de 1000 gerações em relação ao algoritmo de força bruta (FB) (observação: o total de ∑PA é 143810).

Sprint	1 ()	2 (∑PA)	3 (∑PA)	4 (∑PA)	5 (∑PA)	6 (∑PA)
Roleta AG	39280	31360	24520	24000	23940	710
Torneio AG	42810	28830	24000	24000	23480	690
Força Bruta	48300	24000	24000	24000	22840	670

Tabela 2 – Dados obtidos no experimento realizado por método de seleção
Fonte: Elaborado pelos(as) autores(as).

Discussão

Com base nos resultados obtidos, observou-se que à medida que a quantidade de gerações aumentou, o AG melhorou a aptidão dos indivíduos. As métricas de avaliação da aptidão dos indivíduos (tempo de execução, taxa de aproveitamento da provisão e somatório da pontuação de aptidão da melhor solução encontrada) também melhoraram progressivamente, indicando que o algoritmo está escolhendo funcionalidades mais relevantes para a próxima versão e observando as restrições impostas. Isso demonstra a capacidade do algoritmo genético de otimizar a seleção de características para melhorar o desempenho do modelo preditivo.

O AG demonstrou ainda uma otimização significativa na seleção de *features,* valiosas para a próxima versão do software em comparação com execuções feitas por ele com os parâmetros “Gerações” e “População” menores. Isso foi evidenciado pela capacidade do AG de encontrar soluções que otimizam a relação entre os objetivos do projeto, no caso, a maximização do valor das funcionalidades desenvolvidas na próxima *sprint*.

A análise da convergência do AG demonstrou que, ao longo das gerações, ele foi capaz de identificar soluções ótimas ou próximas do ótimo em um número relativamente reduzido de iterações, o que resultou em economia de tempo de processamento. Como evidência, observou-se que mesmo utilizando configurações menos robustas, como população de tamanho 5 e apenas 5 gerações, a melhor solução obtida na primeira *sprint* com o método de seleção por roleta atingiu 86,25% do somatório da pontuação de aptidão alcançada com configurações significativamente superiores, envolvendo população de 100 indivíduos e de 1000 gerações.

Proposta de nova modelagem matemática para o Problema da Próxima Versão

A partir das informações do estudo de Elvassore (2016), diversas melhorias foram propostas para a implementação realizada. O principal motivo para utilizar o estudo de Elvassore como base foi a combinação de AG com o NRP, que é uma área de grande interesse. Além disso, o estudo utilizou o jMetal, um *framework* específico e bem estabelecido nesse campo de investigação. O estudo também fez uso de meta-heurística multiobjetivo de algoritmos genéticos, que são amplamente citadas na literatura atual.

A proposta elaborada envolve uma reformulação do NRP para melhor se adequar aos desafios atuais de pesquisa. Essa reformulação considera as *features* disponíveis e as atribui aos funcionários habilitados. Os principais pontos da proposta incluem considerar o custo como horas humanas e o valor como prioridade, além de implementar um planejamento preciso de *features* a serem desenvolvidas, respeitando restrições de precedência, disponibilidade de recursos humanos, competências e datas finais de desenvolvimento.

O NRP foi modelado considerando diversas variáveis, entre elas: as funcionalidades, entendidas como melhorias demandadas pelos clientes e associadas a custos específicos; as restrições de precedência, que estabelecem dependências entre *features*; os clientes, com seus respectivos valores estratégicos para a organização; os requisitos vinculados a cada *feature*; e um orçamento que não deve ser excedido. A adaptação do modelo proposta neste estudo incluiu: a atribuição de prioridade a cada funcionalidade; a definição de uma lista de funcionários com disponibilidade semanal; e a exigência de que cada *feature* seja executada apenas por profissionais com a habilidade correspondente. Além disso, a data final passou a substituir o orçamento global como restrição principal, com o objetivo de maximizar o número de *features* implementadas (ponderadas por sua prioridade) e, simultaneamente, minimizar a data de conclusão do projeto.

A implementação envolve classes para avaliar objetivos e restrições da solução, contendo *features* planejadas e funcionários. Também utiliza algoritmos de mutação, cruzamento, planejamento de *features* e geração de *features*.

Essa proposta de nova modelagem matemática para o NRP visa otimizar a seleção de *features* a serem implementadas em um software, maximizando o valor de negócio enquanto respeita as restrições de capacidade dos empregados e o esforço total. Nesta pesquisa, um algoritmo genético foi utilizado para explorar o espaço de soluções, modelando *features* e empregados com seus respectivos atributos e relacionamentos. Parâmetros como valor de negócio, esforço necessário, capacidade dos empregados e precedência entre *features* foram considerados, garantindo uma alocação eficiente e priorizada das tarefas.

A modelagem foi formulada com base em elementos matemáticos de otimização combinatória, sendo validada através de uma implementação que integra e executa essas regras. Essa abordagem proporciona uma solução prática e eficaz para a gestão de *releases* de software, permitindo uma melhor alocação de recursos e um planejamento estratégico mais preciso das funcionalidades a serem desenvolvidas em cada versão.

A seguir, apresenta-se a formulação matemática da nova modelagem proposta para o Problema da Próxima Versão.

Parâmetros:

- **Valor de negócio da *feature*:** importância ou benefício que a *feature* proporciona para a organização ou para o cliente.
- **Esforço necessário para implementação:** quantidade de trabalho ou recursos necessários para desenvolver a *feature*.
- **Conjunto de *features*:** conjunto de funcionalidades ou melhorias a serem consideradas para inclusão na próxima versão do produto.
- **Equipe de funcionários:** equipe de desenvolvimento disponível para trabalhar nas *features*.

- **Quantidade máxima de *features*:** limite máximo de funcionalidades que podem ser incluídas na próxima versão.
- **Esforço máximo disponível:** capacidade total de trabalho que a equipe de funcionários pode dedicar ao desenvolvimento das *features*.
- **Capacidade dos funcionários:** quantidade de trabalho que cada funcionário pode realizar no período considerado.
- **Habilidades dos funcionários:** competências e especializações dos membros da equipe que influenciam quais *features* eles podem implementar.
- **Precedência das *features*:** dependências entre as *features*, indicando quais precisam ser concluídas antes de outras.
- **Tipo da *feature*:** categoria ou classificação da *feature*, como *bug fix*, melhoria ou nova funcionalidade.
- **Status da *feature*:** estado atual da *feature*, como planejada, em desenvolvimento ou concluída.
- **Título da *feature*:** nome ou descrição breve da *feature*.
- **Responsável pela *feature*:** membro da equipe responsável por desenvolver a *feature*.
- **Quantidade de serviço da *feature*:** volume de trabalho ou complexidade associado à *feature*.
- **Data de início da *feature*:** data prevista ou efetiva de início do desenvolvimento da *feature*.
- **Data fim da *feature*:** data prevista para a conclusão ou entrega de uma determinada *feature*.
- **Tarefa pai:** tarefa maior ou projeto ao qual a *feature* está relacionada.
- **Quantidade de anexos:** número de documentos ou arquivos anexados à *feature* para suporte ou referência.
- **Prioridade da *feature*:** importância relativa em relação a outras *features*.

VARIÁVEIS DE DECISÃO:

- x_i: binário indicando se a *feature* i será implementada (1 se for implementada, 0 caso contrário).
- y_{ij}: binário indicando se a *feature* será implementada pelo empregado .

FUNÇÃO OBJETIVO:

- $max \sum_{i=1}^{N} V_i x_i$: maximizar o valor de negócio total das *features* selecionadas.

RESTRIÇÕES:

- $y_{ij} \leq x_i \; \forall \, i \in \{1, ..., N\}, \forall j \in \{1, ..., M\}$: capacidade dos empregados.
- $y_{ij} \leq x_i \; \forall \, i \in \{1, ..., N\}, \forall j \in \{1, ..., M\}$: relacionamento entre *features* e empregados.

- $x_i \leq x_{p_i}, \forall\, i \in \{1, \ldots, N\}, \forall\, P_i$: precedência entre *features*.
- $\sum_{i=1}^{N} E_i\, x_i \leq E_{max}$: limite de esforço total.
- $\sum x_i \leq F_{max}$: limite de número de *features*.

Nessa implementação descrita, várias tecnologias modernas e *frameworks* foram utilizadas para otimizar a execução e a eficácia da modelagem. O jMetal, por exemplo, foi utilizado para a implementação de algoritmos meta-heurísticos multiobjetivo, oferecendo uma base sólida para desenvolver soluções robustas. Diversos algoritmos genéticos, como Multi-Objective Cellular Genetic Algorithm (MOCell), Non-dominated Sorting Genetic Algorithm II (NSGA-II), (Pareto Envelope-based Selection Algorithm II (PESA-II) e (Strength Pareto Evolutionary Algorithm II (SPEA-II), foram escolhidos por sua eficácia em explorar o espaço de soluções e encontrar resultados otimizados.

A modelagem matemática é baseada em elementos de otimização combinatória, permitindo a formulação de modelos que consideram múltiplas restrições e objetivos. Técnicas avançadas de meta-heurísticas foram aplicadas para a exploração do espaço de soluções, considerando múltiplos parâmetros de negócios e técnicos.

Para testar a API, o Postman foi utilizado para fazer requisições e o Heroku para hospedar a API, facilitando o acesso e a execução remota. Além disso, o Oracle APEX foi usado para testes adicionais. A API pode ser testada por meio da URL https://apex.oracle.com/pls/apex/r/thiagoddcqg/nrpag/login, com as credenciais de acesso (usuário: demo, senha: demodemo), disponibilizadas para avaliação da API.

A imagem abaixo (Figura 3) mostra a interface da aplicação NRPAG no Oracle APEX.

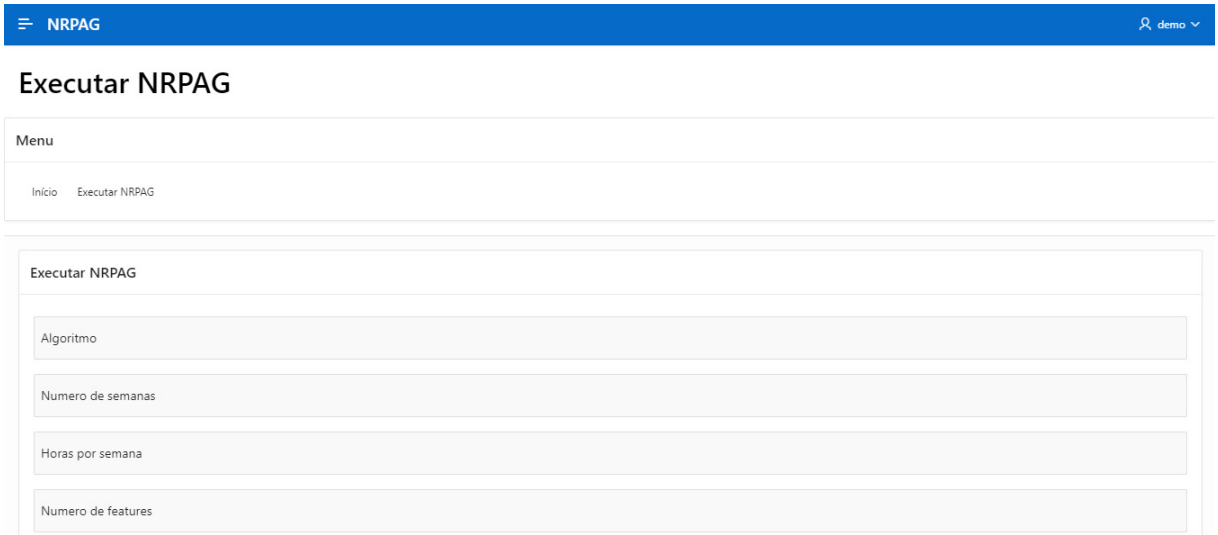


Figura 3 – Interface da aplicação NRPAG na plataforma Oracle APEX
Fonte: Elaborado pelos(as) autores(as).

Análise da nova modelagem para o Problema da Próxima Versão

A implementação aqui apresentada utiliza dados de empresas participantes do projeto Supersede (Projeto […], 2024) para avaliar a qualidade da solução, com pontuações baseadas em prioridade e data final. O projeto Supersede desenvolve ferramentas para a evolução e adaptação contínua de sistemas pervasivos, utilizando feedback de usuários e análise de dados em tempo real. Esses dados fornecem uma base sólida e realista para testar e validar as soluções propostas.

O protocolo do experimento inclui a comparação de diferentes algoritmos com variáveis, como número de *features* e funcionários. Os testes variam o número de *features* e funcionários, considerando restrições de precedência, habilidade requerida e otimização de funcionários. Essas variáveis são fundamentais para entender como cada algoritmo se comporta em cenários diferentes e para identificar as condições sob as quais cada algoritmo é mais eficaz.

Os resultados dos experimentos mostraram que o MOCell foi a solução mais eficaz e rápida para o NRP. Este algoritmo se destacou na capacidade de explorar o espaço de soluções e por encontrar respostas otimizadas rapidamente. Em comparação, NSGA-II e PESA-II apresentaram bons resultados, mostrando-se como alternativas viáveis, dependendo das especificidades do problema.

Por outro lado, o SPEA-II não se mostrou adequado para resolver o NRP proposto. Isso sugere que, apesar de suas capacidades em outras áreas, o SPEA-II pode ter limitações em lidar com as complexidades específicas do NRP quando comparado aos outros algoritmos testados.

Considerações finais

Neste trabalho, constatou-se que a eficácia do algoritmo genético no NRP está diretamente relacionada a uma função de aptidão bem ajustada, que considere métricas-chave como a prioridade e a complexidade das funcionalidades. Com base nos resultados obtidos, são propostas as seguintes melhorias para pesquisas futuras:

- **Utilização de mais dados:** incluir informações adicionais sobre as funcionalidades (como precedência) e sobre os desenvolvedores (como perfil e produtividade), a fim de aferir a aptidão de cada indivíduo de forma mais precisa e alinhada à realidade.
- **Conformidade com restrições técnicas:** incorporar esse critério no cálculo da função objetivo.
- **Gerenciamento de tempo:** aplicar técnicas avançadas de gerenciamento de tempo à solução gerada pelo algoritmo genético.
- **Execuções adicionais:** realizar um maior número de execuções na fase de experimentação para minimizar possíveis vieses nos resultados.
- ***Datasets* consolidados:** utilizar conjuntos de dados consolidados na área e repetir os experimentos diversas vezes, com o intuito de obter resultados estatisticamente significativos.

Essa abordagem proporciona uma solução prática e eficaz para a gestão de releases de software, permitindo uma melhor alocação de recursos e um planejamento estratégico mais preciso das funcionalidades a serem desenvolvidas em cada versão. Outras melhorias incluem a inclusão de mais meta-heurísticas de algoritmos genéticos para explorar diferentes abordagens de otimização.

A eficácia do algoritmo genético no NRP é influenciada por uma função de aptidão bem ajustada, que considera métricas importantes como prioridade e complexidade das *features*. A partir das melhorias propostas, espera-se uma maior precisão na avaliação da aptidão dos indivíduos, levando a soluções mais robustas e aplicáveis na prática.

Para trabalhos futuros, considera-se necessário: ampliar o conjunto de dados sobre as *features* e os desenvolvedores; incluir a conformidade com restrições técnicas; aplicar técnicas avançadas de gerenciamento de tempo; realizar mais execuções experimentais e utilizar *datasets* consolidados; incorporar um maior número de meta-heurísticas de algoritmos genéticos.

Essas melhorias e direções futuras visam aprimorar ainda mais a eficácia e a eficiência da modelagem matemática aplicada ao Problema da Próxima Versão, proporcionando uma ferramenta valiosa para a gestão estratégica de releases de software.

Referências

DEPARTAMENTO ESTADUAL DE TRÂNSITO DE GOIÁS. *Termo de referência* – Anexo I. Goiânia: Detran-GO, 2020. Disponível em: https://goias.gov.br/detran/wp-content/uploads/sites/8/2023/07/20201123-ANEXO-I-TERMO-DE-REFERENCIA.pdf. Acesso em: 30 abr. 2025.

ELVASSORE, Valentin. *Experimenting with generic algorithms to resolve the next release problem*. 2016. Dissertação (Mestrado em Inovação e Pesquisa em Informática - MIRI) — Institut Superieur d'Informatique, de Modelisation et de leurs Applications, Universitat Politècnica de Catalunya, Catalunya, 2016. Disponível em: https://upcommons.upc.edu/entities/publication/e2c6aabe-c3f3-4bef-b8fa-21d3bd45870f. Acesso em: 30 abr. 2025.

GORSCHEK, T.; WOHLIN, C.; ÖSTBERG, O. A staged model for systematic review. *Empirical Software Engineering*, [*s. l.*], v. 11, n. 4, p. 543-562, 2006.

HARMAN, Mark; MANSOURI, Sahar A.; ZHANG, Yuanyuan. Search-based software engineering: Trends, techniques and applications. *ACM Computing Surveys*, [*s. l.*], v. 45, n. 1, p. 1-61, 2012. DOI: https://doi.org/10.1145/2379776.2379787. Acesso em: 30 abr. 2025.

NIU, N.; HUANG, C.; JIN, H. An evolutionary algorithm for feature selection based on mutual information. *Information Sciences*, [*s. l.*], v. 178, n. 14, p. 2799-2813, 2008. DOI: https://doi.org/10.1016/j.ins.2015.02.031. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0020025515001310. Acesso em: 30 abr. 2025.

PROJETO SUPERSEDE. *GitHub*, [*s. l.*], c2025. Disponível em: https://github.com/supersede-project. Acesso em: 10 jun. 2024.

SARRAB, M.; AL SHIBLI, I. Optimization techniques for software release planning: A systematic literature review. *Journal of Software: Evolution and Process*, [*s. l.*], v. 31, n. 11, e2153, 2019.

SOMMERVILLE, I. *Software Engineering.* 9. ed. Boston, EUA: Addison Wesley, 2011.

SOUZA, L. F.; REBOUÇAS FILHO, P. P. Applications of genetic algorithm in software engineering, distributed computing and machine learning. *In*: INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCE AND ITS APPLICATIONS (ICCSA), 2020. *Anais* [...]. Springer, 2020. p. 309-324.

tecnia revista de educação, ciência e tecnologia do ifg

v. 10, Edição Especial 1 | 2025

Submetido 09/06/2024. Aprovado 20/03/2025
Avaliação: revisão duplo-anônimo

Previsão de vazão de rios usando rede perceptron multi-camada otimizada por neural architecture search

RIVER FLOW FORECASTING USING A MULTI-LAYER PERCEPTRON NETWORK. OPTIMIZED BY NEURAL ARCHITECTURE SEARCH

PREDICCIÓN DEL CAUDAL DE RÍOS USANDO UNA RED PERCEPTRÓN MULTICAPA OPTIMIZADA POR NEURAL ARCHITECTURE SEARCH

Eduardo Henrique Próspero Souza
Instituto Federal do Espírito Santo (Ifes)
duvrdx@gmail.com

Vinicius Marques de Oliveira
Petrobras
vinicius.armlock@hotmail.com

Jefferson Oliveira Andrade
Instituto Federal do Espírito Santo (Ifes)
jefferson.andrade@ifes.edu.br

Karin Satie Komati
Instituto Federal do Espírito Santo (Ifes)
kkomati@ifes.edu.br

Resumo

Este estudo propõe um modelo de rede neural otimizada pela abordagem de Neural Architecture Search (NAS) para a previsão da vazão de água de rios, utilizando dados de estações fluviométricas localizadas em seus afluentes. Os dados dessas estações são modelados como séries temporais multivariadas e servem como entrada para uma rede neural do tipo Perceptron Multicamada. Utilizaram-se duas bases de dados de rios: Rio Itapemirim e Rio Acre. Os resultados dos experimentos demonstram a capacidade dessa abordagem em capturar a complexidade e variabilidade dos dados de vazão dos rios, alcançando coeficientes de determinação (R^2) de 0,999 para o Rio Itapemirim e 0,975 para o Rio Acre. Esses resultados evidenciam que a rede neural tem uma alta capacidade de modelar a complexidade e variabilidade dos dados de vazão tanto do Rio Itapemirim quanto do Rio Acre.

Palavras-chave: série temporal multivariada; coeficiente de determinação; Rio Itapemirim; Rio Acre.

Abstract

This study proposes a neural network model optimized through the Neural Architecture Search (NAS) approach for forecasting river water flow using data from fluviometric stations located along tributaries. The data collected from these stations are represented as multivariate time series and serve as input to a Multilayer Perceptron (MLP) neural network. Two river basins were considered: the Itapemirim River and the Acre River. The experimental results demonstrate the effectiveness of the proposed approach in capturing the complexity and variability of river flow data, achieving coefficients of determination (R^2) of

0.999 for the Itapemirim River and 0.975 for the Acre River. These findings highlight the high modeling capacity of the neural network for accurately representing flow dynamics in both basins.

Keywords: multivariate time series; coefficient of determination; Itapemirim River; Acre River.

Resumen

Este estudio propone un modelo de red neuronal optimizado por el enfoque NAS (Neural Architecture Search) para predecir el caudal de agua de los ríos utilizando datos de estaciones hidrométricas ubicadas en sus afluentes. Los datos de estas estaciones se modelan como series temporales multivariadas y sirven como entrada para una red neuronal Perceptrón Multicapa. Se utilizaron dos conjuntos de datos de ríos: el río Itapemirim y el río Acre. Los resultados experimentales demuestran la capacidad de este enfoque para capturar la complejidad y variabilidad de los datos de caudal de los ríos, logrando coeficientes de determinación (R^2) de 0,999 para el río Itapemirim y 0,975 para el río Acre. Esto s resultados indican que el modelo de red neuronal captura eficazmente la complejidad y variabilidad de los datos de caudal. El estudio concluye que el enfoque optimizado de la red neuronal muestra un potencial significativo para la predicción precisa del caudal de los ríos.

Palabras clave: serie temporal multivariada; coeficiente de determinación; Río Itapemirim. Río Acre

Introdução

Diversas cidades brasileiras enfrentam desafios ambientais significativos em virtude da sua proximidade com os rios. Um dos principais problemas é a ocorrência de inundações, que resultam em perdas materiais e, em casos extremos, em perdas de vidas humanas (Saito *et al.*, 2023). Esses eventos têm impactos negativos tanto do ponto de vista socioeconômico quanto na qualidade de vida das populações afetadas (Farias; Mendonça, 2022). Nesse contexto, a capacidade de prever a possibilidade de inundações torna-se essencial para a implementação de medidas preventivas eficazes e para a mitigação dos danos. A utilização de tecnologias de monitoramento e modelagem, aliada a políticas públicas adequadas, pode contribuir significativamente para a redução dos riscos e para a proteção das comunidades vulneráveis.

O monitoramento hidrometeorológico consiste na observação contínua dos diferentes estágios do ciclo da água na superfície terrestre e na atmosfera (Agevap, 2024). A gestão desses dados envolve atividades de coleta, processamento, armazenamento, recuperação e disponibilização de informações históricas sobre as condições atmosféricas e a vazão dos rios. A vazão, por sua vez, é definida como a quantidade de fluido que atravessa uma determinada seção de escoamento. Esse processo é realizado por meio de uma rede de estações hidrométricas, que inclui estações pluviométricas para medir a precipitação, estações fluviométricas para monitorar o fluxo dos rios e estações meteorológicas para registrar dados climáticos.

O conhecimento da vazão de um rio ou de um corpo d'água é fundamental para a elaboração de planos de gestão sustentável, o controle de enchentes, o dimensionamento de barragens, a concessão de outorgas e a resolução de conflitos hídricos entre diferentes usuários de água, como a agricultura, o consumo humano e a dessedentação de animais (Andrade, *et al.*, 2017). Em termos fluviométricos, a vazão corresponde ao volume de água que escoa por uma seção do rio em um determinado intervalo de tempo.

A previsão da vazão da água em um rio, a partir dos dados obtidos por estações fluviométricas, pode ser modelada como um problema de séries temporais multivariadas (Lima; Marques; Orth, 2023). A previsão de séries temporais consiste em analisar

e modelar dados sequenciais registrados ao longo do tempo, com o objetivo de realizar projeções futuras com base nos padrões históricos. Uma série temporal univariada é definida como uma sequência de pontos de dados organizados em ordem sucessiva no tempo (Morettin; Toloi, 2018). Já uma série temporal multivariada envolve múltiplas variáveis dependentes do tempo, nas quais cada variável pode depender não apenas de seus próprios valores anteriores, mas também dos valores anteriores das demais variáveis (Morettin; Toloi, 2020). Em geral, uma série temporal é constituída por amostragens realizadas em pontos sucessivos igualmente espaçados no tempo, caracterizando uma sequência de dados em tempo discreto (Oliveira; Komati; Andrade, 2021).

Modelos de redes neurais têm se mostrado promissores na previsão de séries temporais multivariadas, sendo capazes de capturar relações não lineares entre variáveis que não são facilmente tratáveis por modelos estatísticos tradicionais (Gonzalez-Vidal; Jimenez; Gomez-Skarmeta, 2019). No entanto, o desempenho desses modelos depende da escolha adequada da arquitetura, motivo pelo qual muitos trabalhos recentes se dedicam a comparar e propor novas configurações para diferentes tipos de problemas. Embora seja possível otimizar os parâmetros por meio da experimentação, Miikkulainen *et al.* (2019) argumentam que a abordagem de tentativa e erro é dispendiosa, especialmente em redes neurais que envolvem centenas de milhares de hiperparâmetros.

Uma abordagem para encontrar a melhor arquitetura de rede neural é o AutoML (aprendizado de máquina automatizado), que consiste na automação total ou parcial da aplicação de métodos de aprendizado de máquina a um problema (Telikani *et al.*, 2021). O *AutoML* pode ser implementado por meio de duas tarefas principais: o Neural Architecture Search (NAS) e a otimização de hiperparâmetros (Hutter; Kotthoff; Vanschoren, 2019). O NAS busca automatizar a definição da topologia das redes neurais artificiais, enquanto a otimização de hiperparâmetros refere-se à seleção do conjunto de parâmetros que maximiza o desempenho do modelo de aprendizado.

O objetivo deste trabalho é desenvolver e avaliar modelos de redes neurais artificiais (RNA) do tipo Perceptron Multicamadas (MLP, do inglês *Multilayer Perceptrons*), otimizados por meio da abordagem de NAS, para o problema de predição da vazão de rios. O estudo amplia a pesquisa de Souza *et al.* (2023), que havia proposto um modelo MLP otimizado por NAS com base na implementação de Oliveira, Komati e Andrade (2024), utilizando dados do Rio Itapemirim e avaliação baseada no coeficiente de determinação (R^2). A nova pesquisa expande o escopo do trabalho anterior ao incorporar: (i) um conjunto de dados ampliado para o Rio Itapemirim, (ii) uma nova base de dados para o Rio Acre e (iii) a inclusão da métrica RMSE (*Root Mean Squared Error*, em português, Raiz do Erro Quadrático Médio).

Este artigo está organizado da seguinte forma: na Seção 2, são apresentados os conceitos de NAS e MLP; na Seção 3, são descritos os trabalhos relacionados; a Seção 4 aborda os materiais e métodos utilizados nos experimentos; a Seção 5 apresenta os resultados obtidos; e, por fim, a conclusão é discutida na Seção 6.

Fundamentação Teórica

Esta seção apresenta detalhes sobre o modelo MLP e a abordagem de NAS, dispondo sobre os principais conceitos teóricos que fundamentam o desenvolvimento deste trabalho.

Multilayer Perceptrons (MLP)

O MLP é um tipo de RNA composta por neurônios organizados em uma camada de entrada, uma ou mais camadas ocultas totalmente conectadas e uma camada de saída (Park; Lek, 2016). A camada de entrada recebe os dados iniciais, seguida pelas camadas ocultas, que processam as informações da camada anterior aplicando uma função de ativação à soma ponderada das entradas em cada neurônio, o que introduz não-linearidades essenciais para o aprendizado de padrões complexos. A quantidade de neurônios nas camadas ocultas é um fator importante no MLP, pois influencia a capacidade do modelo de capturar padrões nos dados. Um número maior de neurônios pode aumentar a expressividade da rede, mas também eleva o risco de sobreajuste. Por fim, a camada de saída gera a resposta final do modelo, cujo número de neurônios varia conforme a tarefa específica.

O funcionamento do MLP inicia-se com a propagação direta, na qual os dados de entrada atravessam a rede camada por camada. Cada neurônio calcula a soma ponderada de suas entradas e aplica uma função de ativação, como ReLU, *sigmoid* ou *tanh*, para gerar sua saída. A diferença entre a saída predita pela rede e o valor real é avaliada por meio de uma função de perda, como erro quadrático médio ou entropia cruzada. Esse erro é então propagado para trás na rede durante o processo de retropropagação (*backpropagation*), no qual os pesos das conexões são ajustados para minimizar a função de perda. A propagação direta e a retropropagação são repetidas ao longo de diversas iterações, ou épocas, até que a rede aprenda a mapear adequadamente as entradas para as saídas. O número máximo de iterações define o limite para o treinamento: valores altos podem levar ao sobreajuste, enquanto valores baixos podem impedir a convergência adequada do modelo.

O algoritmo de retropropagação utiliza o gradiente da função de perda em relação aos pesos da rede para atualizá-los de forma iterativa. A taxa de aprendizado controla a magnitude dessas atualizações, determinando o tamanho dos passos que o algoritmo de otimização realiza ao percorrer o espaço da função de erro em busca dos pesos que minimizem a perda. A escolha do otimizador, como o Stochastic Gradient Descent (SGD) ou o Adam, influencia diretamente a eficiência e a eficácia do treinamento. Além disso, o parâmetro *alpha*, também denominado termo de penalidade, é utilizado para restringir o tamanho dos pesos, contribuindo para a prevenção do sobreajuste (*overfitting*).

Haykin (2001) destaca que o MLP se sobressai como um modelo não linear geral, com capacidade de aprender padrões complexos sem apresentar viés indutivo. O viés indutivo ocorre quando um algoritmo de aprendizado de máquina busca, a partir de um conjunto de dados de treinamento, a hipótese que melhor descreva as relações entre os objetos e se ajuste aos dados (Bachmann; Anagnostidis; Hofmann, 2024). No entanto, o MLP apresenta algumas limitações: em geral, requer uma quantidade substancial de dados para generalizar adequadamente, bem como outros modelos de aprendizado profundo. Além disso, seu desempenho pode ser sensível à escolha dos hiperparâmetros.

Neural Architecture Search (NAS)

O Neural Architecture Search (NAS) é uma abordagem no campo do aprendizado de máquina que visa automatizar o processo de criação de arquiteturas de redes neurais profundas. Antes do surgimento dessa técnica, o design de modelos de aprendizado profundo dependia de conhecimento técnico especializado e de um processo manual de experimentação e ajuste de parâmetros. Pesquisadores precisavam

explorar diversas combinações de configurações e ajustar hiperparâmetros para cada aplicação específica (Zoph; Le, 2016). O NAS introduz um método automatizado que permite a exploração sistemática de espaços de arquitetura, identificando configurações adequadas para problemas específicos. Essa abordagem reduz a necessidade de intervenção manual e possibilita a construção de modelos com desempenho competitivo em menos tempo.

O NAS é uma abordagem que automatiza o processo de definição de arquiteturas de redes neurais, operando de maneira iterativa. Algoritmos de busca exploram um espaço predefinido de arquiteturas, testando diferentes configurações e selecionando aquelas que melhor atendem a critérios como precisão, desempenho computacional e eficiência no uso de recursos (Brandão; Correa; Guedes, 2023). A cada iteração, as arquiteturas candidatas são avaliadas, e os resultados orientam a escolha das próximas configurações, permitindo o ajuste até a identificação de uma solução adequada para o problema.

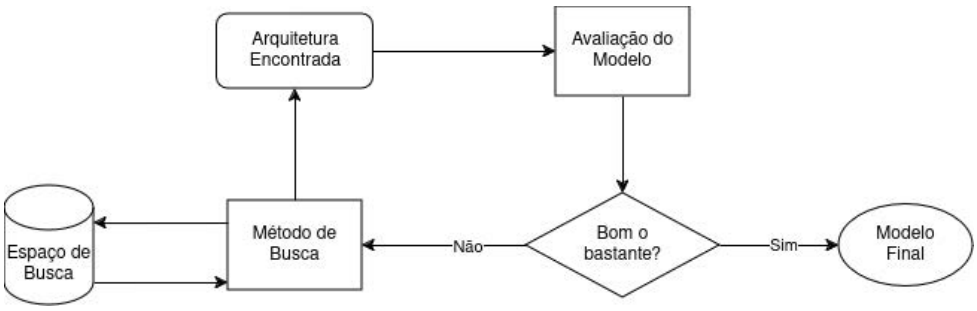


Figura 1 - Diagrama da técnica de NAS
Fonte: Elaborado pelos(as) autores(as).

Conforme ilustrado na Figura 1, o processo de NAS organiza-se em três componentes principais: espaço de busca, estratégia de busca e estratégia de estimativa de desempenho (Elsken; Metzen; Hutter, 2019). O espaço de busca define as arquiteturas possíveis, especificando os hiperparâmetros e suas combinações permitidas. A estratégia de busca é o método utilizado para explorar esse espaço de forma eficiente, evitando a estagnação em ótimos locais e promovendo a diversidade entre as configurações testadas. Já a estratégia de estimativa de desempenho é responsável por avaliar as arquiteturas candidatas e guiar a seleção das melhores alternativas, visando identificar o modelo com maior capacidade preditiva. Finalizada a busca, a arquitetura escolhida é utilizada para treinar o modelo definitivo, cujos resultados de predição são registrados para análise e validação posteriores.

Trabalhos Correlatos

O artigo de Suddul *et al.* (2020) propõe uma abordagem híbrida baseada em otimização evolutiva para a previsão em tempo real de inundações em rios. A metodologia combina variações do modelo MLP com algoritmos de otimização: o Algoritmo Genético (GA) e o Algoritmo dos Morcegos (BA). Foram desenvolvidos e avaliados quatro métodos: MLP com otimização manual, MLP otimizado por GA (MLP-GA), MLP treinado com BA (MLP-BA) e uma abordagem híbrida (MLP-BA-GA). Os experimentos utilizaram dados hidrológicos de uma rede de sensores do Rio Dee, no País de Gales, Reino Unido. Os resultados indicam que o modelo híbrido MLP-BA-GA apresentou o melhor desempenho, alcançando 99,86% de precisão (coeficiente R^2) no conjunto

de validação, superando tanto os modelos tradicionais quanto os otimizados individualmente por GA e BA. Embora o custo computacional na fase de treinamento seja elevado, o modelo final é considerado adequado para implementação em sistemas de alerta em tempo real.

O objetivo do trabalho de Santos-Neto *et al.* (2021) foi apresentar uma proposta de criação de um modelo hidroclimatológico da bacia do Rio Acre, utilizando RNA do tipo MLP. Dados mensais de temperatura da superfície do mar dos oceanos Pacífico e Atlântico tropicais, além do Atlântico Sudoeste, bem como da pressão média mensal em Darwin e Taiti, no período de 1971 a 2016, foram utilizados como entradas para a RNA. A base de dados contém 552 registros, sendo 87% usados para treinamento e 13% para teste. As previsões da cota máxima mensal do Rio Acre, com antecedência de um a quatro meses, foram feitas com simulações variando de 1 até 30 neurônios na camada oculta para cada horizonte de previsão, e métricas de desempenho foram aplicadas para avaliar a eficiência do modelo. As simulações apresentaram resultados satisfatórios, com um R^2 de 0,83 e RMSE de 0,2374 para a previsão com antecedência de um mês, utilizando um modelo MLP com 25 neurônios na camada oculta.

O trabalho de Brandão, Correa e Guedes (2023) apresenta uma análise comparativa de duas abordagens de RNA para a previsão dos níveis dos rios, MLP e LSTM (do inglês Long Short-Term Memory), utilizando dados de séries temporais de quatro estações hidrológicas na bacia do Rio Madeira – que passa pelos estados de RO, MT, AM e AC. As séries temporais da cota do rio em cada uma das quatro estações contemplam registros diários entre 2001 e 2014, totalizando 20.220 registros. Diversas configurações arquiteturais foram exploradas para as MLP, enquanto as LSTM consistiam de uma única camada oculta recorrente. A métrica de desempenho adotada foi o coeficiente de determinação (R^2), que quantifica a capacidade preditiva das RNA em relação aos dados reais. O melhor desempenho foi obtido com a MLP na estação de Porto Velho, alcançando R^2 de 0,963. Os resultados indicam que, em alguns casos, as LSTM superaram as MLP, especialmente em horizontes de previsão mais curtos; porém, para períodos de previsão mais longos, não foram observadas diferenças estatisticamente significativas de desempenho.

O artigo de Zanial *et al.* (2023) propõe um modelo híbrido que combina RNAs com o algoritmo de otimização Cuckoo Search (CS) para prever a vazão do rio na usina hidrelétrica de Terengganu, na Malásia. Utilizando dados históricos de precipitação e vazão (1971–2017) fornecidos pelo Departamento de Irrigação e Drenagem da Malásia, o estudo compara o desempenho de uma RNA tradicional com o do modelo híbrido CSRNA. A integração do Cuckoo Search visa otimizar os pesos e vieses da rede, mitigando problemas como convergência lenta e mínimos locais. Os resultados mostraram que o CS-RNA obteve desempenho superior, alcançando R^2 de 0,935 e RMSE de 10,95m^3/s na fase de teste em comparação aos valores de R^2 de 0,923 e RMSE de 12,7m^3/s da RNA convencional. O estudo conclui que o modelo híbrido é mais preciso e robusto.

O artigo de Souza *et al.* (2023) apresenta um estudo sobre o uso de modelos de séries temporais multivariadas para prever o nível e a vazão da água do Rio Itapemirim. Dois conjuntos de dados foram utilizados nos experimentos: (i) uma base em que a variável alvo é o nível do rio, com 608 registros de dados de vazão de três estações fluviométricas (denominada “Vazão e Nível”), e (ii) uma base em que a variável alvo é a vazão, utilizando 15.800 registros de dados de vazão das mesmas três estações (denominada “Somente Vazão”). Os dados de nível foram coletados manualmente pela Defesa Civil, enquanto os dados de vazão foram obtidos por meio de sensores distribuídos ao longo do percurso do rio ou de seus afluentes. Modelos do tipo

MLP foram otimizados utilizando a abordagem NAS para encontrar a melhor arquitetura para cada base de dados. Em ambos os casos, os dados foram divididos em 75% para treinamento e 25% para teste. A conclusão aponta que os resultados com a base "Somente Vazão" foram superiores, alcançando um R^2 de 0,955, valor semelhante ao do modelo matemático de Bof (2022), que obteve R^2 de 0,959.

O artigo de Mihel, Lerga e Krvavica (2024) apresenta uma revisão sistemática dos métodos de aprendizado de máquina (ML) aplicados à previsão, reconstrução e modelagem de relações nível-vazão em rios e estuários de influência tidal. Os autores destacam que, apesar do avanço no uso de ML em ambientes de água doce, existe uma lacuna significativa no tratamento de ambientes costeiros dinâmicos, como estuários e rios tidais, onde processos complexos e não lineares prevalecem. O artigo analisa 35 estudos, divididos entre abordagens baseadas em modelos estatísticos simples, classificadores e métodos de ensemble, redes neurais rasas, redes neurais recorrentes e modelos híbridos que combinam ML com métodos físicos ou de pré-processamento de dados. As bases de dados utilizadas incluem registros horários de níveis de água, vazões, dados meteorológicos e parâmetros de maré, coletados em diferentes locais ao redor do mundo. Os resultados quantitativos variam, mas indicam que modelos híbridos superam abordagens tradicionais em termos de métricas como RMSE e R^2, especialmente em cenários de previsão multistep e de alta complexidade. A revisão conclui que, embora redes neurais recorrentes e sistemas híbridos apresentem maior potencial para lidar com a variabilidade espaço-temporal dos processos hidrológicos, permanecem desafios como a necessidade de volumes de dados de alta resolução maiores e a dificuldade de generalização dos modelos em condições não observadas.

De modo geral, este trabalho apresenta similaridades e diferenças em relação aos estudos correlatos analisados. Assim como em Suddul *et al.* (2020) e Zanial *et al.* (2023), foi utilizado o modelo MLP em conjunto com técnicas de otimização automática. De maneira semelhante aos estudos de Brandão, Correa e Guedes (2023) e Santos-Neto *et al.* (2021), o MLP também foi empregado como preditor principal. Em particular, o trabalho de Santos-Neto *et al.* (2021) foca na previsão do nível do Rio Acre utilizando dados climatológicos como entrada, enquanto este estudo prevê a vazão, usando dados fluviométricos. Diferentemente de Zanial *et al.* (2023), que empregam o algoritmo de otimização Cuckoo Search, este trabalho adota uma abordagem baseada em NAS. Este artigo ainda expande o estudo de Souza *et al.* (2023), ampliando a base de dados do Rio Itapemirim, incorporando o Rio Acre e utilizando tanto RMSE quanto R^2 como métricas de avaliação. Alinhado aos artigos revisados, este trabalho adota as métricas RMSE e R^2 para avaliar o desempenho dos modelos.

Materiais E Métodos

Esta seção está organizada em quatro partes: a área de estudo dos dois rios analisados, o processo de elaboração das bases de dados utilizadas nos experimentos, as métricas adotadas para avaliação dos modelos e a descrição da técnica de NAS aplicada no treinamento do modelo MLP. A Figura 2 ilustra o fluxo geral do trabalho, abrangendo a coleta e o pré-processamento das bases de dados (destacados pelo retângulo azul), o processo iterativo de seleção do melhor modelo por meio do NAS (delimitado pelo retângulo preto) e, por fim, a execução dos testes com o modelo selecionado, seguida da coleta dos resultados e métricas.

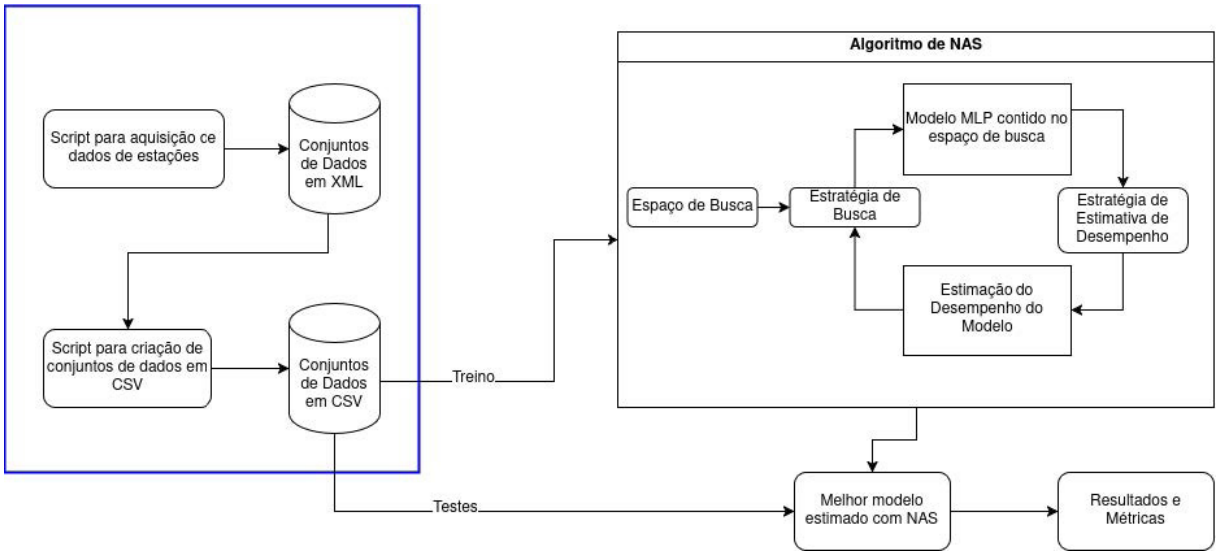


Figura 2 - Fluxograma de coleta de dados, treinamento e teste dos modelos
Fonte: Elaborado pelos(as) autores(as).

Áreas de Estudo

Esta subseção descreve as características geográficas e hidrológicas dos dois rios utilizados nos experimentos: o Rio Itapemirim e o Rio Acre. Serão apresentados dados sobre suas localizações, principais afluentes, extensão dos cursos d'água e aspectos relevantes para o contexto de previsão de vazão, complementados por mapas ilustrativos das bacias hidrográficas correspondentes.

RIO ITAPEMIRIM

O Rio Itapemirim está situado na cidade de Cachoeiro de Itapemirim, no estado do Espírito Santo, uma região historicamente afetada por episódios de alagamento (Souza; Santos, 2018). O rio é formado pela confluência de dois braços, que se unem no município de Alegre: o braço direito, que nasce em Muniz Freire, e o braço esquerdo, que nasce em Ibitirama, na Serra do Caparaó (AGERH, 2024). Seu principal afluente é o Rio Castelo, e a foz do Rio Itapemirim localiza-se no oceano Atlântico, na altura do município de Marataízes, também no Espírito Santo. A bacia hidrográfica do Rio Itapemirim ocupa uma área de aproximadamente 687.000 hectares e abrange 17 municípios.

A Figura 3 apresenta, no canto superior esquerdo, o mapa do Brasil com a localização do estado do Espírito Santo destacada em um retângulo lilás. No canto inferior esquerdo, vê-se uma ampliação desse retângulo, representando o mapa do estado. A parte direita da figura detalha a área de interesse, indicando a localização das quatro estações fluviométricas utilizadas: Francisco Gross (em amarelo), São João (em vermelho), Pacotuba (em azul) e Ilha da Luz (em lilás). A estação Ilha da Luz é a mais próxima da área urbana do município de Cachoeiro de Itapemirim, identificado pela sigla CI no mapa.

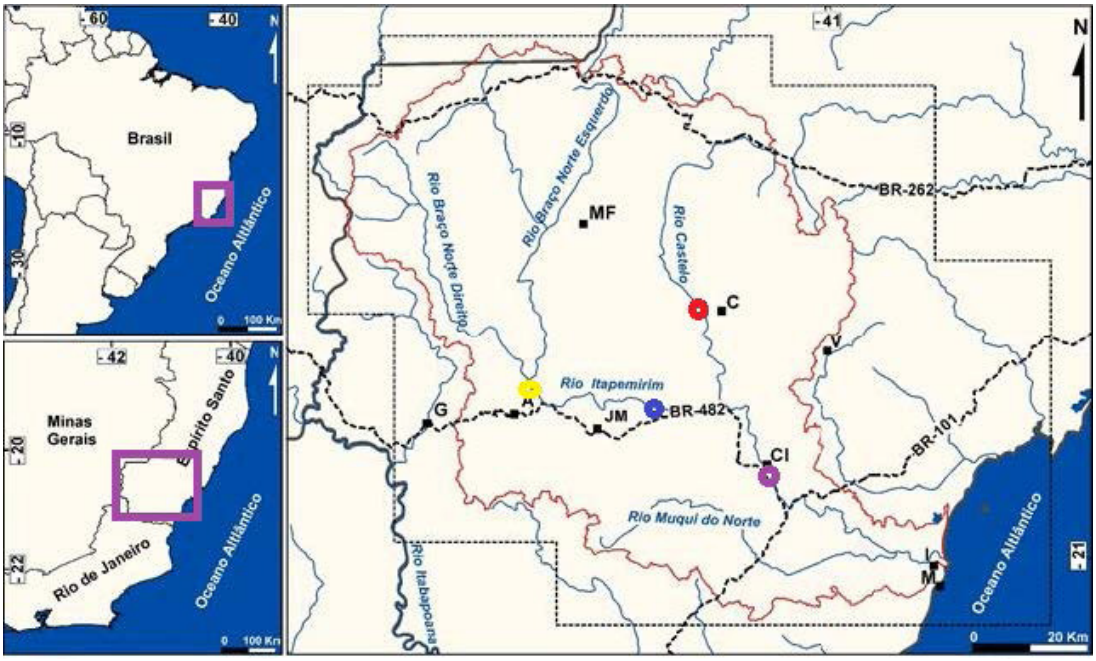


Figura 3 - Localização das estações da base de dados do Rio Itapemirim
Fonte: Adaptado de Neves *et al.* (2016).

RIO ACRE

O Rio Acre é um dos principais rios da Região Norte do Brasil e dá nome ao estado do Acre. Suas cheias já ocasionaram situações de emergência em diversos municípios ao longo de seu percurso (Monteiro; Menezes, 2024). A bacia hidrográfica do Rio Acre é formada por importantes afluentes, como o Rio Xapuri, o Riozinho do Rola e o Arroyo Bahia, além de igarapés, córregos, cursos de escoamento e esgotos urbanos (Sant'anna, 2017).

O rio nasce a uma altitude aproximada de 300 metros. Em seu curso superior, até a localidade de Seringal Paraguaçu (próxima a Assis Brasil), o Rio Acre atua como fronteira natural entre o Brasil e o Peru. Posteriormente, até a cidade de Brasileia, delimita a divisa entre o Brasil e a Bolívia. Após adentrar definitivamente no território brasileiro, o rio atravessa cidades como Porto Acre, segue pelo estado do Amazonas e deságua em sua foz. O Rio Acre percorre mais de 1.190 km desde suas nascentes até a desembocadura (IBGE, 2024).

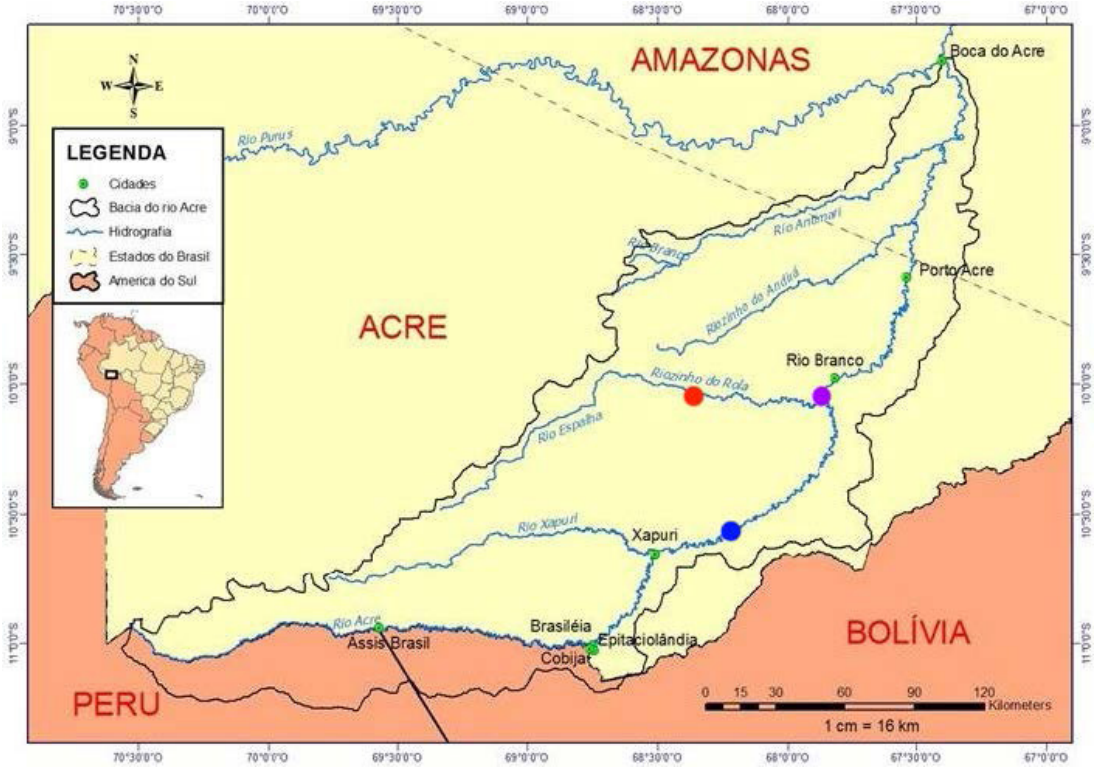


Figura 4 - Localização das estações da base de dados do Rio Acre
Fonte: Adaptado de Neto *et al.* (2017).

A Figura 4 apresenta, no retângulo localizado à esquerda, abaixo da legenda, o mapa da América do Sul com a localização do Rio Acre destacada por um retângulo preto. A ampliação dessa área, mostrada à direita, evidencia as principais estações utilizadas nos experimentos: a estação “Capixaba” (marcada em azul escuro), a estação “Rio Rola” (em vermelho) e a estação “Rio Branco” (em roxo), esta última situada próxima à área urbana da capital acreana, Rio Branco.

Bases de dados

Esta subseção apresenta os procedimentos de coleta e pré-processamento das bases de dados utilizadas nos experimentos. As fontes de dados empregadas foram o ServiceANA[1] e a plataforma HidroWEB.[2]

- O ServiceANA é um serviço web disponibilizado pela Agência Nacional de Águas (ANA) do Brasil, que oferece acesso a dados hidrológicos por meio de APIs. Entre as informações fornecidas, estão registros históricos sobre rios, bacias hidrográficas e estações telemétricas. Neste trabalho, utilizou-se o serviço de estações telemétricas para coletar dados de vazão das estações distribuídas ao longo dos rios estudados. Esses dados, captados automaticamente por sensores, são disponibilizados para acesso e download via API, possibilitando sua integração direta com sistemas de processamento de dados.

1 ServiçoAna (c2025).

2 Ana (c2025).

- O HidroWEB é mantido pela ANA. Trata-se de um portal que oferece acesso aos dados hidrológicos coletados por estações em todo o Brasil e em alguns países vizinhos. Além dos dados, o HidroWEB disponibiliza mapas interativos que mostram a localização de todas as estações, facilitando a visualização e seleção dos sensores mais relevantes para cada estudo. Esse recurso facilita o entendimento da distribuição geográfica dos sensores e a escolha das estações que melhor representam as condições hidrológicas dos rios estudados. Neste trabalho, o HidroWEB foi utilizado para identificar as estações mais bem posicionadas ao longo dos Rios Itapemirim e Acre.

Para a coleta de dados, foi desenvolvido um script que realiza requisições HTTP do tipo POST ao ServiceANA, retornando dados de vazão das estações selecionadas, com especificação de identificador, data de início e fim. As respostas são armazenadas em pastas específicas para cada rio. Todas as bases de dados foram construídas a partir do (i) download dos dados históricos via WebService e (ii) execução de um script Python, que converte e concatena os arquivos, gerando um único arquivo CSV. O código está disponível no GitHub.[3]

Para o Rio Itapemirim, foram utilizados os dados de vazão das três estações afluentes (Francisco Gross, São João e Pacotuba) e da estação alvo, Ilha da Luz. A Tabela 1 apresenta uma amostra de cinco registros, compostos por data, hora, vazões das três estações e a vazão da estação alvo no período entre 19-02-2021 à zero hora e 01-05-2024 às 23h.

Para o Rio Acre, os dados de entrada são das estações Rio Rola e Capixaba, com a variável alvo sendo a vazão da estação Rio Branco. A Tabela 2 também apresenta cinco registros, com informações de data, hora e vazões das três estações no mesmo período de coleta. Essa base foi denominada “Rio Acre”.

Data e Hora	Francisco Gross (m³/s)	Pacotuba (m³/s)	São João (m³/s)	Ilha da Luz (m³/s)
2023-10-02 22:00:00	86,63	103,38	13,48	260,72
2023-10-02 21:00:00	88,21	107,19	13,70	265,70
2023-10-02 20:00:00	92,99	113,03	13,70	271,98
2023-10-02 19:00:00	97,83	117,01	14,15	279,60
2023-10-02 18:00:00	101,08	121,05	14,15	286,04

Tabela 1 - Amostra dos dados da base “Rio Itapemirim”
Fonte: Elaborado pelos(as) autores(as).

Data e Hora	Rio Rola (m³/s)	Capixaba (m³/s)	Rio Branco (m³/s)
2022-09-04 17:45:00	0,80	63,70	29,10
2022-09-04 17:30:00	0,80	63,70	28,70
2022-09-04 17:15:00	0,80	63,70	29,40
2022-09-04 17:00:00	0,80	63,70	30,50
2022-09-04 16:45:00	0,80	63,70	30,10
2022-09-04 16:30:00	0,80	63,70	30,10

Tabela 2 - Amostra dos dados da base “Rio Acre”
Fonte: Elaborado pelos(as) autores(as).

3 Conjunto [...] (c2025).

Métricas

Para avaliar o desempenho dos modelos, foram utilizadas duas métricas: o coeficiente de determinação R^2 e o RMSE (Root Mean Squared Error) (Zanial *et al.*, 2023). O coeficiente R^2 é uma medida estatística que indica a proporção da variabilidade da variável dependente (variável alvo) explicada pelas variáveis independentes (variáveis preditoras) em um modelo de regressão, conforme apresentado na Equação 1.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \quad (1)$$

Nesse caso, n é o numero de observações, Y_i é o valor observado, $\bar{Y}$ é a média das observações e $\hat{Y}_i$ é o valor estimado (previsão) de Y_i. O R^2, sendo usado para avaliar quão bem um modelo se ajusta aos dados. Seus valores variam de 0 a 1, um valor elevado (próximo de 1) sugere que o modelo é capaz de capturar grande parte da variação nos dados, enquanto valores próximos de 0 indicam baixo desempenho explicativo. De acordo com Prairie (1996), valores de superiores a 0,9 são considerados indicativos de alta capacidade preditiva, enquanto valores iguais ou inferiores a 0,65 são insatisfatórios.

A métrica RMSE mede a diferença entre os valores previstos por um modelo e os valores reais de um conjunto de dados, sendo calculada como a raiz quadrada da média dos quadrados dos erros. Essa métrica quantifica a dispersão dos erros, indicando o grau de proximidade das previsões dos valores reais. Valores menores de RMSE sugerem melhor desempenho do modelo. O cálculo do RMSE é apresentado na Equação 2.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \quad (2)$$

Sendo n é o numero de observações, Y_i é o valor observado e $\hat{Y}_i$ é o valor estimado (previsão) de Y_i. A unidade do RMSE é a mesma unidade de Y_i.

MLP Otimizado pela Técnica de NAS

Neste trabalho, adotou-se a metodologia de NAS baseada na proposta de Oliveira, Komati e Andrade (2024). Nesse estudo, os autores desenvolveram um modelo de previsão de consumo de gás em uma planta de pelotização utilizando técnicas de aprendizado de máquina (AM), com o objetivo de comparar seu desempenho em relação a métodos estatísticos tradicionais. A avaliação dos modelos foi realizada com a métrica RMSE, e os resultados indicaram que as técnicas de AM superaram os métodos estatísticos convencionais. Entre os modelos avaliados, o destaque foi o MLP otimizado por NAS, que não apenas apresentou menor erro de previsão, mas também um tempo de predição significativamente reduzido, sendo considerado o mais promissor para aplicação em ambientes de produção.

A abordagem de NAS foi implementada utilizando um algoritmo de otimização baseado na combinação de *Hill Climbing* e *greedy search* (Elsken; Metzen; Hutter, 2017). O *Hill Climbing* é um método heurístico que parte de uma solução inicial e

realiza pequenas alterações, aceitando apenas as mudanças que melhoram o desempenho, até que não sejam encontradas novas melhorias. Já o *greedy search* é uma estratégia que seleciona, a cada etapa, a opção localmente mais promissora, buscando alcançar uma solução global ótima. No contexto de NAS, a combinação dessas técnicas permite explorar o espaço de arquiteturas de maneira eficiente, incrementando gradativamente a qualidade das soluções encontradas.

A Tabela 3 apresenta o conjunto de valores considerados para cada hiperparâmetro explorado. Os hiperparâmetros da arquitetura MLP analisados incluem: quantidade de neurônios na camada oculta, funções de ativação, valor de Alpha (termo de regularização), algoritmos otimizadores, taxa de aprendizado e número máximo de iterações.

Hiperparâmetros	Valores
Quantidade de neurônios na camada oculta	[10, 20, 30, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500]
Função de Ativação	[identity, logistic, relu, tanh]
Alpha	[0,001, 0,0001, 0,00001, 0,000001, 0,0000001, 0.00000001]
Otimizador	[SGD, RMSprop, Adam]
Taxa de aprendizado	[0,1, 0,01, 0,001, 0,0001, 0,00001, 0,000001]
Quantidade máxima de iterações	[50, 100, 150, 200, 250, 300, 350, 400, 450, 500]

Tabela 3 - Parâmetros e valores via NAS com AutoKeras
Fonte: Elaborado pelos(as) autores(as).

Experimentos, Resultados e Discussão

As bases de dados foram divididas em 75% para treino e 25% para teste, seguindo o procedimento adotado por Oliveira, Komati e Andrade (2024). O número de épocas para os experimentos foi fixado em 10. A Tabela 4 apresenta os hiperparâmetros dos melhores modelos encontrados pelo algoritmo NAS para as duas bases de dados. Observa-se que os dados do Rio Acre parecem apresentar maior complexidade, refletida na necessidade de uma arquitetura com um número significativamente maior de neurônios na camada oculta (150) em comparação ao modelo do Rio Itapemirim (10).

Hiperparâmetros	Rio Itapemirim	Rio Acre
Qtde de neurônios na camada oculta	10	150
Função de Ativação	relu	relu
Alpha	1e-08	1e-03
Otimizador	Adam	Adam
Taxa de Aprendizado	0.001	0.1
Qtde máxima de iterações	500	350

Tabela 4 - Tabela com os hiperparâmetros da arquitetura dos melhores modelos por base de dados
Fonte: Elaborado pelos(as) autores(as).

A Tabela 5 apresenta os resultados das métricas obtidas nos conjuntos de teste. A primeira coluna identifica o rio e a referência ao artigo correspondente, a segunda coluna mostra o valor do R^2, a terceira exibe o RMSE e a quarta informa a quantidade total

de registros da base de dados. A variável alvo para o Rio Itapemirim é a vazão (m^3/s) na estação Ilha da Luz, e para o Rio Acre é a vazão (m^3/s) na estação Rio Branco.

Rio e artigo	R2	RMSE	registros
Rio Itapemirim (este artigo)	0,999	0,032	25.280
Rio Itapemirim (SOUZA *et al.*, 2023)	0,955	0,032	15.800
Rio Acre (este artigo)	0,975	92,963	40.177
Rio Acre (SANTOS-NETO *et al.*, 2021)	0,83	0,2374	552

Tabela 5 - Tabela comparativa com as métricas dos experimentos deste trabalho e trabalhos correlatos
Fonte: Elaborado pelos(as) autores(as).

Lembrando que o estudo de Prairie (1996) indicou que valores de R^2 superiores a 0,9 podem ser considerados como representativos de alta capacidade preditiva. Observa-se que o modelo desenvolvido para o Rio Itapemirim obteve um valor de R^2 de 0,999 e um erro RMSE de apenas 0,032 m^3/s. Além disso, o R^2 de 0,999 superou o resultado obtido por Souza *et al.* (2023), cujo modelo apresentou R^2 de 0,955. A Figura 5 apresenta a curva dos dados reais de vazão do Rio Itapemirim (em metros cúbicos por segundo) em laranja, juntamente com a curva de predição em azul. As curvas são tão próximas que, em muitos trechos, a linha azul torna-se imperceptível sobre a linha laranja. Destaca-se que o modelo conseguiu prever corretamente o pico de vazão, indicado pela seta vermelha, ocorrido em março de 2024, após as fortes chuvas no estado do Espírito Santo (Oliveira, 2024).

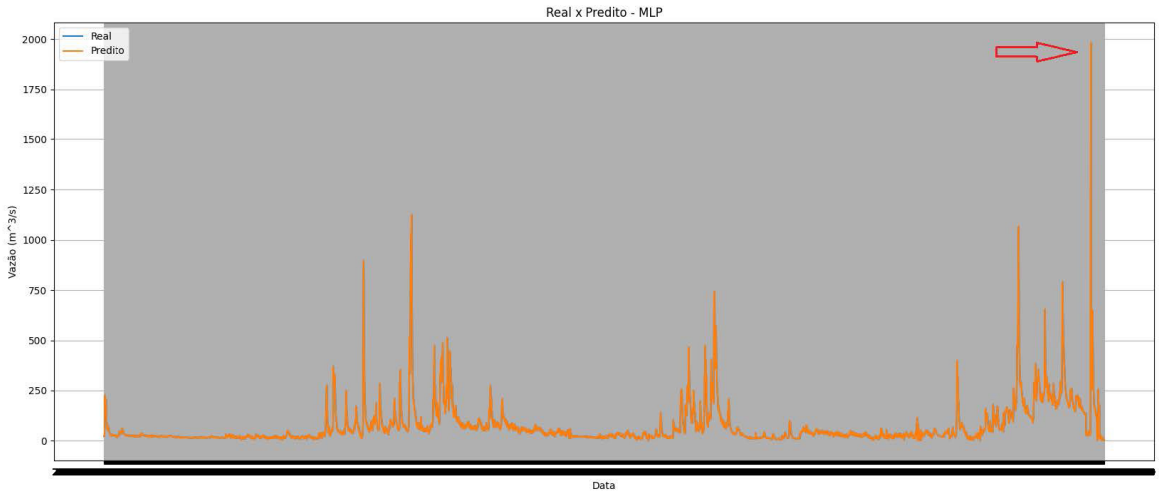


Figura 5 - Gráfico comparando a curva de dados de vazão do Rio Itapemirim e o valor predito pelo melhor modelo
Fonte: Elaborado pelos(as) autores(as).

Os valores previstos são extremamente próximos aos valores reais, o que pode ser um indicativo de possível overfitting. No entanto, esse comportamento também pode refletir o fato de que o uso de uma base de dados maior (25.280 registros em comparação aos 15.800 registros do estudo anterior) contribuiu para uma modelagem mais precisa dos dados, o que está de acordo com a literatura, que aponta a necessidade de grandes volumes de dados para que MLPs consigam generalizar adequadamente. Resultados semelhantes são observados na literatura: o artigo de Suddul *et al.* (2020) obteve um R^2 de 0,9986 utilizando um MLP para previsão no Rio Dee, enquanto o estudo de Brandão,

Correa e Guedes (2023) atingiu R2 de 0,963 ao prever níveis do Rio Madeira no período de 2001 a 2014, utilizando uma base com 20.220 registros.

Para a base de dados Rio Acre, o melhor modelo atingiu um R^2 de 0,975 e um RMSE de 92,963 m3/s. O valor de R^2 pode ser considerado bom, indicando uma alta capacidade preditiva; entretanto, o valor do RMSE é relativamente elevado, correspondendo a aproximadamente 17% do valor médio da vazão (521 m^3/s). Esse erro pode ser percebido de forma qualitativa nas curvas da Figura 6, em que a curva dos dados reais da vazão do rio em metros cúbicos por segundo é representada em laranja, e a curva da predição em azul. Observa-se que, embora existam diferenças entre as curvas, a predição acompanha, de maneira geral, as variações da curva real. Dois picos importantes são destacados: o primeiro, indicado pela seta vermelha, corresponde à inundação ocorrida em abril de 2023 (Rodrigues, 2023); e o segundo, indicado pela seta verde, refere-se à inundação de março de 2024 (Monteiro; Menezes, 2024).

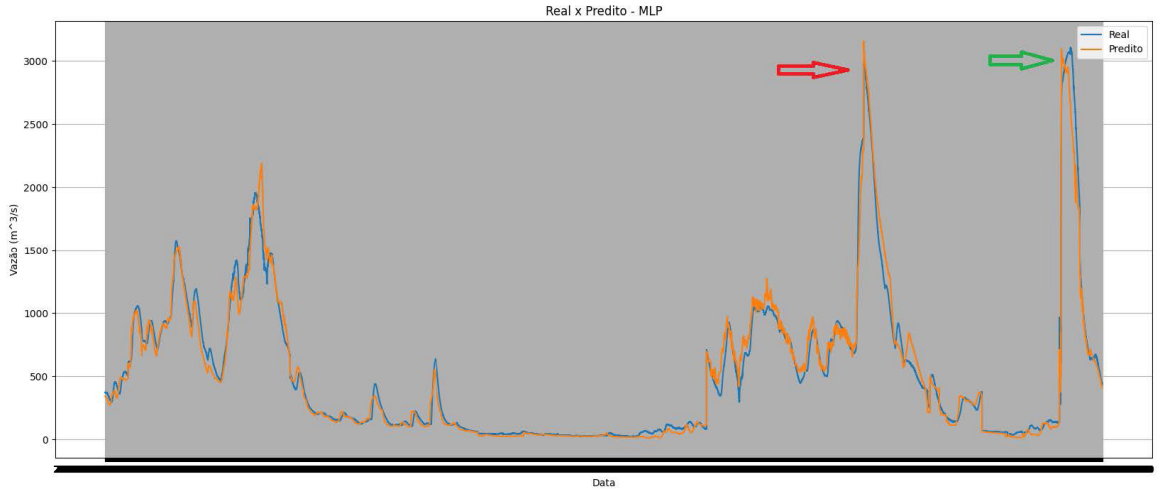


Figura 6 - Gráfico comparando a curva de dados de vazão do Rio Acre e o valor predito pelo melhor modelo
Fonte: Elaborado pelos(as) autores(as).

O bom desempenho obtido na previsão da vazão para o Rio Itapemirim pode ser parcialmente atribuído à sua configuração hidrográfica mais simples e bem definida. O Rio Itapemirim é formado pela confluência de dois principais braços – o braço direito e o braço esquerdo. Essa formação relativamente linear e a estrutura de drenagem menos complexa favorecem a modelagem preditiva, especialmente porque foram utilizadas estações fluviométricas posicionadas estrategicamente em cada um dos braços (Francisco Gross e São João). Essa escolha de estações permite captar a maior parte da variabilidade hidrológica relevante para a previsão da vazão na área de interesse.

Em contraste, o Rio Acre apresenta características hidrológicas e geográficas mais desafiadoras. Com uma extensão superior a 1.190 km, o Rio Acre atravessa regiões de fronteira entre Brasil, Peru e Bolívia antes de adentrar definitivamente no território brasileiro. Sua bacia hidrográfica é composta por uma rede densa e complexa de afluentes, igarapés e escoamentos urbanos. Nos experimentos realizados neste trabalho, foram utilizadas apenas estações associadas aos afluentes Rio Rola e Rio Xapuri, sem cobrir toda a complexidade da bacia. Essa limitação na abrangência dos dados pode ter impactado negativamente a capacidade do modelo de capturar todas as dinâmicas hidrológicas do sistema, resultando em um erro de predição mais elevado (RMSE de 92,963 m^3/s) em comparação ao Rio Itapemirim. Essa limitação na representatividade dos

dados para o Rio Acre deve ser considerada como um viés relevante nos experimentos realizados.

Conclusão

Este trabalho teve como objetivo realizar a predição da vazão de rios utilizando dados de estações fluviométricas de seus afluentes por meio de modelos MLP otimizados pela abordagem de NAS. Os resultados obtidos demonstraram que a rede neural foi capaz de capturar a complexidade e a variabilidade dos dados de vazão tanto do Rio Itapemirim quanto do Rio Acre. A adoção de técnicas de AutoML, especificamente o NAS, evidenciou a relevância da automação na busca eficiente por arquiteturas adequadas, reduzindo a necessidade de intervenção manual intensiva – um aspecto particularmente importante em aplicações dinâmicas como a previsão hidrológica.

As diferenças de desempenho observadas entre os dois rios reforçam a importância da representatividade espacial das estações na modelagem de sistemas hidrológicos, indicando que a escolha e a distribuição dos pontos de coleta têm impacto direto na qualidade das previsões. Assim, como um dos trabalhos futuros, propõe-se a ampliação das bases de dados com informações de outras estações fluviométricas, bem como a extensão da abordagem para outros rios, a fim de validar a generalização e a adaptabilidade do modelo.

Além disso, a incorporação de variáveis adicionais, como dados meteorológicos, pode enriquecer a modelagem e melhorar ainda mais a precisão das previsões. Pretende-se também investigar o uso de outros modelos, como abordagens híbridas, sugeridas por Mihel, Lerga e Krvavica (2024), explorar outras técnicas de AutoML, comparar o tempo de execução das diferentes arquiteturas e integrar a opinião de especialistas em hidrologia ou gestão de recursos hídricos.

Agradecimentos

A professora Komati agradece ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa DT-2 (n. 302726/2023-3) e pelo projeto n. 407742/2022- 0; também agradece à Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (Fapes) pelo projeto n. 1023/2022 P:2022-8TZV6.

Referências

AGÊNCIA ESTADUAL DE RECURSOS HÍDRICOS. *Diagnóstico e Prognóstico das Condições de Uso da Água na Bacia Hidrográfica do Rio Itapemirim como Subsídio Fundamental ao Enquadramento e Plano de Recursos Hídricos*. Espírito Santo: AGERH, 2024. Disponível em: https://agerh.es.gov.br/Media/agerh/Documenta%C3%A7%C3%A3o%20CBHs/Itapemrim/RT_Levantamento_Dados_CBH%20Itapemirim.pdf. Acesso em: 1 maio 2024.

AGÊNCIA NACIONAL DAS ÁGUAS. Rede Hidrometeorológica Nacional. Brasília, DF: ANA, c2025. Disponível em: https://www.snirh.gov.br/hidroweb/apresentacao. Acesso em: 9 set. 2025.

ANDRADE, P. L.; SOUZA, B. M.; SOUZA, J. W. F.; PARNAÍBA, M. A. Otimização do método dos molinetes com ajuste dos perfis hidrodinâmicos para a estimação da descarga líquida em corpos hídricos por meio de técnicas de interpolação e integração numérica. *In*: 31º CONGRESSO BRASILEIRO DE MATEMÁTICA (CBM), 2017, João Pessoa. *Anais* […]. João Pessoa: CBM, 2017.

ASSOCIAÇÃO PRÓ-GESTÃO DAS ÁGUAS DA BACIA HIDROGRÁFICA DO RIO PARAÍBA DO SUL. *Monitoramento*. Seropédica: Agevap, 2024. Disponível em: https://comiteguandu.org.br/monitoramento/. Acesso em: 1 maio 2024.

BACHMANN, G.; ANAGNOSTIDIS, S.; HOFMANN, T. Scaling MLPs: A tale of inductive bias. *Advances in Neural Information Processing Systems*, [*s. l.*], v. 36, 2024.

BOF, L. *SARI* - Sistema de alerta do Rio Itapemirim. Espírito Santo: Agerh, 2022. Disponível em: https://servicos.agerh.es.gov.br/sari/. Acesso em: 10 out. 2023.

BRANDÃO, J.; CORREA, F.; GUEDES, E. A comparative analysis of artificial neural networks on river level forecasting for the Rio Madeira Basin. *In*: XX ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL, 2023, Porto Alegre. *Anais* […]. Porto Alegre: SBC, 2023. Disponível em: https://sol.sbc.org.br/index.php/eniac/article/view/25699. Acesso em: 8 set. 2025.

CONJUNTO de dados rio. Github, [*s. l.*], c2025. Disponível em: https://github.com/duvrdx/dataset_rio. Acesso em: 9 set. 2025.

ELSKEN, T.; METZEN, J. H.; HUTTER, F. Neural architecture search: A survey. *The Journal of Machine Learning Research*, [*s. l.*], v. 20, n. 1, p. 1997-2017, 2019.

ELSKEN, T.; METZEN, J.-H.; HUTTER, F. Simple and efficient architecture search for convolutional neural networks. *arXiv*, [*s. l.*], 2017.

FARIAS, A.; MENDONÇA, F. Riscos socioambientais de inundação urbana sob a perspectiva do sistema ambiental urbano. *Sociedade & Natureza*, Uberlândia, v. 34, n. 1, 2022.

GONZALEZ-VIDAL, A.; JIMENEZ, F.; GOMEZ-SKARMETA, A. F. A methodology for energy multivariate time series forecasting in smart buildings based on feature selection. *Energy and Buildings*, Elsevier, v. 196, p. 71-82, 2019.

HAYKIN, S. S. Redes neurais: princípios e prática. [*S. l.: s. n.*], 2001.

HUTTER, F.; KOTTHOFF, L.; VANSCHOREN, J. *Automated machine learning*: methods, systems, challenges. 1. ed. Switzerland: Springer Nature, 2019. ISBN 978-3-030-05318-5.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. *Biblioteca IBGE*. Rio Branco, AC: IBGE, 2024. Disponível em: https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=4991. Acesso em: 1 maio 2024.

LIMA, P.; MARQUES, F.; ORTH, S. Predição do nível de água utilizando os modelos ARIMA e Random Forest: Um estudo de caso da barragem eclusa do São Gonçalo. *In*: L SEMINÁRIO INTEGRADO DE SOFTWARE E HARDWARE, 2023, Porto Alegre. *Anais* […]. Porto Alegre: SBC, 2023. ISSN 2595-6205. Disponível em: https://sol.sbc.org.br/index.php/semish/article/view/25059. Acesso em: 8 set. 2025.

MIHEL, A. M.; LERGA, J.; KRVAVICA, N. Estimating water levels and discharges in tidal rivers and estuaries: Review of machine learning approaches. *Environmental modelling & software*, Elsevier, p. 106033, 2024.

MIIKKULAINEN, R.; LIANG, J.; MEYERSON, E.; RAWAL, A.; FINK, D.; FRANCON, O.; RAJU, B.; SHAHRZAD, H.; NAVRUZYAN, A.; DUFFY, N. Evolving deep neural networks. *In*: ARTIFICIAL Intelligence in the Age of Neural Networks and Brain Computing. 1. ed. United Kingdom: Elsevier, 2019. p. 293-312. ISBN 9780128162507.

MONTEIRO, H.; MENEZES, R. Mais de 11 mil pessoas estão fora de casa por enchente no AC e governo federal reconhece emergência. *GI Globo*, [*s. l.*], 2024. Disponível em: https://g1.globo.com/ac/acre/noticia/2024/02/26/mais-de-11-mil-pessoassaoafetadaspor-enchente-no-ac-e-governo-federal-reconhece/-situacaodeemergencia.ghtml. Acesso em: 1 maio 2024.

MORETTIN, P. A.; TOLOI, C. M. Análise de séries temporais – Vol. 1: modelos lineares univariados. 3. ed. São Paulo: Editora Blucher, 2018. 474 p.

NETO, L. S.; MANIESI, V.; SILVA, M. J.; SILVA, D.; QUERINO, C.; REIS, V. Análise da precipitação mensal e pentadal durante a cheia de 2015 no Rio Acre usando o produto 3B43 do TRMM. *In*: VII SIMPÓSIO INTERNACIONAL DE CLIMATOLOGIA, 2017, Petrópolis. *Anais* […]. Petrópolis: [*S. n.*], 2017.

NEVES, M. A. N.; MIRANDA, R. F.; TRIGO, M. S.; OLIVEIRA, M. S. M.; PESSOA, A. D.; MANCINI, L. H. Assinatura isotópica das águas pluviais e subterrâneas na bacia hidrográfica do Rio Itapemirim, estado do Espírito Santo. *Revista Águas Subterrâneas*, Belo Horizonte, 2016. Disponível em: https://aguassubterraneas.abas.org/asubterraneas/article/view/28688. Acesso em: 8 set. 2025.

OLIVEIRA, S. Rio Itapemirim sobe com velocidade e prefeitura de Cachoeiro faz alerta. *A Gazeta*, [*s. l.*], 2024. Disponível em: https://www.agazeta.com.br/agora/rioitapemirim-sobe-com-velocidade-e-prefeitura-decachoeiro-faz-alerta-0324. Acesso em: 1 maio 2024.

OLIVEIRA, V. M.; KOMATI, K. S.; ANDRADE, J. O. Implementing neuroevolution for gas consumption forecasting in the steel industry. *In*: 2024 L Latin American Computer Conference (CLEI). [*S. l.: s. n.*], 2024. p. 1-10.

OLIVEIRA, V. M.; KOMATI, K. S.; ANDRADE, J. O. Seleção de características de séries temporais multivariadas do consumo de gás na pelotização de minério de ferro. *In*: XXVIII SIMPÓSIO BRASILEIRO DE ENGENHARIA DE PRODUÇÃO, 2021, Rio de Janeiro. *Anais* […]. Rio de Janeiro: Unesp, 2021.

PARK, Y.-S.; LEK, S. Artificial neural networks: Multilayer perceptron for ecological modeling. *In*: DEVELOPMENTS in environmental modelling. [*S. l.*]: Elsevier, 2016, v. 28, p. 123-140.

PRAIRIE, Y. T. Evaluating the predictive power of regression models. *Canadian Journal of Fisheries and Aquatic Sciences*, Canada, v. 53, n. 3, p. 490-492, 1996.

RODRIGUES, I. Sobe para 56 mil o número de moradores atingidos pela enchente do Rio Acre, aponta Defesa Civil. *G1 Globo*, Acre, 2023. Disponível em: https://g1.globo.com/ac/acre/noticia/2023/04/01/sobe-para-56-mil-o-numerodemoradores-atingidos-pela-enchente-do-rio-acre-aponta/-defesa-civil.ghtml. Acesso em: 1 maio 2024.

SAITO, S. M.; DIAS, M. C. A.; ALVALÁ, R. C. S.; STENNER, C.; FRANCO, C. O.; RIBEIRO, J. V. M.; SOUZA, P. A. d.; SANTANA, R. A. S. d. M. População urbana exposta aos riscos de deslizamentos, inundações e enxurradas no Brasil. *Sociedade & Natureza*, Uberlândia, v. 31, p. e46320, 2023.

SANT'ANNA, F. M. Bacia do rio Acre: a formação de uma governança transnacional. *In*: GOVERNANÇA multiescalar dos recursos hídricos transfronteiriços na Amazônia. São Paulo: Editora Unesp, 2017. p. 199-269. ISBN 978-85-9546-180-2.

SANTOS-NETO, L. A.; MANIESI, V.; QUERINO, C. A. S.; SILVA, M. J. G.; BROWN, V. R. Modelagem hidroclimatologica utilizando redes neurais multi layer perceptron em bacia hidrográfica no sudoeste da amazônia. *Revista Brasileira de Climatologia*, Paraná, v. 26, 2021. Disponível em: https://ojs.ufgd.edu.br/index.php/rbclima/article/view/14238. Acesso em: 8 set. 2025.

SERVICEANA. Página Principal. [*S. l.*], c2025. Disponível em: https://telemetriaws1.ana.gov.br/ServiceANA.asmx. Acesso em: 9 set. 2025.

SOUZA, A. C. D.; SANTOS, C. T. Identificação dos problemas de alagamento na cidade de Cachoeiro de Itapemirim-ES: estudo de caso sobre a área central. *In*: 16º CONGRESSO BRASILEIRO DE GEOLOGIA DE ENGENHARIA E AMBIENTAL, 2018, São Paulo. *Anais* […]. São Paulo: CBGE, 2018.

SOUZA, E.; OLIVEIRA, V.; ANDRADE, J.; KOMATI, K. Previsão de nível e vazão de água de um rio usando rede perceptron multi-camada: um estudo de caso do Rio Itapemirim. *In*: XI ESCOLA REGIONAL DE INFORMÁTICA DE GOIÁS, 2023, Porto Alegre. Anais […]. Porto Alegra: SBC, 2023. Disponível em: https://sol.sbc.org.br/index.php/erigo/article/view/27244. Acesso em: 8 set. 2025.

SUDDUL, G.; DOOKHITRAM, K.; BEKAROO, G.; SHANKHUR, N. An evolutionary multilayer perceptron algorithm for real time river flood prediction. *In*: IEEE. Zooming innovation in consumer technologies conference (ZINC). [*S. l.*]: IEEE, 2020. p. 109-112.

TELIKANI, A.; TAHMASSEBI, A.; BANZHAF, W.; GANDOMI, A. H. Evolutionary machine learning: *A survey. ACM Comput. Surv., Association for Computing Machinery*, New York, USA, v. 54, n. 8, 2021. ISSN 0360-0300. Disponível em: https://doi.org/10.1145/3467477. Acesso em: 8 set. 2025.

ZANIAL, W. N. C. W.; MALEK, M. B. A.; REBA, M. N. M.; ZAINI, N.; AHMED, A. N.; SHERIF, M.; ELSHAFIE, A. River flow prediction based on improved machine learning method: Cuckoo search-artificial neural network. *Applied Water Science*, Springer, v. 13, n. 1, p. 28, 2023.

ZOPH, B.; LE, Q. V. Neural architecture search with reinforcement learning. *arXiv*, [*s. l.*], 2016.

tecnia revista de educação, ciência e tecnologia do ifg

v. 10, Edição Especial 1 | 2025

Submetido 09/06/2024. Aprovado 08/03/2025
Avaliação: revisão duplo-anônimo

Modelos de aprendizado profundo aplicados à detecção de pólipo colorretal

THE APPLICATION OF DEEP LEARNING MODELS IN THE DETECTION OF COLORECTAL POLYPN

MODELOS DE APRENDIZAJE PROFUNDO APLICADOS A LA DETECCIÓN DE PÓLIPOS COLORRECTALES

Álisson Assis Cardoso
Universidade Federal de Goiás (UFG)
alsnac@ufg.br

Diene Xie
Universidade Federal de Goiás (UFG)
diene12374@gmail.com

Larissa Silva Xavier Rosa
Universidade Federal de Goiás (UFG)
larissarosa@discente.ufg.br

Ricardo Augusto Pereira Franco
Universidade Federal de Goiás (UFG)
ricardo@inf.ufg.br

Vilmar Cardoso Prestes Filho
Universidade Federal de Goiás (UFG)
vilmarfilho@discente.ufg.br

Resumo

O câncer colorretal, uma das principais causas de mortalidade no mundo, pode ser prevenido com a detecção precoce de pólipos. Este artigo propõe realizar um estudo de detecção de pólipos em imagens de colonoscopia utilizando modelos de redes neurais profundas de detecção de objetos. Para tanto, uma revisão de trabalhos na literatura é realizada com o objetivo de selecionar modelos avançados de detecção de objetos para serem utilizados. Além disso, são apresentados conjuntos de dados públicos de imagens de exames de colonoscopia, usados nos experimentos para detecção de pólipos. Também são apresentadas análises utilizando a técnica de pré-processamento de equalização de histograma, a fim de melhorar o contraste das imagens e, consequentemente, espera-se melhora no desempenho dos modelos. Resultados em termos de precisão, recall e mean average precision (mAP) são apresentados e usados para fins de comparação entre os modelos implementados. Os resultados obtidos indicam que os modelos treinados para a detecção de pólipos apresentaram resultados superiores em relação aos relatados na literatura, evidenciando que esses modelos podem ser um poderoso aliado da medicina no auxílio à detecção de pólipos e na prevenção do câncer colorretal.

Palavras-chave: detecção de objetos; pólipos; aprendizado profundo; câncer colorretal.

Abstract
The following text presents a synopsis of the abstract. Colorectal cancer, a leading cause of mortality worldwide, is preventable. The primary objective is to facilitate the timely identification of polyps. The present article puts forth a proposal for a study on the detection of polyps in colonoscopy images. The utilization of deep neural network models for the purpose of object detection is a subject of considerable interest. In order to this end, a literature review will identify and select advanced object detection models. The object is to be utilized. Furthermore, public datasets of colonoscopy images will be utilized. The following presentation will outline the utilization of various methodologies in experimental settings aimed at detecting polyps. Analyses employing the histogram equalization preprocessing technique are also presented with the objective of enhancing the quality of the results. The contrast of the images has been demonstrated to enhance the performance of the models. The following section presents the results of the study. Precision, recall, and mean average precision (mAP) are presented and utilized for the purpose of a comparison of the implemented models. The findings suggest that the models were successfully trained. The results obtained from the polyp detection process exhibited a higher degree of precision and accuracy when compared with the results reported in the existing literature. It has been demonstrated that these models have the capacity to serve as a valuable instrument in the field of medicine, particularly in the context of polyp detection. The following text will provide a comprehensive overview of colorectal cancer prevention.

Keywords: Object detection is a process that involves identifying the presence of an object in an image or video. polyps. deep learning. colorectal cancer.

Resumen
El cáncer colorrectal, una de las principales causas de mortalidad en el mundo, se puede prevenir com la detección temprana de pólipos. Este artículo propone realizar un estudio de detección de pólipos em imágenes de colonoscopia utilizando modelos de redes neuronales profundas para la detección de objetos. Para esto, se realiza una revisión de trabajos en la literatura con el objetivo de seleccionar modelos avanzados de detección de objetos a utilizar. Además, se presentan conjuntos de datos públicos de imágenes de exámenes de colonoscopia, utilizados en experimentos para detectar pólipos. También se presentan análisis utilizando la técnica de preprocesamiento de ecualización de histogramas, con el fin de mejorar el contraste de las imágenes y, en consecuencia, se espera una mejora en el rendimiento de los modelos. Se presentan los resultados en términos de precisión, recuperación y precisión promedio media (mAP) y se utiliza para fines de comparación entre los modelos implementados. Los resultados obtenidos indican que los modelos entrenados para la detección de pólipos presentaron resultados superiores en relación a los obtenidos en la literatura, demostrando que estos modelos pueden ser un poderoso aliado en medicina para ayudar en la detección de pólipos y en la prevención del cáncer colorrectal.

Palabras clave: Detección de objetos. Pólipos. Aprendizaje profundo. Cáncer colonrectal .

Introdução

O câncer ainda é uma doença complexa e que representa uma ameaça significativa à saúde global. De acordo com a Agência Internacional de Pesquisa sobre o Câncer (IARC), em 2020, foram estimados cerca de 19,3 milhões de novos casos de câncer e 10 milhões de mortes relacionadas ao câncer em todo o mundo (IARC Global Cancer Observatory, 2020). São vários os fatores de risco associados a essa doença, tais como genética, estilo de vida e exposição ambiental (ACS, 2020).

O câncer é uma doença que pode se manifestar de diferentes formas em diversas partes do corpo humano. Uma dessas formas de câncer que merece destaque

é o câncer de intestino, também conhecido como câncer de cólon e reto ou câncer colorretal. Ele é o terceiro câncer mais comum em homens e o segundo câncer mais comum em mulheres. De acordo com o Instituto Nacional de Câncer, no Brasil, o número estimado de novos casos para cada ano entre 2023 e 2025 é de 45.630 novos casos (Inca, 2022). Já em termos de mortalidade, de acordo com o Instituto Nacional de Câncer José Alencar Gomes da Silva, em 2020, ocorreram 20.245 óbitos por câncer colorretal, sendo 9.889 óbitos entre homens e 10.356 entre mulheres (Inca, 2022).

O câncer colorretal é um crescimento desordenado de células no cólon (intestino grosso) ou no reto. A primeira etapa observada geralmente é a formação de pólipos no revestimento interno do cólon ou reto. Os pólipos são pequenos crescimentos de tecido anormal que se projetam para dentro do lúmen intestinal. Somente certos tipos de pólipos, chamados de adenomas, transformam-se em câncer no decorrer do tempo. Com o crescimento, os pólipos adenomatosos podem eventualmente invadir camadas mais profundas e, se entrarem na corrente sanguínea ou no sistema linfático, podem se espalhar em outras partes do corpo, podendo gerar tumores secundários (Fearon; Vogelstein, 1990).

A evolução para o câncer colorretal é um processo lento, e a detecção precoce e a intervenção médica podem salvar vidas, pois aumentam significativamente as chances de um tratamento bem-sucedido (Krishnendu; Geetha; Gopakumar, 2020). Por essa razão, é importante fazer o rastreamento dos pólipos por meio da realização constante de consultas médicas e exames, permitindo a remoção deles antes que se tornem malignos. Um dos exames é o de colonoscopia, que ajuda no diagnóstico e na detecção precoce do câncer colorretal, bem como na avaliação de pólipos e outras condições do cólon e reto. Durante o exame, um tubo flexível, chamado colonoscópio, com uma câmera na sua extremidade, é inserido no reto e avança lentamente pelo cólon. A câmera transmite imagens em tempo real para um monitor, permitindo que o médico examine o revestimento interno do cólon em busca de anomalias, como os pólipos (Levin; Winawer, 2008).

Detecção automática de pólipos utilizando Inteligência Artificial

A análise dos resultados de exames de colonoscopia é demorada e pode apresentar falhas de detecção de pólipos ou identificação de falsos pólipos, resultando em procedimentos incorretos de tratamento. A detecção automática dos pólipos com Inteligência Artificial (IA) surge como uma das abordagens para auxiliar nessa tarefa (Krishnendu; Geetha; Gopakumar, 2020).

A utilização da IA tem se destacado na área da saúde como uma ferramenta de apoio à tomada de decisões. Ela permite analisar grandes quantidades de dados médicos e identificar padrões complexos, fornecendo observações valiosas aos profissionais de saúde (Obermeyer; Emanuel, 2016). A aprendizagem profunda (*deep learning*) é um ramo da IA que se baseia em redes neurais profundas e tem a capacidade de extrair automaticamente características relevantes em imagens médicas. Dessa forma, estudos recentes têm surgido sobre o uso de modelos de redes neurais convolucionais para a detecção de pólipos em imagens de exames de colonoscopia (Elkarazle *et al.*, 2023).

A detecção de pólipo colorretal no contexto da IA corresponde ao processo de treinamento de um modelo de aprendizado de máquina (machine learning) para aprender um conjunto de características específicas que representem os pólipos e, assim, detectá-los por meio de uma caixa delimitadora (bounding box), a partir de uma imagem ou vídeo de entrada (Elkarazle *et al.*, 2023).

São várias as etapas envolvidas na construção de um sistema de detecção de pólipos usando IA. Primeiramente, a seleção de um conjunto de dados para treinamento e validação, em seguida um pré-processamento dos dados, depois aplicação de técnicas de aumento de dados para aumentar a variedade de imagens, seguido do treinamento do modelo e finalizando com a sua validação. Existem diversas formas de diversificar cada etapa, desde a escolha de parâmetros até a escolha do modelo a ser treinado (Lecun; Bengio; Hinton, 2015).

Além disso, há diversos modelos de detecção de objetos, de arquiteturas diferentes com parâmetros distintos e aplicados com objetivos variados, em conjunto de dados diferentes. Um modelo pode não ter o mesmo valor de resultados obtidos ao alterar o tipo de objeto a ser detectado. Dessa forma, o objetivo deste artigo é apresentar uma análise de modelos existentes, selecionando os mais atuais (estado da arte) para realizar suas implementações práticas, avaliar e validar a eficácia na detecção de pólipos em imagens de colonoscopia.

Equalização de Histograma

Histograma de imagem é uma representação gráfica de como estão distribuídas as intensidades dos pixels da imagem (Queiroz; Gomes, 2011). Por exemplo, o histograma de uma imagem na escala cinza mostra a quantidade de pixels de cada valor possível, ou seja, valores variando entre 0 (preto) e 255 (branco).

Em imagens coloridas,tem-se um histograma para cada canal de cor. Por exemplo, no espaço RGB (espaço de cores Red, Green e Blue), teria um histograma para vermelho (red), um para verde (green) e um para azul (blue). Outro tipo de espaço bastante conhecido é o YCrCb, sendo Y o brilho e Cr e Cb, os componentes cromáticos de diferença de azul e de diferença de vermelho (Gonzalez; Woods, 2007).

O histograma é uma ferramenta importante e muito utilizada na área de processamento de imagens, já que ele pode mostrar diversas informações, a exemplo do contraste da imagem obtida a partir da largura do histograma. Quanto mais ampla for a distribuição, maior será o contraste. Outra informação importante que se obtém analisando o histograma é a exposição: muitos pixels com intensidades próximas a 255 representam uma imagem superexposta, enquanto muitos pixels com intensidades próximas a 0 representam uma imagem subexposta (Thomaz, 2020).

A equalização do histograma é uma técnica bastante usada e consiste em redistribuir as intensidades dos pixels de forma mais uniforme durante o intervalo de valores disponíveis, com o objetivo de melhorar o contraste da imagem (Souza; Correia, 2007). Um exemplo é apresentado na Figura 1: a imagem à esquerda possui pixels com valores concentrados em uma pequena faixa, enquanto a imagem à direita apresenta um histograma com valores distribuídos ao longo de toda a faixa possível e, consequentemente, exibe um contraste mais bem distribuído no intervalo considerado.

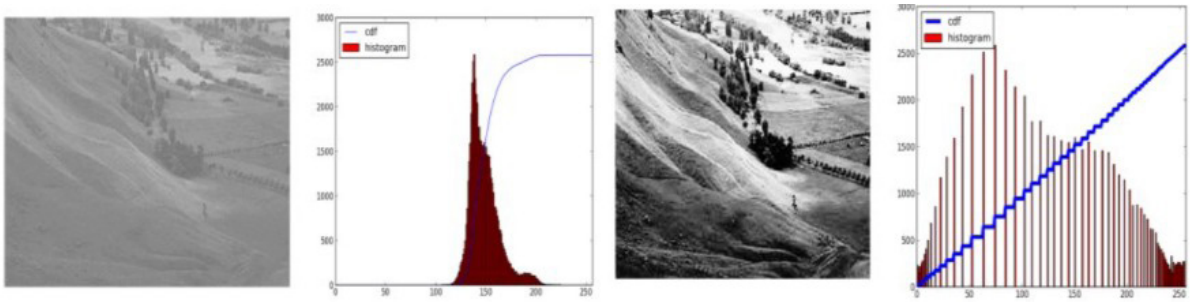


Figura 1 - À esquerda, tem-se uma imagem e seu histograma, com valores concentrados em uma pequena faixa do eixo x. À direita, observa-se a mesma imagem com histograma equalizado e, portanto, com melhor contraste, já que os valores no eixo x estão mais distribuídos no intervalo analisado
Fonte: Bradski e Kaehler (2000).

Organização do artigo

O artigo está organizado da seguinte forma: a Seção 2 apresenta a revisão bibliográfica de trabalhos relacionados; a metodologia incluindo coleta de dados, modelos de treinamento e inferências, equalização de histograma de imagens e as métricas utilizadas para análises é apresentada na Seção 3; os resultados obtidos são descritos na Seção 4; e a conclusão é apresentada na Seção 5.

Trabalhos relacionados

Grande parte dos estudos recentes sobre o uso de técnicas de inteligência artificial na área da saúde baseiam-se em métodos de aprendizagem profunda. Algoritmos de aprendizado profundo têm sido aplicados com sucesso em muitos campos da saúde, especialmente na análise de imagens médicas (Krishnendu; Geetha; Gopakumar, 2020). Dessa forma, esta seção apresenta alguns trabalhos desenvolvidos na literatura para realizar a tarefa de detecção de pólipos utilizando imagens de colonoscopia e algoritmos de detecção de objetos.

Region-based Convolutional Neural Network (R-CNN) é um algoritmo de detecção de objetos desenvolvido em 2014 (Zhang *et al.*, 2023).O R-CNN usa redes neurais convolucionais para localizar objetos de interesse e realizar a extração de características de forma independente de cada região de interesse para processamento posterior. Elkarazle *et al.* (2023) apresentaram bons resultados na detecção de pólipos utilizando modelos baseados no modelo R-CNN, detectando pólipos de formas e tamanhos distintos.

O modelo Single-Shot Detector (SSD) depende de uma única rede neural profunda para detectar um objeto (Zhang *et al.*, 2023). O modelo SSD discretiza o espaço de saída das caixas delimitadoras em várias caixas com diferentes proporções. Depois que as caixas delimitadoras são discretizadas, a rede é dimensionada por localização do mapa de características. A localização de um objeto de interesse é prevista usando vários mapas de características com diferentes resoluções. Os trabalhos em (Zhang *et al.*, 2023) propuseram utilizar o modelo SSD para realizar a detecção de pólipos colorretais.

A primeira versão do You Only Look Once (YOLO) foi apresentada em 2016 e é um dos algoritmos mais utilizados recentemente para a tarefa de detecção de objetos. O algoritmo YOLO foi modificado diversas vezes, gerando novas versões no decorrer

dos últimos anos. O YOLO também é um dos modelos de detecção de objetos que vem sendo utilizado na detecção de pólipos (Redmon; Farhadi, 2016).

O modelo Real-Time Detection Transformer (RT-DETR) foi apresentado em julho de 2023 (Zhang et al., 2023). Uma de suas características é o uso de transformers e consiste em usar um backbone, um codificador híbrido e um decodificador transformador com cabeçotes auxiliares de predição. Os autores apresentaram ainda os resultados que superaram os resultados com outros modelos, como a YOLO e suas diversas versões, utilizando o conjunto de dados da COCO 2017.

Zhang *et al.* (2023) também projetaram e treinaram uma Convolutional Neural Network (CNN) profunda para realizar a detecção de pólipos usando um conjunto diversificado e representativo de imagens de colonoscopias de triagem coletadas de mais de 2.000 pacientes. Foram treinadas diferentes arquiteturas CNN nesse estudo.

Além da escolha de um modelo para treinamento, é de suma importância considerar quais dados de treinamento serão utilizados. Como os bancos de dados usados na maioria dos trabalhos são conjuntos de dados limitados em tamanho e diversidade, Ma *et al.* (2021) propuseram um conjunto de dados em grande escala,chamado de LDPolypVideo, com uma variedade de pólipos e ambientes intestinais. Os testes realizados apresentaram quedas tanto na métrica recall quanto na precisão. De acordo com os autores, as quedas demonstram os desafios em trabalhar com conjuntos de dados de grande escala e com imagens de pólipos diversificados.

Nesse contexto, Tanwar *et al.* (2022) apresentaram um trabalho envolvendo técnicas de pré-processamento aplicadas em imagens de colonoscopia, tais como uso de filtros e de equalização de histograma, antes de realizarem um treinamento usando o modelo Single-Shot Detector (SSD). O modelo proposto pelos autores pode detectar e classificar pólipos colorretais com 92% de precisão.

Metodologia

Esta seção tem como objetivo descrever os métodos adotados na implementação prática dos modelos de detectores analisados. Os modelos implementados foram versões do YOLO e do modelo RT-DETR. Os experimentos foram conduzidos com três conjuntos de dados, realizando testes com e sem equalização de histograma.

A coleta de dados desempenha um papel fundamental em qualquer pesquisa baseada em aprendizado profundo, pois influencia diretamente na qualidade de resultados, mas sabe-se que há uma certa dificuldade para disponibilização de imagens médicas, por questões éticas, e em virtude da Lei Geral de Proteção de Dados Pessoais. Diante disso, foi realizado um levantamento de conjuntos de dados públicos (*datasets*) contendo imagens de colonoscopia.

Conjunto de dados

Após o levantamento, foi selecionado o conjunto de dados Kvasir-SEG, contendo 1.000 imagens de pólipos e suas anotações. A resolução das imagens varia de 332x487 a 1920x1072 pixels, e a divisão para treinamento e teste foi feita na proporção de 95/5 (Jha *et al.*, 2020).

Foi utilizado também o conjunto de dados CVC-ClinicDB (Bernal *et al.*, 2015), com 612 imagens com resolução de 384×288 pixels. Para o treinamento, validação e teste a divisão foi feita na proporção de 80/10/10.

Por fim, utilizou-se o conjunto de dados LDPolyp, que possui 40.186 imagens com resolução de 560×480 pixels, obtidas por meio de frames de vídeos (Ma *et al.*, 2021). Este último teve a divisão baseada na separação dos vídeos, sendo 100 vídeos para o conjunto de treinamento e 60 vídeos para o de teste e a validação.

Entre os conjuntos de dados escolhidos, os datasets Kvasir e o CVC-ClinicDB continham originalmente anotações de segmentação. Dessa forma, foi necessário realizar a conversão das anotações para o formato de detecção de objetos.

Modelos

Em sua primeira versão, a YOLO apresentou uma abordagem inovadora com mais velocidade em relação a outros modelos da época, como a Fast-RCNN, Faster-RCNN e DPM. Além disso, a velocidade de inferência era consideravelmente mais rápida, sendo possível realizar a inferência em tempo real (Redmon *et al.*, 2016).

A YOLO versão 3 apresenta como uma de suas principais melhorias: novo modelo de backbone (DarkNet53), com mais camadas convolucionais; adição de conexões residuais; previsões em múltiplas escalas, melhorando a detecção de objetos pequenos; e a possibilidade de múltiplas classes em uma mesma caixa delimitadora, em razão da troca da função softmax para entropia binária cruzada(Redmon; Farhadi, 2018).

A YOLOv5 foi o primeiro modelo a utilizar a biblioteca Pytorch, sendo disponibilizado pela Ultralytics em 2020 (Terven; Cordova-Esparza, 2023; Jocher *et al.*, 2022). Suas melhorias mais relevantes foram: um algoritmo de âncora automática, que aplica um algoritmo k-means nos dados e evolui as posições das caixas predefinidas para serem usadas nas predições; o modelo de backbone utilizado foi a CSPDarkNet, uma versão modificada que começa com uma camada de convolução ampliada, reduzindo custos computacionais; uma série de transformações para aumento de dados, que melhora o aprendizado; há também uma camada SPPF (Spatial Pyramid Pooling Fast), que acelera o cálculo agrupando características de diferentes escalas em um mapa de características de tamanho fixo.

A versão YOLOv6 foi publicada em 2022 pelo Meituan Vision AI Department, e superou os modelos anteriores em métricas de precisão e velocidade (Terven; Cordova-Esparza, 2023; Li *et al.*, 2022). As novidades destacadas dessa versão foram: um *backbone* baseado em RepVGG, chamado EfficientRep, que usa maior paralelismo; e uma estratégia de auto-destilação (*self-distillation*), que permite transferir o aprendizado de camadas profundas para camadas superficiais por meio de módulos de atenção (Ding *et al.*, 2021).

Assim como a versão 6, a YOLOv7 superou os detectores da época em precisão e velocidade na faixa de 5 FPS a 160 FPS (Terven; Cordova-Esparza, 2023; Wang; Bochkovskiy; Liao, 2022). Sua precisão foi incrementada sem afetar a velocidade de inferência, sobretudo em virtude das suas inovações: o uso da estratégia E-ELAN (Extended Efficient Layer Aggregation Network), que permite que um modelo profundo aprenda e convirja de forma mais eficiente, controlando o caminho de gradiente; normalização dos lotes (*batch normalization*) durante a convolução, que regula as ativações intermediárias, melhorando a estabilidade e o desempenho geral da rede (Terven; Cordova-Esparza, 2023).

A versão YOLOv8 foi lançada em janeiro de 2023, é considerada um modelo de última geração para tarefas de detecção de objetos (Terven; Cordova-Esparza, 2023). Ela deriva diretamente da YOLOv5 e também foi disponibilizada pela Ultralytics. Quanto a suas melhorias, citam-se: alteração no backbone (CSPDarkNet), permitindo combinar recursos na convolução, melhorando, consequentemente, a precisão;

funções de ativação separadas para objetividade, probabilidade de a caixa delimitadora conter um objeto, e classificação, probabilidade de o objeto ser de determinada classe, utilizando a função sigmóide para a primeira tarefa e a função softmax para a segunda tarefa; as funções de perda (*loss function*) de caixa delimitadora e de classificação também foram separadas, o que auxilia na detecção de objetos pequenos (Terven; Cordova-Esparza, 2023).

O modelo RT-DETR foi lançado em julho de 2023, e apresenta como principal característica o uso de *transformers*. Mais detalhadamente, o modelo utiliza um codificador híbrido eficiente para processar recursos em várias escalas e propõe uma seleção de consultas com reconhecimento de IoU (Intersection over Union), fornecendo consultas de objetos iniciais de maior qualidade ao decodificador (Zhao *et al.*, 2023).

Para fins de comparação de resultados, os seguintes modelos pré-treinados, realizando a transferência de aprendizado (*transfer learning*) no conjunto de dados Kvasir- SEG: YOLOv5 (modelos Nano, Small, Medium, Large e Extra-large); YOLOv6 (modelos Nano, Small, Medium e Large); YOLOv7 (modelos Tiny e X); YOLOv8 (modelos Nano, Small, Medium, Large e Extra-large); e RT-DETR (modelos Large e Extra-large).

Todos os modelos foram treinados no Google Colab, utilizando o GPU (Graphic Processing Unit) NVIDIA Tesla T4, exceto o treinamento, aplicando o modelo RT-DETRX. Em razão das restrições de memória RAM da GPU do Google Colab, foi utilizada uma GPU Tesla V100 da DGX-1 NVIDIA para realizar o treinamento com o dataset LDPolyp completo com o modelo RT-DETR-X. Os parâmetros dos treinamentos utilizados foram de 300 épocas, utilizando um batch size, com variação de 8 a 32. Os resultados obtidos serão discutidos na Seção 4.

Equalização de histograma

No contexto desta pesquisa, é comum que as imagens de colonoscopia obtidas em exames apresentem baixo contraste em virtude das condições de iluminação, como sombras e reflexos. Isso ocorre porque o colonoscópio, o equipamento utilizado, é inserido no intestino, o que pode dificultar a obtenção de imagens de alta qualidade.

A equalização de histograma aplicada em imagens de colonoscopia pode ajudar a realçar o contraste, consequentemente destacando melhor as características da mucosa intestinal, como vasos sanguíneos e pólipos. As imagens resultantes da aplicação da equalização de histograma podem, então, ser utilizadas para aprimorar a detecção de pólipos (Thomaz, 2020).

Os testes de equalização de histograma foram conduzidos utilizando imagens do dataset LDPolyp. Foi realizada a equalização de histograma no espaço YCrCb, sendo cada imagem foi convertida do espaço RGB para YCrCb, e a equalização do histograma foi aplicada apenas na banda Y – que representa o brilho da imagem – e, finalmente, a imagem foi convertida de volta para o espaço RGB. Na Figura 2, é possível ver a imagem antes da aplicação da equalização do histograma e após a aplicação da técnica.

Com o objetivo de analisar a equalização de histograma, utilizou-se uma amostra do dataset LDPolyp (1.000 imagens avaliadas) e o dataset completo Kvasir-SEG. Para o dataset LDPolyp, a análise foi realizada com o treinamento do modelo RT-DETR-X e para o dataset Kvasir-SEG a análise foi realizada com o treinamento dos modelos YOLOv8-N, YOLOv8-S, YOLOv8-M, YOLOv8-L, YOLOv8-X e RT-DETR-X.

Métricas de Avaliação

As métricas utilizadas para comparar os resultados entre os modelos de treinamento RT-DETR e YOLO nas versões 5 a 8 foram: Precisão, Recall, IoU, mAP@50 e mAP@50-95.

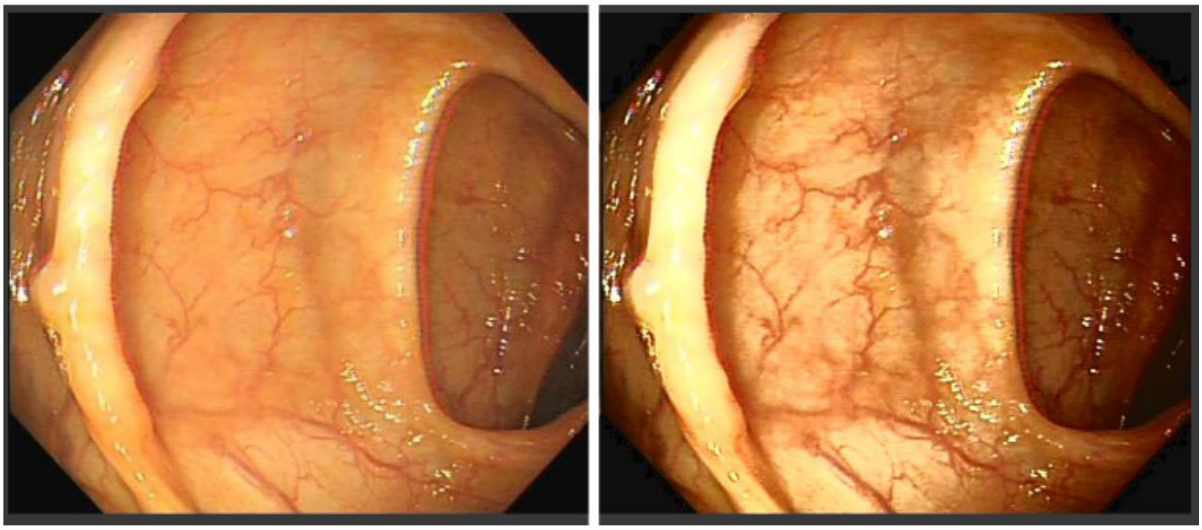


Figura 2 - A imagem à esquerda é a imagem original do dataset LDPolyp, e à direita observa-se a mesma imagem após a equalização do histograma da banda Y (brilho da imagem) no espaço YCrCb
Fonte: Elaborado pelos(as) autores(as).

IoU (Intersection over Union) é a métrica que representa a interseção sobre a união, comparando as caixas delimitadoras anotadas e previstas. A equação 1 representa o IoU, sendo A a caixa delimitadora verdadeira e B a caixa delimitadora prevista.

$$IoU(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

Dado um valor limite de IoU, um resultado de previsão de modelo pode ser considerado como: Verdadeiro Positivo (VP) - classificação certa da classe positivo; Falso Negativo (FN) - modelo previu como classe negativo, sendo o valor real a classe positivo; Falso Positivo (FP) - modelo previu como classe positivo, sendo o valor real a classe negativo; ou Verdadeiro Negativo (VN) - classificação certa da classe negativo.

A Precisão corresponde à métrica que avaliará, dentre todas as classificações de classe positivo que o modelo fez, quantas de fato têm esse valor real. Ela é representada pela equação 2.

$$Preciso = \frac{VP}{VP + FP} \quad (2)$$

Recall é a métrica que verificará quantas classificações corretas foram feitas da classe positivo em relação ao total esperado, ou seja, em relação a todas as situações de classe positivo como valor esperado. Essa métrica está representada na equação 3.

$$Recall = \frac{VP}{VP + FN} \quad (3)$$

A métrica Average Precision (AP) representa a área debaixo da curva de um gráfico das métricas Precisão e Recall, conhecida também como curva AP.

O mAP@50 (Mean Average Precision) é a média da AP dado um limiar de sobreposição de pelo menos 50% de IoU e o mAP@50-95 é a média dos valores de AP para cada limiar de sobreposição no intervalo de 50% a 95%, com incrementos de 5%.

Resultados

Foram realizados primeiramente experimentos sem equalização de histograma. Após isso, realizaram-se experimentos com os modelos que obtiveram melhores valores de métricas utilizando a técnica de equalização de histograma, visando proporcionar um ganho no desempenho dos modelos. A seguir, são apresentados os resultados obtidos em ambos os experimentos, seguindo a ordem em que foram realizados.

Resultados sem aplicação de Equalização de Histograma

No processo de treinamento no conjunto de dados Kvasir-SEG (Jha *et al.*, 2020), foram registradas as métricas exibidas na Tabela 1 para os modelos YOLO nas versões 5 a 8 e RT-DETR.

Modelo	Precisão	Recall	mAP@50	mAP@50-95
YOLOv5-N	0,884	0,830	0,910	0,726
YOLOv5-S	0,923	0,871	0,913	0,743
YOLOv5-M	0,889	0,872	0,916	0,744
YOLOv5-L	0,889	0,927	0,932	0,755
YOLOv5-X	0,918	0,815	0,906	0,725
YOLOv6-N	0,872	0,866	0,917	0,707
YOLOv6-S	0,840	0,873	0,897	0,718
YOLOv6-M	0,854	0,850	0,887	0,673
YOLOv6-M	0,815	0,855	0,877	0,650
YOLOv7-Tiny	0,896	0,782	0877	0,604
YOLOv7	0,877	0,909	0,917	0,720
YOLOv7-X	0,938	0,836	0,915	0,692
YOLOv8-N	0,873	0,875	0,923	0,723
YOLOv8-S	0,876	0,903	0,931	0,751
YOLOv8-M	0,922	0,855	0,931	0,757
YOLOv8-L	0,920	0,873	0,930	0,774
YOLOv8-X	0,887	0,891	0,915	0,734
RT-DETR L	0,882	0,855	0,885	0,736
RT-DETR X	0,905	0,855	0,872	0,737

Tabela 1 - Métricas de desempenho dos modelos treinados para o conjunto de dados Kvasir (os melhores resultados de cada métrica estão sublinhados)
Fonte: Elaborado pelos(as) autores(as).

Os modelos com melhores métricas estão reunidos na Tabela 2. O critério de seleção foi o map@50-95. Considerando esse critério, destaca-se o modelo YOLOv8L.

Modelo	Precisão	Recall	mAP@50	mAP@50-95
YOLOv8-N	0,873	0,875	0,923	0,723
YOLOv8-S	0,876	0,903	0,931	0,751
YOLOv8-M	0,922	0,855	0,931	0,757
YOLOv8-L	0,920	0,873	0,930	0,774
YOLOv8-X	0,887	0,891	0,915	0,734

Tabela 2 - Desempenho dos melhores modelos treinados para o conjunto de dados Kvasir (os melhores resultados de cada métrica estão sublinhados)
Fonte: Elaborado pelos(as) autores(as).

A Figura 3 exibe algumas imagens de pólipos detectados pelo modelo que obteve as melhores métricas, isto é, o modelo Yolov8L, bem como os níveis de precisão associados a cada detecção. Pode-se observar uma compreensão visual e quantitativa do quão preciso e confiável é o modelo na tarefa de detecção de pólipos em imagens de colonoscopia.

Após os resultados obtidos, também foi realizada uma análise comparativa, investigando na literatura os resultados de outros modelos que utilizaram o mesmo conjunto de dados. Essa análise foi realizada considerando a metodologia apresentada em Jha *et al.* (2020). A comparação está ilustrada na Tabela 3, e a métrica avaliada foi o Mean Average Precision 50 (mAP@50).

Método	Rede Principal	mAP@50
EfficientDet-D0	EfficientNet-b0, biFPN	0,5047
Faster R-CNN	ResNet50	0,8418
RetinaNet	ResNet50	0,9095
RetinaNet	ResNet101	0,9095
YOLOv3+spp	Darknet53	0,8532
YOLOv4	Darknet53, CSP	0,8234
ColonSegNet	-	0,8166
YOLOv8-L (proposto)	-	0,9300

Tabela 3 - Métricas de desempenho dos modelos obtidos na literatura e do melhor modelo treinado (YOLOv8-L) para o conjunto de dados Kvasir
Fonte: Elaborado pelos(as) autores(as).

Além do conjunto de dados Kvasir, também foram feitos treinamentos com os melhores modelos para o conjunto de dados CVC Clinic, cujos resultados são exibidos na Tabela 4.

Para o CVC Clinic, observa-se que os modelos YOLOv8-N e YOLOv8-M exibem boas pontuações de Precisão, Recall e mAP@50, aproximando-se de 1, o que evidencia uma capacidade de identificação precisa dos objetos de interesse nas imagens. Esses resultados sustentam a confiabilidade e robustez desses modelos na tarefa de detecção de objetos.

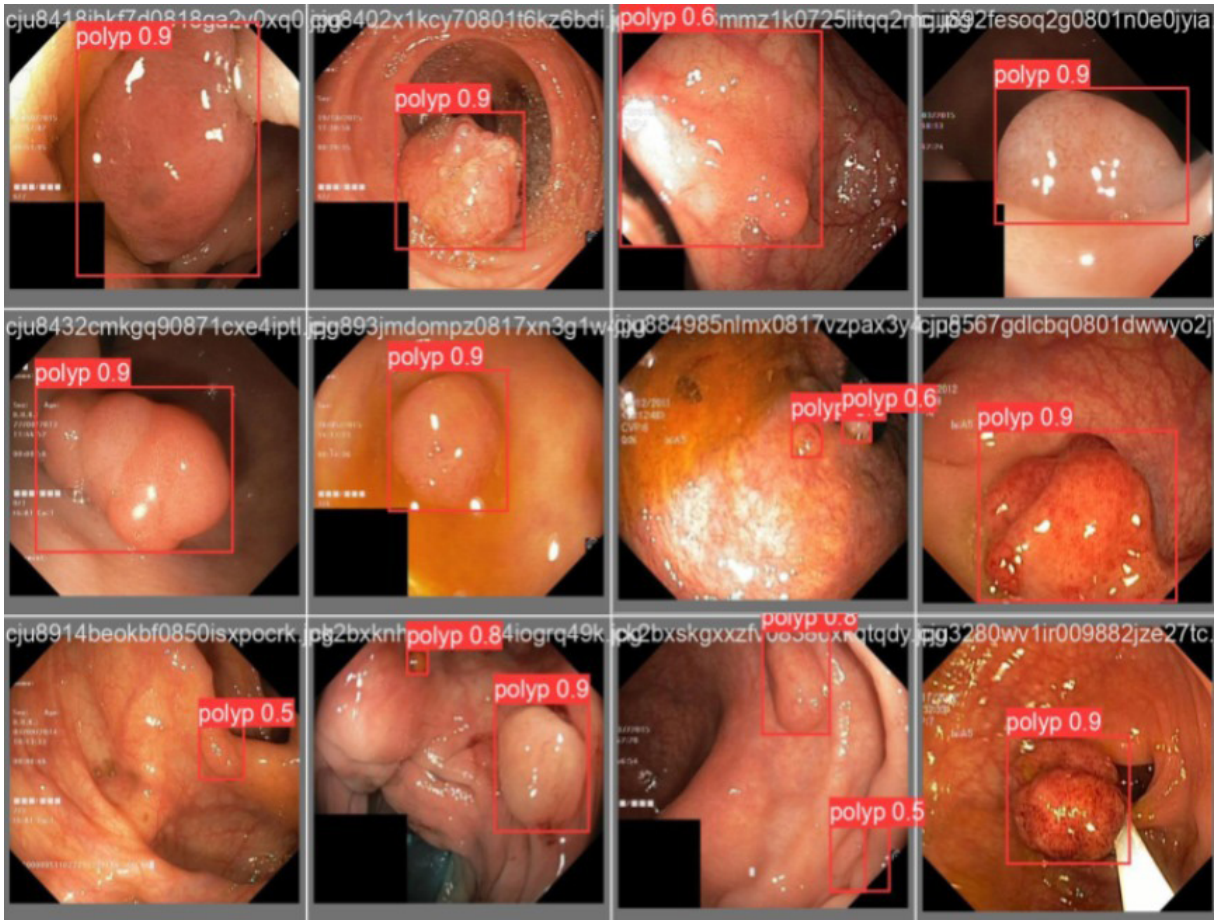


Figura 3 - Pólipos detectados na validação do modelo YOLOv8L
Fonte: Elaborado pelos(as) autores(as).

Modelo	Precisão	Recall	mAP@50	mAP@50-95
YOLOv8-N	0,996	1,000	0,995	0,857
YOLOv8-M	0,983	1,000	0,995	0,856
YOLOv8-L	0,984	0,998	0,995	0,842
YOLOv9-C	0,969	1,000	0,993	0,848
RT-DETR-L	0,998	1,000	0,995	0,845

Tabela 4 - Métricas de desempenho dos modelos validados para o CVC Clinic (os melhores resultados de cada métrica estão sublinhados)
Fonte: Elaborado pelos(as) autores(as).

Em contrapartida, o desempenho do modelo YOLOv8-L também se destaca, embora se posicione abaixo dos modelos YOLOv8-N e YOLOv8-M em termos de precisão e mAP@50-95. No entanto, sua pontuação de Recall, próximo a 1, sugere uma alta habilidade em recuperar todos os objetos relevantes nas imagens, tornando-o uma escolha sólida para aplicações que demandam mais sensibilidade na detecção.

O modelo YOLOv9-C apresenta resultados competitivos, apesar de uma precisão inferior em comparação com os modelos YOLOv8. No entanto, suas métricas de Recall e mAP@50-95 indicam que ainda é capaz de identificar a maioria dos objetos com uma precisão aceitável.

Por fim, pode-se observar que o modelo RT-DETR-X se destaca por sua precisão excepcionalmente alta, indicando uma habilidade quase perfeita em evitar falsos positivos

para esse conjunto de dados. Contudo, sua pontuação de mAP@50-95 é ligeiramente menor em comparação com os modelos YOLOv8, sugerindo uma precisão um pouco reduzida na detecção de objetos mais desafiadores de localizar.

Para o conjunto de dados LDPolyp, em virtude da sua grande quantidade de imagens, o treinamento foi realizado apenas do modelo RT-DETR, cujos resultados estão apresentados na Tabela 5.

Modelo	Precisão	Recall	mAP@50	mAP@50-95
RT-DETR-X	0,1339	0,7428	0,3875	0,4163

Tabela 5 - Desempenho do RTDETR-X treinado com o conjunto de dados LDPolyp
Fonte: Elaborado pelos(as) autores(as).

Resultados com aplicação de Equalização de Histograma

Em seguida, foram realizados os testes com equalização de histograma, a fim de verificar o ganho nas métricas de desempenho dos modelos ao utilizar essa técnica. Na Tabela 6, pode-se observar a comparação das métricas dos melhores modelos antes e depois da aplicação da técnica de equalização do histograma no conjunto de dados Kvasir.

A análise comparativa dos modelos treinados para o conjunto de dados Kvasir, tanto antes quanto após a equalização de histograma, revela descobertas que enfatizam os benefícios dessa técnica de pré-processamento para alguns dos modelos.

Um dos aspectos mais notáveis é a melhoria na métrica de precisão do modelo RTDETR- X após a aplicação da equalização. Isso significa que o modelo está conseguindo detectar corretamente as áreas de interesse nas imagens, reduzindo, assim, o número de falsos positivos. Essa precisão aprimorada é vital, especialmente em tarefas críticas como a detecção de anomalias, onde cada erro pode ter consequências significativas.

No entanto, é importante ressaltar que a implementação da equalização de histograma não é uma solução universal. Embora tenha apresentado melhorias significativas em termos de precisão, em alguns casos, pode haver uma leve redução no Recall ou em outras métricas de desempenho. Portanto, é necessário realizar uma avaliação abrangente e equilibrada dos resultados, considerando não apenas métricas isoladas, mas também o impacto geral no desempenho do modelo.

Modelo	Equalização de Histograma	Precisão	Recall	mAP@50	mAP@50-95
YOLOv8-N	Não	0,873	0,875	0,923	0,723
YOLOv8-N	SIm	0,873	0,855	0,915	0,747
YOLOv8-S	Não	0,876	0,903	0,931	0,751
YOLOv8-S	Sim	0,888	0,863	0,896	0,736
YOLOv8-M	Não	0,922	0,855	0,931	0,757
YOLOv8-M	Sim	0,848	0,909	0,910	0,724
YOLOv8-L	Não	0,920	0,873	0,930	0,774
YOLOv8-L	Sim	0,854	0,873	0,895	0,700
YOLOv8-X	Não	0,887	0,891	0,915	0,734
YOLOv8-X	Sim	0,884	0,830	0,904	0,720

RT-DETR-X	Não	0,905	0,855	0,872	0,737
RT-DETR-X	Sim	0,958	0,873	0,917	0,762

Tabela 6 - Desempenho dos modelos treinados para o conjunto de dados Kvasir antes e após a equalização de histograma (os melhores resultados de cada métrica estão sublinhados)
Fonte: Elaborado pelos(as) autores(as).

Com o objetivo de avaliar a utilização da técnica de equalização de histogramas em imagens de vídeos, utilizou-se uma parcela de 1.000 amostras do conjunto de dados LDPolyp. A Tabela 7 apresenta a comparação das métricas antes e depois da aplicação da referida técnica em uma amostra de 1.000 imagens do conjunto de dados LDPolyp.

Modelo	Equalização de Histograma	Precisão	Recall	mAP@50	mAP@50-95
RT-DETR-X	Não	0,757	0,417	0,425	0,193
RT-DETR-X	Sim	0,857	0,500	0,516	0,238

Tabela 7 - Desempenho do RTDETR-X treinado com o conjunto de dados LDPolyp (os melhores resultados de cada métrica estão sublinhados)
Fonte: Elaborado pelos(as) autores(as).

Nota-se, observando a Tabela 7, a melhoria em todas as métricas avaliadas do modelo RT-DETR-X após a equalização de histograma na amostra de 1.000 imagens do conjunto de dados LDPolyp. A precisão registrou um aumento de aproximadamente 20%, saltando de 0,757 para 0,857. Da mesma forma o Recall, com cerca de 19%, passando de 0,417 para 0,500. O mAP@50 e o mAP@50-95 registraram melhorias de aproximadamente 21,4% e 23,3% respectivamente, refletindo um aumento na precisão média e na detecção em uma variedade mais ampla de cenários e condições. Isso sugere um progresso na capacidade do RT-DETR-X de identificar objetos com maior precisão e confiança após a aplicação da equalização de histograma.

Conclusão

Este estudo abordou a detecção precoce de pólipos colorretais por meio do uso de IA, com um foco significativo na metodologia adotada. A seleção criteriosa de conjuntos de dados públicos, compostos por uma variedade de imagens de colonoscopia, foi essencial para garantir a representatividade da amostra.

Os treinamentos dos modelos YOLO (versões 5 até a 8) e RT-DETR (Large e Extra-Large) foram conduzidos de maneira sistemática e controlada, com ajustes de hiperparâmetros para otimizar o desempenho. A análise comparativa com outros modelos, como o RetinaNet, destaca a superioridade dos modelos YOLO em termos de detecção de pólipos (Lin *et al.*, 2017).

Além disso, foram feitos testes com a aplicação da equalização de histograma para melhorar o contraste das imagens e verificar se, consequentemente, haveria melhora no desempenho dos modelos. Os resultados dos testes indicaram que a equalização de histograma teve um impacto positivo no desempenho dos modelos, melhorando ligeiramente as métricas. Contudo, essa técnica não foi um fator decisivo para o desempenho final dos modelos, sugerindo que outras técnicas de pré-processamento podem ser exploradas em futuros trabalhos.

Referências

AMERICAN CANCER SOCIETY. *Cancer Facts & Figures 2020*. [*S. l.*]: ACS, 2020.

BERNAL, J.; SÁNCHEZ, F. J.; FERNÁNDEZ-ESPARRACH, G.; GIL, D.; RODRÍGUEZ, C.; VILARIñO, F. Wm-dova maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, [*s. l.*], v. 43, p. 99-111, 2015.

BRADSKI, G.; KAEHLER, A. OpenCV. *Dr. Dobb's journal of software tools*, [*s. l.*], v. 3, n. 2, 2000.

DING, X.; ZHANG, X.; MA, N.; HAN, J.; DING, G.; SUN, J. Repvgg: Making vgg- style convnets great again. *IEEE*, [*s. l.*], 2021. Disponível em: https://ieeexplore.ieee.org/document/9577516. Acesso em: 9 set. 2025.

ELKARAZLE, K.; RAMAN, V.; THEN, P.; CHUA, C. Detection of colorectal polyps from colonoscopy using machine learning: A survey on modern techniques. *National Library of Medicine*, [*s. l.*], 2023. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9953705/. Acesso em: 9 set. 2025.

FEARON, E. R.; VOGELSTEIN, B. A genetic model for colorectal tumorigenesis. *National Library of Medicine*, [*s. l.*], 1990. Disponível em: https://pubmed.ncbi.nlm.nih.gov/2188735/. Acesso em: 9 set. 2025.

GONZALEZ, R. C.; WOODS, R. E. *Processamento Digital de Imagens*. 5. ed. São Paulo: Pearson Education do Brasil, 2007.

IARC GLOBAL CANCER OBSERVATORY. *Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries*. [*S. l.*]: IARC Global Cancer Observatory, 2020. Disponível em: https://pubmed.ncbi.nlm.nih.gov/33538338/. Acesso em: 9 set. 2025.

INSTITUTO NACIONAL DE CÂNCER. *Estimativa 2023*: incidência de câncer no Brasil. Rio de Janeiro: Inca, 2022. Disponível em: https://www.inca.gov.br/publicacoes/livros/estimativa-2023-incidencia-de-cancer-nobrasil. Acesso em: 9 set. 2025.

JHA, D.; SMEDSRUD, P. H.; RIEGLER, M. A.; HALVORSEN, P.; LANGE, T.; JOHANSEN, D.; JOHANSEN, H. D. Kvasir-seg: A segmented polyp dataset. *arXiv*, [*s. l.*], 2020. Disponível em: https://arxiv.org/abs/1911.07069. Acesso em: 9 set. 2025.

JOCHER, G.; CHAURASIA, A.; STOKEN, A.; BOROVEC, J. ultralytics/yolov5: v7.0 yolov5 sota realtime instance segmentation. *Zenodo*, [*s. l.*], 2022. DOI: https://doi.org/10.5281/zenodo.7347926.

KRISHNENDU, S.; GEETHA, S.; GOPAKUMAR, G. A review on polyp detection and segmentation in colonoscopy images using deep learning. *International Journal of Engineering Research & Technology (IJERT)*, [*s. l.*], v. 9, n. 10, 2020. Disponível

em: https://www.ijert.org/a-review-on-polyp-detection-and-segmentation-in-colonoscopyimages-using-deep-learning. Acesso em: 9 set. 2025.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *nature*, [*s. l.*], 2015. DOI: https://doi.org/10.1038/nature14539.

LEVIN, B.; LIEBERMAN, D. A.; MCFARLAND, B.; ANDREWS, K. S. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the american cancer society, the us multi-society task force on colorectal cancer, and the american college of radiology. *National Library of Medicine*, [*s. l.*], 2008. Disponível em: https://pubmed.ncbi.nlm.nih.gov/18384785/. Acesso em: 9 set. 2025.

LI, C.; LI, L.; JIANG, H.; WENG, K.; GENG, Y.; LI, L.; KE, Z.; LI, Q.; CHENG, M.; NIE, W.; LI, Y.; ZHANG, B.; LIANG, Y.; ZHOU, L.; XU, X.; CHU, X.; WEI, X.; WEI, X. Yolov6: A single-stage object detection framework for industrial applications. *arXiv*, [*s. l.*], 2022. Disponível em: https://arxiv.org/abs/2209.02976. Acesso em: 9 set. 2025.

LIN, T.-Y.; GOYAL, P.; GIRSHICK, R.; HE, K.; DOLLÁR, P. Focal loss for dense object detection. *In*: PROCEEDINGS of the IEEE International Conference on Computer Vision. [*S. l.: s. n.*], 2017. p. 2980-2988.

MA, Y.; CHEN, X.; CHENG, K.; LI, Y.; SUN, B. Ldpolypvideo benchmark: A largescale colonoscopy video dataset of diverse polyps. *Medical Image Computing and Computer Assisted Intervention Society*, [*s. l.*], 2021, p. 387-396.

OBERMEYER, Z.; EMANUEL, E. J. Predicting the future - big data, machine learning, and clinical medicine. *National Library of Medicine*, [*s. l.*], 2016. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5070532/. Acesso em: 9 set. 2025.

QUEIROZ, J. E. R.; GOMES, H. M. *Introdução ao Processamento Digital de Imagens*. Campina Grande: UFCG, 2011. Disponível em: http://www.dsc.ufcg.edu.br/~hmg/disciplinas/graduacao/vc-2011.2/RitaTutorialPDI.pdf. Acesso em: 9 set. 2025.

REDMON, J.; DIVVALA, S.; GIRSHICK, R.; FARHADI, A. You only look once: Unified, real-time object detection. *arXiv*, [*s. l.*], 2016. Disponível em: https://arxiv.org/abs/1506.02640. Acesso em: 9 set. 2025.

REDMON, J.; FARHADI, A. Yolov3: An incremental improvement. *arXiv*, [*s. l.*], 2018. Disponível em: https://arxiv.org/abs/1804.02767. Acesso em: 9 set. 2025.

SOUZA, T.; CORREIA, S. Estudo de técnicas de realce de imagens digitais e suas aplicações. *In*: II CONGRESSO DE PESQUISA E INOVAÇÃO DA REDE NORTE NORDESTE DE EDUCAÇÃO TECNOLÓGICA, 2007, Paraíba. *Anais* […]. Paraíba: Connepi, 2007.

TANWAR, S.; VIJAYALAKSHMI, S.; SABHARWAL, M.; KAUR, M.; ALZUBI, A. A.; LEE, H.-N. Detecção e classificação de pólipo colorretal usando aprendizado profundo. *BioMed Research International*, [*s. l.*], v. 2024, n. 1, 2022. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1155/2022/2805607. Acesso em: 9 set. 2025.

TERVEN, J.; CORDOVA-ESPARZA, D. A comprehensive review of yolo: From yolov1 and beyond. *arXiv*, [*s. l.*], 2023. Disponível em: https://arxiv.org/abs/2304.00501. Acesso em: 9 set. 2025.

THOMAZ, V. A. *Avaliação de Aumento de Dados via Geração de Imagens Sintéticas para Segmentação e Detecção de Pólipos em Imagens de Colonoscopia Utilizando Aprendizado de Máquina*. 2020. Tese (Doutorado em Ciências da Computação) — Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2020.

WANG, C.-Y.; BOCHKOVSKIY, A.; LIAO, H.-Y. M. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv*, [*s. l.*], 2022. Disponível em: https://arxiv.org/abs/2207.02696. Acesso em: 9 set. 2025.

ZHANG, A.; LIPTON, Z. C.; LI, M.; SMOLA, A. J. Dive into Deep Learning. *GitHub*, [*s. l.*], 2023. Disponível em: http://d2l.ai/index.html. Acesso em: 10 set. 2025.

ZHAO, Y.; LV, W.; XU, S.; WEI, J.; WANG, G.; DANG, Q.; LIU, Y.; CHEN, J. Detrs beat yolos on real-time object detection. *arXiv*, [*s. l.*], 2023. Disponível em: https://arxiv.org/abs/2304.08069. Acesso em: 9 set. 2025.

tecnia revista de educação, ciência e tecnologia do ifg

v. 10, Edição Especial 1 | 2025

Submetido 31/05/2024. Aprovado 18/12/2024
Avaliação: revisão duplo-anônimo

Análise quantitativa e qualitativa preliminar dos efeitos dos algoritmos de reamostragem no registro de imagens utilizando a detecção de cantos

PRELIMINARY QUANTITATIVE AND QUALITATIVE ANALYSIS OF THE EFFECTS OF RESAMPLING ALGORITHMS ON IMAGE REGISTRATION USING CORNER DETECTION

ANÁLISIS PRELIMINAR CUANTITATIVO Y CUALITATIVO DE LOS EFECTOS DE LOS ALGORITMOS DE REMUESTREO EN EL REGISTRO DE IMÁGENES MEDIANTE LA DETECCIÓN DE BORDES

Carlos Eduardo Falandes
Instituto Nacional de Pesquisas Espaciais (Inpe)
carlos.falandes@inpe.br

Fabrício Galende Marques de Carvalho
Instituto Nacional de Pesquisas Espaciais (Inpe)
fabricio.galende@inpe.br

Resumo

Este trabalho apresenta uma análise preliminar, quantitativa e qualitativa, dos efeitos de diferentes algoritmos de reamostragem no processo de registro de imagens por meio da detecção de cantos. O registro consiste em alinhar e combinar múltiplas imagens de uma mesma cena, capturadas por sensores distintos ou em diferentes momentos, para formar uma representação completa da cena. Esse processo frequentemente requer a reamostragem das imagens, redimensionando-as para padronizar a base de comparação durante a extração de características. No entanto, a reamostragem pode introduzir distorções que comprometem etapas subsequentes, como a detecção de cantos e a correspondência de padrões. A avaliação foi realizada comparando como os métodos clássicos de reamostragem – Vizinho Mais Próximo, Bilinear e Bicúbica – afetam a correspondência de padrões, logo o registro de imagens. O estudo utiliza imagens de satélite que são inicialmente reduzidas a uma razão de reamostragem específica e, em seguida, ampliadas novamente às suas dimensões originais para a realização do registro. Os resultados obtidos são avaliados por meio de métricas como Erro Quadrático Médio, Coeficiente de Correlação, Relação Sinal-Ruído de Pico e Raiz do Erro Quadrático Médio aplicada à correspondência dos pontos de controle. Os resultados indicam que a reamostragem Bicúbica é a mais eficaz, apresentando os menores índices de erro no registro. O método Vizinho Mais Próximo, por sua vez, insere menos erros que a reamostragem Bilinear, que apresentou os maiores índices de erro e variabilidade.

Palavras-chave: processamento de imagens; reamostragem de imagens; registro de imagens; detecção de características; correspondência de padrões.

Abstract
This study presents a preliminary quantitative and qualitative analysis of the impact of different resampling algorithms on the image registration process using corner detection. Image registration refers to the alignment and integration of multiple images of the same scene, acquired by different sensors or at different times, to generate a more comprehensive representation. This process often requires image resampling and resizing to establish a standardized basis for comparison during feature extraction. However, resampling may introduce distortions that affect subsequent stages, such as corner detection and pattern matching. The evaluation was conducted through a comparative analysis of classical resampling methods. For this purpose, satellite images were initially reduced according to a specific resampling ratio and subsequently enlarged back to their original dimensions for registration. The results were assessed using metrics such as Mean Square Error, Correlation Coefficient, Peak Signal-to-Noise Ratio and Root Mean Square Error, applied to the matching of control points. The findings indicate that bicubic resampling provides the best performance, yielding the lowest registration error rates. By contrast, bilinear resampling presented the highest error values and variability, while the nearest neighbor method outperformed bilinear resampling, although it remained less effective than bicubic resampling.

Keywords: image processing; image resampling; image registration; feature detection; pattern matching.

Resumen
Este trabajo, de enfoque cualitativo y cuantitativo, presenta un análisis preliminar de los efectos de diferentes algoritmos de remuestreo en el proceso de registro de imágenes mediante la detección de bordes. El registro consiste en alinear y combinar múltiples imágenes, capturadas por sensores distintos o en momentos diferentes, para formar la representación completa de una escena. A menudo, este proceso requiere el remuestreo de las imágenes, redimensionándolas para estandarizar la base de comparación durante la extracción de sus características. Sin embargo, el remuestreo puede introducir distorsiones que comprometen las etapas posteriores del proceso de registro, como la detección de esquinas y la correspondencia de patrones. Para evaluar los efectos de los diferentes algoritmos de remuestreo sobre la correspondencia de patrones y, en consecuencia, sobre la calidad del registro de imágenes, se realizó una comparación entre los métodos más clásicos: Vecino Más Cercano, Bilineal y Bicúbico. En el estudio se utilizaron imágenes de satélite, inicialmente reducidas a una relación de remuestreo específica y, a continuación, ampliadas a sus dimensiones originales para realizar el registro. Los resultados obtenidos se evaluaron mediante métricas como el Error Cuadrático Medio, el Coeficiente de Correlación, la Relación Señal-Ruido de Pico y la Raíz del Error Cuadrático Medio aplicada a la correspondencia de puntos de control. El análisis evidenció que el remuestreo Bicúbico es el más eficaz, ya que presenta los menores índices de error en el registro. El método del Vecino Más Cercano, por su parte, introduce menos errores que el remuestreo Bilineal, que presentó los índices de error y variabilidad más altos.

Palabras clave: procesamiento de imágenes; remuestreo de imágenes; registro de imágenes; detección de características; emparejamiento de patrones.

Introdução

Em diversas áreas, como astronomia, medicina e sensoriamento remoto, o registro de imagens desempenha papel crucial (Lin, 2023; Porwal; Katiyar, 2014). Esse processo envolve alinhar e combinar várias imagens da mesma cena, capturadas por diferentes sensores ou em diferentes momentos, para formar uma representação completa da cena. Essa representação é útil para correlacionar imagens de diferentes

sensores, facilitando análises de fenômenos celestes, monitoramento de doenças e observação de mudanças ambientais, como queimadas e desmatamento.

Para que o registro de imagens seja realizado, é necessário que as imagens compartilhem regiões comuns, permitindo a extração de características e a busca por correspondências. Frequentemente, isso requer a reamostragem das imagens, ou seja, seu redimensionamento para criar uma base comum de comparação. A extração de características e a detecção de cantos, etapas essenciais nesse processo, são abordadas por Falandes, Carvalho e Morelli (2024) e Zitová e Flusser (2003). Contudo, métodos de reamostragem podem introduzir distorções indesejáveis, como desfoque, descontinuidade e serrilhamento nas bordas, comprometendo a qualidade das imagens, especialmente em áreas críticas como em cantos e bordas (Falandes; Carvalho, 2023).

Sabendo que os métodos de reamostragem modificam as características da imagem, este estudo analisa como os métodos clássicos de reamostragem, como Vizinho Mais Próximo, Bilinear e Bicúbica, afetam os processos de detecção e descrição de características, que são fundamentais para o registro de imagens. Para essa análise, foram avaliados os resultados da correspondência de pontos de controle e do mosaico gerado pelo registro por diferentes métricas, associadas a diferentes razões percentuais de reamostragem. A avaliação do mosaico de registro foi feita com métricas como Erro Médio Quadrático, Relação Sinal-Ruído de Pico e Coeficiente de Correlação, que permitem comparar a qualidade dos resultados, conforme abordado por Prasantha, Shashidhara e Balasubramanya (2009) e Falandes e Carvalho (2023). Para a avaliação da correspondência de pontos de controle, utilizaram-se duas métricas: a Raiz do Erro Quadrático Médio das correspondências e o Erro Absoluto comparado com a correspondência correta.

Diferentemente de outros estudos, como em Falandes e Carvalho (2023), que abordou o impacto das técnicas de reamostragem nos contornos de formas geométricas, a principal contribuição deste estudo é a análise quantitativa dos efeitos dessas técnicas na qualidade do registro de imagens. O objetivo é investigar como os métodos de reamostragem afetam o registro de imagens, apresentando resultados quantitativos que proporcionem uma melhor compreensão.

Revisão da Literatura

Nesta seção, apresentam-se os principais conceitos e métodos que fundamentam o desenvolvimento deste trabalho, com ênfase nos procedimentos de reamostragem, no registro de imagens e nas métricas de avaliação.

Métodos de reamostragem

Em diversos trabalhos é comum encontrar os métodos clássicos de reamostragem, em virtude da sua fácil implementação e do baixo custo computacional. Esses métodos consistem na utilização de dados conhecidos para estimar valores em pontos desconhecidos com base em conceitos matemáticos. Nessa etapa, estabelece-se uma relação entre as dimensões originais da imagem e as dimensões desejadas por meio da razão entre elas, assim permitindo analisar as proximidades entre os pixels originais e a grade de interpolação (Gonzalez; Woods, 2019).

Cada método tem abordagens diferentes, por exemplo, a Interpolação por Vizinho Mais Próximo (NN - do inglês, *nearest neighbor*) baseia-se na atribuição da intensidade do pixel mais próximo na imagem original a cada ponto da grade de interpolação. Sua implementação é simples e requer baixo tempo computacional. No entanto, esse método pode resultar em artefatos indesejados, como perda de detalhes finos e serrilhamento nos contornos (Gonzalez; Woods, 2019).

Esse método pode ser descrito pela Equação 1, sendo $P_0(x', y')$ o pixel a ser interpolado, $P(x,$ y) o pixel da imagem original e d_x e d_y são as distâncias entre os pontos originais e os interpolados em x e y respectivamente: $d_x = x - x$ e $d_y = y - y$.

$$P_0(x', y') = \begin{cases} P(x, y) & \text{para } d_x < 0,5 \text{ e } d_y < 0,5 \\ P(x+1, y) & \text{para } d_x \geq 0,5 \text{ e } d_y < 0,5 \\ P(x, y+1) & \text{para } d_x < 0,5 \text{ e } d_y \geq 0,5 \\ P(x+1, y+1) & \text{para } d_x \geq 0,5 \text{ e } d_y \geq 0,5 \end{cases} \quad (1)$$

Já a Interpolação Bilinear apresenta semelhanças com o método anterior, entretanto considera não apenas o pixel mais próximo, mas também os quatro pixels mais perto da imagem original. As intensidades dos pixels são ponderadas com base nas distâncias d_x e d_y, que são lineares e complementares. Essa ponderação resulta em uma suavização das transições em regiões de alto contraste. A Equação 2 descreve a forma como essas distâncias influenciam nos valores dos pixels.

$$(x', y') = \begin{bmatrix} (1-d_x)(1-d_y) \cdot P(x+1, y+1) + d_x(1-d_y) \cdot P(x, y+ \\ + d_x d_y \cdot P(x, y) + d_y(1-d_x) \cdot P(x+1, y) \end{bmatrix} \quad (2)$$

Por fim, a Interpolação Bicúbica utiliza os 16 pixels mais próximos da imagem original. A ideia desse método é considerar a distância geométrica entre os pixels, atribuindo pesos com base em uma spline cúbica. Diferentemente dos métodos anteriores, esse reduz o serrilhado sem causar um excesso de suavização. Contudo, esse método é mais exigente em termos de recursos computacionais (KEYS, 1981). A Equação 3 descreve o cálculo da intensidade do novo pixel $P_0(x', y')$, utilizando uma média ponderada $P(x+m, y+n)$. Os pesos atribuídos a cada pixel vizinho são definidos pela função $R(t)$ (Equação 4), que representa uma spline cúbica suavizada dependente da distância entre os pixels.

$$P_0(x', y') = \sum_{m=-1}^{2} \sum_{n=-1}^{2} P(x+m, y+n) \cdot R(m - d_x) \cdot R(d_y - n) \quad (3)$$

$$R(t) = \frac{1}{6} \left[(t+2)^3 - 4(t+1)^3 + 6t^3 - 4(t-1)^3 \right] \quad (4)$$

Assim, como os diferentes métodos afetam as características da imagem de maneiras distintas, a qualidade da reamostragem tem grande influência na etapa seguinte de registro de imagens, que depende fortemente da preservação dos contornos e detalhes após o processo.

Métodos de registro

O registro de imagens consiste no alinhamento de imagens que compartilham trechos de uma mesma cena, tiradas em momentos, ângulos ou sensores diferentes (Zitová; Flusser, 2003). Para a realização do registro é necessário passar pelas seguintes etapas: filtro binário (Gonzalez; Woods, 2019), detecção de pontos de controle (Saharan, 2016), descrição de características, casamento de pontos de controle, estimação do modelo de transformação e refinamento do modelo de transformação.

A primeira etapa, de filtragem binária, tem como objetivo remover ruídos e eliminar informações irrelevantes que possam dificultar ou distorcer a identificação de pontos importantes para o alinhamento. Isso favorece o desempenho das etapas seguintes, como a detecção e a extração de características.

Na sequência, a detecção dos pontos de controle é fundamental, visto que estes estão em regiões específicas de uma imagem e podem ser identificados por conter características distintivas, sendo, portanto, usados como base para estimar a transformação aplicada no registro de imagens. Existem várias técnicas de detecção de pontos de controle, cada uma com suas especificidades, entre elas: o Detector de Cantos Harris, que identifica cantos com alta precisão, além de ter um baixo tempo computacional e ser fácil de implementar (Harris; Stephens, 1988); o Detector ORB (do inglês, *Oriented FAST and Rotated BRIEF*), notável por sua velocidade e eficiência em ambientes com pouca iluminação (Rublee *et al.*, 2011); o Detector SURF (do inglês - *Speeded-Up Robust Features*), conhecido por sua robustez e rapidez em grandes escalas (Bay; Tuytelaars; Van Gool, 2006); e o Detector SIFT (do inglês - *Scale-Invariant Feature Transform*), reconhecido por sua capacidade de detectar e descrever pontos de interesse locais invariáveis à escala, rotação e iluminação (Lowe, 1999). Essas técnicas diferem quanto à complexidade de implementação, robustez frente a distorções e eficácia em diferentes cenários e aplicações.

Em particular, o Detector de Cantos Harris destaca-se por sua precisão na identificação de cantos e por sua robustez contra ruídos. Sua implementação simples e de baixo custo computacional torna-o ideal para aplicações de visão computacional que exigem rapidez, além de ser resistente a transformações afins, como translações e rotações. O método opera calculando o gradiente de intensidade de cada pixel nas direções *x* e *y*, ou seja, o quanto a imagem muda em cada direção. A partir desses gradientes, é construída uma matriz de autocorrelação *M*, que resume a variação da intensidade em uma vizinhança do pixel. Esse processo é descrito na Equação 5.

$$M = \sum_{v=-1}^{1} \sum_{u=-1}^{1} P(x+v, y+u) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \tag{5}$$

Nesse caso, I_x e I_y são as derivadas da intensidade da imagem nas direções horizontal e vertical respectivamente. Essa matriz ajuda a entender como a imagem varia ao redor de cada ponto.

A Equação 6 define uma função de resposta *R*, que, quando alta, indica que o pixel está em um canto. Essa função utiliza o determinante e o traço da matriz *M* para

identificar regiões onde a intensidade varia simultaneamente em diversas direções. O valor *k* é uma constante empírica, geralmente entre 0.04 e 0.06.

$$R = \det(M) - k\,(\mathrm{tr}(M))^2 \tag{6}$$

Por meio da detecção de cantos (pontos de controle) em duas imagens é possível estabelecer relações entre pontos correspondentes, geralmente realizando-se a descrição de características ao redor desses pontos. Neste trabalho foi usada a área das regiões onde os pontos foram detectados (Zhang *et al.*, 2021), abordagem semelhante ao uso dos Momentos de Zernike, que também se baseia na área para descrever as características regionais (Mahi; Isabaten; Serief, 2014). Com os pontos de controle devidamente correspondentes, identificados por meio das áreas semelhantes, é possível estipular uma transformação adequada para o registro de imagens.

Na literatura, diversas transformações são utilizadas no registro, como a Procrustes, que ajusta escala, rotação e translação (Gong *et al.*, 2022); a Afim, que considera a não ortogonalidade entre os eixos (Chandrappa; Anil, 2021); e a Projetiva, que, além dos parâmetros anteriores, corrige distorções de perspectiva (Gong *et al.*, 2022). Mais recentemente, têm ganhado destaque abordagens baseadas em regiões, em características e em aprendizado profundo, especialmente aplicadas ao registro de imagens de sensoriamento remoto (Zhang *et al.*, 2021).

Em particular, a transformação Afim se destaca por sua simplicidade de implementação e o baixo custo computacional, além de oferecer boa precisão com apenas três pares de pontos correspondentes. Essa transformação é amplamente empregada em aplicações de processamento de imagens e visão computacional, uma vez que permite operações como translação, rotação, escala e cisalhamento (Gong *et al.*, 2022). A forma geral da transformação é expressa na Equação 7. Contudo, neste trabalho, optou-se por uma versão simplificada da transformação Afim, restrita à correção de translações, conforme descrito na Equação 8, o que se mostra suficiente para o tipo de desalinhamento observado nos dados utilizados.

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} a & b & t_x \\ c & d & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ 1 \end{bmatrix} \tag{7}$$

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ 1 \end{bmatrix} \tag{8}$$

Sendo:
(x, y): coordenadas transformadas;
(x_o, y_o): coordenadas originais;
a, b, c e d: parâmetros de rotação, escala e cisalhamento; e
t_x e t_y: parâmetros de translação.

Após a obtenção dos modelos de transformação com base nos pontos de controle previamente correspondidos, é essencial aplicar um algoritmo de remoção de *outliers*. Essa etapa elimina correspondências que destoam significativamente das demais. O algoritmo RANSAC (do inglês - *Random Sample Consensus*) é frequentemente

escolhido para essa finalidade, em razão da sua fácil implementação e eficiência, sendo especialmente útil quando ao menos 50% das correspondências são confiáveis. Com os *outliers* removidos, é possível aplicar o modelo de transformação mais adequado para o processo de registro entre as imagens.

Métricas de avaliação

Avaliar a qualidade de uma imagem é uma tarefa complexa, com diversas técnicas propostas, mas nenhuma universal. Entre as abordagens mais comuns, estão a análise de diferenças pontuais, a correlação de imagens, a detecção de bordas, as redes neurais (RN), a análise de regiões de interesse (ROI) e o sistema visual humano (HVS). Desse modo, para avaliar quantitativamente a qualidade dos resultados, utilizaram-se ferramentas baseadas principalmente em medições pontuais, que, apesar de sua implementação simples, são eficientes e amplamente exploradas, como abordado por Pappas, Safranek e Chen (2005). Nesse contexto, Najjar (2024), por sua vez, realizou uma análise comparativa entre os diferentes métodos de avaliação.

A quantificação da qualidade final do mosaico foi conduzida com base na similaridade entre as intensidades da imagem processada e de uma imagem de referência. Além disso, o desempenho do registro também foi avaliado por meio da comparação entre o modelo de transformação obtido e as correspondências esperadas, permitindo a identificação de possíveis erros no alinhamento.

Dentre as métricas utilizadas: o Erro Quadrático Médio (MSE – do inglês, Mean Squared Error) calcula a média dos quadrados das diferenças entre os valores dos pixels da imagem original e da imagem registrada – quanto menor o MSE, melhor a qualidade da interpolação, indicando maior semelhança entre as imagens. Já o MSE elevado indica uma grande discrepância entre as imagens, evidenciando baixa qualidade da interpolação. A seguir, apresenta-se a equação para o MSE por meio da Equação 9.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(i_o - i)^2 \tag{9}$$

Sendo:
i_o: pixel na imagem de referência;
i: pixel na imagem analisada; e
n: número total de pixels nas imagens.

Adicionalmente, o Coeficiente de Correlação (CC) é uma métrica que quantifica como as variações de intensidade dos pixels em uma imagem se relacionam com as variações de intensidade na imagem de referência. O valor do CC varia de -1 (correlação negativa perfeita) a 1 (correlação positiva perfeita), com 0 indicando ausência de correlação. A equação para o CC é dada por meio da Equação 10.

$$\mathrm{CC} = \frac{\sum_{i=1}^{n}(i_o - \bar{i_o})(i - \bar{i})}{\sqrt{\sum_{i=1}^{n}(i_o - \bar{i_o})^2}\sqrt{\sum_{i=1}^{n}(i - \bar{i})^2}} \tag{10}$$

Sendo:
$\bar{i_0}$ e $\bar{i}$: média das intensidades na imagem analisada e de referência respectivamente.

Complementando essa análise, a Relação Sinal-Ruído de Pico (PSNR - do inglês, Peak Signal-to-Noise Ratio) é uma métrica que utiliza os resultados do Erro Médio Quadrático para quantificar a relação entre o sinal (informação útil na imagem) e o ruído (distorções ou erros na imagem) em relação à imagem de referência. Quanto maior o valor do PSNR, maior a quantidade de sinal na imagem. A equação para o PSNR é dada por meio da Equação 11, que relaciona o quadrado de MAX (intensidade máxima do pixel) e o MSE (ruído).

$$\mathrm{PSNR} = 10 \cdot \log_{10} \left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}} \right) \tag{11}$$

Para avaliar o erro introduzido pelo registro, adotou-se a Raiz do Erro Quadrático Médio (RMSE - do inglês, Root Mean Square Error), que é uma versão normalizada do MSE (Equação 12). Essa métrica facilita a interpretação dos erros, pois apresenta as mesmas unidades dos dados de saída, tornando mais claro o impacto do erro. O RMSE estima o erro, em pixels, do modelo de transformação em relação aos pontos de controle, que são aqueles que deveriam coincidir entre a imagem registrada e a imagem de referência, fornecendo, assim, uma indicação do quão precisos são os resultados do registro em unidades de pixels.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (p_o - p)^2} \tag{12}$$

Sendo:
n: número de pontos de controle casados;
p: posição calculada; e
p_o: posição esperada.

Por fim, o erro de registro reflete a discrepância entre os valores obtidos e os esperados, calculada para as coordenadas xxx e yyy a fim de avaliar o impacto em ambas as direções, conforme as Equações 13 e 14. Para uma visão geral, calculou-se a raiz quadrada da soma dos quadrados dos erros nas duas direções, resultando na distância euclidiana entre as posições esperadas e encontradas, como se observa na Equação 15. Essas métricas indicam a diferença entre as posições esperadas e as encontradas durante o registro.

$$E_x = x_i - x$$
$$E_y = y_i - y \tag{13}$$

$$E_x = x_i - x \tag{14}$$
$$E_y = y_i - y$$

$$SQ = \sqrt{E_x^2 + E_y^2} \tag{15}$$

Sendo:
x: posição calculada;
x_i: posição ideal em x;
y: posição calculada; e
y_i: posição ideal em y.

Metodologia para avaliação

Os testes foram conduzidos para quantificar e comparar os efeitos das diferentes técnicas de reamostragem no registro de imagens. Foram utilizadas imagens de satélite com diversas formas geométricas: e.g., reservatórios de água, montanhas, baías, plantações. Essas regiões são relevantes, pois os pontos de controle selecionados para a correspondência de padrões são baseados nelas e as características de contorno são usadas para descrever as regiões no processo de casamento de padrões. No entanto, ao reamostrar essas regiões, ocorre uma perda significativa de detalhes nos contornos (Falandes; Carvalho, 2023).

Diante disso, para esta análise, selecionou-se a Figura 1 em razão das suas características distintivas, que são cruciais para a definição de pontos de controle. Ademais, a imagem abrange uma área situada sob a Serra da Canastra, região de grande relevância em virtude da vasta extensão de plantações. Portanto, compreender o impacto das diferentes técnicas de reamostragem no registro de imagens em áreas como essa é fundamental, considerando-se a necessidade de fiscalização e monitoramento ambiental por meio de sensoriamento remoto.

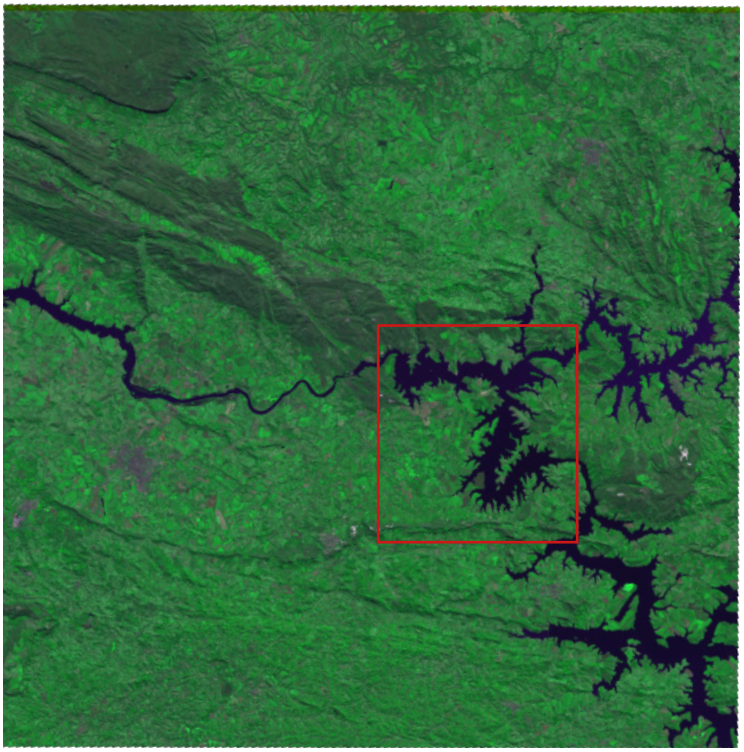


Figura 1 - Imagem do satélite CBERS-04A, obtida pelo sensor MUX com resolução espacial de 16,5 m, destacando a região usada nos testes
Fonte: Inpe (2025).

Em virtude das grandes dimensões da Figura 1, optou-se por realizar as análises em um recorte específico, apresentado na Figura 2. Esse recorte foi selecionado por conter as características fundamentais anteriormente mencionadas, permitindo uma avaliação mais rápida. Além disso, o recorte teve as suas dimensões definidas como 1013 x 1013 pixels, a escolha de um número primo assegura que a razão entre as dimensões originais e as novas dimensões nunca seja um número inteiro, evitando, assim, qualquer alinhamento regular. Essa precaução é significativa ao considerar o processo de arredondamento da posição do pixel durante a reamostragem das dimensões da imagem, destacando, desse modo, a sensibilidade do método em determinar as novas intensidades. Dessa forma, evidenciam-se as características de cada técnica, o que evita qualquer viés no processo de análise dos resultados. A imagem utilizada nos testes foi da banda NIR, que oferece melhores contrastes em áreas de vegetação em razão da alta reflectância das plantas nessa faixa espectral. Melo e Ribeiro (2022) também usaram a banda NIR para mapear a vegetação em uma floresta densa, ressaltando sua importância para identificar padrões vegetais.

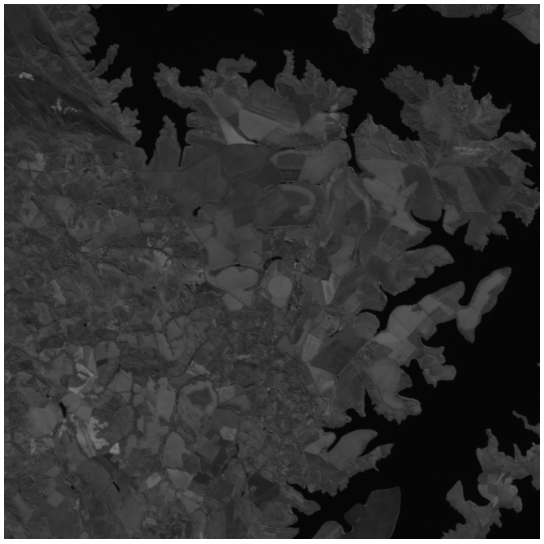


Figura 2 - Recorte da banda NIR da Figura 1 com dimensões de 1013 x 1013 pixels
Fonte: Inpe (2025).

Para entender como diferentes técnicas de reamostragem afetam o processo de registro de imagens, foi adotado o seguinte procedimento: primeiramente, a Figura 2 é reduzida em incrementos de 10%, até alcançar 90% do tamanho original, posteriormente, é ampliada de volta às suas dimensões originais, como exibido na Figura 3. A imagem resultante é então submetida a uma série de processos: filtragem binária, detecção de cantos e identificação de áreas de interesse. A seguir, busca-se estabelecer correspondências entre a imagem reamostrada e a original por meio dos cantos e das áreas descritas, identificando pontos correspondentes em regiões similares das duas imagens, como demonstrado na Figura 4.

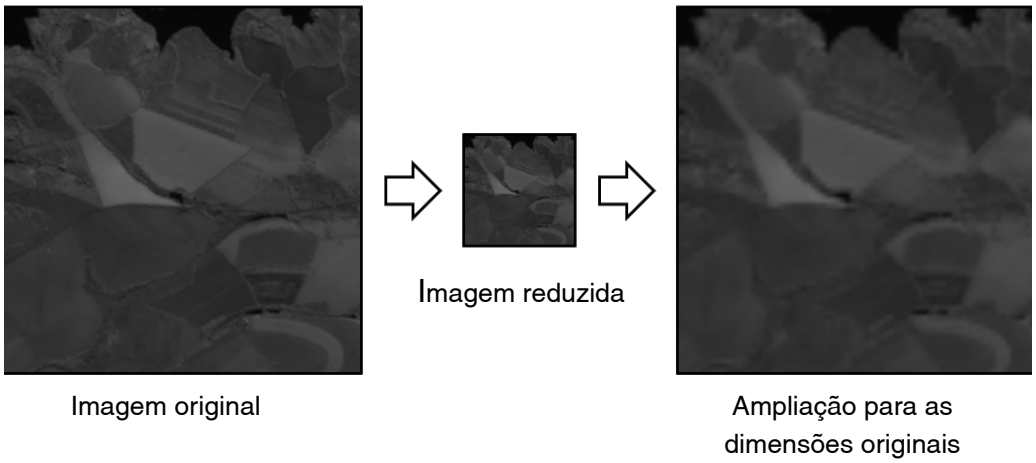


Figura 3 - Ilustração do processo de reamostragem utilizando o método Bilinear com razão percentual de redução em 70% da original
Fonte: Elaborado pelo(as) autores(as).

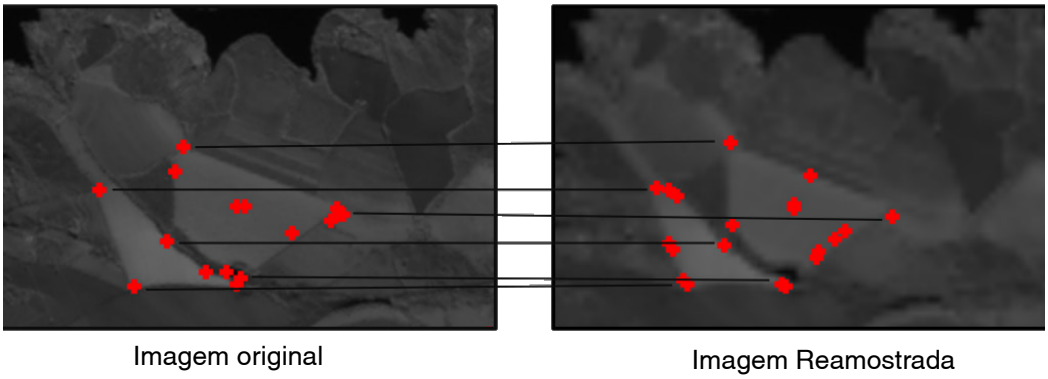


Figura 4 - Casamento de pontos de controle entre a imagem original e a Reamostrada
Fonte: Elaborado pelos(as) autores(as).

Uma vez definidas as correspondências entre os pontos de controle das imagens, modelos de transformações afins são estimados. Esse processo inclui a aplicação do algoritmo RANSAC para eliminação de *outliers* e escolha do melhor modelo de transformação, em seguida o erro de correspondência de padrões é calculado usando a métrica RMSE, e também são calculados os erros de correspondência nas coordenadas *x* e *y*. Além disso, o erro de intensidade no resultado do mosaico do registro em relação à imagem original também é avaliado por meio das métricas MSE, CC e PSNR, com objetivo de avaliar a qualidade visual final da imagem.

Esse conjunto de etapas é repetido, variando a razão de redução da reamostragem em incrementos de 10%, até alcançar 90% do tamanho original. Esse processo é aplicado a cada método de reamostragem, gerando dados estatísticos específicos para cada técnica. Essa metodologia permite avaliar como cada técnica de reamostragem impacta o processo de registro, ressaltando características essenciais de cada método, particularmente em termos de perda de detalhes durante as etapas de redução e ampliação da imagem.

Resultados e discussão

Após o processo de avaliação, os resultados foram apresentados em gráficos com linhas curvas para facilitar a visualização e proporcionar uma melhor representação geométrica dos dados. Para obter uma interpretação mais precisa, a magnitude dos gráficos foi ajustada para eliminar registros claramente incorretos (com erros altíssimos), facilitando a visualização dos dados. Esses erros ocorreram principalmente porque o alto percentual de redução fez com que as características distintivas se perdessem, o que gerou correspondências incorretas e comprometeu o registro das imagens.

Os impactos da reamostragem no registro de imagens em relação aos pontos de controle é exibido no Gráfico 1, o qual mostra que a reamostragem Bicúbica apresenta uma tendência mais previsível de aumento no erro à medida que a taxa de redução cresce. Em relação aos demais métodos, a reamostragem NN apresenta menor previsibilidade em relação à Bicúbica, porém apresentam resultados análogos entre 30% e 60% de redução. A reamostragem Bilinear introduziu as maiores taxas de erro e a maior variação entre os resultados. Vale enfatizar que as razões percentuais de redução de 70% para NN, 90% e 80% para Bilinear e 90% para Bicúbica apresentaram erro de 0 pixel. Isso ocorre porque, para calcular o erro nos pontos de controle, é necessário eliminar as correspondências incorretas. Nesses casos, como a maior parte das correspondências foi incorreta, o algoritmo simplesmente as ignora, considerando-as inutilizáveis.

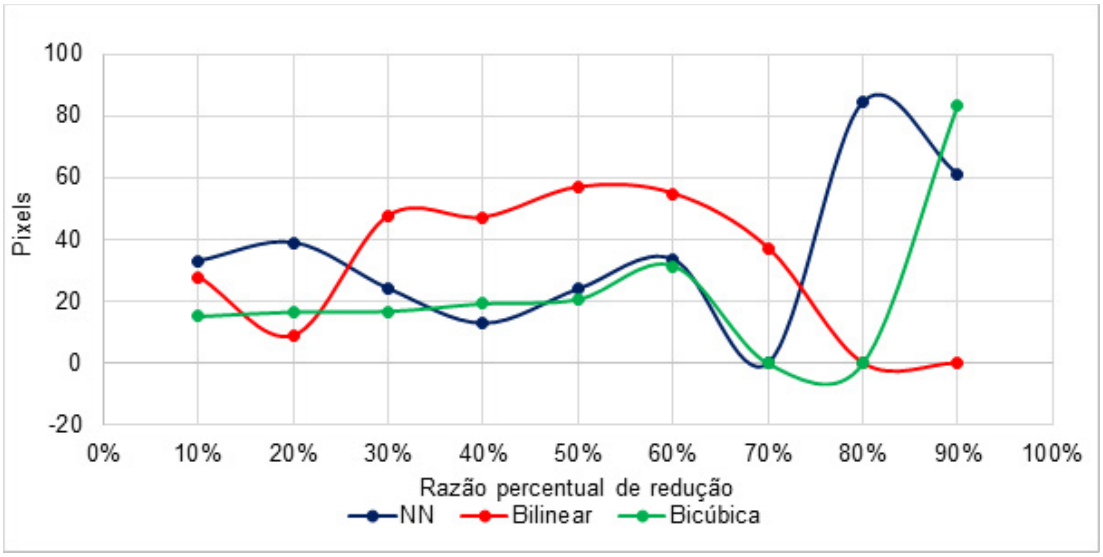


Gráfico 1- Raiz do erro quadrático médio no registro em relação aos pontos de controle
Fonte: Elaborado pelos(as) autores(as).

Os Gráficos 2 e 3 destacam o erro de registro, que representa o deslocamento perceptível na sobreposição das imagens. Nesses gráficos, os erros nas direções *x* e *y* são analisados, mostrando que os métodos de reamostragem Bicúbica e NN apresentam resultados semelhantes até 60% da razão percentual de redução, no entanto a Bicúbica, em média, introduz menos erros. Por outro lado, a reamostragem bilinear exibe os maiores erros e a maior variabilidade. Ressalta-se que os gráficos têm magnitude limitada, pois alguns pontos apresentam erros extremos, superiores a centenas de pixels.

Para compreender melhor os dados, é importante levar em consideração a resolução espacial da imagem, ou seja, 16,5 metros. Desse modo, os métodos de reamostragem mais previsíveis permitem calcular o erro introduzido em termos de distâncias reais, auxiliando a tomada de decisões em aplicações reais. Como exemplo, pode-se observar a redução percentual de 50% para a reamostragem Bilinear, identificando, com isso, um erro no registro de cerca de -45 pixels no eixo *x* e cerca de 15 pixels no eixo *y*. Isso representa um deslocamento de 742,5 m para a esquerda e 247 m para cima da posição ideal. Por isso, nessa situação, o método não é o ideal, visto que sua utilização comprometeria o posicionamento dos locais exibidos na imagem final.

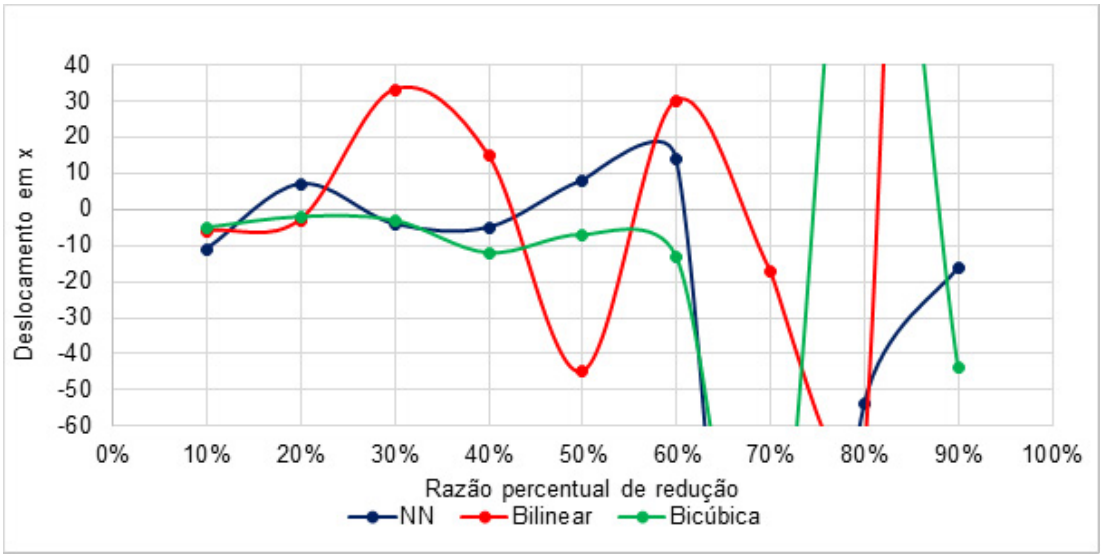


Gráfico 2- Erro na posição da imagem registrada no eixo x
Fonte: Elaborado pelos(as) autores(as).

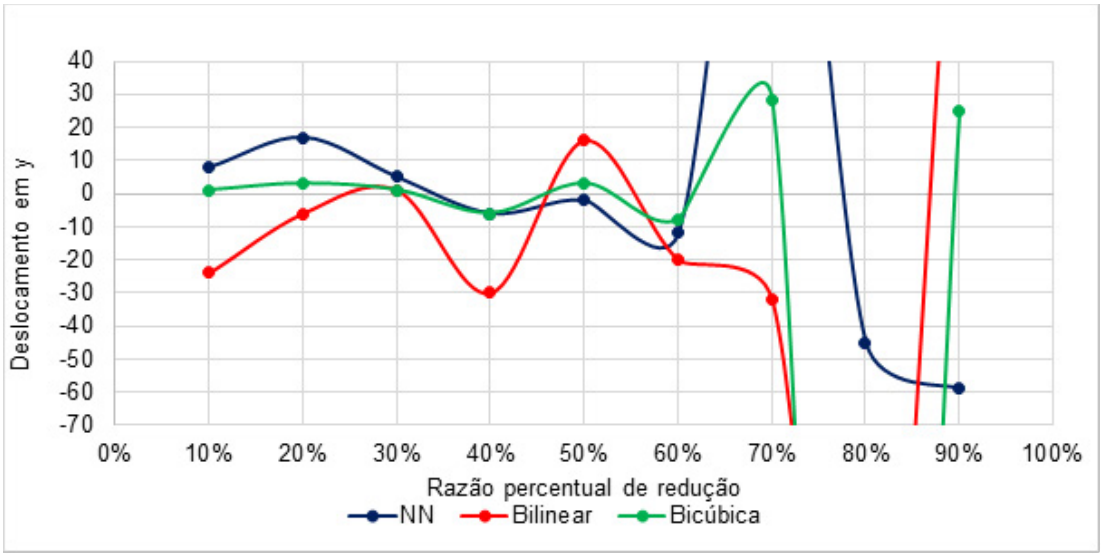


Gráfico 3- Erro na posição da imagem registrada no eixo y
Fonte: Elaborado pelos(as) autores(as).

Para facilitar a compreensão do erro introduzido em *x* e *y* mostrado anteriormente, foi realizada a raiz quadrada da soma dos quadrados dos erros, ilustrando como diferentes níveis de redução percentual afetam a correspondência de regiões no registro.

Como mostrado no Gráfico 4, a reamostragem Bilinear gera os maiores erros, enquanto as demais inserem menos erros. A Bicúbica e a NN apresentam resultados semelhantes, entre 30% e 60% da razão percentual de redução. Em particular, a reamostragem Bicúbica demonstra menor variação nos erros conforme se varia a razão percentual de redução, indicando maior previsibilidade em relação aos demais métodos.

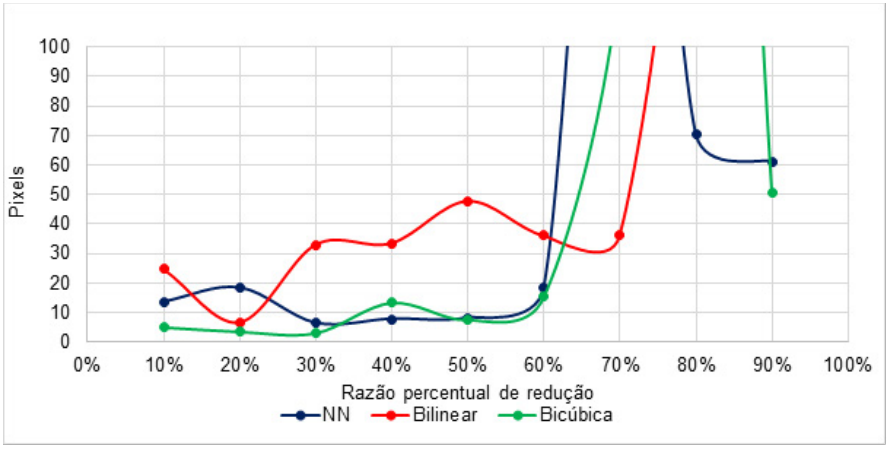


Gráfico 4- Soma do módulo do erro nos eixos x e y
Fonte: Elaborado pelos(as) autores(as).

Em todas as discussões anteriores, observa-se que a reamostragem Bilinear apresenta taxas de erro maiores em comparação aos outros métodos. Esses erros estão associados ao fato de que esse método suaviza as áreas de alto contraste, o que resulta na atenuação dos contornos das formas (Falandes; Carvalho, 2023). Isso faz com que as áreas, que são cruciais para a descrição e identificação de características, sejam reduzidas como demonstrado na Figura 5, que, ao comparar as áreas destacadas na imagem (g) com as outras nas imagens (f e h), evidencia-se que as áreas em (g) são menores.

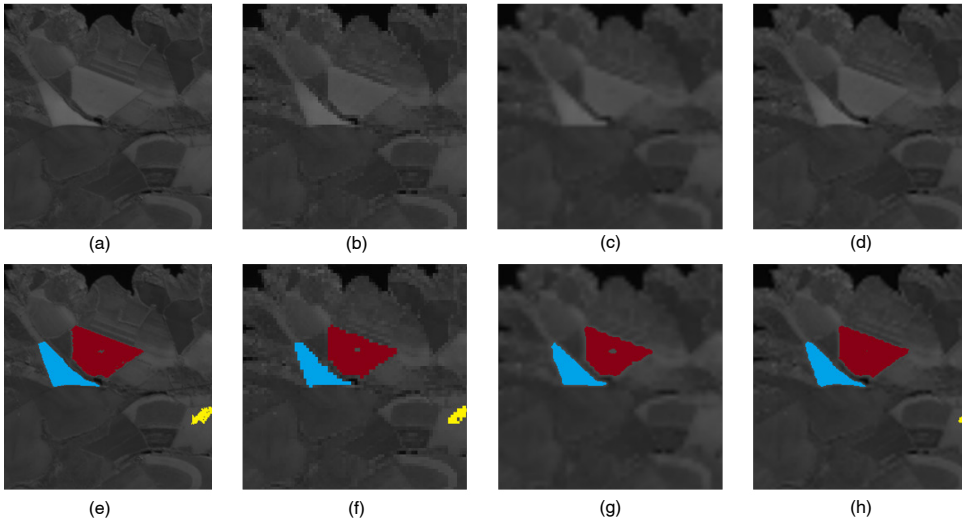


Figura 5 - Comparação da degradação de bordas em diferentes métodos: (a) imagem original sem reamostragem; (b), (c), (d) imagem após reamostragem, utilizando os métodos NN, Bilinear e Bicúbica respectivamente; (e) áreas destacadas na imagem original para comparação; (f), (g), (h) áreas destacadas após reamostragem pelos métodos NN, Bilinear e Bicúbica respectivamente, ilustrando a degradação das áreas das figuras e o impacto sobre os contornos
Fonte: Elaborado pelos(as) autores(as).

Essa redução nas áreas em virtude da perda de contraste nas bordas afeta a precisão na correspondência de padrões, já que a correspondência entre os pontos de controle é calculada com base nas áreas dessas formas. Portanto, a perda de contraste resulta em correspondências incorretas durante a busca por correspondência entre os pontos de controle, causando erros maiores no processo de registro. A reamostragem NN não reduz o contraste, uma vez que não estipula novas intensidades, contudo deixa os contornos com aspecto serrilhado, como é visível na Figura 5 (b). Por fim, a reamostragem Bicúbica, apesar de reduzir o contraste, o faz de maneira mais sutil, pois é baseada em splines cúbicas, que, mesmo alterando características dos contornos, mantêm o alto contraste entre as regiões, resultando em erros menores.

As avaliações mostradas a seguir consideraram a qualidade final do registro, ou seja, o quão fidedigno é o resultado em relação à imagem original, para isso os dados foram representados graficamente. Assim como nos gráficos anteriores, a magnitude foi restrita, uma vez que os pontos que extrapolam esse limite são resultados com erros extremamente altos, que não apresentam informações relevantes para a qualidade visual, já que são inutilizáveis por causa da quantidade de erros na correspondência do registro.

Ao observar os resultados dos Gráficos 5 e 6, é evidente a similaridade entre ambos, entretanto um mostra a quantidade de correlação entre as intensidades, e o outro a relação entre sinal-ruído. Em ambos, a reamostragem Bicúbica apresenta, em média, os piores resultados, já que essa reamostragem é menos eficiente no processo de correspondência de padrões, levando a resultados visuais mais distorcidos, ou seja, com ruídos maiores. O registro após a reamostragem por NN teve resultados superiores, porém o aspecto visual da imagem apresenta serrilhado nas regiões de contorno. Desse modo, observa-se que a reamostragem Bicúbica apresenta as menores taxas de erro, provavelmente em razão da sua característica de manter os contrastes nos contornos e a suavidade nos interiores.

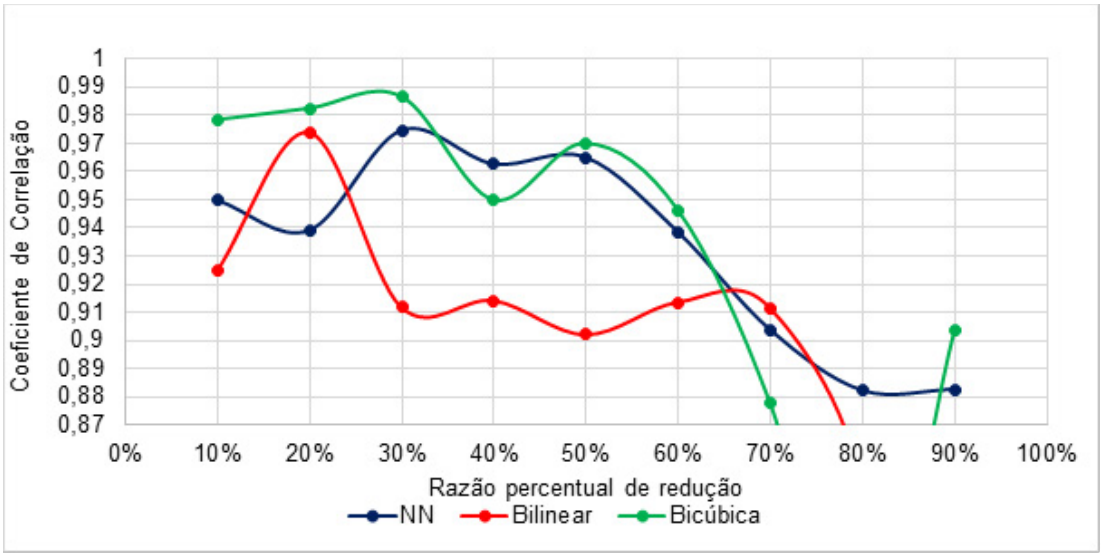


Gráfico 5 - Coeficiente de correlação da imagem registrada em relação à original

Fonte: Elaborado pelos(as) autores(as).

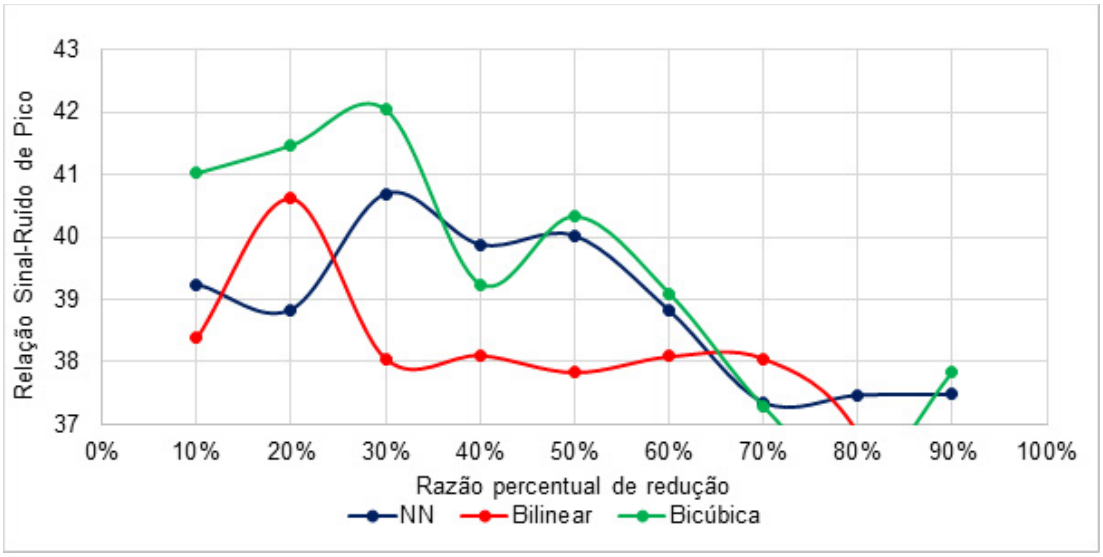


Gráfico 6- Relação sinal-ruído de pico da imagem registrada em relação à original
Fonte: Elaborado pelos(as) autores(as).

O Gráfico 7 demonstra a quantidade de erros entre as intensidades da imagem reamostrada com as da original, o que deixa evidente que os métodos por NN e Bicúbica apresentam, em média, os melhores resultados. Entretanto, a reamostragem Bicúbica tende a apresentar um desempenho mais estável e com menor erro ao longo das diversas taxas de redução, o que sugere uma preservação mais eficaz dos detalhes da imagem original. Em contraste, o método Bilinear exibe, em média, mais erros, especialmente em taxas de redução intermediárias – isso possivelmente decorre de sua menor capacidade de manter os contrastes da imagem, resultando em uma maior degradação.

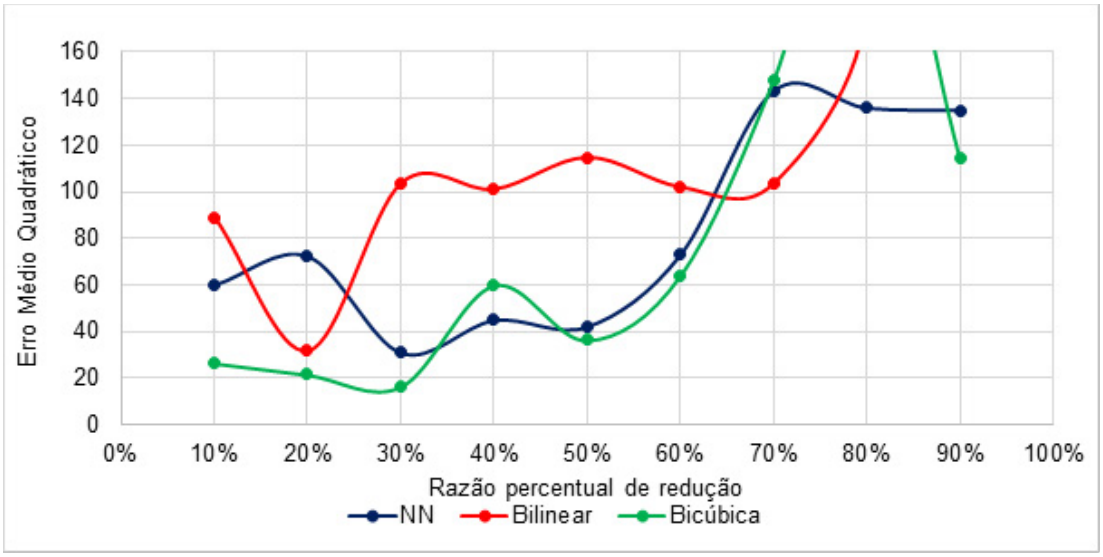


Gráfico 7- Erro médio quadrático da imagem registrada em relação à original
Fonte: Elaborado pelos(as) autores(as).

Considerações finais

A precisão na etapa de reamostragem desempenha um papel crítico no registro de imagens, uma vez que a integridade dos contornos e detalhes após o processo de reamostragem é essencial para que as características distinguíveis sejam preservadas, tornando possível o alinhamento correto durante o registro.

A análise comparativa dos métodos evidencia que a interpolação bicúbica, embora tenha um custo computacional mais elevado, destaca-se por inserir menos erros no processo de registro, em virtude da sua característica de preservação dos detalhes, além disso tem, em média, melhor previsibilidade de erro ao longo de diferentes taxas de redução, o que facilita a dedução dos prováveis erros presente na imagem final. A interpolação Bilinear, por outro lado, apresenta desempenho superior ao do método NN em baixas razões de redução da escala original, porém, nas demais situações, tende a suavizar excessivamente os contornos, o que prejudica a precisão da correspondência de padrões e afeta diretamente o processo de registro, tornando-se o método com as maiores taxas de erro. Por conseguinte, o método de reamostragem por NN, apesar de ser mais rápido e preservar melhor a qualidade da correspondência, é propenso a produzir resultados visuais inferiores em razão do efeito de serrilhado, que persiste mesmo após a conclusão do registro.

Os resultados dos testes indicam que a reamostragem Bicúbica é a mais eficaz, produzindo taxas de erro menores durante o processo de registro. Enquanto isso, o método do NN se destaca em relação à reamostragem Bilinear, por inserir menos erros no registro, sobretudo em diversas razões de redução. Portanto, a interpolação Bilinear apresenta o maior impacto, juntamente com várias oscilações de erro.

Além disso, os erros introduzidos no registro em razão da reamostragem impactam significativamente a precisão das localizações das imagens em relação à realidade. Isso pode causar diversos problemas, como o monitoramento de regiões, pois diferentes resoluções espaciais podem gerar discrepâncias notáveis nas posições geográficas, comprometendo o controle e a identificação de focos de incêndio e desastres naturais.

Com base nos resultados obtidos, propõe-se a continuidade da pesquisa voltada para a fusão de imagens de diferentes satélites, utilizando técnicas de reamostragem e registro. O objetivo é explorar a combinação de dados provenientes de satélites com diferentes resoluções espaciais, o que permitirá a obtenção de imagens mais detalhadas e com maior cobertura temporal. A fusão de imagens pode, assim, oferecer uma abordagem mais robusta e dinâmica para o sensoriamento remoto, com aplicações diretas no monitoramento de fogo ativo.

Referências

BAY, H.; TUYTELAARS, T.; VAN GOOL, L. *SURF*: speeded up robust features. European Conference on Computer Vision, Graz. Lecture Notes in Computer Science. Berlin: Springer, 2006. DOI: http://dx.doi.org/10.1007/11744023_32.

CHANDRAPPA, D. N.; ANIL, N. S. Satellite image matching and registration using affine transformation and hybrid feature descriptors. *International Journal of Advanced Intelligence Paradigms*, [*s. l.*], v. 1, n. 1, 2021.

DUNG, P. T.; CHUC, M. D.; THANH, N. T. N.; HUNG, B. Q.; CHUNG, D. M. Comparison of Resampling Methods on Different Remote Sensing Images for Vietnam's Urban Classification. *Research and Development on Information and Communication Technology*, [*s. l.*], v. 2, n. 15, p. 8-20, 2018.

FALANDES, C. E.; CARVALHO, F. G. M. Análise Quantitativa Preliminar de Métodos de Reamostragem de Imagens Digitais Aplicáveis a Diferentes Tipos de Formas Geométricas. *In*: ESCOLA REGIONAL DE INFORMÁTICA DE GOIÁS (ERI-GO), 11., 2023, Goiânia. *Anais* [...]. Porto Alegre: SBC, 2023.

FALANDES, C. E.; CARVALHO, F. G. M.; MORELLI, F. Algoritmo de detecção de cantos aplicado ao problema de registro de imagens de satélites. *In*: SCIENCE & BUSINESS CONNECTION: CONGRESSO CIENTÍFICO E TECNOLÓGICO, 2024, São José dos Campos. *Anais* […]. São José dos Campos: PIT, 2024. Disponível em: 10.29327/2-science-business-connection-407088.811823. Acesso em: 12 set. 2025.

GONG, X.; YAO, F.; MA, J.; JIANG, J.; LU, T.; ZHANG, Y.; ZHOU, H. Feature matching for remote-sensing image registration via neighborhood topological and affine consistency. *Remote Sensing*, [*s. l.*], v. 14, n. 11, p. 2606, 2022.

GONZALEZ, R.; WOODS, R. E. *Digital Image Processing*. 4. ed. [*S. l.*]: Pearson, 2019.

GOSHTASBY, A. A. *2-D and 3-D image registration*: for Medical, Remote Sensing, and Industrial Applications. Nova Jersey: John Wiley & Sons, 2005.

HAN, D. Comparison of Commonly Used Image Interpolation Methods. *In*: 2ND INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE AND ELECTRONICS ENGINEERING, 2013, [*s. l.*]. *Anais* […]. [*S. l.: s. n.*]. DOI: http://dx.doi.org/10.2991/iccsee.2013.391.

HARRIS, C.; STEPHENS, M. *A Combined Corner and Edge Detector*. Plessey Research Roke Manor. United Kingdom: The Plessey Company pic, 1988.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS. Catálogo de Imagens. Disponível em: https://www.dgi.inpe.br/catalogo/explore. Acesso em: 10 set. 2025.

KAI, P. M. *et al.* Effects of resampling image methods in sugarcane classification and the potential use of vegetation indices related to chlorophyll. In: IEEE ANNUAL COMPUTERS, SOFTWARE, AND APPLICATIONS CONFERENCE, 45., 2021, Madrid. *Anais* […]. Madrid: [*S. n.*]. DOI: 10.1109/COMPSAC51774.2021.00227.

KEYS, R. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics*, Speech, and Signal Processing, [*s. l.*], v. 29, n. 6, p. 1153-1160, 1981. DOI: http://dx.doi.org/10.1109/TASSP.1981.1163711.

LIN, B. *et al.* A registration algorithm for astronomical images based on geometric constraints and homography. *Remote Sensing*, [*s. l.*], v. 15, n. 1921, p. 1-25, 2023. DOI: https://doi.org/10.3390/rs15071921.

LOWE, D. G. Object recognition from local scale-invariant features. *In*: INTERNATIONAL CONFERENCE ON COMPUTER VISION, 1999, Corfu, Grécia. *Anais* […]. Corfu, Grécia: [*S. n.*]. Disponível em: https://ieeexplore.ieee.org/document/790410. Acesso em: 12 set. 2025.

MAHI, H.; ISABATEN, H.; SERIEF, C. Z. Zernike Moments and SVM for Shape Classification in Very High-Resolution Satellite Images. *The International Arab Journal of Information Technology*, [*s. l.*], v. 11, n. 1, p. 43-51, 2014.

MEDHA, V. W.; PRADEEP, M. P.; HEMANT, K. A. Image Registration Techniques: An overview. *International Journal of Signal Processing*, [*s. l.*], v. 2, n. 3, p. 11-28, 2009.

MELO, G. K.; RIBEIRO, E. A. W. Mapeamento exploratório da vegetação em uma escala local de paisagem: banda NIR como dado de partida. *Geografia*, [*s. l.*], v. 47, n. 1, 2022.

NAJJAR, Y. A. Comparative analysis of image quality assessment metrics: MSE, PSNR, SSIM and FSIM. *Journal of Science and Research*, [*s. l.*], v. 13, n. 3, p. 1-8, 2024. DOI: http://dx.doi.org/10.21275/SR24302013533.

PAPPAS, T. N.; SAFRANEK, R. J.; CHEN, J. Perceptual Criteria for Image Quality Evaluation. *Handbook of Image and Video Processing*, [*s. l.*], 2005.

PORWAL, S.; KATIYAR, S. K. Performance evaluation of various resampling techniques on IRS imagery. *(IC3), Noida*, Índia, p. 489-494, 2014.

PRASANTHA, H. S.; SHASHIDHARA, H. L.; BALASUBRAMANYA, M. K. N. Image scaling comparison using universal image quality index. In: Int. Conf. Adv. *Computing Control and Telecommunication Technologies*, Bangalore, p. 859-863, 2009.

RUBLEE, E.; RABAUD, V.; KONOLIGE, K.; BRADSKI, G. ORB: an efficient alternative to SIFT or SURF. *In*: INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2011, Barcelona, Espanha. *Anais* […]. Barcelona, Espanha: [*S. n.*], 2011. Disponível em: https://ieeexplore.ieee.org/document/6126544.

SAHARAN, R. A Review paper on Image Registration Techniques. *Journal of New Innovations in Engineering and Technology*, [*s. l.*], v. 4, n. 4, p. 34-39, 2016.

ZHANG, X.; LENG, C.; HONG, Y.; PEI, Z.; CHENG, I.; BASU, A. Multimodal Remote Sensing Image Registration Methods and Advancements: A Survey. *Remote Sensing*, [*s. l.*], v. 13, n. 5128, 2021.

ZITOVÁ, B.; FLUSSER, J. Image Registration Methods: A Survey. *Image and Vision Computing*, [*s. l.*], v. 21, p. 977-1000, 2003.

tecnia revista de educação, ciência e tecnologia do ifg

v. 10, Edição Especial 1 | 2025

Submetido 30/05/2024. Aprovado 21/03/2025
Avaliação: revisão duplo-anônimo

Expansão automática de léxico para Análise de Sentimentos de textos no domínio do Mercado Financeiro Brasileiro

AUTOMATIC LEXICON EXPANSION FOR SENTIMENT ANALYSIS: OF TEXTS IN THE BRAZILIAN FINANCIAL MARKET DOMAIN

AMPLIACIÓN AUTOMÁTICA DEL LÉXICO PARA EL ANÁLISIS DE SENTIMIENT O DE TEXTOS EN EL DOMÍNIO DEL MERCADO FINANCIERO BRASILEÑO

Thiago Monteles de Sousa
Universidade Federal de Goiás (UFG)
thiagomonteles@discente.ufg.br

Deborah Silva Alves Fernandes
Universidade Federal de Goiás (UFG)
deborah@inf.ufg.br

Kéthlyn Campos Silva
Universidade Federal de Goiás (UFG)
kethlyncampos@discente.ufg.br

Márcio Giovane C. Fernandes
Universidade Estadual de Goiás (UEG)
marcio.giovane@ueg.br

Resumo

Este artigo explora a geração de léxicos especializados para o Mercado Financeiro Brasileiro (MFB), adotando uma abordagem híbrida que combina a criação de um léxico em português com a análise de sentimentos em *tweets* e notícias do MFB. A metodologia consiste em uma série de etapas que expandem um léxico semente por meio de técnicas como Word2Vec, sinônimos/antônimos e Pointwise Mutual Information (PMI). Os resultados demonstram que a abordagem lexical alcançou um *F1-Score* de 71,5% na classificação de *tweets* e 68,4% em notícias, enquanto a combinação do léxico com o modelo de aprendizagem de máquina support vector machine (SVM) resultou em um *F1-Score* de 80% para *tweets*. Além disso, o estudo destaca a eficácia da lematização no pré-processamento para melhorar a precisão e cobertura do léxico como também a oportunidade da abordagem demonstrada na criação de léxicos específicos.

Palavras-chave: expansão lexical; mercado financeiro brasileiro; processamento de linguagem natural; redes sociais.

Abstract

This article investigates the generation of specialized lexicons for the Brazilian Financial Market (MFB) through a hybrid approach that integrates the construction of a Portuguese lexicon with sentiment analysis

applied to *tweets* and financial news. The proposed methodology involves a sequence of steps aimed at expanding a seed lexicon by employing techniques such as Word2Vec, synonym and antonym extraction, and Pointwise Mutual Information (PMI). Experimental results indicate that the lexical approach achieved an F1-score of 71.5% in tweet classification and 68.4% in news classification. Furthermore, when combined with the Support Vector Machine (SVM) learning model, the lexicon attained an F1-score of 80% for tweets. The study also underscores the effectiveness of lemmatization in preprocessing, both for improving the accuracy and expanding the coverage of the lexicon, as well as the potential of the proposed methodology for developing domain-specific lexical resources.

Keywords: lexical expansion; brazilian financial market; natural language processing; social networks.

Resumen

Este artículo explora la generación de léxicos especializados para el Mercado Financiero Brasileño (MFB), adoptando un enfoque híbrido que combina la creación de un léxico en Portugués con el análisis de sentimientos en *tweets* y noticias del MFB. La metodología consiste en una serie de pasos que expanden um léxico semilla mediante técnicas como Word2Vec, sinónimos/antónimos y la Información Mutua Puntual (PMI). Los resultados demuestran que el enfoque léxico alcanzó un f1-score de 71.5 % en la clasificación de *tweets* y 68.4% en noticias, mientras que la combinación del léxico con el modelo de aprendizaje automático máquina de vectores de soporte (SVM) resultó en un f1-score de 80% para *tweets*. Además, el estudio destaca la eficacia de la lematización en el preprocesamiento para mejorar la precisión y cobertura del léxico, así como la oportunidad de la aproximación demostrada en la creación de léxicos específicos.

Palabras clave: expansión léxica; mercado financiero brasileño; procesamiento de lenguaje natural; redes sociales.

Formatações Gerais

Com a popularização das plataformas de redes sociais online, como o Twitter (conhecido como X a partir de 2023), Facebook, LinkedIn e outras, milhares de usuários têm interagido com postagens e mensagens que abordam uma grande variedade de tópicos. Essas plataformas tornaram-se espaços onde os usuários expressam suas opiniões cada vez mais e também as utilizam como instrumento para a tomada de decisões (Bos; Frasincar, 2022). O Twitter, em particular, é uma das redes sociais mais populares no mundo. A rede permitia que cada usuário publicasse mensagens chamadas *tweets*, com limite de 4 mil caracteres para a versão paga e 280 para a gratuita. Esses *tweets* são visualizados por outros usuários por meio do compartilhamento de publicações (*retweets*) e interações, tornando-se uma fonte relevante para acompanhar tendências e opiniões (Carosia; Coelho; Silva, 2020).

No contexto do mercado financeiro, o Twitter é uma plataforma amplamente utilizada por investidores para expressar suas opiniões em razão da simplicidade e influência midiática nas dinâmicas de preços das ações. No entanto, dada a enorme quantidade de publicações, análises manuais se tornam inviáveis. Nesse cenário, a Análise de Sentimentos (AS), uma abordagem de Processamento de Linguagem Natural (PLN), é empregada para extrair indicadores automáticos das opiniões. A AS divide as tarefas em identificação da polaridade (positiva ou negativa) e da emoção associada, como felicidade ou tristeza (Pereira, 2021). Existem duas abordagens principais: Aprendizagem de Máquina (AM), que oferece resultados promissores, mas requer grande quantidade de dados rotulados, tornando o processo trabalhoso e custoso; e a abordagem lexical, que

se baseia na Orientação Semântica das palavras nos textos, proporcionando facilidade de construção, seja de forma automática, seja por meio de textos relacionados ao mercado financeiro (Mahmood *et al.*, 2020).

Dessa forma, este trabalho tem como objetivo abordar as possibilidades de geração de vocabulários especializados, examinando uma perspectiva híbrida para criar um léxico em Português voltado ao domínio do Mercado Financeiro Brasileiro (MFB). O intuito é identificar palavras que indiquem graus de otimismo ou pessimismo em textos relacionados ao campo-lavo, contribuindo para a aplicação de PLN nesse contexto da língua portuguesa, área que ainda apresenta escassez de estudos publicados (Januário *et al.*, 2021; Pereira, 2021). Assim, este trabalho visa contribuir para o avanço da área de PLN e fornecer recursos para a criação de léxicos com domínios específicos. Nesse contexto, será elaborada uma estratégia para validar os vocabulários obtidos em tarefas de AS no âmbito do MFB.

As principais contribuições deste trabalho estão resumidas a seguir.

- Elaborar diferentes configurações para a geração de léxicos do domínio-alvo, resultando na criação de léxicos do campo especializado.
- Testar o desempenho dos léxicos por meio da análise de sentimentos em *tweets* e notícias no campo do Mercado Financeiro Brasileiro.
- Comparar o desempenho entre abordagem lexical, aprendizagem de máquina supervisionado e uma proposta que mescle as duas abordagens na tarefa de classificação de sentimentos.

Trabalhos Relacionados

A abordagem lexical é um recurso presente em várias atividades de processamento de linguagem natural, como análise de sentimentos, classificação de textos, recuperação de opinião, identificação de temas, entre outras. Quando elaborados de forma adequada, os léxicos podem fornecer uma boa capacidade de classificação, além de poderem ser utilizados como recursos adicionais aos modelos de aprendizagem de máquina (Oliveira; Cortez; Areal, 2016). Detectar subjetividades em sentenças e classificá-las em uma classe é um desafio, especialmente em domínios específicos, como o mercado de ações (Das *et al.*, 2022), doenças (Jung *et al.*, 2021), documentos jurídicos (Smywinski-Pohl *et al.*, 2019) e outros que exigem corpora especializado.

A construção de um dicionário de léxicos pode seguir diferentes abordagens. Uma delas é totalmente manual, como em Loughran e McDonald (2011), que apresenta uma popular coleção de palavras rotuladas para o domínio do mercado financeiro. Para isso, foram utilizados documentos de textos extraídos do portal U.S *Securities and Exchange Commission* entre 1994 e 2008, resultando em seis grupos de palavras. Outra abordagem é de forma automática, como o realizado por Smywiński-Pohl *et al.* (2019). Neste, é proposta a construção de um dicionário polonês, que mapeia a relação entre os termos jurídicos e extrajurídicos. Para isso, os pesquisadores compilaram documentos judiciais e extrajudiciais e realizaram etapas de pré-processamento para a redução de ruídos. Posteriormente, foram elaborados dois dicionários que combinam n-gramas obtidos por meio da ferramenta SRILM toolkit e a semelhança de cosseno entre os vetores dos termos dos dois dicionários com o auxílio do modelo *Word2Vec.*

Além das abordagens de construção mencionadas, existe uma abordagem híbrida, que utiliza um conjunto de palavras como semente para um contexto específico e um processo de expansão desse vocabulário. No estudo de Bos e Frasincar (2022), foram avaliadas três abordagens para a expansão automática de léxicos relacionados ao mercado financeiro: uma baseada na probabilidade de pertencimento das palavras a conjuntos positivos ou negativos, utilizando a medida *Pointwise Mutual Information* (PMI); outra que usa uma adaptação da medida *Term Frequency-Inverse Document Frequency* (TF-IDF), considerando documentos como categorias e avaliando a frequência das palavras em várias categorias; e uma terceira que emprega o Word2Vec como embedding de palavras para definir a proximidade entre conjuntos de palavras e termos da vizinhança para classificar em categorias apropriadas.

O processo de avaliar a qualidade do léxico em tarefas de AS pode ser entendido por meio da abordagem de um analisador lexical que faz a soma das pontuações dos termos-alvo, também conhecido como *Sentiment Orientation* (SO), como é usado em Oliveira, Cortez e Areal (2016), Carosia, Coelho e Silva (2020), Shan, Jiang e Wang (2021) e Wang *et al.* (2020). Uma opção com aprendizagem de máquina supervisionada consiste em utilizar SO para incrementar essas informações como entrada para um classificador de sentimentos. Um exemplo de aplicação dessa abordagem é o estudo realizado por Bos e Frasincar (2022), que utilizou support vector machine (SVM) com Bag-Of-Words (BOW) para codificação de texto na validação de um léxico de mercado financeiro americano e obteve uma acurácia de 75,1%.

Como mencionado em Pereira (2021), poucos trabalhos focam a análise dos textos na língua portuguesa. Assim, pensando nesse panorama apresentado pela revisão bibliográfica, em que se observa um déficit de propostas que adotam léxicos especializados no contexto da língua portuguesa, neste artigo será adotada uma estratégia para a construção automática de léxicos específicos para o domínio do mercado financeiro brasileiro.

Metodologia

Este capítulo fornece uma descrição detalhada dos procedimentos adotados neste estudo. No início, foi apresentado o conjunto de dados utilizado. Em seguida, foi apresentado o protocolo de processamento de texto aplicado em todas as etapas. Por último, foi detalhada a proposta de construção do léxico-alvo, incluindo a criação do léxico semente e suas variações.

Conjunto de Dados	Otimistas	Pessimistas	Total
Conjunto de *tweets* (AUTOR/A, 2019)	2048	1180	3228
Conjunto de Notícias (Januário *et al.*, 2021)	555	273	828

Quadro 1 - Informações dos conjuntos de dados para avaliar o desempenho final dos léxicos
Fonte: Elaborado pelos(as) autores(as).

Base de dados

Um dos conjuntos de textos adotadas é composto de 1.031.419 *tweets* distintos. As mensagens foram coletadas no ano de 2019, e foi possível utilizar uma API[1] fornecida pelo Twitter para esse fim. Foram utilizados os nomes de empresas e seus *tickers*[2] como filtros para a seleção das publicações. A coleta foi realizada conforme descrito em (AUTOR/A, 2019).

No contexto da avaliação de léxicos para a classificação de sentimentos em *tweets* sobre o MFB, utilizou-se um conjunto de teste composto de 3228 *tweets* rotulados, dos quais 2048 foram categorizados como otimistas e 1180 como pessimistas. Adicionalmente, na classificação de notícias do MFB, empregou-se um conjunto de teste composto de 828 notícias rotuladas, com 555 classificadas como otimistas e 273 como pessimistas, conforme produzido por Januário *et al.* (2021).

Pré-processamento dos textos

Durante os experimentos, todos os textos foram submetidos a etapas pré-processamento, o que é crucial para assegurar a qualidade dos dados utilizados nos próximos experimentos. O processo envolve a normalização das palavras para minúsculas, a remoção de stopwords usando a lista para a língua portuguesa disponível no Natural Language Toolkit (NLTK) (Bird, 2006), a eliminação de menções a usuários e a exclusão de URLs, hashtags, números, emoticons e pontuações. Além disso, o processo inclui a geração de tokens, que consiste na separação dos textos em palavras individuais.

Construção lexical

O fluxo do método principal, denominado como a primeira configuração do léxico, é apresentado na Figura 1. O método de construção e expansão automática do léxico é composto distintas. A primeira é a criação de um conjunto de palavras que represente o domínio, e a segunda etapa é a expansão das palavras.

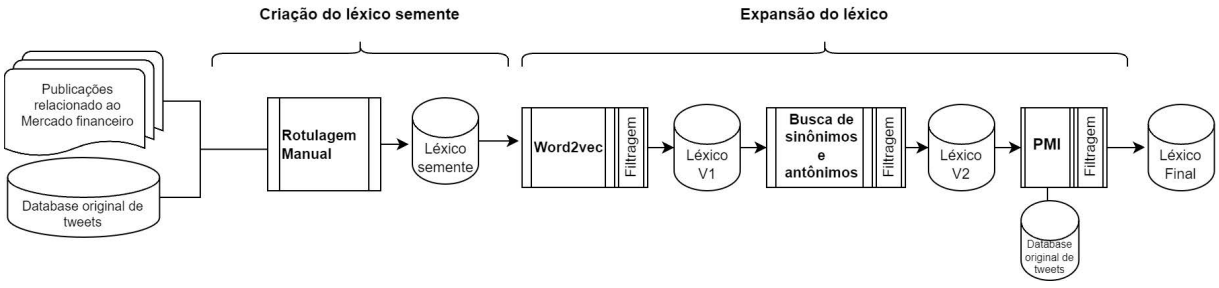


Figura 1 - Fluxo da construção lexical da principal configuração proposta. Semente (S) + Word2Vec (W2V) + Sinônimos e Antônimos (S/A) + Pointwise Mutual Information (PMI)
Fonte: Elaborado(a) pelos(as) autores(as).

1 Application Programming Interface.

2 Rótulos utilizados para identificar ações de uma empresa.

LÉXICO SEMENTE

O processo da criação do conjunto semente inicia-se com uma análise exploratória do corpora, buscando identificar as palavras que potencialmente exercem maior influência nos textos do domínio-alvo por conta da repetição observada. A primeira etapa para o domínio do MFB foi a seleção de conjuntos de textos que incluíram *tweets* e notícias coletadas no ano de 2019, submetidos ao pré-processamento dos textos, conforme descrito anteriormente, a fim de obter um conjunto de palavras sem possíveis ruídos. Dessa forma, foi possível gerar uma nuvem de palavras por meio do uso da biblioteca WordCloud, cujo resultado é apresentado na Figura 2.

Com base na WordCloud, as principais palavras identificadas são selecionadas e rotuladas manualmente. Exemplos dessas palavras incluem “lucro”, “sobe”e “alta”, as quais receberam uma classificação **otimista**. Por outro lado, palavras como “baixa”, “perda”e “queda” receberam uma classificação **pessimista**. Além da observação inicial da nuvem de palavras, também foi realizada uma leitura visual de alguns *tweets* e notícias recolhidos de sites de conteúdo relacionados ao MFB, como Informoney e Estadão Investimentos, a fim de catalogar suas palavras-chave.

EXPANSÃO LEXICAL

Com o léxico semente estabelecido, o processo de expansão lexical automático começa com o uso de *word embeddings*, que criam representações de palavras em espaços n-dimensionais, capturando relações sintáticas e semânticas. Foi utilizado um modelo Word2vec de 600 dimensões para a língua portuguesa, previamente treinado por Hartmann *et al.* (2017).

A expansão pelo Word2vec (W2V) começa com a divisão do léxico semente (S) em otimista e pessimista. Para cada léxico com n palavras, são criados lotes (batch) com três palavras em ordem de inserção do conjunto semente, por exemplo, $batch_1(palavra_1, palavra_2, palavra_3)$, até que as n palavras estejam em lotes com no máximo três palavras. Posteriormente, o $batch_i$ é então entregue ao modelo pré-treinado Word2vec. Dessa forma, busca-se, dentro do espaço de representação, as palavras com os K vizinhos mais próximos, com base no cosseno de similaridade entre as palavras de entrada e seus vizinhos. Os três primeiros termos são selecionados para serem considerados como palavras candidatas, passando pela filtragem mencionada e, por fim, incorporados ao conjunto da rotulagem em expansão atualmente.

Figura 2 - Nuvem de palavras dos termos mais frequentes nos corpora de *tweets* do domínio do Mercado Financeiro Brasileiro
Fonte: Elaborado pelos(as) autores(as).

A segunda extensão envolve a expansão por Sinônimos e Antônimos (S/A), utilizando uma técnica de *web scraping* no site de dicionário online DICIO. Para isso, foi utilizada a biblioteca Python chamada Beautiful Soup.

O processo começa com a extração das palavras do léxico a ser estendido. Para cada palavra, é acessado o endereço virtual da página correspondente no site do dicionário, e as informações sobre sinônimos e antônimos são extraídas. Cada sinônimo é rotulado com a mesma orientação da palavra original, enquanto os antônimos recebem uma orientação oposta.

O passo para a expansão (S/A) é realizado extraindo-se as palavras do léxico a ser estendido. Para cada palavra, é feito o acesso à URL correspondente no site do dicionário e, por meio da ferramenta Beautiful Soup, são extraídas as informações sobre sinônimos e antônimos. Cada termo candidato sinônimo é rotulado com a mesma orientação da palavra que está sendo estendida. Por outro lado, caso o termo candidato seja um antônimo, ele será associado a uma orientação oposta. Por fim, os candidatos são submetidos a uma filtragem para verificar se os termos já estão presentes no léxico-alvo.

A terceira extensão, conhecida como “expansão por PMI”, utiliza a medida probabilista *Pointwise Mutual Information* (PMI) para quantificar o sentimento de uma palavra com base em sua probabilidade de ocorrência em um conjunto de dados. Essa abordagem amplamente explorada em trabalhos anteriores, como Oliveira, Cortez e Areal (2016), Losada e Gamallo (2020), Bos e Frasincar (2022), avalia a força de uma palavra em ser considerada positiva ou negativa em relação ao domínio em questão.

A medida estatística PMI é definida pela Equação 1:

$$PMI(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

Onde x e y são variáveis ou conjuntos de variáveis, *P*(x, y) representa a probabilidade conjunta de x e y ocorrerem, e *P*(x) e *P*(y) representam as probabilidades marginais de ocorrência de x e y no conjunto de variáveis.

A Orientação Semântica (OS) de uma nova palavra x é calculada como a diferença entre a força associada ao conjunto de palavras positivas (*setPositivo*) e a força associada ao conjunto de palavras negativas (*setNegativo*), conforme a Equação 2.

$$OS(x) = PMI(x, setPositivo) - PMI(x, setNegativo) \tag{2}$$

Essa diferença reflete a intensidade da associação da palavra com cada conjunto, permitindo inferir seu sentimento em relação ao domínio de interesse.

Com isso, é possível realizar uma série de passos com o objetivo de estender um conjunto de palavras. As etapas do procedimento estão ilustradas na Figura 3.

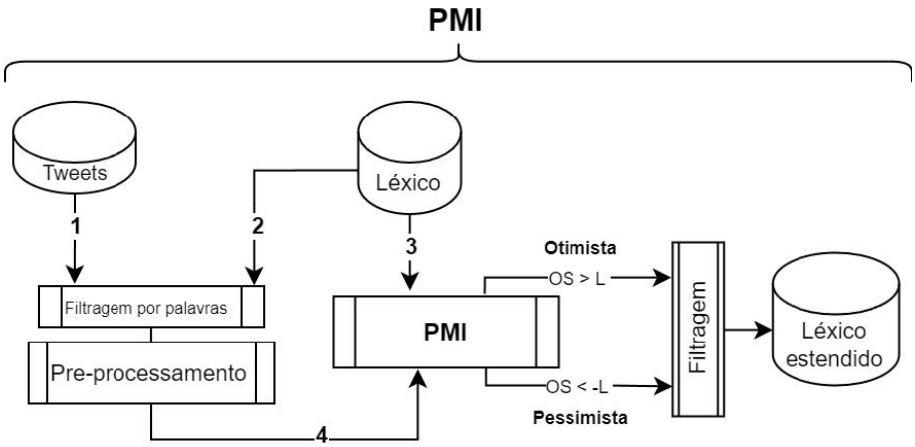


Figura 3 - Fluxo da extensão do léxico utilizando a medida Pointwise Mutual Information.
Fonte: Elaborado pelos(as) autores(as).

O processo começa filtrando *tweets* que contenham palavras do léxico a ser ampliado (etapas 1 e 2 da Figura 3), considerando que o PMI necessita relacionar as ocorrências dessas palavras-chave com outras, indicando maior probabilidade de co-ocorrência com sentimentos semelhantes. Em seguida, os *tweets* filtrados passam por pré-processamento e geram uma sequência de *tokens*. Esses *tokens*, juntamente com as palavras do léxico a ser ampliado, são usados para calcular a Orientação Semântica (OS) da nova palavra (etapas 3 e 4).

Por fim, a orientação semântica (OS) da nova palavra x será definida pela diferença entre o PMI com o léxico otimista e com o léxico pessimista. Para definir a polaridade da nova palavra, utiliza-se um limiar *L*, onde, caso o resultado de *O* seja maior que *L*, será considerado otimista, e caso seja menor que *–L*, será considerado pessimista:

$$OS(x) = \begin{cases} Otimista; & \text{se } OS \geq L; \\ Pessimista; & \text{se } OS \leq -L; \end{cases} \tag{3}$$

Dessa forma, a construção do léxico final para a configuração, que começou com o léxico semente e foi ampliada pelas etapas Word2Vec, busca por sinônimos/antônimos e, finalmente, a busca por novos termos por meio da medida PMI é concluída.

Alguns trabalhos como Carosia, Coelho e Silva (2020), Das *et al.* (2022) e Shan, Jiang, Wang (2021) rotulam o peso das palavras do léxico com +1 para positivas e -1 para negativas. Entretanto, essa abordagem considera que todos os termos possuem o mesmo grau de importância para a definição do sentimento associado ao texto-alvo. Uma abordagem contrastante é o uso personalizado de pesos associados a cada termo, como é feito em trabalhos como Oliveira, Cortez e Areal (2016), Bos e Frasincar (2022), Wang *et al.* (2020) e Losada e Gamallo (2020), em que o processo de rotulagem é utilizado para definir um peso para as palavras. Neste trabalho, será utilizada a pontuação PMI para determinar os pesos associados aos termos dos léxicos, conforme as etapas a seguir.

- Peso para palavras das etapas Semente, W2V e S/A: os termos obtidos nessas etapas, por tratar-se de palavras obtidas por relação semântica direta da rotulagem manual feita no conjunto semente, receberão pontuação máxima (+1 para otimistas e -1 para pessimistas).

- Peso para palavras da etapa PMI: será utilizada a função Min-Max Scaling, disponível na biblioteca *scikit-lear*[3], para normalizar entre +1 e -1 a pontuação PMI obtida para cada palavra estendida nessa etapa. A palavra com a maior pontuação PMI otimista será normalizada para +1, e a palavra com a menor pontuação PMI pessimista será normalizada para -1.

O conjunto final com os pesos personalizados segue o modelo apresentado no Quadro 2.

Dessa forma, a construção do léxico final, que iniciou-se com o léxico semente e foi ampliada pelas etapas Word2Vec, busca por sinônimos/antônimos e, finalmente, a busca por novos termos por meio da medida PMI (S+W2V+S/A+PMI) foi finalizada. Com o objetivo de verificar a melhor configuração e o impacto das etapas de expansão no léxico final, foram implementadas variações de configurações do léxico para serem avaliadas em experimentos na classificação de *tweets* e Notícias. Todas as configurações são apresentadas no Quadro 3.

Palavra	Etapa de ingresso	Peso	Rótulo
positiva	Semente(S)	+1	Otimista
recuar	Word2Vec	-1	Pessimista
perder	S/A	-1	Pessimista
conseguindo	PMI	+0.814	Otimista
greve	PMI	-1	Pessimista

Quadro 2 - Exemplo de pesos para palavras vindas de diferentes etapas
Fonte: Elaborado pelos(as) autores(as).

Cosntrução	Etapas
1	S
2	S+PMI
3	S+S/A+PMI
4	S+W2V+S/A+PMI

Quadro 3 - Configurações dos léxicos finais
Fonte: Elaborado pelos(as) autores(as).

Configurações dos experimentos de classificação

Os experimentos realizados neste estudo têm como objetivo testar diferentes configurações de léxicos gerados pelo processo descrito anteriormente. Inicialmente, aplicou-se uma abordagem que utiliza a técnica de soma das pontuações dos termos do léxico para a classificação de textos do MFB. Essa abordagem consiste em calcular uma pontuação total para cada texto, somando as pontuações individuais dos termos presentes no léxico, e utilizando essa pontuação para determinar a classificação do texto.

Em um segundo experimento, foram implementadas duas técnicas de aprendizado supervisionado: *Naive Bayes* (NB) e *Support Vector Machine* (SVM). Essas técnicas foram escolhidas devido à sua eficácia em tarefas de classificação de texto e foram implementadas com a biblioteca *scikit-learn* em Python. Para a representação dos textos, utilizou-se a abordagem *bag-of-words* (BOW), que transforma os textos em

3 https://scikit-learn.org

vetores de frequência de palavras, permitindo que os algoritmos de aprendizado de máquina processem os dados de forma eficiente.

No terceiro experimento, procurou-se enriquecer a representação dos textos utilizando informações adicionais fornecidas pelo analisador lexical. Essas informações foram integradas à representação matricial do texto, com o intuito de melhorar o desempenho dos modelos de classificação. A finalidade foi explorar se a inclusão de características léxicas adicionais poderia proporcionar um aumento na precisão dos modelos.

O treinamento em todas as tarefas de aprendizado supervisionado foi conduzido utilizando a técnica de validação cruzada *K-Fold*. Essa técnica envolve a divisão dos dados em K subconjuntos (ou "*folds*"), realizando múltiplos treinamentos e validações cruzadas.

Otimização do limiar para a etapa PMI

Para definir o limiar adequado para a rotulagem das palavras na etapa de extensão por PMI, será empregada a otimização bayesiana visando maximizar a métrica F-score na classificação de sentimentos de *tweets* e notícias, ao mesmo tempo que se minimiza a porcentagem de textos não classificados para ambos os corpus. Para isso, foi proposta uma métrica que consolida tais informações. A Equação 4 apresenta a função S(L), que pode ser entendida como uma combinação linear das métricas:

$$S(L) = \alpha \cdot (f1\_score\_T + f1\_score\_N) - \beta \cdot (Unclassified\_T - Unclassified\_N) \quad (4)$$

Onde $f1_{score}$ representa as pontuações F1 para os conjuntos de dados de *tweets* (T) e de notícias (N), e *Unclassified* representa a porcentagem de classificações não identificadas nesses conjuntos de dados. Os valores α e β são pesos atribuídos a cada métrica, refletindo sua importância relativa.

De acordo com o trabalho de Gardner *et al.* (2014), a Equação 5 apresenta a função fundamental da otimização bayesiana, representada da seguinte forma:

$$\min_{x \in X} f(x) \quad (5)$$

Onde $f(x)$ é a função objetivo que se deseja minimizar, e X é o espaço de busca. A otimização bayesiana envolve dois componentes principais: o Processo Gaussiano (GP) e a Função de Aquisição.

Um Processo Gaussiano é usado para prever o valor de uma função com base em observações anteriores. Ele fornece:

- **Média (μ):** Representa a estimativa média da função no ponto x.
- **Desvio padrão (σ):** Representa a incerteza dessa estimativa.

Como explicado por Snoek, Larochelle e Adams (2012), esses dois elementos modelam a função objetivo que queremos otimizar. Com base no GP, a função de aquisição decide onde olhar em seguida para encontrar o valor mínimo da função. Ela equilibra entre explorar novas áreas (onde a incerteza é alta) e explorar áreas promissoras (onde os valores conhecidos são bons). A Equação 6 representa a função de aquisição ***Expected Improvement*** **(EI)**:

$$EI(x) = \sigma(x)[z\Phi(z) + \phi(z)] \quad (6)$$

- $\sigma(x)$ é a incerteza no ponto x.
- z é uma medida de quão bom x parece ser, calculada como $z = \frac{\mu(x) - f_{\text{melhor}}}{\sigma(x)}$, onde f_{melhor} é o melhor valor encontrado até agora.
- Φ e ϕ são funções da distribuição normal.

O processo termina ao atingir uma convergência quando a função de aquisição deixa de sugerir pontos significativamente diferentes ou quando a quantidade determinada de repetições é alcançada. Neste experimento, foram utilizadas 100 iterações, com o intervalo de valores *L* variando entre 0 e 25.

Construção	Otimistas	Pessimistas	Total
S	75	75	150
S+PMI	507	1153	1660
S+S/A+PMI	1492	1518	3010
S+W2V+S/A+PMI	1685	1946	3631

Quadro 4 - Quantidade de palavras dos dicionários
Fonte: Elaborado pelos(as) autores(as).

Resultados e discussões

Neste capítulo, serão expostos os resultados dos experimentos realizados neste trabalho. Primeiramente, serão apresentados os resultados referentes à quantidade de termos decorrentes da expansão lexical proposta neste estudo, assim como os resultados da otimização de limiar para definição da polaridade dos termos. Em seguida, serão discutidos os desempenhos dos léxicos em tarefas de classificação de sentimentos em *tweets* e notícias, ambas sobre o domínio do mercado financeiro brasileiro. Por fim, será feita uma comparação entre o desempenho do método lexical e o método supervisionado.

Expansão lexical

Os resultados da expansão lexical apresentados no Quadro 4 é derivado da proposta de construção apresentada anteriormente.

A expansão foi realizada inicialmente por meio da criação de uma semente (S). A coleta da semente resultou em um conjunto de 75 palavras consideradas otimistas para o contexto proposto, assim como outras 75 palavras consideradas pessimistas.

Em seguida, foi realizada a primeira variação no pipeline de expansão (S+PMI), na qual se observou o impacto da expansão buscando apenas as palavras por meio da medida probabilística PMI após o léxico semente. Isso resultou em 507 palavras otimistas e 1153 palavras pessimistas.

Uma segunda variação foi realizada, na qual foram buscados sinônimos e antônimos (S/A) do conjunto semente, seguidos pelo uso do PMI. Essa abordagem (S+S/A+PMI) resultou em 1492 palavras otimistas e 1518 palavras pessimistas.

Por fim, uma última variante da expansão lexical foi realizada, buscando a similaridade entre as palavras por meio da word embedding Word2Vec (W2V). Em seguida, foi feita uma busca por sinônimos e antônimos, finalizando com a busca por termos mais específicos em relação aos já expandidos utilizando o PMI. Essa abordagem (S+W2V+S/A+PMI) resultou em um conjunto otimista de 1685 palavras e 1946 palavras pessimistas.

O limiar selecionado para o processo de expansão do léxico foi determinado com base nos resultados da otimização bayesiana, conforme ilustrado na Figura 1. Na Figura 4, podemos observar o comportamento das métricas ao longo de diferentes limiares. Para facilitar a visualização das métricas, foi utilizada a função de desempenho $S(L)$, que combina o *F1-Score* e a taxa de classificação incorreta. Essa função foi normalizada para variar entre 0 e 1, sendo 1 o ponto máximo de desempenho e 0 o mínimo.

Com isso, o valor máximo de 1 acontece quando $L = 3, 67$. Nesse ponto, as métricas individuais registram os seguintes valores: *F1-Score* para *tweets* (*F1_tweets*) é 0,7154, a proporção de *tweets* não classificados (Não classificado(T)) é 0,02, *F1-Score* para notícias (*F1_Noticias*) é 0,684, e a proporção de notícias não classificadas é zero.

Figura 4 - Variação das métricas normalizadas em função do limiar *T*. A linha sólida representa *S*(*T*), enquanto as linhas tracejadas representam as métricas *F*1_*tweets*, *F*1_*Noticia*s, Não classificado(*T*) e Não classificado(*N*)
Fonte: Elaborado pelos(as) autores(as).

Desse modo, a Figura 4 destaca a importância de selecionar um limiar adequado para balancear a maximização das métricas de desempenho e a minimização das classificações não identificadas. Escolher um limiar apropriado é crucial para garantir a eficácia da abordagem de expansão lexical apresentada neste trabalho.

Desempenho do léxico na classificação de *tweets* e notícias

Para avaliar o desempenho nas tarefas de classificação de sentimentos em *tweets* e notícias, foram consideradas as métricas de precisão, revocação (*recall*), *F1-Score* e acurácia geral. Além disso, utilizou-se a métrica denominada não classificados, que consiste na verificação da porcentagem de textos sem inferência da classe (Bos; Frasincar, 2022).

Na avaliação do desempenho das diferentes configurações de construção lexical, foram analisados os sentimentos de um conjunto de 3228 *tweets* categorizados manualmente. Esses resultados podem ser visualizados na Tabela 1.

Construção	Acurácia	Precisão	Recall	F1	Não classificadas
S+PMI	61,9%	65,1%	61,9%	63,4%	10,5%
S+PMI (lematizado)	66,5%	66,9%	66,5%	65,9%	4%
S+S/A+PMI	65,7%	68,6%	65,7%	67%	8,6%
S+S/A+PMI (lematizado)	71,7%	72,2%	71,7%	71,5%	2,8%
S+W2V+S/A+PMI	65%	66,7%	65%	65,7%	6,4%
S+W2V+S/A+PMI (lematizado)	71,1%	71,9%	71,1%	70,5%	2%

Tabela 1 - Avaliação dos léxicos de sentimento financeiro na classificação de *tweets* no conjunto de dados relacionados ao Mercado Financeiro Brasileiro (em %, melhores valores em negrito)
Fonte: Elaborado pelos(as) autores(as).

Todas as variações do léxico foram comparadas com uma abordagem de pré-processamento dos termos, em que as palavras, tanto no dicionário proposto quanto nos *tweets* a serem classificados, foram lematizadas, reduzindo-as ao seu lema raiz (Jung *et al.*, 2021). O melhor resultado foi obtido na proposta S+S/A+PMI (lematizado), com *F1-Score* de 71,5%, e uma precisão de 71,7%. Em contraste, sua versão não lematizada apresentou uma diferença negativa de até 6% na acurácia e 4,5% no F1. Isso se deve à facilidade de identificação dos termos uma vez normalizado, o que reduz a dimensionalidade dos termos, simplificando a comparação e o reconhecimento das palavras nos textos. No entanto, em termos de porcentagem de *tweets* que não foram classificados em razão da soma dos termos zerados ou à falta de cobertura das palavras do léxico nos tweets-alvo, a proposta mais adequada foi a S+W2V+S/A+PMI (lematizado), com um resultado de 2%, sendo assim considerando o léxico com a maior cobertura das palavras no conjunto de teste. Isso se deve principalmente as 621 palavras a mais nessa construção em comparação com a melhor abordagem geral.

O uso do léxico não se limita a textos no nível de sentença, mas também pode ser aplicado em documentos com textos mais extensos. Para testar o dicionário que obteve o melhor resultado previamente apresentado na Tabela 1, foi realizada uma avaliação do desempenho na classificação de um conjunto de notícias sobre o MFB, produzido por Januário *et al.* (2021). Essas 828 notícias possuem rótulos que indicam se o sentimento é otimista (555 notícias), refletindo uma alta expectativa de um determinado investidor em relação a uma ação, ou negativo, considerando um contexto pessimista (273 notícias).

	Acurácia	*F1-Score*
(1) Baseline (original) (JANUÁ- RIO *et al.*, 2021)	57,1%	57,4%
(2) Baseline (com stemming) (JANUÁRIO *et al.*, 2021)	58,2%	58,8%
(3) S+S/A+PMI	63,1%	63,4%
(4) S+S/A+PMI (com stemming)	58,5%	59,2%
(5) S+S/A+PMI (com lematização)	68,3%	68,4%

Tabela 2 - Desempenho médio das acurácias e *F1-Score* de notícias rotuladas usando o léxico de melhor pontuação (S+S/A+PMI) em comparação com o baseline (em %, melhores valores em negrito)
Fonte: Elaborado pelos(as) autores(as).

A Tabela 2 compara diferentes métodos de classificação, incluindo a linha de base original e variações do léxico com várias técnicas de processamento de texto. A linha de base original alcançou 57,1% de acurácia e um valor de *F1-Score* de 57,4%. Com a aplicação de *stemming* no pré-processamento, houve uma leve melhoria para 58,2% de acurácia e 58,8% de valor de *F1-Score*. No entanto, o uso do léxico proposto neste trabalho teve um impacto ainda mais significativo. A abordagem S+S/A+PMI alcançou 63,1% de acurácia e 63,4% de *F1-Score*. Com lematização, por sua vez, resultou em melhorias adicionais, elevando a acurácia para 68,3% e o valor de *F1-Score* para 68,4%.

Comparação com método supervisionado

Uma comparação foi realizada entre o desempenho do léxico de melhor resultado demonstrado anteriormente com os métodos SVM e NB, além de uma abordagem mista com analisador lexical. Os experimentos utilizaram um subconjunto de 2000 *tweets* do conjunto original, criado por meio da técnica de *Random Undersampling*.

Métrica	Léxico	SVM	NB	SVM+Léxico	NB+Léxico
F1-Score	67,8%	78,9% ± 0,7	76,4% ± 0,7	80% ± 0,6	77,7% ± 0,8

Tabela 3 - Comparação com método supervisionado treinando usando validação cruzada *K-Fold* K=50 em um sub-conjunto de 2000 *tweets*
Fonte: Elaborado pelos(as) autores(as).

Conforme ilustrado na Tabela 3, observou-se que o método de aprendizado de máquina SVM alcançou um valor de *F1-Score* de 78,9% com um desvio padrão de 0,7%, enquanto o NB atingiu 76,4% com um desvio padrão de 0,7%. Por outro lado, o léxico proposto alcançou um valor de *F1-Score* de 67,8%. Já quando é utilizado o léxico com as abordagens de AM, os melhores resultados foram alcançados, tendo como destaque SVM + Léxico com 80% de *F1-Score*. Essa comparação indica que os métodos supervisionados tiveram um desempenho superior em comparação ao uso único do vocabulário utilizado na classificação de *tweets* relacionados ao MFB, como também visto em Januário *et al.* (2021) e Das *et al.* (2022). Além disso, ao incluir informações adicionais do vocabulário como parte do treinamento, a abordagem supervisionada resulta em ganhos de desempenho.

Na abordagem lexical, a variação dos resultados está ligada à formulação do léxico utilizado e seu domínio. Exemplos disso são o estudo de Jung *et al.* (2021), que cobriu 41% dos termos de vocabulário conhecidos em triagens de câncer de mama. Já Wang *et al.* (2020) obteve 69,6% de acurácia na análise de sentimentos de comentários de filmes. Resultados semelhantes ocorrem no contexto financeiro, como acurácia de 70% em publicações sobre o sistema financeiro americano Das *et al.* (2022) e a pontuação *F1-Score* de 58,2% Januário *et al.* (2021).

Conclusão

Este artigo comparou distintas abordagens para a criação e expansão automática de léxicos em língua portuguesa, focando a aplicação ao cenário do Mercado Financeiro Brasileiro, que apresenta poucos estudos relacionando tanto a língua portuguesa quanto o uso desses conjuntos de palavras especializados em tarefas de suporte na tomada de decisão por meio da análise de mensagens (Pereira, 2021).

Os resultados alcançados destacaram um desempenho promissor na avaliação de sentimentos presentes em *tweets* e notícias relacionadas ao mercado, o que potencialmente poderia oferecer informações valiosos para a orientação de decisões e a análise do panorama desse contexto.

Foram apresentadas três abordagens de construção lexical com variações de pré-processamento, resultando em seis configurações finais para léxicos no contexto do Mercado Financeiro Brasileiro. Os experimentos abrangeram análise de sentimentos em mensagens curtas, como *tweets* relacionados ao mercado brasileiro, e em textos maiores, como notícias do mesmo domínio. A configuração S+S/A+PMI (com lematização) obteve o melhor desempenho, alcançando um *F1-Score* de 71,5% para a classificação de *tweets* e 68,4% para notícias, superando o baseline para notícias (JANUÁRIO *et al.*, 2021). Além disso, a abordagem lexical, combinada com o modelo *Support Vector Machine*, alcançou um *F1-Score* de 80%.

Dessa forma, o método proposto permite a criação de léxicos personalizados que podem ser ajustados de acordo com o contexto temporal dos dados, se adaptando às variações nas nuances da linguagem ao longo do tempo. Essa flexibilidade é particularmente relevante, pois permite a criação de léxicos específicos para diferentes áreas ou períodos, refletindo melhor as variações na linguagem utilizada. No contexto do mercado financeiro, por exemplo, isso possibilita uma análise de sentimentos mais precisa e dinâmica, levando em consideração as atualizações do mercado e os eventos que possam afetar as decisões dos investidores. Esse tipo de abordagem também pode ser aplicado em outras áreas, como a análise de tendências em redes sociais ou o monitoramento de crises, em que a atualização constante do vocabulário é crucial para uma compreensão eficaz das mensagens e sentimentos em evolução.

Referências

BIRD, S. *NLTK*: The Natural Language Toolkit. Barcelona: Association for Computational Linguistics, 2006.

BOS, T.; FRASINCAR, F. Automatically building financial sentiment lexicons while accounting for negation. *Cognitive Computation*, [*s. l.*], v. 14, p. 442-460, 2022.

CAROSIA, A. E.; COELHO, G. P.; SILVA, A. E. Analyzing the brazilian financial market through portuguese sentiment analysis in social media. *Applied Artificial Intelligence*, London, v. 34, p. 1-19, 2020.

DAS, S. R.; DONINI, M.; ZAFAR, M. B.; HE, J.; KENTHAPADI, K. Finlex: An effective use of word embeddings for financial lexicon generation. *The Journal of Finance and Data Science*, Elsevier, v. 8, p. 1-11, 2022.

GARDNER, J. R.; KUSNER, M. J.; XU, Z. E.; WEINBERGER, K. Q.; CUNNINGHAM, J. P. Bayesian optimization with inequality constraints. *ICML*, [*s. l.*], p. 937-945, 2014.

HARTMANN, N.; FONSECA, E.; SHULBY, C.; TREVISO, M.; RODRIGUES, J.; ALUISIO, S. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *Proceedings of Symposium in Information and Human Language Technology*, Uberlândia, p. 122-131, oct. 2017.

JANUÁRIO, B. A.; CAROSIA, A. E. d. O.; SILVA, A. E. A. da; COELHO, G. P. Sentiment analysis applied to news from the brazilian stock market. *IEEE Latin America Transactions*, [*s. l.*], v. 20, n. 3, p. 512-518, 2021.

JUNG, E.; JAIN, H.; SINHA, A. P.; GAUDIOSO, C. Building a specialized lexicon for breast cancer clinical trial subject eligibility analysis. *Health Informatics Journal*, [*s. l.*], v. 27, 2021.

LOSADA, D. E.; GAMALLO, P. Evaluating and improving lexical resources for detecting signs of depression in text. *Language Resources and Evaluation*, [*s. l.*], v. 54, p. 1-24, 2020.

LOUGHRAN, T.; MCDONALD, B. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance*, [*s. l.*], v. 66, p. 35-65, 2011.

MAHMOOD, A. T.; KAMARUDDIN, S. S.; NASER, R. K.; NADZIR, M. M. A combination of lexicon and machine learning approaches for sentiment analysis on facebook. *Journal of System and Management Sciences*, Pequim, v. 10, p. 140-150, 2020.

OLIVEIRA, N.; CORTEZ, P.; AREAL, N. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, Elsevier, v. 85, p. 62-73, 2016.

PEREIRA, D. A. A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, [*s. l.*], v. 54, p. 1087-1115, 2021.

SHAN, R.; JIANG, T.; WANG, Y. Research on the construction of domain sentiment lexicon based on label propagation algorithm. *ACM International Conference Proceeding Series*, [*s. l.*], p. 1024-1029, 2021.

SMYWIŃSKI-POHL, A. *et al*. Automatic construction of a polish legal dictionary with mappings to extra-legal terms established via word embeddings. *ICAIL*, [*s. l.*], p. 234-238, 2019.

SNOEK, J.; LAROCHELLE, H.; ADAMS, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems*, [*s. l.*], v. 25, p. 2951-2959, 2012.

WANG, Y. *et al*. Automatic construction of domain sentiment lexicon for semantic disambiguation. *Multimedia Tools and Applications*, [*s. l.*], v. 79, n. 31-32, p. 22355-22373, 2020.

tecnia revista de educação, ciência e tecnologia do ifg

v. 10, Edição Especial 1 | 2025

Submetido 31/05/2024. Aprovado 14/03/2025
Avaliação: revisão duplo-anônimo

Fake news: a brief tertiary review through health, deep learning, and emerging perspectives

NOTÍCIAS FALSAS: UMA REVISÃO TERCIÁRIA RÁPIDA DAS PERSPECTIVAS DE SAÚDE, APRENDIZADO PROFUNDO E EMERGENTES.

NOTICIAS FALSAS: UN REVISIÓN TERCIARIA RÁPIDA A TRAVÉS DE LA SALUD, EL APRENDIZAJE PROFUNDO Y LAS PERSPECTIVAS EMERGENTES.

Juliana R. S. Gomes
Universidade Federal de Goiás (UFG)
juliana.resplande@discente.ufg.br

Valdemar Vicente Graciano Neto
Universidade Federal de Goiás (UFG)
valdemar@inf.ufg.br

Jacson Rodrigues Barbosa
Universidade Federal de Goiás (UFG)
jacson@inf.ufg.br

Eliomar Araújo de Lima
Universidade Federal de Goiás (UFG)
eliomar@inf.ufg.br

Arlindo Rodrigues Galvão Filho
Universidade Federal de Goiás (UFG)
arlindo@inf.ufg.br

Abstract

Context: The proliferation of fake news represents a significant social threat, especially regarding health information, a problem exacerbated by the COVID-19 pandemic. Deep Learning (DL) techniques are central to detection efforts, with increasing focus on health-related misinformation. **Objective:** This paper extends our previous work, synthesizing secondary studies (SS) on fake news detection, focusing on DL roles, the health domain, and recent trends (2022-2023). **Method:** A rapid tertiary review was conducted, analyzing 15 SS published between 2013 and August 2023, categorized by emphasis: DL applications, health misinformation, or recent publications. **Results:** A consistent dependence on DL and Natural Language Processing for text classification and fabricated media detection was identified. Health-focused or recent trend studies addressed challenges using specific datasets. Key challenges include echo chambers, cross-domain applications, early detection needs, and threats from generative models. Demands for transparency, blocking mechanisms, and Explainable Artificial Intelligence were highlighted. **Conclusion:** This review provides a synthesized view of research on fake news detection, emphasizing intersections with DL and health contexts, confirming the prevalence of core techniques despite diverse methodologies, and pointing to challenges requiring urgent attention.

Keywords: fake news; tertiary review; health; deep learning.

Resumo

Contexto: A proliferação de notícias falsas representa uma ameaça social significativa, especialmente em informações de saúde, problema agravado pela pandemia de Covid-19. Técnicas de Aprendizado Profundo (DL) são centrais nos esforços de detecção, com foco crescente em desinformação relacionada à saúde. **Objetivo:** Este artigo estende o trabalho anterior dos autores, sintetizando estudos secundários (ES) sobre detecção de notícias falsas, focando nos papéis do DL, no domínio da saúde e tendências recentes (2022-2023). **Método:** Foi realizada uma revisão terciária rápida analisando 15 ES publicados entre 2013 e agosto de 2023, categorizados por ênfase: aplicações de DL, desinformação em saúde ou publicação recente. **Resultados:** Identificou-se dependência consistente em DL e Processamento de Linguagem Natural para classificação de texto e detecção de mídia fabricada. Estudos em saúde ou tendências recentes abordaram desafios usando conjuntos de dados específicos. Principais desafios incluem câmaras de eco, aplicações interdomínio, necessidade de detecção precoce e ameaças de modelos generativos. Demandas por transparência, mecanismos de bloqueio e Inteligência Artificial Explicável foram destacadas. **Conclusão:** Esta revisão fornece uma visão sintetizada da pesquisa em detecção de notícias falsas, enfa- tizando interseções com DL e contextos de saúde, confirmando a prevalência de técnicas centrais apesar de metodologias diversas, e apontando desafios que requerem atenção urgente.

Palavras-chave: notícias falsas; revisão terciária; saúde; aprendizado profundo.

Resumen

Contexto: La proliferación de noticias falsas representa una amenaza social significativa, especialmente en información de salud, problema exacerbado por la pandemia de Covid-19. Las técnicas de Aprendizaje Profundo (DL) son centrales en los esfuerzos de detección, con enfoque creciente en la desinformación relacionada con la salud. **Objetivo:** Este artículo extiende el trabajo previo de los autores, sintetizando estudios secundarios (ES) sobre detección de noticias falsas, centrándose en los roles del DL, el dominio de la salud y tendencias recientes (2022-2023). **Método:** Se realizó una revisión terciaria rápida anal- izando 15 ES publicados entre 2013 y agosto de 2023, categorizados por énfasis: aplicaciones de DL, desinformación en salud o publicación reciente. **Resultados:** Se identificó una dependencia consistente en DL y Procesamiento de Lenguaje Natural para clasificación de textos y detección de medios fabrica- dos. Estudios enfocados en salud o tendencias recientes abordaron desafíos usando conjuntos de datos específicos. Los principales desafíos incluyen cámaras de eco, aplicaciones entre dominios, necesidad de detección temprana y amenazas de modelos generativos. Se destacaron demandas de transparencia, mecanismos de bloqueo e Inteligencia Artificial Explicable. **Conclusión:** Esta revisión proporciona una visión sintetizada de la investigación en detección de noticias falsas, enfatizando intersecciones con DL y contextos de salud, confirmando la prevalencia de técnicas centrales a pesar de metodologías diversas, y señalando desafíos que requieren atención urgente.

Palabras clave: noticias falsas; revisión terciaria; salud; aprendizaje profundo.

Introduction

The post-2015 period has witnessed an unprecedented use of social media, an information ecosystem often lacking the quality criteria associated with traditional journalism (Aimeur; Amri; Brassard, 2023). However, this digital landscape has also provided a fertile ground for the proliferation of fake news, a phenomenon that transcends mere misinformation. Fake news has evolved into a powerful tool of manipulation, capable of inflicting damage on the reputations of corporations, governments, and ethnic groups (Meel; Vishwakarma, 2020; Schlicht *et. al.*, 2023).

Concurrently, deep Learning has emerged as a popular technique since the 2010s. It bypasses manual handcrafting features, which are a laborious and time-consuming but necessary part of traditional machine learning approaches. Thus, deep Learning allows performing complex tasks, such as computer vision, speech recognition, health care monitoring, etc. (Islam *et. al.*, 2020a; Hangloo; Arora, 2022).

Additionally, concerns about health misinformation are significant. The internet has become a widely used and accessible source for health information, often serving as an initial resource for medical advice before consulting a doctor (Schlicht *et. al.*, 2023a; Wang *et. al.*, 2019). As the internet and social media enable diffuse health-related information, they also lower the cost of generating fake news, which started in the pre-COVID-2019 era. For instance, the anti-vaxxer movement, by encouraging individuals not to vaccinate their children, contributed to measles outbreaks in the UK, the US, Germany, and Italy in 2017 (wang, 2017).

This issue increased during the Covid-19 pandemic, in which quarantines increased internet usage (Varma *et. al.*, 2021). The intense flood of real and fake information about Covid-19 through all sources, including social media, was coined as "information epidemic" or "infodemic" (Kim *et. al.*, 2021; Schlicht *et. al.*, 2023). Health authorities even announced that preventing the creation and propagation of fake news about the virus is as essential as alleviating the contagious power of Covid-19 (Kim *et. al.*, 2021).

In response to this challenge, there has been a remarkable growth in the interest in the field, as evidenced by the volume of research with a diverse array of technologies (Meel; Vishwakarma, 2020). In the context of that research area, secondary studies (such as surveys, systematic mapping, or systematic research) have synthesized primary studies and approached the topic under multiple perspectives (Petersen; Vakkalanka; Kuzniarz, 2015), which has generated numerous publications.

This research expands on our previous publication, 'A Rapid Tertiary Review at the Fake News Domain (Gomes *et. al.*, 2023), presented at the XI Escola Regional de Informática de Goiás. In that initial work, we proposed the methodology of rapid tertiary research within the fake news domain, characterizing it as a tertiary review conducted according to rapid review (RR) protocols (Cartaxo; Pinto; Soares, 2020). The goal of RRs is to accelerate evidence synthesis compared to traditional systematic reviews, thereby delivering timely insights.

The present study delves deeper by investigating specific dimensions, namely health-related disinformation, the role of deep learning, and recent advancements from 2022 to 2023. We observe a divergence in methodologies used by researchers focusing on different sub-topics ('interest groups'). Despite this methodological variety, the research outcomes show considerable similarity, often concentrating on deep Learning solutions and various facets of the information lifecycle.

Nevertheless, research addressing the primary challenges highlighted in our original review (Gomes *et. al.*, 2023) places significant emphasis on health topics and the most recent literature. Furthermore, our analysis indicates that the development of detailed taxonomies is more common in broader secondary studies than in those focused on specific research niches.

This paper is structured as follows. Section 2 covers the related work and fake news definition; Section 3 presents the research method; Section 4 presents data extraction and results reporting; and Section 5 concludes the paper and points to future work.

Background and related work

This section defines fake news and its related synonyms using (Sharma *et. al.*, 2019). We delineate the main technologies and their taxonomy based on (Hangloo; Arora, 2022) and provide a brief overview of related work.

Definition

Fake news can be defined as fabricated content that mimics real news (Wu *et. al.*, 2022). It is important to note that "fake news" lacks a universally accepted definition, and its interpretation can vary widely (Aimeur; Amri; Brassard, 2023).

Sharma *et. al.* (2019) mention that fake news is news or messages published and spread through the media, containing false information, regardless of the means and motivations that led to its dissemination. This definition allows for capturing the different types of fake news identified in various studies. It allows distinguishing them as fabricated content (completely false), deceptive content (deceptive use of information to frame an issue), imposter content (genuine sources represented by false sources), manipulated con- tent (genuine information or images manipulated to deceive), false connection (headlines, visuals, or captions that do not support the content), and false context (genuine content shared with false contextual information) (Sharma *et. al.*, 2019).

It is possible to subdivide fake information by intention, such as misinformation and disinformation. The former refers to the involuntary dissemination of false information that may result from distortion or misunderstanding due to cognitive biases or lack of comprehension or attention, while the latter refers to false information created and explicitly disseminated to deceive (Aimeur; Amri; Brassard, 2023; sharma *et. al.*, 2019).

Technologies

The main technologies related to fake news are also varied, concentrating in terms of artificial intelligence (AI), natural language processing (NLP), fact-checking (AFC), crowdsourcing (CDS), blockchain (BKC), and graph neural networks (GNN) (Aimeur; Amri; Brassard, 2023).

Artificial Intelligence (AI)

Techniques mainly employ machine learning (ML) or deep learning (DL). ML techniques refer to classical methods in which features are manually extracted for the AI model, such as the number of words or the length of sentences. Methods guided by DL often have the training features extracted. automatically; on the other hand, they are methods that require greater computational capacity (Hangloo; Arora, 2022).

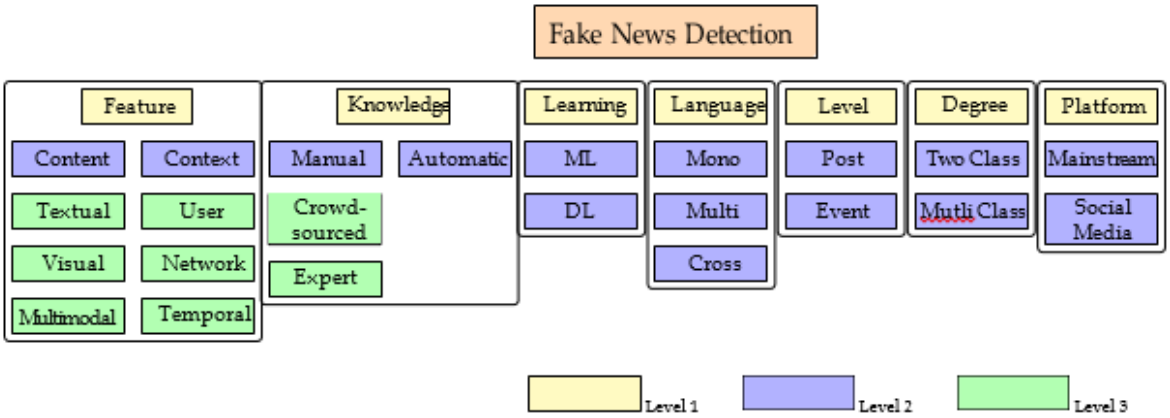


Figure 1. Taxonomy of fake news detection techniques
Source: Hangloo; Arora, 2022.

Natural Language Processing (NLP) techniques are text-related methods. Fake news can be examined by its textual content, for instance, syntactic, lexical, psycholinguistic, and semantic. In the syntactic domain, the use of word categories such as nouns, verbs, and adjectives is examined using the Pos-tagging technique. In the syntactic domain, various characteristics such as the number of unique words and their frequencies, the number of sentences, and grammatical errors are extracted to detect suspicious texts. Psycholinguistic information can be obtained using the LIWC statistical system (Reis *et. al.*, 2019). Traditional ML classifiers, such as Logistic Regression, can be paired with NLP techniques such as term frequency-inverse document frequency (TF-IDF) for text analysis.

Fact-checking (FC) techniques verify the textual news through external information, for example, by trying to counter-evidence from a reliable source, which can be carried out manually or automatically. When done manually, the technique can be carried out by **Crowd-sourcing (CDS)** or by experts in the field (Ali *et. al.*, 2022; Kondamudi *et. al.*, 2023; Hangloo; Arora, 2022; Aimeur; Amri; Brassard, 2023).

Blockchain (BKC) is a decentralized ledger that operates on an append-only basis, meaning data is stored across multiple servers without a central authority and can only be added, not deleted or altered. The security of this system is ensured through the verification and cryptographic sealing of blockchain units (blocks) during insertion (Dhall *et. al.*, 2021). Features such as immutability, decentralization, tamper resistance, consensus, record-keeping, and the non-repudiation of transactions are vital aspects that render blockchain technology valuable, not only for cryptocurrencies but also for verifying the authenticity and integrity of digital assets (Ahmad; Aliaga Lazarte; Mirjalili, 2022).

These technologies can be grouped according to their purpose (Ali *et. al.*, 2022; Kondamudi *et. al.*, 2023; Hangloo; Arora, 2022). In this article, we follow (Hangloo; Arora, 2022) as shown in Figure 1, dividing the taxonomy into four main categories: Feature-based (containing Content-based or Social context-based), Knowledge-Based, and Learning-based, with three additional aspects such as language level, detection level, degree of fakeness, and platform.

Feature-based approaches examine characteristics or patterns to detect fake news. For example, sentiment analysis techniques search for a negative sentiment bias in the text. These techniques can be subdivided into content-based or context-based (Hangloo; Arora, 2022).

Content-based technologies analyze the media present in the news, which can be textual, visual, or both (called multimodal). Textual content can be extracted from

video and audio through automatic speech recognition systems (ASR), also known as *Speech-to-text*, and from images and videos through OCR.

Context-based techniques, on the other hand, examine the context of the en- environment in which the news is inserted, such as the reactions expressed in comments, responses, and reports, as well as user behavior, such as the time between posts and checking whether the user's behavior resembles that of social network boots. User and posting interactions create a network architecture that can be analyzed using techniques such as blockchain and graphs. It also includes methods considering temporal data on how rumors resurface and credibility data on the news source (Aimeur; Amri; Brassard, 2023; Ali *et. al.*, 2022; Kondamudi *et. al.*, 2023).

Knowledge-based techniques relate to fact-checking techniques, which involve cross-referencing claims with information from trusted sources. As explained previously, fact-checking can be divided into Automatic Fact-Checking and Manual Fact-checking, which can be done through crowdsourcing or expert-based (Ali *et. al.*, 2022; Kondamudi *et. al.*, 2023; Hangloo; Arora, 2022).

Learning-based techniques denote artificial intelligence methods, typically delineated as machine learning (ML) or deep learning (DL) approaches. ML-guided methods are characterized by simplicity, facilitating users' interpretation of intermediate steps, and necessitating minimal data for training. However, they involve manual manipulation of input characteristics, commonly referred to as *feature engineering*. Conversely, DL-guided methods demand a larger volume of data and are often perceived as black boxes due to their complexity and the challenges associated with interpreting intermediate results. Nevertheless, they require less manual data processing and, when a substantial amount of data is available, produce outcomes that exceed those achieved by machine learning (ML) techniques (Garg; Khan; Alam, 2020).

When it comes to **language**, the detection mechanism has the option to consider either one language or multiple languages, whether it be monolingual or multilingual. This is because news that pertains to global events tends to be spread in more than one language and can benefit from being classified across different languages. Another mentioned technique is cross-language learning, where a model trained in a language with rich data, such as English, is used to detect fake news in other languages like Portuguese (Hangloo; Arora, 2022).

Regarding the **detection level**, classification can be performed *a posteriori*, that is, after news circulation. In this case, the claim is analyzed based on the rumor that was spread. Most detection methods are a *posteriori*, not considering this level of related events (Hangloo; Arora, 2022).

Concerning the **degree of fakeness**, a news item can be verified as true or false, with two possible classes, or true, false, and *half false*, with multiple possible classes (Hangloo; Arora, 2022). Liar, the dataset that stands out among the most mentioned in the secondary studies analyzed, has six truth scales: false, not very true, half true, almost true, and true.

From the **platform** perspective, fake news can be analyzed on a specific social network or in the mainstream news media. Facebook, YouTube, WhatsApp, and Instagram lead the way in terms of global user numbers (Dixon, 2024). As we will point out, the population tends to stay in information bubbles and believe the news and journalistic sources that most closely align with the user's worldview. Therefore, the greatest difficulty would be to point out *fake news* associated with the environment in which the individual is inserted on the social network and in the mainstream media.

Related work. While there is a lack of dedicated tertiary studies specifically focused on the phenomenon of fake news, the closest existing tertiary research pertains to

sentiment analysis (Ligthart; Catal; Tekinerdogan, 2021). Notably, sentiment analysis techniques hold relevance in the context of fake news detection, as they can be harnessed to identify emotional cues, biases, and hate speech, all of which may indicate the presence of fake news. The findings from this research indicate several key trends and challenges in sentiment analysis, including a growing preference for complex Deep Learning techniques capable of detecting intricate patterns, the necessity for adapting techniques to different domains, and the persistent challenges associated with domain and language dependencies.

Research Method

In this section, we describe our approach to advancing current knowledge and practices in the field of fake news by examining it from three distinct perspectives: deep learning (DL), health, and general. We also compare it with the latest studies from the years 2022 and 2023.

Research questions. To guide our investigation through various phases of the fake news lifecycle and different research angles, we pose the following research questions (RQ):

RQ1: How do different perspectives within peer-reviewed literature define fake news?
RQ2: What techniques, tools, and methods are documented across different perspectives?
RQ3: What are the primary challenges faced in each perspective when addressing fake news?

Search, selection, and extraction. We utilize the search mechanism and studies se- selection from the rapid review of (Gomes *et. al.*, 2023), identifying 15 pertinent secondary studies from the top 50 results of a Google Scholar[1] search conducted on August 29, 2023.

The search string was created using population and intervention (PI). The population is the study area, which comprises fake news, its synonyms, and the intervention is the method intended to be applied in the population, which are secondary studies. We employ the following query:

("fake news" OR misinformation OR rumor OR disinformation) AND
("systematic review" OR "systematic mapping" OR "literature review" OR "survey").

Table 2 presents a comprehensive list of the studies that were identified for review. A total of 13 articles were considered in this study. Of these, a subset of 10 articles was selected based on specific criteria, as detailed in Table 1. Each article selected for inclusion is associated with deep learning (DL), pertains to health, or was published between 2022 and 2023. The subset in question comprises a total of six articles published between the years 2022 and 2023, of which four are specifically oriented towards deep learning (DL) and three pertain to health-related applications.

1 https://scholar.google.com.br/

Reference	Deep Learning	Health	2022-2023
(S1)		✓	✓
(S2)			✓
(S3)			✓
(S5)	✓		
(S6)	✓		
(S9)			✓
(S10)	✓		✓
(S12)		✓	✓
(S13)	✓	✓	

Table 1 - Subset from selected studies that are focused on DL or on health or are from 2022 and 2023
Source: Elaborated by the author.

There are instances of overlapping research areas: (S13) is the only study from the health and DL area; (Ahmad; Aliaga Lazarte; Mirjalili, 2022b; Schlicht *et. al.*, 2023) are both related to the health sector and date from 2022 to 2023, while (S10) focuses on DL technologies within the health field. Nevertheless, none encompasses all three characteristics.

Among the 15 results, 8 (53.3%) were systematic, with an average of 65.5 primary studies being analyzed. (Kim et. *al.*, 2021b; kondamudi *et. al.*, 2023b) forms an outlier with 182 and 218 examined works, leaving these out, the average drops to 49.8 studies.

All three health studies were systematic, forming an average of 64.33 primary studies analyzed. In addition, all three 2023 studies were systematic, making up an average of 107.3 considered research. In the realm of Deep Learning, only (S13) is systematic, which is also health research.

Figure 2 visually represents the citation connections between the chosen secondary studies. It should be noted that the studies by (S10) and (S1) are not cited or cited by the other selected articles.

Within each interest group listed in Table 1, there are no direct or indirect citations between the works. The selected health studies (S1; S12; S13) do not reference each other.

Reference	Year	Sys.	No.	Title
(S1)	2022	Yes	80	A Systematic Literature Review on Fake News in the Covid-19 Pandemic: Can AI Propose a Solution?
(S2)	2023	Yes	61	Fake news, disinformation, and misinformation in social media: a review
(S3)	2022	No		Fake News Detection Techniques on social media: A Survey
(S4)	2020	Yes	35	Approaches to Identify Fake News: A Systematic Literature Review
(S5)	2022	No		Deep learning for fake news detection: A comprehensive survey
(S6)	2020	No		Deep learning for misinformation detection on online social networks: a survey and new perspectives
(S7)	2021	Yes	10	A systematic literature review on disinformation: Towards a unified taxonomical framework
(S8)	2021	Yes	182	A systematic review on fake news research through the lens of news creation and consumption: Research efforts, challenges, and future directions

(S9)	**2023**	Yes	218	A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches
(S10)	**2022**	No		A Brief Survey for Fake News Detection via **Deep Learning Models**
(S11)	2020	No		Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-art, challenges and opportunities
(S12)	**2023**	Yes	43	Automatic detection of **health** misinformation: a systematic review
(S13)	2021	Yes	70	A systematic survey on **deep learning** and machine learning approaches of fake news detection in the pre- and post-**Covid-19** pandemic
(S14)	2021	No		Minimizing the spread of misinformation in online social networks: A survey
(S15)	2020	No		A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities

Table 2 - Selected secondary studies and bolded related keywords: deep learning, health. Recent studies, from 2022 and 2023, have the related year bolded in blue. For systematic secondary studies (Sys.), we provide the number of primary studies (No.)
Source: Elaborated by the author.

Similarly, DL articles (S5; S6; S10; S13) do not cite one another. This pattern holds for studies from 2022 and 2023. Only the articles (S3; S9), published in 2022 and 2023, respectively, extend their bibliographies to include works from 2021, in addition to those from 2020, either directly or indirectly.

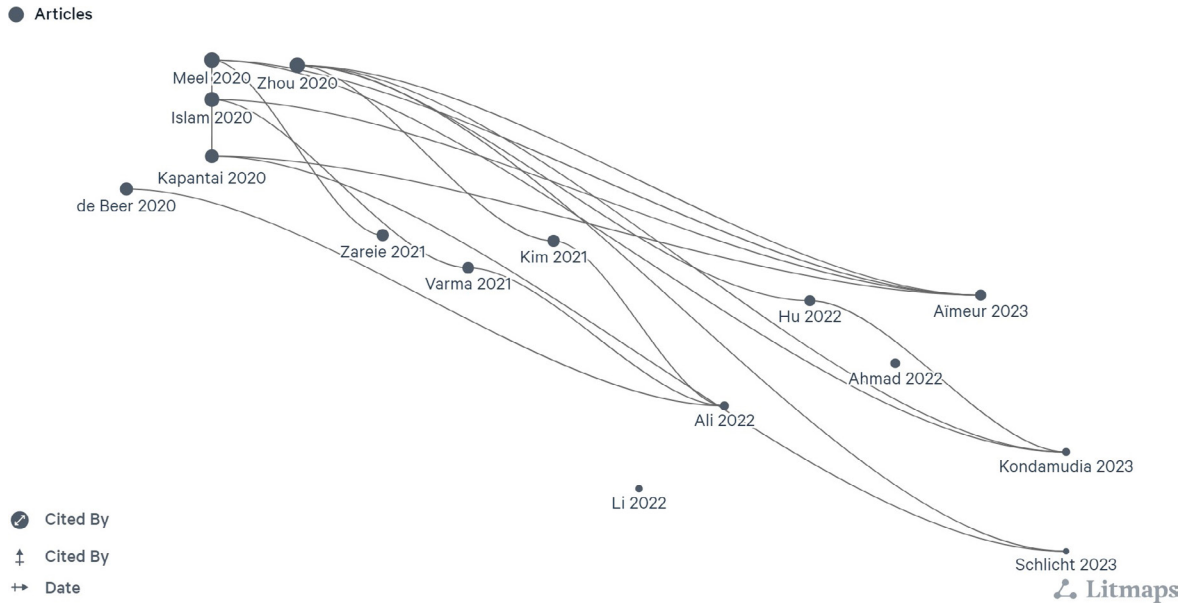


Figure 2. Visualization of secondary studies, their citation counts, and the temporal citation relationships. Each data point represents a study, with the vertical position indicating the number of citations received, and the horizontal position indicating the publication date. The lines connecting the data points represent the citations between the studies
Source: Elaborated by the author using Litmaps (LTD, 2023).

Research questions (RQs) reported in the secondary studies were also extracted, as shown in Table 3. Five works explicitly stated their research questions (S1; S2; S6; S7; S10; S13). Most of the studies included similar research questions to those raised herein. In the deep learning (DL) domain, most studies identify specific RQs, except for (S5). For each category of RQ (Definition, Techniques, Challenges, and Research Methods), there is one work from DL area and one work from 2022-2023.

In the health sector, most research papers state their research questions, except for the work by (S12), which is an outlier. This observation extends to recent studies from 2022-2023, where only (S3; S9) do not explicitly state their RQs.

Most secondary studies have focused on RQs about technologies, tools, and strategies for detecting fake news (S1; S2; S6; S10; S13): three works were published in 2022- 2023, three related to DL techniques, and two related to health-related studies. Secondly, the RQs affiliated with the primary challenges and future research are stated by three works, two from 2022-2023 (one from the health sector) and another from DL.

Concern with definition is that it focuses; no secondary studies selected from 2022 to 2023 explicitly identified the definition as a specific research question. The elements pertain to the definition of the concept in question (Table 3). Of relevance are two aspects, namely, “health” and “deep learning.” In the context of health and deep learning, the S10 study was distinguished by its incorporation of an additional query, a feature that is not present in another research (S10). The inquiry into the employed research methodologies within the extant literature is of paramount importance.

Research question	Secondary study	Intereset subset
Definition	(S2) (S6) (S7)	2022-2023 DL-
Techniques, tools, and methods	(S1) (S2) (S6) (S10) (S13)	**Health, 2022-2023** 2022-2023 DL **DL, 2022-2023** **DL, Health**
Main challenges/Future research	(S1) (S2) (S6)	**Health, 2022-2023** 2022-2023 DL
Research Methods	(S10)	**DL, 2022-2023**

Table 3 - Research questions in reported secondary studies
Source: Elaborated by the author.

In terms of data extraction, from the variables in (Gomes *et. al.*, 2023), we make a view from the studies between the considered domains: General, Health, Deep Learning, and the most recent studies (2022-2023). The variables extracted for each study are listed below:

V1 Year
V2 Complete reference
V3 Research questions of the study
V4 Is it systematic?
V5 If it is systematic, the number of primary studies
V6 Focus of the study
V7 Aspects of the ecosystem and information lifecycle covered
V8 Tasks covered
V9 Covered techniques
V10 Architectural solutions (and technologies) mentioned
V11 Mentioned techniques, tools, and methods
V12 Public datasets
V13 Public models

V14 Explicit research gaps
V15 Implicit research gaps

Data Analysis

The subsequent section is an exposition of the results of the investigation into the research questions of this tertiary study, with an examination of the research questions from three angles: deep learning (DL), health, and general. Comparisons have also been made with studies from 2022 and 2023.

RQ1: HOW DO DIFFERENT PERSPECTIVES WITHIN PEER-REVIEWED LITERATURE DEFINE FAKE NEWS?

Generalist fake news studies tend to define fake news approaches according to its technology and content, as explained in Section 2 (S2; S3; S4; S9; S11; S15).

As noted in (Gomes *et. al.*, 2023), fake news issues are often examined in the literature regarding health and deep learning techniques.

Within the domain of deep learning (DL), (S5) classifies fake news based on characteristics such as news content, social context, and external knowledge, categorizing DL techniques into supervised, weakly supervised, and unsupervised methods. (S13) Assess DL techniques in terms of the pre- and post-COVID-19 pandemic.

After the pandemic, research on detecting fake news related to health took off. (S1) and (S13) are two studies on this. (S1) looks at misinformation about COVD-19 on social media. (S13) uses deep learning techniques to compare pre- and post-pandemic scenarios. (S12) looks at health misinformation more broadly. It identifies analogies and differences between COVD-19 and other health datasets.

(S11; S7; S12; S14) research on taxonomy. Three of them (S11; S7; S14) are out of the interest group in Table 1, making up half of the six studies that are not from 2022-23 and do not focus on health or DL. (S12) is the only paper from the interest group, a health and recent study, that searches upon taxonomy.

As noted in (Gomes *et. al.*, 2023), the fake news information cycle is distributed in key phases: Propagation (37%), Creation (33.3%), and Consumption (29.7%). The distribution of fake news information cycle is like interest subsets: Health, DL, and 2022-2023. However, consumption is explored more in secondary work related to health.

Table 3 illustrates the fake news task mentioned in the secondary work. Studies are divided into four groups: all, Health, DL, and recent (2022-2023 publication year). The distribution is similar between all groups. The most cited tasks are textual classification, followed by fabricated media detection. However, technologies to deal with the detection of fabricated media are rarely mentioned (S11). Intervention and prevention of fake news are less frequently discussed in the literature, with no mentions in the Health, DL, and recent subgroups.

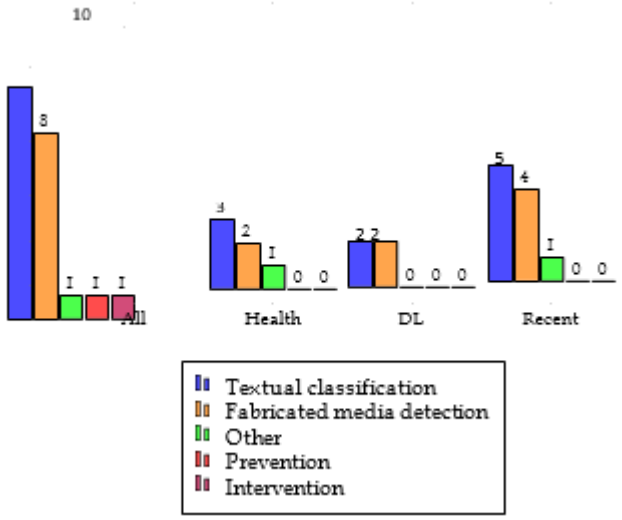


Figure 3. Fake news task stage covered
Source: Elaborated by the author.

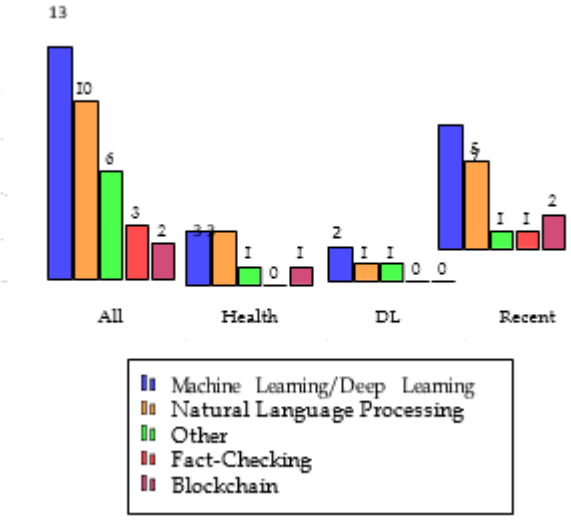


Figure 4. Mentioned technologies
Source: Elaborated by the author.

RQ2: WHAT TECHNIQUES, TOOLS, AND METHODS ARE DOCUMENTED ACROSS DIFFERENT PERSPECTIVES?

The technologies mentioned are listed in Table 4[2], which is divided into four groups: Health, DL, and recent (published between 2022 and 2023). Machine Learning/Deep Learning has been cited most frequently, followed by Natural Language Processing in all its subgroups. (S12) reports that 33% of medical studies used Deep Learning, 28% used Machine Learning, and 37% used a combination of the two. In studies involving deep learning methods, pretrained Transformers generally surpassed other models in performance, while ensemble techniques such as random forests were more effective than most conventional machine learning approaches.

2 In this work, we consider Crowdsourcing as Fact-checking.

Dataset	Mentioned by	Size	Source
FakeNewsNet (SHU *et. al.*, 2018)	(S2; S3; S5; S6; S11; S10; S15)	23k Statements'	GossipCop Polifact
LIAR (WANG, W. Y., 2017)	(S2; S5; S3; S6; S11; S15)	12,8k Statements'	Polifact
CREDBANK (MITRA; GILBERT, 2015)	(S3; S5; S6; S11; S15)	6M Tweets'	Twitter/X
FacebookHoax (TACCHINI *et. al.*, 2017)	(S3; S5; S6; S11; S15)	15,5 Posts	Facebook

Table 4 - Most mentioned public datasets by secondary studies
Source: Elaborated by the author.

The most frequently referenced public datasets, as indicated in Table 4, include FakeNewsNet, which is cited in 7 secondary studies, LIAR, noted in 6 studies, Credbank, mentioned in 5 studies, and FacebookHoax, also cited 5 times. Credbank, FakeNewsNet, LIAR, and FacebookHoax were published between 2015 and 2018.

Fake News Net contains 23K Gossip Cop (About […], c2019) and Politi Fact (Polifact, c2020) fact-checking websites (Shu *et. al.*, 2020). LIar composes 12.8K human-labeled short statements from the fact- checking website PolitiFact (Wang, 2017). Each news is labeled with six-grade truthfulness: true, false, half-true, part-true, barely--true, and mostly-true. Credbank is a large crowd-sourced dataset of 6M tweets over 96 days starting from October 2015 (Mitra; Gilbert, 2015). FacebookHoax contains information about posts on Facebook pages associated with scientific news (non-hoax) and conspiracy pages (hoax), gathered using the Facebook API. The data collection includes 15.5K postings from 32 pages (Tacchini *et. al.*, 2017).

(S5; S6; S11; S15) cite all four datasets, CREDBANK, FakeNewsNet, LIAR, and FacebookHoax. (S5; S6) are DL-focused works, and only (S5) is an article from 2022 and 2023.

It is worth noting that only (S12; S5) list six datasets more recent than 2018, all related to Covid-19. Among these datasets, four are in English and have more than 100 citations as of October 10th, 2023: CoAID, FakeCovid, FakeHealth, and ReCOVery (Cui; Lee, 2020; Shahi; Nandini, 2020; Dai; Sun; Wang, 2020; Zhou *et. al.*, 2020).

Dataset	Mentioned by	Size	Source
CoAID (Cui; Lee, 2020)	(S12; S5)	5K News 297K interactions 1K posts	Multiple media outlets
FakeCovid (Shahi; Nandini, 2020)	(S12; S5)	7.6K news articles	Poynter
FakeHealth (Dai; Sun; Wang, 2020)	(S12; S5)	6M Tweets'	HealthNewsReviews
ReCOVery (Zhou; Mulay, *et. al.*, 2020)	(S12; S5)	2K news articles	NewsGuard21

Table 5 - Health mentioned public datasets
Source: Elaborated by the author.

CoAID, which stands for "Covid-19 Healthcare Misinformation Dataset", includes 5,216 news articles, 296,752 related user engagements, and 958 posts on social networks (Cui; Lee, 2020). FakeCovid consists of 7,623 fact-checked news articles from 105 countries, sourced from Poynter-listed fact-checkers (The Corona […], 2025)

and Snopes ([2025?]) , covering Covid-19 and collected between April 1st, 2020, and January 7th, 2020 (Shahi; Nandini, 2020).

FakeHealth encompasses two datasets: HealthStory and HealthRelease, representing- ing news stories and news releases data from HealthNewsReview (c2022). HealthStory contains 1.69K news articles, while Health Release contains 606 news articles (Dai; Sun; Wang, 2020). ReCOVery is a Covid-19-correlated multimodal fake news detection dataset, including news-related images, textual content, and social context, sourced from the News- Guard21 website, comprising 2,000 news articles and 1.4 million tweets (Zhou *et. al.*, 2020).

RQ3: WHAT ARE THE PRIMARY CHALLENGES FACED IN EACH PERSPECTIVE WHEN ADDRESSING FAKE NEWS?

Six secondary studies intersect with the primary challenges, including two within the health sector (S1; S12) and three that were published between 2022 and 2023 (S1; S2; S12). The literature identifies these primary challenges as outlined by (Gomes *et. al.*, 2023):

Bridging echo chambers (S1; S2; S8; S11): People typically seek out information that confirms their existing beliefs and ignore evidence that contradicts them (S1). Such tendencies lead to the creation of information bubbles or echo chambers (S11).

Cross-domain, cross-platform, multilingual datasets and frameworks (S11; S12; S15): The nature of fake news is characterized by its perpetual evolution and intricacy. It manifests in diverse forms, encompassing a myriad of content, linguistic variations, thematic nuances, methodologies of dissemination, and geographical origins. A significant aspect of its sophistication is the ability to appear authentic, thereby evading immediate detection and recognition. Nevertheless, numerous current methods for its detection tend to concentrate solely on single facets like content, dissemination, style, or language (S2; S11; S12; S15). **Real-time learning and Early detection of fake news** (S2; S11; S12): The social networks allow fast spread of content, and fake news, due to its structure and social bots, hence there is a need for early detection (S2). A potential search direction is user profiling, in which the capture of contextual information on user behavior derived from social media users and the network can provide additional useful information to increase detection accuracy (S2; S12). Another direction is real--time detection, utilizing web ap- applications for fact-checking that can continuously learn from newly fact-checked articles, providing real-time identification of fraudulent information (S11).

We also extracted other challenges, such as:

Generative models (S11; S6; S15): Six studies characterize generative content, particularly deepfakes (S2; S6; S7; S8; S11; S15), describing these as manufactured images or videos. However, only two of these studies (S11; S15) emphasize the challenge of detecting such fabricated media. Among these, (S6) is one of the earliest secondary studies in the deep learning domain and the only one that discusses text generators, citing GPT-3 as a tool capable of automating the creation of fake news.

Transparency and blocking (S8; S15): The limited coverage of fact-checking websites and regulatory approach necessitates the provision of a more transparent communicative interface for news consumers to access and comprehend the algorithm results. News consumers often rely on algorithmic decision-making instead of their own judgment due to the lack of transparency in the regulations (e.g., warning labels) (S8). To effectively prevent the proliferation of false information, it is necessary to implement

new policies and regulations. Furthermore, a successful strategy to block and mitigate the spread of fake news can be based on the structure of the network or users (S15).

Explainable detection (S12; S15). There is also a need to make the fake news detection explainable to users, or via model interpretability or mining social feedback (S12; S15). Biomedical claims, which are usually taken from scientific literature, can be difficult for regular users to understand. Text simplification is a promising area of research that could provide simplified explanations of these claims (S12).

Conclusion

This paper is a Rapid Tertiary study of fake news, in which we built upon our previous work, analyzing health, deep learning, and emerging perspectives (Gomes *et. al.*, 2023). A review of the selected studies reveals that two-thirds of them pertain to health or deep learning, with the remainder published between 2022 and 2023. Findings from research questions across all studies and specific interest groups in general were consistent, as illustrated in Figures 3 and 4. The tasks most frequently referenced included textual classification and fabricated media detection, regardless of whether the focus was solely on Deep Learning, Health, recent studies, or a broader scope. Similarly, Machine Learning/Deep Learning and Natural Processing techniques were predominantly cited not only in general and deep learning-focused research but also in recent and health-related studies.

Variations between groups are primarily observed in the research methodologies discussed in Section 3. All health studies were systematic. These studies were unique in referencing datasets newer than 2018, particularly those concerning COVID-19, although some of these studies were among the least recent overall. The sole systematic study specific to Deep Learning also pertains to the health sector.

In the studies from 2022-2023, all those from 2023 were systematic, and as illustrated in 2, there is a noticeable decrease in citations among these studies. We speculate that the growing volume of publications in recent years, as identified in our preliminary findings, reduces the likelihood of authors citing one another, given the increased number of available works. Also, a significant issue for taxonomy is evident in general secondary studies, which do not belong to any specific interest group.

Health and recent papers (S1; S2; S12) play a major role in the interest groups in the main challenges stated in (Gomes *et. al.*, 2023). In this study, we augment the scope of our inquiry by addressing three additional challenges: The following topics will be discussed: first, generative models; second, transparency and blocking; and third, explainable detection. Among these studies, (S6) is noteworthy as one of the earliest secondary studies in the deep learning domain, and it is the only study that discusses text generators for automating the creation of fake news. Citing GPT-3 as a tool capable of automating the creation of fake news. **Limitations and Future Work.** As an extension of (Gomes *et. al.*, 2023), our work inherits its limitations. Notably, this rapid review may have omitted relevant secondary studies. To mitigate this concern, we prioritized the most pertinent secondary studies identified through the Google Scholar algorithm. Moving forward, future endeavors should focus on (i) refining the procedures for conducting rapid reviews within the context of tertiary studies and (ii) replicating this study with an expanded pool of secondary studies. Enhancements in the search string are necessary to minimize irrelevant results, and the inclusion of additional scientific databases could further enrich our findings.

Acknowledgement

This work has been supported by the Agência Nacional de Telecomunicações (Anatel) and Fundação de Amparo à Pesquisa do Estado de Goiás (Fapeg). Este trabalho também foi apoiado pelo Centro de Excelência em Inteligência Artificial (Ceia) e pelo Centro de Competência EmbrapII em Tecnologias Imersivas (Advanced Knowledge Center for Immersive Technologies - Akcit) do Instituto de Informática da Universidade Federal de Goiás (INF-UFG).

References

ABOUT Us. *Gossip Cop*, [*s. l.*], c2019. Available from: https://web.archive.org/web/20190807002653/https://www.gossipcop.com/about/. Access from: 24 aug. 2025.

AHMAD, T.; ALIAGA LAZARTE, E. A.; MIRJALILI, S. A Systematic Literature Review on Fake News in the Covid-19 Pandemic: Can AI Propose a Solution? *Applied Sciences*, [*s. l.*], v. 12, n. 24, 2022. ISSN 2076-3417. DOI: 10.3390/app122412727. Available from: https://www.mdpi.com/2076-3417/12/24/12727. Access from: may 22, 2025.

AIMEUR, E.; AMRI, S.; BRASSARD, G. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, [*s. l.*], v. 13, n. 1, p. 30, 2023. DOI: 10.1007/s13278-023-01028-5. Available from: https://pubmed.ncbi.nlm.nih.gov/36789378/. Access from: may 22, 2025.

ALI, I.; AUYB, M. N.; SHIVAKUMARA, P.; NOOR, N. F. B. M. Fake News Detection Techniques on Social Media: A Survey. *Wireless Communications and Mobile Computing*, [*S.l.*], v. 2022, 2022. https://doi.org/10.1155/2022/6072084. Available from: https://onlinelibrary.wiley.com/doi/10.1155/2022/6072084. Access from: may 22, 2025.

CARTAXO, B.; PINTO, G.; SOARES, S. Rapid Reviews in Software Engineering. *In*: FELDERER, M.; TRAVASSOS, G. H. (ed.). *Contemporary Empirical Methods in Software Engineering.* Cham: Springer International Publishing, 2020, p. 357–384. ISBN 978-3-030-32489-6. DOI: 10.1007/978-3-030-32489-6_13. Available from: https://content.e-bookshelf.de/media/reading/L-13432442-5babef3091.pdf. Access from: may 22, 2025.

CUI, L.; LEE, D. CoAID: COVID-19 Healthcare Misinformation Dataset, 2020. arXiv: 2006.00885 [cs.SI].

DAI, E.; SUN, Y.; WANG, S. Ginger Cannot Cure Cancer: Battling Fake Health News with a Comprehensive Data Repository. *Proceedings of the International AAAI Conference on Web and Social Media*, [*s. l.*], v. 14, n. 1, p. 853–862, may 2020. DOI: 10.1609/icwsm.v14i1.7350. Available from: https://ojs.aaai.org/index.php/ICWSM/article/view/7350. Access from: may 22, 2025.

DHALL, S.; DWIVEDI. A. D.; PAL, S. k; SRIVASTAVA, G. Blockchain-based Framework for Reducing Fake or Vicious News Spread on Social Media/Messaging

Platforms. *ACM Trans. Asian Low-Resour. Lang. Inf. Process*, New York, NY, USA, v. 21, n. 1, Nov. 2021. ISSN 2375-4699. DOI: 10.1145/3467019. Available from: https://doi.org/10.1145/3467019. Access from: may 22, 2025.

DIXON, S. J. Most popular social networks worldwide as of January 2024, ranked by number of monthly active users. *Statista*, [*s. l.*], 2024. Available from: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. Access from: may 22, 2025.

GARG, D.; KHAN, S.; ALAM, M. Integrative use of IoT and deep learning for agricultural applications. In: SINGH, P. K.; SHARMA, S.; SHARMA, R.; SHARMA, S.; SHARMA, A. (ed.). *Proceedings of ICETIT 2019*. Cham: Springer International Publishing, 2020, p. 521–531. ISBN 978-3-030-30577-2. DOI: 10.1007/978-3-030-30577-2_46. Available from: https://link.springer.com/chapter/10.1007/978-3-030-30577-2_46. Access from: 24 aug. 2025.

HANGLOO, S.; ARORA, B. Combating multimodal fake news on social media: methods, datasets, and future perspective. *Multimedia systems*, *Springer*, [*s. l.*], v. 28, n. 6, p. 2391–2422, 2022. DOI: https://doi.org/10.1007/s00530-022-00966-y. Available from: https://dl.acm.org/doi/10.1007/s00530-022-00966-y. Access from: may 22, 2025.

HEALTH NEWS REVIEW, [*s. l.*], 2022. Available from: https://web.archive.org/web/20220707131651/https://www.healthnewsreview.org/. Access from: 24 aug. 2025.

ISLAM, M. R.; LIU, S.; WANG, X.; XU, G. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, [*s. l.*], v. 10, dec. 2020. DOI: 10.1007/s13278-020-00696-x. Available from: https://www.researchgate.net/publication/344437686_Deep_learning_for_misinformation_detection_on_online_social_networks_a_survey_and_new_perspectives. Access from: may 22, 2025.

KIM, B.; XIONG, A.; LEE, D.; HAN, K. A systematic review on fake news research through the lens of news creation and consumption: Research efforts, challenges, and future directions. *PLOS ONE*, Public Library of Science, [*s. l.*], v. 16, n. 12, p. 1–28, Dec. 2021. DOI: 10.1371/journal.pone.0260080. Available from: https://doi.org/10.1371/journal.pone.0260080. Access from: may 22, 2025.

KONDAMUDI, M. R.; SAHOO, S. R.; CHOUHAN, R.; YADAV, N. A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches. *Journal of King Saud University - Computer and Information Sciences*, [*s. l.*], v. 35, n. 6, p. 101571, 2023. ISSN 1319-1578. DOI: https://doi.org/10.1016/j.jksuci.2023.101571. Available from: https://www.sciencedirect.com/science/article/pii/S1319157823001258. Access from: may 22, 2025.

LIGTHART, A.; CATAL, C.; TEKINERDOGAN, B. Systematic Reviews in Sentiment Analysis: A Tertiary Study. *Artif. Intell. Rev.*, Kluwer Academic Publishers, USA, v. 54, n. 7, p. 4997–5053, Oct. 2021. ISSN 0269-2821. DOI: 10.1007/s10462-021-09973-3. Available from: https://doi.org/10.1007/s10462-021-09973-3. Access from: may 22, 2025.

LTD, L. *Litmaps:* Your Literature Review Assistant. [*S.l.: s.n.*], 2023. Available from: https://www.litmaps.com/. Access from: may 22, 2025.

MEEL, P.; VISHWAKARMA, D. K. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, [*s. l.*], v. 153, p. 112986, 2020. ISSN 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2019.112986. Available from: https://www.sciencedirect.com/science/article/pii/S0957417419307043. Access from: may 22, 2025.

MITRA, T.; GILBERT, E. Credbank: A Large-Scale Social Media Corpus With Associated Credibility Annotations. *Proceedings of the International AAAI Conference on Web and Social Media*, [*s. l.*], v. 9, n. 1, p. 258–267, 2015. DOI: 10.1609/icwsm.v9i1.14625. Available from: https://ojs.aaai.org/index.php/ICWSM/article/view/14625. Access from: may 22, 2025.

PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, [*s. l.*], v. 64, p. 1–18, 2015. ISSN 0950-5849. DOI: https://doi.org/10.1016/j.infsof.2015.03.007. Available from: https://www.sciencedirect.com/science/article/pii/S0950584915000646. Access from: may 22, 2025.

POLITIFACT. The Poynter Institute, [*s. l*], c2020. Available from: https://www.politifact.com/. Access from: 24 aug. 2025.

REIS, J. C. S; CORREIA, A.; MURAI, F.; VELOSO, A. Supervised Learning for Fake News Detection. *IEEE Intelligent Systems*, [*s. l.*], v. 34, n. 2, p. 76–81, 2019. DOI: 10.1109/MIS.2019.2899143. Available from: https://www.researchgate.net/publication/332952191_Supervised_Learning_for_Fake_News_Detection. Access from: may 22, 2025.

SCHLICHT, I. B.; FERNADEZ, E.; CHULVI, B.; ROSSO, P. Automatic detection of health misinformation: a systematic review. *Journal of Ambient Intelligence and Humanized Computing*, [*s. l.*], p. 1–13, 2023. Available from: https://link.springer.com/article/10.1007/s12652-023-04619-4. Access from: may 22, 2025.

SHAHI, G. K.; NANDINI, D. *FakeCovid - A Multilingual Cross-domain Fact Check News Dataset for COVID-19*. [*S. l.*]: ICWSM, June 2020. DOI: 10.36190/2020.14. Available from: https://doi.org/10.36190/2020.14. Access from: may 22, 2025.

SHARMA, K.; OIAN, F.; JIANG, H.; RUCHANSKAY, N.; LIU, Y. Combating Fake News: A Survey on Identification and Mitigation Techniques. ACM Trans. Intell. Syst. Technol**.** *Association for Computing Machinery*, New York, NY, USA, v. 10, n. 3, Apr. 2019. ISSN 2157-6904. DOI: 10.1145/3305260. Available from: https://doi.org/10.1145/3305260. Access from: may 22, 2025.

SHU, K.; MAHUDESWARAN, D.; WANG, S.; LEE, D.; LIU, H. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *PubMed Disclaimer*, [*s. l.*], v. 8, n. 3, p.171-188, jun. 2020. DOI: 10.1089/big.2020.0062. Available from: https://pubmed.ncbi.nlm.nih.gov/32491943/. Access from: may 22, 2025.

SNOPES. [*s. l.*], [2025?]. Available from: https://www.snopes.com/. Access from: 24 aug. 2025.

TACCHINI, E.; BALLARIN, G.; DELLA VEDOVA, M. L.; MORET, S.; DE ALFARO, L. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*, [*s. l.*], 2017. DOI: 10.48550/arXiv.1704.07506. Available from: https://arxiv.org/abs/1704.07506. Access from: aug. 24, 2025.

THE CORONA Virus Facts/ Datos Corona Virus Alliance Database. *Pointer.50*, [*s. l.*], 2025. Available from: https://www.poynter.org/ifcn-covid-19-misinformation/. Access from: 24 aug. 2025.

VARMA, R.; VERMA, Y.; VIJAYVARGIYA, P.; CHURI, P. A systematic survey on deep learning and machine learning approaches of fake news detection in the pre-and post-COVID-19 pandemic. *International Journal of Intelligent Computing and Cybernetics*, Emerald Publishing Limited, [*s. l.*], v. 14, n. 4, p. 617–646, 2021. Available from: https://www.researchgate.net/publication/355057574_A_systematic_survey_on_deep_learning_and_machine_learning_approaches_of_fake_news_detection_in_the_pre-_and_post-COVID-19_pandemic. Access from: aug. 24, 2025.

WANG, W. Y. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. *In*: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vancouver, Canada: Association for Computational Linguistics, July 2017, p. 422–426. DOI: 10.18653/v1/P17-2067. Available from: https://aclanthology.org/P17-2067. Access from: aug. 24, 2025.

WANG, Y.; McKEE, M.; TORBICA, A.; STUCKLER, D. Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. *Social Science & Medicine*, [*s. l.*], v. 240, p. 112552, 2019. ISSN 0277-9536. DOI: https://doi.org/10.1016/j.socscimed.2019.112552. Available from: https: //www.sciencedirect.com/science/article/pii/S0277953619305465. Access from: aug. 24, 2025.

WU, Y.; NGAI, E. W. T.; WU, P.; WU, C. Fake news on the internet: a literature review, synthesis and directions for future research. *Internet Research*, [*s. l.*], v. 32, p. 1662–1699, 5 jan. 2022. ISSN 1066-2243. DOI: 10.1108/INTR-05-2021-0294. Available from: https://doi.org/10.1108/INTR-05-2021-0294. Access from: aug. 24, 2025.

ZHOU, X.; MULAY, A.; FERRARA, E.; ZAFARANI, R. Recovery: a Multimodal Repository for COVID-19 News Credibility Research. *In*: PROCEEDINGS of the 29th ACM International Conference on Information & Knowledge Management. Virtual Event, Ireland: Association for Computing Machinery, 2020 (CIKM '20). New York, USA: Association for Computing Machinery, 2020, p. 3205–3212. ISBN 9781450368599. DOI: 10.1145/3340531.3412880. Available from: https://doi.org/10.1145/3340531.3412880. Access from: aug. 24, 2025.

Secondary Studies Included in this Work

S1 AHMAD, T.; ALIAGA LAZARTE, E. A.; MIRJALILI, S. A Systematic Literature Review on Fake News in the COVID-19 Pandemic: Can AI Propose a Solution? *Applied Sciences*, [*s. l.*], v. 12, n. 24, 2022. ISSN 2076-3417. DOI: 10.3390/app122412727. Available from: https://www.mdpi.com/2076-3417/12/24/12727. Access from: aug. 24, 2025.

S2 AIMEUR, E.; AMRI, S.; BRASSARD, G. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, [*s. l.*], v. 13, n. 1, p. 30, 2023. DOI: 10.1007/s13278-023-01028-5. Available from: https://pubmed.ncbi.nlm.nih.gov/36789378/. Access from: aug. 24, 2025.

S3 ALI, I.; AYUB, N. B.; SHIVAKUMARA, P.; NOOR, N. F. M. Fake News Detection Techniques on Social Media: A Survey. *Wireless Communications and Mobile Computing*, Hindawi, [*s. l.*], v. 2022, 2022. Access from: aug. 24, 2025. DOI: 10.1155/2022/6072084. Available from: https://www.onlinelibrary.wiley.com/doi/10.1155/2022/6072084. Access from: Aug. 24, 2025.

S4 BEER, D. de; MATTHEE, M. Approaches to Identify Fake News: A Systematic Literature Review. HOESEL, S. van; GÓMEZ, J. M.; ZHU, Q. (ed.). *Information and communication technologies in education, research, and industrial applications*. Cham: Springer, 2021, p. 13–22. ISBN 978-3-030-49263-2. DOI: 10.1007/978-3-030-49264-9_2. Available from: https://doi.org/10.1007/978-3-030-49264-9_2. Access from: Aug. 24, 2025.

S5 HU, L.; WEI, S.; ZHAO, Z.; WU, B. Deep learning for fake news detection: A comprehensive survey. *AI Open*, [*s. l.*], v. 3, p. 133–155, 2022. ISSN 2666-6510. DOI: https://doi.org/10.1016/j.aiopen.2022.09.001. Available from: https: //www.sciencedirect.com/science/article/pii/S2666651022000134. Access from: aug. 24, 2025.

S6 ISLAM, M. R.; LIU, S.; WANG, X.; XU, G. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining*, [*s. l.*], v. 10, dec. 2020. DOI: 10.1007/s13278-020-00696-x. Available from: https://link.springer.com/article/10.1007/s13278-020-. Access from: aug. 24, 2025.

S7 KAPANTAI, E.; CHRISTOPOULOU, A.; BERBERIDIS, C.; PERISTERAS, V. A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society*, [*s. l.*], v. 23, n. 5, p. 1301–1326, 2021. DOI: 10.1177/1461444820959296. eprint: https://doi.org/10.1177/1461444820959296. Available from: https://doi.org/10.1177/1461444820959296. Access from: aug. 24, 2025.

S8 KIM, B.; XIONG, A.; LEE, D.; HAN, K. A systematic review on fake news research through the lens of news creation and consumption: Research efforts, challenges, and future directions. *PLOS ONE*, Public Library of Science, [*s. l.*], v. 16, n. 12, p. 1–28, dec. 2021. DOI: 10.1371/journal.pone.0260080. Available from: https://doi.org/10.1371/journal.pone.0260080. Access from: aug. 24, 2025.

S9 KONDAMUDI, M. R.; SAHOO, S. R.; CHOUHAN, L.; YADAV, N. A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches. *Journal of King Saud University - Computer and Information Sciences*, [*s. l.*], v. 35, n. 6, p. 101571, 2023. ISSN 1319-1578. DOI: https://doi.org/10.1016/j.jksuci.2023.101571. Available from: https: //www.sciencedirect.com/science/article/pii/S1319157823001258. Access from: aug. 24, 2025.

S10 LI, J.; LEI, M. A Brief Survey for Fake News Detection via Deep Learning Models. Procedia Computer Science, [*s. l.*], v. 214, p. 1339–1344, 2022. *In*: 9TH INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY AND QUANTITATIVE MANAGEMENT. ISSN 1877-0509. DOI: https://doi.org/10.1016/j.procs.2022.11.314. Available from: https: //www.sciencedirect.com/science/article/pii/S1877050922020269. Access from: aug. 24, 2025.

S11 MEEL, P.; VISHWAKARMA, D. K. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, [*S. l.*], v. 153, p. 112986, 2020. DOI: 10.1016/j.eswa.2020.112986. Available from: https://prohic.nl/wp-content/uploads/2020/11/2020-10-26-FakeNewsOverviewMeta.2020.pdf. Access from: 24 aug. 2025.

S12 SCHLICHT, I. B.; FERNANDEZ, E.; CHULVI, B.; ROSSO, P. Automatic detection of health misinformation: a systematic review. *Journal of Ambient Intelligence and Humanized Computing*, [*S. l.*], v. 15, n. 3, p. 2009–2021, 2024. DOI: 10.1007/s12652-023-04619-4. Available from: https://link.springer.com/article/10.1007/s12652-023-04619-4. Access from: 24 aug. 2025.

S13 VARMA, R.; VERMA, Y.; VIJAYVARGIYA, P.; CHURI, P. P. A systematic survey on deep learning and machine learning approaches of fake news detection in the pre-and post-COVID-19 pandemic. *International Journal of Intelligent Computing and Cybernetics*, [*S. l.*], v. 14, n. 4, p. 617–646, 2021. DOI: 10.1108/IJICC-04-2021-0069. Available from: https://www.emerald.com/insight/content/doi/10.1108/IJICC-04-2021-0069/full/html. Access from: 24 aug. 2025.

S14 ZAREIE, A.; SAKELLARIOU, R. Minimizing the spread of misinformation in online social networks: A survey. *Journal of Network and Computer Applications*, [*S. l.*], v. 186, p. 103094, 2021. ISSN 1084-8045. DOI: https://doi.org/10.1016/j.jnca.2021.103094. Available from: https: //www.sciencedirect.com/science/article/pii/S1084804521001168. Access from: aug. 24, 2025.

S15 ZHOU, X.; ZAFARANI, R. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.,* Association for Computing Machinery, New York, NY, USA, v. 53, n. 5, sept. 2020. ISSN 0360-0300. DOI: 10.1145/3395046. Available from: https://doi.org/10.1145/3395046. Access from: aug. 24, 2025.

tecnia revista de educação, ciência e tecnologia do ifg

v. 10, Edição Especial 1 | 2025

Submetido 31/05/2024. Aprovado 12/04/2025
Avaliação: revisão duplo-anônimo

Redes Perceptron Multicamadas para modelar efeitos de distorção em sinais de guitarra elétrica

MULTILAYER PERCEPTRON NETWORKS TO MODEL DISTORTION EFFECTS IN ELECTRIC GUITAR SIGNALS

REDES DE PERCEPTRONES MULTICAPA PARA MODELAR EFECTOS DE DISTORSIÓN EN SEÑALES DE GUITARRA ELÉCTRICA

Alisson Assis Carvalho
Universidade Federal de Goiás (UFG)
alsnac@ufg.br

Murilo Guimarães Correia
Universidade Federal de Goiás (UFG)
murilogcorreia@discente.ufg.br

Ricardo Augusto Pereira Franco
Universidade Federal de Goiás (UFG)
ricardo@inf.ufg.br

Samuel Carvalho de Almeida
Universidade Federal de Goiás (UFG)
carvalho.carvalho@discente.ufg.br

Resumo

Este trabalho investiga a modelagem de efeitos de distorção em sinais de guitarra elétrica utilizando redes perceptron multicamadas (MLPs). Na proposta, a MLP é responsável por transformar o sinal elétrico original da guitarra em um sinal distorcido. Para o treinamento da rede, foram gerados/coletados dados sintéticos de efeito de distorção a partir de sinais elétricos reais aplicados em um simulador SPICE de circuitos eletrônicos. Os resultados do estudo são apresentados em termos de Erro Quadrático Médio (EQM) e testes estatísticos de aderência de Kolmogorov-Smirnov. As análises demonstram que a proposta apresentada para modelar o sistema dinâmico de um pedal elétrico de distorção com o uso da MLP tem bom desempenho, representando os sinais distorcidos com fidelidade.

Palavras-chave: modelagem de distorção de áudio; redes MLP; processamento de sinal; sistemas dinâmicos.

Abstract

This paper investigates the modeling of distortion effects in electric guitar signals using multilayer perceptron (MLP) networks. In the proposed approach, the MLP is employed to transform the original guitar signal into a distorted output. For network training, synthetic distortion data were generated from real electrical signals processed through a SPICE electronic circuit simulator. The results are evaluated using Mean Square Error (MSE) and Kolmogorov–Smirnov statistical adherence tests. The findings

indicate that the proposed MLP-based model effectively captures the dynamics of an electric guitar distortion pedal, providing a faithful representation of the distorted signals.

Keywords: audio distortion modeling; MLP neural networks; guitar signal processing; dynamic systems.

Resumen

Este trabajo presenta uma investigación del modelado de efectos de distorsión en señales de guitarra eléctrica mediante el uso de redes de perceptrones multicapa (MLP). Para el entrenamiento de la red, se generaron/recogieron datos sintéticos del efecto de distorsión a partir de señales eléctricas reales, estas aplicadas a un simulador de circuito electrónico SPICE. Los resultados del estudio se presentan en términos de error cuadrático medio (MSE) y pruebas de adherencia estadística de Kolmogorov-Smirnov. Los análisis demostraron que la propuesta presentada para modelar el sistema dinámico de un pedal de distorsión eléctrico com el uso de MLP tiene buen desempeño, representando fielmente las señales distorsionadas.

Palabras clave: modelado de distorsión de áudio; redes neuronales MLP; procesamiento de señales de guitarra; sistemas dinámicos.

Introdução

As guitarras elétricas e os pedais são elementos fundamentais no contexto do rock como gênero musical. Os pedais desempenham um papel crucial na manipulação e na criação de uma ampla gama de timbres, variando de tons agressivos e excêntricos a nuances mais suaves e delicadas. Essa diversidade sonora não apenas permite que os músicos desenvolvam sua própria identidade musical, mas também contribui significativamente para a identidade sonora de uma banda (Reiss; Mcpherson, 2014).

Um modelo icônico de pedal de distorção é o "*MXR Distortion+*", conhecido também como "*Distortion Plus*" ou "D+", classificado como um *Fuzz* de distorção mais suave. Lançado pela empresa MXR entre 1978 e 1979, utiliza amplificador operacional 741, potenciômetros de volume e ganho, além de diodos germânicos para recortar tensões acima de um limiar definido, resultando no efeito de distorção característico (Self, 2023).

Compreender o funcionamento dos circuitos eletrônicos é essencial, dada sua natureza não linear. Essa não linearidade origina-se do emprego de componentes não lineares em sua construção, os quais são indispensáveis à formação da sonoridade característica proporcionada pela distorção. Esse desafio se intensifica pela necessidade das tecnologias de distorção digital de reproduzir com fidelidade os sons produzidos pelos circuitos analógicos, que estão intrinsecamente atrelados à sonoridade histórica de diversas bandas e artistas.

Uma alternativa prática e motivadora para lidar com essa complexidade reside na utilização de inteligências artificiais para modelar sistemas não lineares. A ideia central consiste em fornecer à inteligência artificial sinais de entrada de guitarras sem distorção e aplicá-los a um processo de distorção. Empregando técnicas de aprendizado de máquina, a inteligência artificial é capaz de gerar um modelo do efeito de distorção aplicado ao sinal de guitarra elétrica. Entre essas técnicas destaca-se o emprego de redes neurais como solução promissora, conforme proposto em Purwins *et al*. (2019).

As redes neurais, inspiradas no funcionamento dos neurônios do cérebro humano, configuram-se como uma ferramenta poderosa para o processamento de dados

e a execução de tarefas complexas (Brunetto; Schmidt; Dalla'rosa, 2023). A unidade fundamental de uma rede neural, o neurônio artificial, estabelece conexões com outros neurônios por meio de sinapses artificiais, análogas aos dendritos biológicos. Os valores recebidos por cada neurônio são ponderados e combinados por meio de uma função matemática, gerando a saída do neurônio. Esses neurônios se organizam em camadas interconectadas, em que a informação é processada e refinada sucessivamente, culminando na saída final da rede neural (Faceli *et al.*, 2011).

Este trabalho tem como objetivo investigar a capacidade de uma rede MLP (*Multilayer Perceptron*) em modelar efeitos de distorção em sinais de guitarra elétrica. Além disso, busca-se avaliar a fidelidade das modelagens geradas pela MLP em comparação com os sinais de distorção originais. Os resultados obtidos são apresentados e discutidos ao longo do texto, evidenciando o potencial das redes MLP na reprodução de efeitos de áudio complexos.

Referencial teórico

Na música, a distorção é um efeito frequentemente utilizado com a finalidade de alterar a forma do áudio e gerar um som saturado e expressivo na guitarra elétrica. Modelar e controlar essa distorção com precisão é fundamental para músicos e engenheiros de áudio. Redes Neurais Artificiais (RNAs), especialmente Perceptrons Multicamadas (MLPs), apresentam-se como uma abordagem promissora para essa modelagem. Este referencial teórico explora os fundamentos da distorção de áudio, a aplicação de RNAs no processamento de áudio e a relevância das MLPs nesse contexto.

Distorção de áudio

A distorção de áudio pode ser classificada em dois tipos principais: harmônica e não harmônica. A distorção harmônica ocorre quando os harmônicos de um tom fundamental são produzidos de maneira controlada, resultando em um som mais rico e musical. Em contraste, a distorção não harmônica produz componentes de frequência não relacionados ao som original, muitas vezes resultando em uma cacofonia áspera (Reiss; Mcpherson, 2014).

Pedal e simulação

- Elementos para obtenção dos dados

A escolha arbitrária do pedal não afeta no resultado da pesquisa, bem como a forma que o sinal de saída do pedal é coletado. Como o sinal de entrada e o sinal de saída do pedal são fornecidos à rede neural, o que importa é se a rede neural consegue uma aproximação da função de transferência, mas qual é a função de transferência em si não importa para este escopo.

Dessa forma, também se justifica a utilização de um software de simulação de circuitos para simular o pedal, pois não cabe a esta pesquisa avaliar a qualidade de simulação do software, mas sim a capacidade da rede neural de relacionar dois sinais. Esse fato ressalta a aplicabilidade do processo em todos os sistemas dinâmicos não lineares, como os pedais de distorção. Assim, um pedal de distorção genérico foi escolhido por sua relativa simplicidade e ampla utilização no gênero do rock.

Semelhantemente, o software para simulação, o LTSpice (“Analog Devices”, 2023), foi escolhido pela simplicidade e pelo fato de ser gratuito (LTspice, 2023).

- Elementos para obtenção dos dados

O pedal está organizado em 4 partes que realizam funções individuais no processamento do sinal. A primeira dessas partes é o estágio de entrada de sinal (do inglês, *input signal stage*). O sinal escolhido para a simulação estava com uma tensão alta, mais alta do que costuma ser fornecido de entrada por uma guitarra. Por esse motivo, foi colocado uma fonte dependente com ganho de 0.2, valor esse que trouxe o sinal de entrada para um valor mais próximo da realidade.

Na montagem do circuito analógico, o sinal da guitarra deve ser aplicado no nó sinalizado pela palavra *input*. A segunda parte do pedal é o estágio de alimentação de energia (do inglês, *power supply stage*). O uso dos altos valores de resistência contribui para uma alta impedância de entrada do circuito, além de uma alta impedância para o terra do curto-circuito virtual. O capacitor tem o papel de eliminar a variação residual conhecida como ondulação (do inglês, *ripple*), que é um sinal AC indesejado. Na Figura 1, pode-se visualizar o diagrama elétrico do pedal de efeito de distorção.

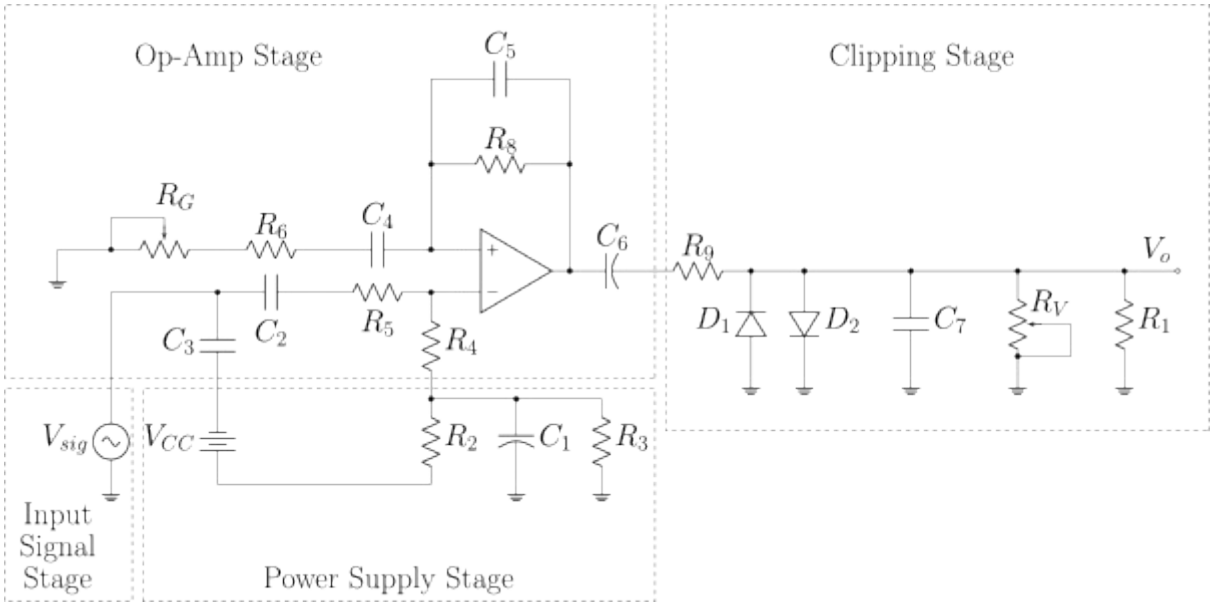


Figura 1 - Diagrama elétrico do pedal
Fonte: elaborado pelo(as) autores(as).

Ademais, a terceira parte do pedal é o estágio de amplificação do Amp Op (do inglês, *Op Amp Amplifier Stage*, ou somente *Op Amp Stage* como referido no esquemático). O amplificador se encontra na configuração não inversora, o que gera um sinal de saída como relação dos resistores R8, R6 e RG, sendo este último um potenciômetro que permite ao usuário controlar o ganho. Além disso, nesse estágio, há o capacitor C3, cuja função é evitar ruído de rádio frequência, bem como ajudar em descargas eletrostáticas e oscilações. O capacitor C6 tem a função de remover a tensão DC para o próximo estágio. Os outros capacitores contribuem para a resposta em frequência com pico em torno de 1,5 KHz, característica comum em outros pedais que ajuda a guitarra a ter destaque em relação a outros instrumentos na música ao ter maior ganho na faixa de frequência audível para os seres humanos.

O último estágio é o de recorte (do inglês, *Clipping Stage*). Nessa etapa, o resistor R9 limita a corrente que chega aos diodos, protegendo-os. Os diodos possuem uma tensão de saturação e, pela disposição deles, saturam com tensão aproximadamente menor que -0,6V e valores maiores que 0,7V. Com isso, o sinal é cortado/limitado além desses limites, dando a sonoridade característica da distorção. Quanto mais abrupto o corte, mais distorcido e com excessos de harmônicos o som fica (Reiss *et al.*, 2014). Finalmente, há um potenciômetro para a regulagem do volume, bem como a impedância de saída do pedal – esta última não está presente na construção do pedal, mas é importante para a simulação por representar uma carga para ser realizada a máxima transferência de potência entre o pedal e a carga.

Redes neurais artificiais em processamento de áudio

Redes neurais perceptron multicamadas (MLPs) são um tipo de RNA comumente usado em problemas de aprendizagem supervisionada. Eles consistem em camadas de neurônios (perceptrons) organizadas em uma estrutura de camadas, incluindo uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. A MLP é conhecida por sua capacidade de modelar funções não lineares complexas (Zhang *et al.*, 2023).

A capacidade de modelar funções não lineares complexas da MLP é explicada ao entender que uma rede com uma camada intermediária pode implementar qualquer função contínua. A utilização de duas camadas intermediárias permite a aproximação de qualquer função (Faceli *et al.*, 2011). Sendo a função que se deseja modelar não linear, usar o método com múltiplas camadas permite uma aproximação dessa função, por isso a MLP foi escolhida como método de aprendizado de máquina aplicado.

Métodos de avaliação dos resultados

A utilização das técnicas de teste de erro quadrático médio (EQM) e Kolmogorov-Smirnov (KS) é muito importante na avaliação de resultados de projetos de processamento de áudio, como a modelagem dos efeitos de distorção de um sinal de guitarra elétrica utilizando uma rede MLP.

- Erro Quadrático Médio (EQM)

O erro quadrático médio é uma métrica comumente usada nos campos de processamento de sinais e aprendizado de máquina para quantificar a diferença entre um sinal estimado e um sinal de referência (Bishop, 2006). Este projeto usa EQM para medir a fidelidade de uma rede MLP a um sinal de áudio de uma guitarra elétrica com som distorcido como referência. Quanto menor o valor do EQM, mais próxima a saída MLP está do sinal de referência, indicando melhor capacidade do modelo em reproduzir o efeito de distorção desejado.

- Teste de Kolmogorov-Smirnov (KS)

O teste Kolmogorov-Smirnov é um método estatístico usado para comparar duas distribuições de probabilidade e avaliar se elas vêm da mesma população (Stephens, 1974). Neste projeto, o KS foi aplicado para verificar se a saída MLP segue a mesma distribuição estatística do sinal distorcido da guitarra elétrica. Um valor alto para a estatística KS ou um valor- menor que o valor significativo de 0.05 indica que as duas

distribuições são diferentes, sugerindo que a saída da MLP não reproduz adequadamente o efeito de viés desejado.

O Teste de Kolmogorov-Smirnov (KS) é usado para verificar a semelhança entre duas distribuições de dados. A estatística de teste KS é calculada como:

onde é a função de distribuição acumulada da primeira distribuição, é a função de distribuição acumulada da segunda distribuição e denota o supremo sobre todos os valores possíveis de . O valor- associado ao KS é usado para determinar a significância estatística da diferença entre as duas distribuições.

- Função de Distribuição Acumulada (FDA)

Além das métricas EQM e KS, a avaliação dos resultados incorpora a análise visual das Funções de Distribuição Acumulada (FDA), derivadas das saídas da MLP e dos sinais de referência. As Figuras 5 e 6 ilustram as FDA para os ganhos de 10k e 30k, respectivamente. Esses gráficos proporcionam uma representação visual das distribuições acumuladas das amplitudes, permitindo uma comparação intuitiva entre a saída da MLP e os sinais de referência.

A função de distribuição acumulada (FDA) é uma ferramenta fundamental em estatística e probabilidade. Ela é definida como a integral da função de densidade de probabilidade (PDF). No contexto deste estudo, a FDA é aplicada às amplitudes dos sinais, proporcionando uma visão acumulativa das probabilidades associadas a diferentes valores de amplitude. A sobreposição próxima entre as curvas FDA da saída da MLP e dos sinais de referência indica consistência estatística, validando a capacidade da MLP em reproduzir os efeitos de distorção esperados nos sinais de guitarra elétrica (Ross, 2020; Wasserman, 2004; Zhang *et al.*, 2023).

Trabalhos relacionados

A área de processamento de áudio e inteligência artificial conta com diversas contribuições notáveis, algumas das quais são apresentadas no Quadro 1. Esses trabalhos anteriores exploram o uso de redes neurais em processamento de áudio, abordando técnicas e aplicações em síntese musical.

Título	Referência	Correlação	Diferencial
"Deep Learning for Audio Signal Processing"	(Purwins *et al.*, 2019)	Explora o uso de redes neurais em processamento de áudio	Destaca-se pela revisão abrangente de técnicas e aplicações em processamento de áudio
"Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders"	(Engel *et al.*, 2017)	Modelagem de síntese neural de áudio, incluindo efeitos	Aplica autoencoders WaveNet na síntese de áudio musical, destacando-se pela aplicação específica em notas musicais
"Audio Super Resolution using Neural Networks"	(Kuleshov *et al.*, 2017)	Apresenta uma abordagem de super-resolução de áudio utilizando redes neurais profundas	Aplicação de *deep learning* para melhorar a qualidade de áudio de baixa resolução

Quadro 1 - Trabalhos relacionados
Fonte: elaborado pelos(as) autores(as).

Diferencial deste trabalho

Este trabalho diferencia-se ao concentrar-se na análise específica dos efeitos de distorção em sinais de guitarra elétrica. Enquanto estudos anteriores podem ter uma visão mais abrangente do processamento de áudio, esta pesquisa aprofunda-se em uma aplicação particular, buscando compreender e reproduzir com fidelidade os efeitos de distorção desejados.

CONTRIBUIÇÕES ESPECÍFICAS

A análise detalhada dos efeitos de distorção em sinais de guitarra elétrica, usando uma abordagem baseada em redes neurais, mais precisamente, uma Multilayer Perceptron (MLP), é uma contribuição única deste trabalho. Enquanto estudos anteriores podem oferecer uma visão mais ampla, nossa pesquisa proporciona uma compreensão aprofundada e aplicada em um contexto musical específico.

POTENCIAIS IMPACTOS

As implicações deste trabalho não se limitam à esfera acadêmica, apresentando também potenciais impactos práticos. A capacidade de modelar de forma eficaz os efeitos de distorção em sinais de guitarra elétrica pode ter aplicações diretas na indústria musical, contribuindo para o desenvolvimento de processadores de efeitos mais avançados e realistas. Além disso, essa pesquisa pode servir como base para explorações mais aprofundadas em áreas relacionadas, como síntese de áudio, processamento de sinais musicais e até mesmo em campos mais amplos, como reconhecimento de padrões sonoros.

FUTURAS DIREÇÕES DE PESQUISA

O enfoque específico deste trabalho abre portas para futuras pesquisas que podem explorar diferentes instrumentos musicais, ampliando a aplicabilidade da abordagem proposta. Além disso, a incorporação de técnicas mais avançadas de aprendizado de máquina, como redes neurais recorrentes ou redes generativas, pode enriquecer ainda mais a modelagem de efeitos de distorção em contextos musicais diversos.

Metodologia

Obtenção e preparação de dados

Na fase inicial de obtenção e preparação dos dados, foi adotada uma abordagem meticulosa para garantir a qualidade das informações coletadas. Foi empregado um software de simulação em conjunto com um circuito eletrônico modelado para realizar simulações de áudio. O sinal de uma guitarra genérica, sem efeitos, foi selecionado como entrada principal. A plataforma de simulação, munida da funcionalidade *waveform*, possibilitou a incorporação desse sinal como entrada no circuito simulado. Após a definição dos parâmetros apropriados, o software gerou o sinal de saída em formato de áudio, fornecendo, assim, os dados cruciais para as próximas etapas da pesquisa.

Os dados passaram por uma fase de preparação cuidadosa, visando assegurar a qualidade das entradas fornecidas à MLP. Os áudios da guitarra elétrica, tanto com

quanto sem distorção, foram registrados a uma taxa de amostragem de 48kHz, com duração de 30 segundos, totalizando 1.440.000 amostras e capturando nuances sutis. A distorção foi introduzida por meio da aplicação de ganhos específicos, criando pares de áudios correspondentes: um com distorção gerada por um resistor de 10 kΩ e outro com distorção devido a um resistor de 30kΩ. Esse procedimento possibilitou a comparação direta dos efeitos da distorção na saída da MLP em relação ao áudio não distorcido.

A Figura 2 ilustra um trecho breve (0,005 segundos) comparando os áudios com distorção, considerando ganhos dos resistores de 10kΩ e 30kΩ, com o áudio sem distorção.

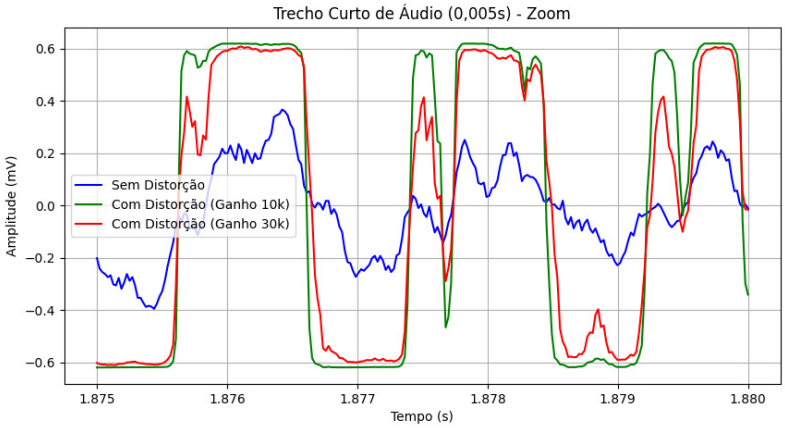


Figura 2 - Comparação entre os áudios com distorção (ganho de 10kΩ e 30kΩ) e o áudio sem distorção com tempo de 0,005 segundos
Fonte: elaborado pelo(as) autores(as).

Arquitetura da Perceptron Multicamadas (MLP)

A arquitetura da MLP desempenha um papel fundamental na modelagem de efeitos de distorção em sinais de guitarra elétrica. A perceptron multicamadas foi projetada para aprender representações complexas de sinais de áudio, permitindo a reprodução de efeitos de distorção realistas. Esta seção detalha a arquitetura MLP usada neste estudo.

A MLP utilizada neste projeto consiste em três camadas principais: uma camada de entrada, duas camadas ocultas e uma camada de saída. A camada de entrada é responsável por receber o sinal de áudio bruto, e as camadas ocultas aprendem uma representação intermediária do sinal. Já a camada de saída produz uma representação estimada do sinal de áudio distorcido. A arquitetura da MLP é representada na Figura 3.

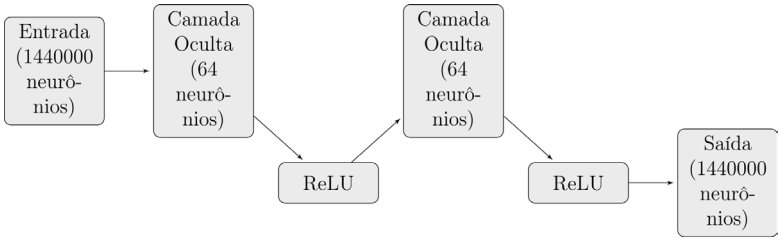


Figura 3 - Diagrama da arquitetura da MLP.
Fonte: elaborado pelos(as) autores(as).

Pode-se observar as seguintes informações sobre a arquitetura dessa MLP.

- **Camada de entrada**: a camada de entrada possui 1.440.000 neurônios, correspondendo ao número de amostras dos áudios em cada entrada. Isso ocorre devido à frequência de amostragem de 48kHz e ao uso dos áudios de entrada de 30 segundos. A frequência de amostragem determina o número de amostras por segundo, resultando em 48.000 amostras/segundos vezes 30 segundos, o que é igual a 1.440.000 amostras no total por áudio com cada neurônio recebendo uma amostra de cada áudio, ou seja, duas amostras por neurônio.
- **Camadas ocultas**: a MLP possui duas camadas ocultas, cada uma com 64 neurônios. As camadas ocultas são responsáveis por aprender representações intermediárias dos sinais de áudio. A função de ativação da unidade de retificação unificada, traduzida de *rectified linear unit* (ReLU), é aplicada após cada camada oculta para introduzir não linearidades na rede.
- **Camada de saída**: a camada de saída possui 1.440.000 neurônios, correspondendo ao número de amostras do áudio na saída. Ela produz a representação estimada do sinal de áudio distorcido.

A função de ativação ReLU desempenha um papel fundamental nas camadas ocultas da MLP. O comportamento da função ReLU consiste em mapear qualquer valor negativo para zero, enquanto mantém os valores positivos inalterados. Esse processo introduz não linearidades nas camadas ocultas da MLP, possibilitando a aprendizagem de relações não lineares complexas nos dados de entrada (Faceli *et al.*, 2011). A presença da função ReLU é crucial para capacitar a rede a discernir padrões e características mais intrincadas nos sinais de áudio.

A derivada da ReLU apresenta uma característica peculiar: o gradiente é zero para valores negativos e igual a um para valores não negativos, sendo crucial durante o treinamento da MLP. A propriedade de permitir a passagem irrestrita de gradientes positivos (durante a retropropagação do erro) é fundamental para evitar o problema do desaparecimento do gradiente, contribuindo para uma convergência mais eficaz durante o processo de otimização. Em contraste com funções de ativação que introduzem saturação, a ReLU facilita o treinamento de redes mais profundas, tornando-a uma escolha prevalente em arquiteturas de MLP (Faceli et al., 2011).

Treinamento MLP

O treinamento da MLP (Perceptron Multicamadas) desempenhou um papel crucial neste projeto, possibilitando que a rede neural modele de maneira eficaz os efeitos da distorção em sinais de guitarra elétrica. Detalhamos a seguir os principais aspectos desse processo.

O otimizador Adam (Kingma; Ba, 2017; Zhang *et al.*, 2023) foi escolhido para ajustar os parâmetros da MLP durante o treinamento. O algoritmo do otimizador Adam é apresentado a seguir.

Entrada:

- θ (parâmetros da rede neural)
- α (taxa de aprendizado)
- $\beta1$, $\beta2$ (hiperparâmetros de momento)
- ε (termo de estabilização numérica)
- $f(\theta)$ (função de perda)

Inicialização:

- $m \leftarrow 0$ (vetor inicial de primeiras ordens do momento)
- $v \leftarrow 0$ (vetor inicial de segundas ordens do momento)
- $t \leftarrow 0$ (contador de iterações)

Passo a passo:

1. Para cada iteração:

1. $t \leftarrow t + 1$
2. $g\_t \leftarrow \nabla\_\theta f(\theta\_t)$ (gradiente da função de perda)
3. $m\_t \leftarrow \beta1 * m\_\{t-1\} + (1 - \beta1) * g\_t$ (primeiro momento do gradiente)
4. $v\_t \leftarrow \beta2 * v\_\{t-1\} + (1 - \beta2) * g\_t\text{^}2$ (segundo momento do gradiente)
5. $\hat{m}\_t \leftarrow m\_t / (1 - \beta1\text{^}t)$ (correção do viés no primeiro momento)
6. $\hat{v}\_t \leftarrow v\_t / (1 - \beta2\text{^}t)$ (correção do viés no segundo momento)
7. $\theta\_t \leftarrow \theta\_\{t-1\} - \alpha * \hat{m}\_t/(sqrt(\hat{v}\_t) + \varepsilon)$ (atualização dos parâmetros)

2. Retorne θ (parâmetros otimizados)

Figura 4 – Algoritmo de treinamento utilizando para obter os melhores pesos possíveis da MLP
Fonte: elaborado pelos(as) autores(as).

A função de perda escolhida foi a função de perda do Erro Quadrático Médio (EQM), adequada para problemas de regressão, como é o caso deste projeto. O objetivo é minimizar a discrepância entre a saída da MLP e o áudio de referência distorcido por um pedal de guitarra.

Além desses elementos fundamentais, foram aplicadas técnicas adicionais para aprimorar o treinamento da MLP. Uma dessas técnicas consistiu na redução da taxa de aprendizado a cada 10 épocas. A taxa de aprendizado é um hiperparâmetro crítico que determina a magnitude dos ajustes dos pesos durante o treinamento. A redução periódica da taxa de aprendizado pode ajudar a estabilizar o treinamento e a evitar oscilações indesejadas.

O treinamento da MLP foi conduzido ao longo de 500 épocas, conforme apresentado nas Tabelas 1 e 2. Embora os resultados apresentados nas figuras da seção 4 mostrem resultados até a época 300, é importante destacar que o treinamento foi estendido até a época 500. Após a análise do comportamento da rede durante o treinamento, observou-se que, a partir da época 300, não há mudanças visíveis nas curvas, exceto com um zoom muito alto (intervalo de tempo muito pequeno).

Essa escolha foi feita para otimizar a apresentação dos resultados, eliminando informações redundantes e concentrando a atenção nas fases iniciais do treinamento, onde as mudanças são mais proeminentes. A Figura 2 mostra um trecho curto (0,005 segundos) comparando o áudio com distorção com ganho do resistor de (10 kΩ), com distorção com ganho do resistor de (30 kΩ) e o áudio sem distorção.

Avaliação dos resultados

A avaliação dos resultados neste projeto foi conduzida por meio de métricas quantitativas, visando mensurar a qualidade das previsões da MLP e sua similaridade com a fala de referência. As métricas adotadas foram o Erro Quadrático Médio (EQM),

teste de Kolmogorov-Smirnov (KS) e a análise da Função de Distribuição Acumulada (FDA).

O EQM foi empregado como uma medida de proximidade entre as saídas previstas pela MLP e os áudios de referência distorcidos. Especificamente, quanto menor o EQM, mais próximas as previsões da MLP estão dos áudios de referência, indicando melhor desempenho da rede.

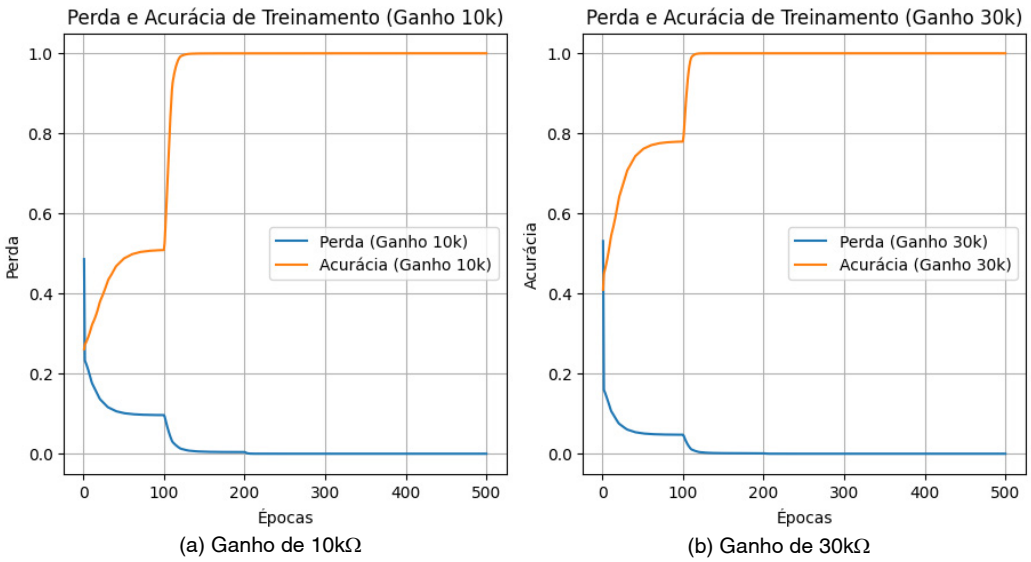
O teste de Kolmogorov-Smirnov (KS) foi aplicado para avaliar a similaridade entre as distribuições acumuladas das amplitudes dos áudios de referência e das saídas geradas pela MLP. Uma estatística KS próxima de zero e um valor- associado (*p-value*) próximo de 1 indicam uma boa concordância entre as distribuições, enquanto valores mais elevados sugerem divergências significativas.

Além disso, para uma compreensão mais aprofundada, comparamos diretamente as saídas das MLPs com o áudio de referência distorcido pelo pedal, como mostrado nas Figuras 8 e 9. Essas figuras destacam visualmente a similaridade entre as previsões da MLP e os áudios de referência.

Para uma análise estatística adicional, examinamos as Funções de Distribuição Acumulada (FDA), geradas a partir das saídas da MLP e dos sinais de referência. As Figuras 5 e 6 apresentam visualmente as FDA para os ganhos de 10kΩ e 30kΩ, respectivamente. Esses gráficos proporcionam uma representação visual das distribuições acumuladas das amplitudes, permitindo uma comparação intuitiva entre a saída da MLP e os sinais de referência.

Os resultados dessas métricas foram resumidos nas Tabelas 1 e 2 para os diferentes períodos de treinamento (100, 200, 300, 400 e 500 épocas), considerando os dois diferentes ganhos do resistor (10kΩ e 30kΩ).

Essa abordagem permitiu uma análise abrangente da performance da MLP em diferentes condições de treinamento, proporcionando insights sobre o impacto do ganho do resistor na qualidade das previsões da rede.

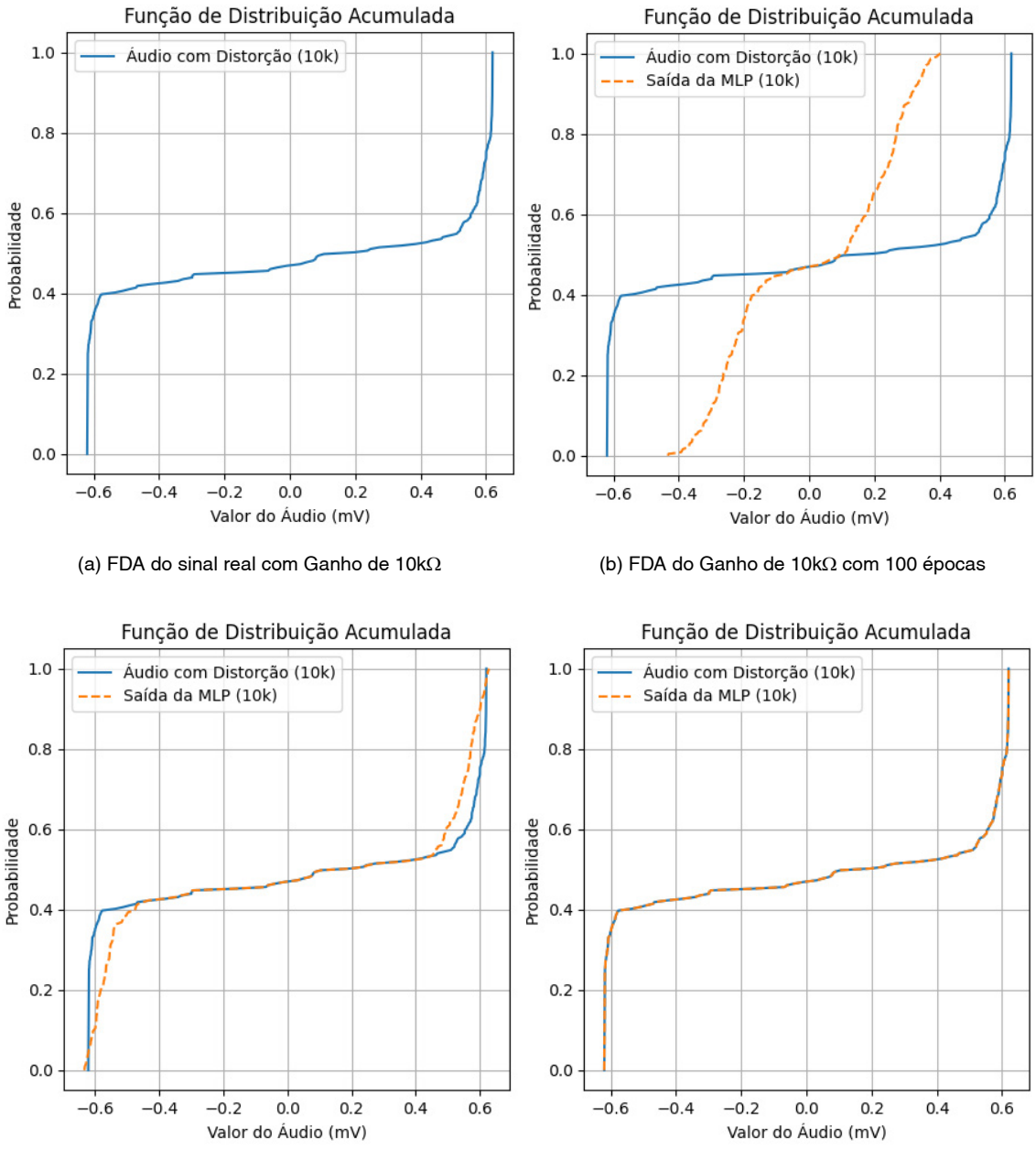


(a) Ganho de 10kΩ (b) Ganho de 30kΩ

Figura 5 - Variação da perda e acurácia durante o treinamento
Fonte: elaborado pelo(as) autores(as).

Simulações e resultados

Nesta seção, são apresentados os resultados das simulações conduzidas para avaliar a eficácia do modelo MLP em reproduzir a distorção de áudio em um sinal de guitarra elétrica. Inicialmente, o processo de treinamento do modelo é detalhado, seguido pela análise do desempenho do modelo treinado e pela comparação entre a saída da MLP e os áudios distorcidos.



(a) FDA do sinal real com Ganho de 10kΩ

(b) FDA do Ganho de 10kΩ com 100 épocas

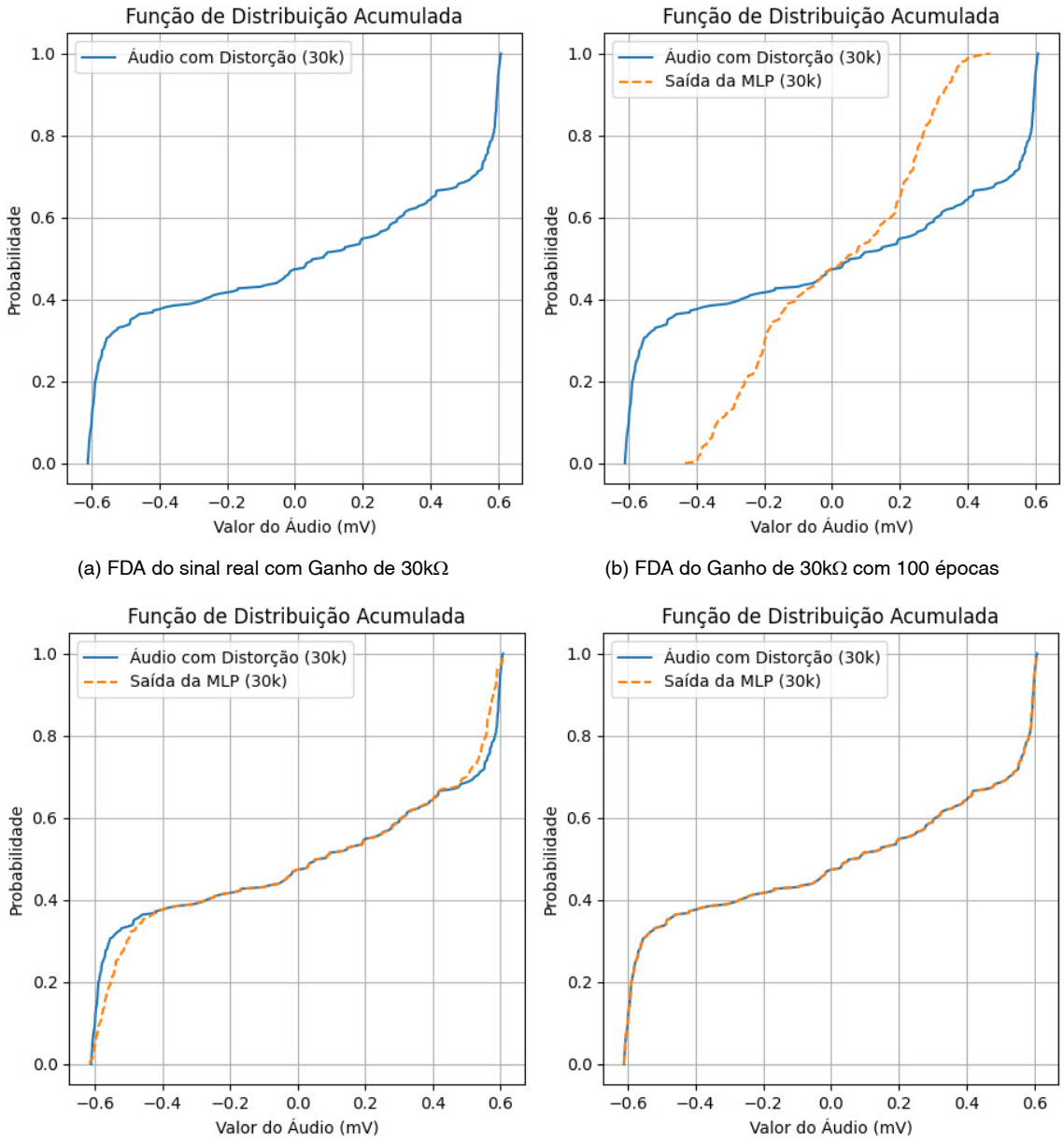
(c) FDA do Ganho de 10kΩ com 200 épocas

(d) FDA do Ganho de 10kΩ com 300 épocas

Figura 6 - Função de distribuição acumulado para o sinal de áudio com ganho de 10kΩ real e a saída da MLP com 100, 200 e 300 épocas
Fonte: elaborado pelo(as) autores(as).

Treinamento do modelo

O modelo MLP foi treinado utilizando o otimizador Adam (Kingma; Ba, 2017; Zhang *et al.*, 2023) por 500 épocas, seguindo um procedimento de treinamento robusto. É importante salientar que o treinamento do modelo foi realizado executando o código cinco vezes, cada uma com duração de 100 épocas. Em cada execução, os valores da taxa de aprendizado foram redefinidos, garantindo uma abordagem mais ampla na exploração do espaço de parâmetros e promovendo a convergência do modelo em diferentes cenários.



(a) FDA do sinal real com Ganho de 30kΩ

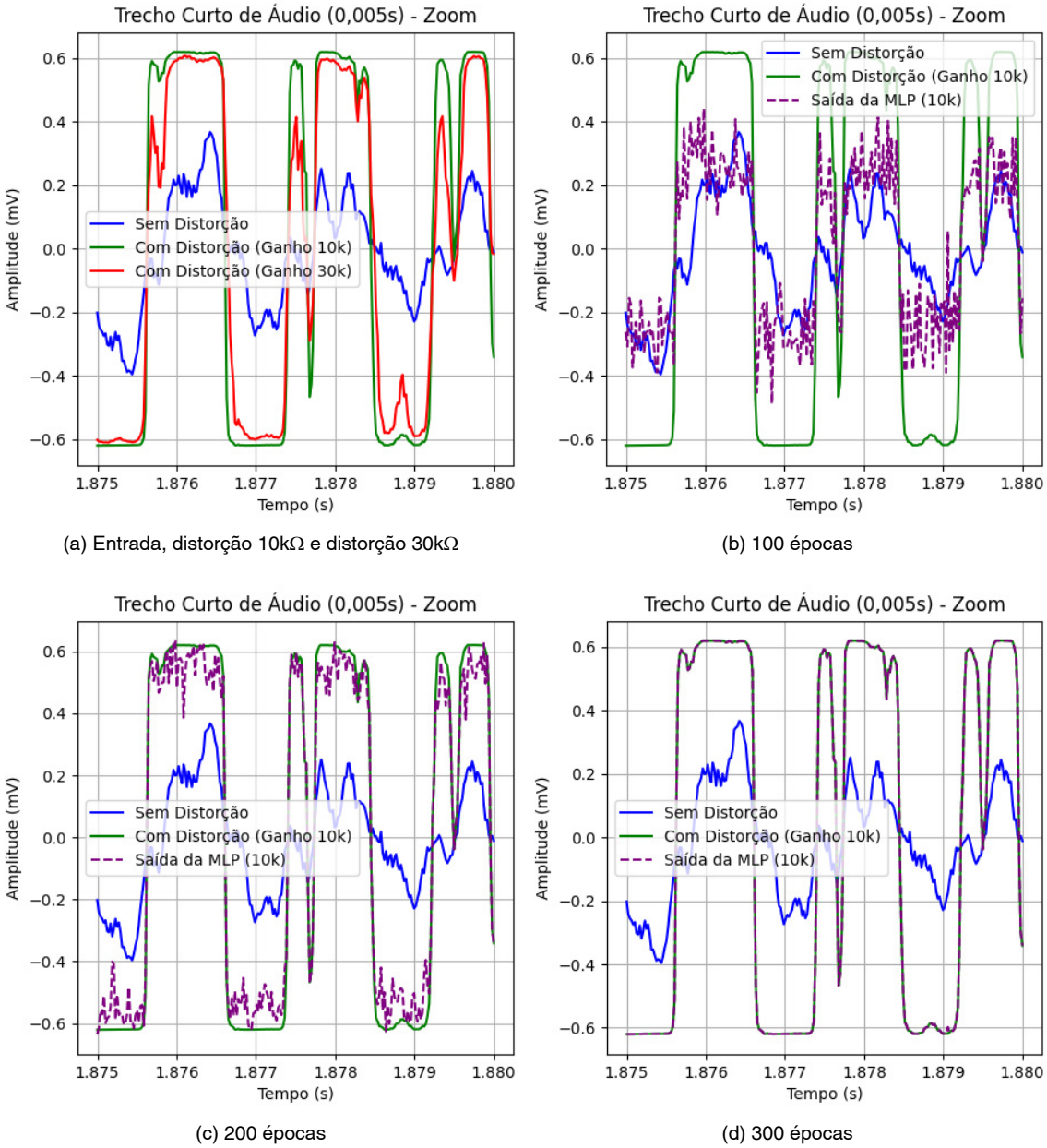
(b) FDA do Ganho de 30kΩ com 100 épocas

(c) FDA do Ganho de 30kΩ com 200 épocas

(d) FDA do Ganho de 30kΩ com 300 épocas

Figura 7 - Função de distribuição Acumulado para o sinal de áudio com ganho de 30kΩ real e a saída da MLP com 100, 200 e 300 épocas
Fonte: elaborado pelo(as) autores(as).

Durante o treinamento, a função de perda escolhida foi o Erro Quadrático Médio (EQM), a qual foi continuamente minimizada para ajustar os parâmetros do modelo. Na Figura 5, pode-se visualizar os valores de perda e de acurácia no decorrer das épocas de treinamento para os ganhos de 10 kΩ e 30 kΩ. Observa-se uma convergência gradual do modelo, indicada pela estabilização da perda e pelo aumento consistente da acurácia com o progresso das épocas.



(a) Entrada, distorção 10kΩ e distorção 30kΩ

(b) 100 épocas

(c) 200 épocas

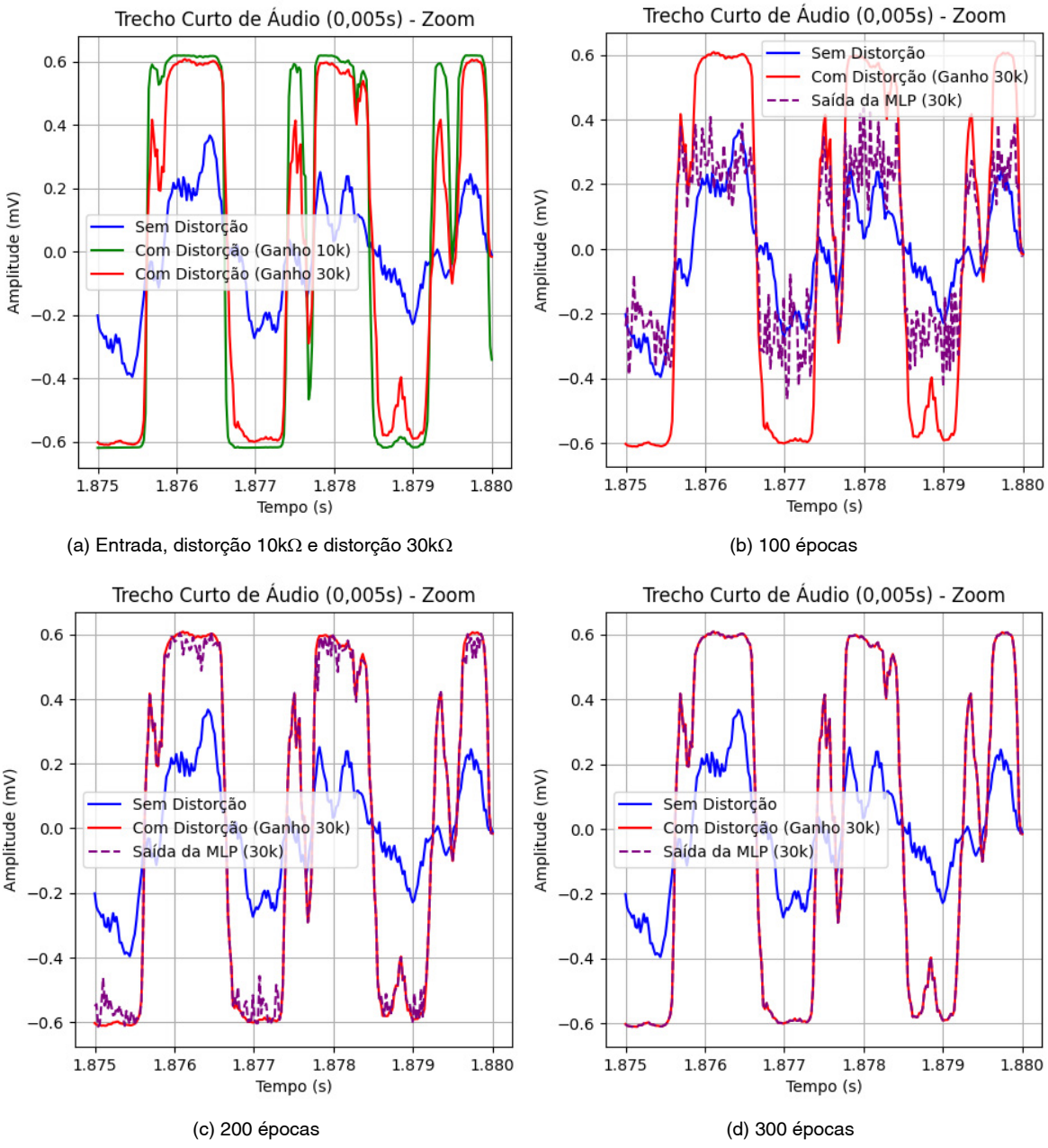
(d) 300 épocas

Figura 8 - Comparação entre as saídas da MLP, o áudio com distorção e o áudio sem distorção, para o ganho de 10kΩ
Fonte: elaborado pelo(as) autores(as).

Avaliação do modelo

A avaliação do desempenho do modelo treinado é um passo crítico para garantir sua capacidade de generalização e aplicabilidade em cenários do mundo real. Neste

contexto, foram empregadas métricas objetivas para quantificar a qualidade da saída da MLP em relação ao áudio de entrada distorcido.



(a) Entrada, distorção 10kΩ e distorção 30kΩ

(b) 100 épocas

(c) 200 épocas

(d) 300 épocas

Figura 9 - Comparação entre as saídas da MLP, o áudio com distorção e o áudio sem distorção, para o ganho de 30kΩ

Fonte: elaborado pelo(as) autores(as).

Primeiramente, calculamos a função de perda EQM (Erro Quadrático Médio) entre a saída da MLP e o áudio distorcido. Essa métrica fornece uma medida direta da diferença entre os sinais, permitindo uma avaliação quantitativa do desempenho do modelo em minimizar a distorção.

Além disso, realizamos o Teste Kolmogorov-Smirnov (KS), uma técnica estatística amplamente utilizada para comparar duas distribuições de dados. Neste caso, comparamos a distribuição da saída da MLP com a distribuição do áudio distorcido. O teste KS

fornece uma medida objetiva da semelhança entre as distribuições, permitindo avaliar a fidelidade da saída do modelo em replicar as características do áudio original.

As Figuras 6 e 7 apresentam as funções de distribuição acumulada para os ganhos de 10 kΩ e 30 kΩ, respectivamente. Esses gráficos oferecem insights sobre a correspondência entre as distribuições, auxiliando na análise da fidelidade da saída da MLP em relação ao áudio distorcido.

Comparação entre saída da MLP e áudios com distorção

A comparação entre a saída da MLP e os áudios com distorção é essencial para verificar a capacidade do modelo em replicar com precisão as características do sinal de entrada. Com base nos resultados obtidos no teste KS, observa-se uma diferença estatisticamente significativa entre a distribuição da saída da MLP e a distribuição do áudio distorcido. No entanto, uma análise visual mais aprofundada das Figuras 8 e 9 revela que a saída da MLP é altamente semelhante ao áudio distorcido.

Essa semelhança visual entre a saída da MLP e o áudio distorcido sugere que o modelo foi eficaz na captura e reprodução das distorções presentes no sinal de entrada. As figuras apresentam as comparações entre as saídas da MLP, os áudios com distorção e os áudios sem distorção para os ganhos de 10 kΩ e 30 kΩ, respectivamente.

As análises visuais dessas figuras oferecem uma perspectiva valiosa sobre a capacidade do modelo em preservar as características essenciais do sinal de entrada, mesmo após o processo de distorção. Esses resultados corroboram a eficácia do modelo MLP na tarefa de modelagem de distorção de áudio.

Resultados de avaliação

Os resultados obtidos na avaliação do modelo para os ganhos de 10kΩ e 30kΩ são apresentados nas Tabelas 1 e 2, respectivamente. Esses resultados são fundamentais para compreender o desempenho do modelo em diferentes condições de treinamento.

Épocas	Teste KS	Valor-p	MSE
100	0.322905	0	0.080315
200	0.116924	0	0.002267
300	0.003869	0	$3.169582 \cdot 10^{-7}$
400	0.003216	0.0000006	$6.741544 \cdot 10^{-9}$
500	0.003034	0.0000034	$6.235293 \cdot 10^{-9}$

Tabela 1 - Resultados de Avaliação para Ganho de 10 kΩ
Fonte: elaborado pelos(as) autores(as).

Épocas	Teste KS	Valor-p	MSE
100	0.2045111	0	0.0313119
200	0.0211680	0	0.0001642
300	0.0004034	0.99980123	$6.259833 \cdot 10^{-9}$
400	0.0004291	0.99936664	$6.068709 \cdot 10^{-9}$
500	0.0004416	0.99896655	$6.044989 \cdot 10^{-9}$

Tabela 2 - Resultados de Avaliação para Ganho de 30 kΩ
Fonte: elaborado pelos(as) autores(as).

As Tabelas fornecem uma análise detalhada dos valores de teste KS, valor- e EQM para diferentes números de épocas de treinamento. Observa-se que, à medida que o número de épocas aumenta, o desempenho do modelo melhora, como evidenciado pela diminuição dos valores de teste KS e EQM.

Esses resultados destacam a importância do treinamento adequado do modelo, bem como a necessidade de monitorar seu desempenho ao longo das épocas para garantir resultados confiáveis e precisos.

Considerações finais

Este estudo investigou a modelagem de efeitos de distorção em sinais de guitarra elétrica por meio de uma rede MLP (*Multilayer Perceptron*), desde a conceituação teórica da distorção de áudio até a implementação prática da arquitetura MLP e a análise dos resultados obtidos.

Os experimentos realizados demonstraram que a MLP é capaz de reproduzir o efeito de distorção em sinais de guitarra elétrica com um bom desempenho. Essa constatação foi corroborada pela comparação direta entre os resultados gerados pela MLP e os sinais de referência. A análise do erro quadrático médio (EQM) também corrobora esse bom desempenho, com valores próximos de zero, indicando boa correspondência entre os sinais gerados pela rede e os originais.

Os resultados obtidos no teste KS revelaram diferenças estatisticamente significativas entre as distribuições dos sinais de saída da MLP e os áudios distorcidos. Contudo, os valores de KS foram extremamente baixos, sugerindo uma alta semelhança entre essas distribuições e, consequentemente, indicando que o modelo foi capaz de capturar efetivamente as características das distorções presentes nos sinais de entrada.

Adicionalmente, os valores de EQM obtidos para ambos os conjuntos de dados de treinamento foram bastante reduzidos, indicando precisão considerável na reprodução dos sinais de saída desejados pela MLP. A análise das Funções de Distribuição Acumulada (FDAs) também corroborou esses resultados, revelando alta similaridade entre as distribuições dos sinais de saída da MLP e dos áudios distorcidos originais.

Assim, este trabalho representa um avanço significativo no campo da modelagem de efeitos de distorção em áudio, com potenciais aplicações que abrangem desde a indústria musical até a produção de áudio. Possíveis direções futuras de pesquisa incluem explorar as capacidades de outras arquiteturas de redes neurais na modelagem de efeitos de áudio complexos e avaliar a metodologia proposta em outros contextos que envolvam sistemas dinâmicos não lineares.

Referências

BISHOP, C. M. *Pattern recognition and machine learning*. Cambridge: Springer, 2006. v. 4.

BRUNETTO, L. F. M.; SCHMIDT, C. E.; DALLA'ROSA, A. *Seleção de hiperparâmetros para uma rede multi-layer perceptron aplicada na predição do preço da soja*. Revista Tecnia, Goiânia, v. 8, n. 2, p. 1-15, 2023.

LTSPICE. [*S. l.*]: Analog, 2023. Disponível em https://www.analog.com/en/design-center/design-tools-and-calculators/ltspice-simulator.html. Acesso em: 31 maio 2023.

ENGEL, J.; RESNICK, C.; ROBERTS, A.; DIELEMAN, S.; NOROUZI, M.; ECK, D.; SIMONYAN, K. *Neural audio synthesis of musical notes with wavenet autoencoders*. PMLR. International Conference on Machine Learning, p. 1068-1077, 2017.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F. *Inteligência artificial*: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. Arxiv, [*s. l.*], v. 1, p. 1-15, 2017.

KULESHOV, V.; ENAM, S. Z.; ERMON, S. *Audio super resolution using neural networks*. Arxiv, [*s. l.*], p. 1-8, 2017. Disponível em: https://arxiv.org/abs/1708.00853. Acesso em: 5 set. 2025.

PURWINS, H.; LI, B.; VIRTANEN, T.; SCHLÜTER, J.; CHANG, S.-Y.; SAINATH, T. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, [*s. l.*], v. 13, n. 2, p. 206-219, 2019.

REISS, J. D.; MCPHERSON, A. *Audio effects*: theory, implementation and application. Florida: CRC Press, 2014.

ROSS, S. M. *Introduction to probability and statistics for engineers and scientists.* [Estados Unidos]: Academic Press, 2020.

SELF, D. *Small signal audio design*. 4 th. Waltham: Focal Press, 2023.

STEPHENS, M. A. *Edf statistics for goodness of fit and some comparisons*. Journal of the American statistical Association, Taylor & Francis, v. 69, n. 347, p. 730-737, 1974.

WASSERMAN, L. *All of statistics: a concise course in statistical inference*. Berlim: Springer Science & Business Media, 2004.

ZHANG, A.; LIPTON, Z. C.; LI, M.; SMOLA, A. J. *Dive into Deep Learning*. Cambridge: Cambridge University Press, 2023.

tecnia revista de educação, ciência e tecnologia do ifg

v. 10, Edição Especial 1 | 2025

Submetido 30/05/2024. Aprovado 24/04/2025
Avaliação: revisão duplo-anônimo

Adoção da Inteligência Artificial no Schema Matching: Um Levantamento Sistemático do Estado da Arte

ADOPTION OF ARTIFICIAL INTELLIGENCE IN SCHEMA MATCHING: A SYSTEMATIC SURVEY OF THE STATE OF THE AR

ADOPCIÓN DE INTELIGENCIA ARTIFICIAL EN LA COINCIDENCIA DE ESQUEMAS SCHEMA MATICHING: UN ESTUDIO SISTEMÁTICO DEL ESTADO DEL ARTE

Ricardo Henricki Dias Borges
Universidade Federal de Goiás
ricardoborges@ufg.br

Valdemar Vicente Graciano Neto
Universidade Federal de Goiás
valdemarneto@inf.ufg.br

Leonardo Andrade Ribeiro
Universidade Federal de Goiás
laribeiro@inf.ufg.br

Resumo

Com a crescente complexidade da integração de dados em razão do aumento em sua quantidade e diversidade, o *Schema Matching* desempenha um papel fundamental. Nesse cenário desafiador, a Inteligência Artificial (IA) surge como uma solução promissora para aprimorar a eficiência do *Schema Matching*. Este artigo apresenta os resultados de um mapeamento sistemático da literatura, investigando as técnicas e os algoritmos de IA mais utilizados em aplicações de *Schema Matching*. Os insights obtidos oferecem orientação valiosa para pesquisadores e profissionais que buscam aprimorar a integração de dados por meio do *Schema Matching*.

Palavras-chave: *Schema Matching*; inteligência artificial; mapeamento sistemático.

Abstract

The following text presents a synopsis of the abstract. Given the increasing intricacy of data integration, attributable to both the proliferation of data and the diversification of its characteristics, Schema Matching is a critical component of this process. In the context of this challenging scenario, the application of Artificial Intelligence. The advent of Artificial Intelligence (AI) has emerged as a promising solution to enhance the efficiency of Schema Matching. The present article This paper presents the findings of a systematic literature review that investigated Artificial Intelligence (AI) techniques and algorithms. This is the most common usage in Schema Matching applications. The insights obtained provide valuable guidance. This text is intended for researchers and professionals who are seeking to improve data integration through Schema Matching.

Keywords: Schema Matching; Artificial Intelligence; systematic review.

Resumen
Con la creciente complejidad de la integración de datos debido al aumento de su cantidad y diversidad, *Schema Matching* juega un papel clave. En este desafiante escenario, la Inteligencia Artificial (IA) emergecomo una solución prometedora para mejorar la eficiencia del *Schema Matching* . Este artículo presenta los resultados de un mapeo sistemático de la literatura, investigando las técnicas y algoritmos de IA más utilizados en aplicaciones de *Schema Matching* . Los conocimientos adquiridos proporcionan una valiosa orientación para los investigadores y profesionales que buscan mejorar la integración de datos a través de *Schema Matching* .

Palabras clave: Coincidencia de Esquemas; inteligencia artificial; mapeo sistemático.

Introdução

A integração de dados é crucial para muitas organizações que precisam combinar dados de diferentes fontes para transformá-los em informações. No entanto, a integração de dados pode ser um desafio, especialmente quando os dados provêm de fontes com esquemas diferentes. Realizada de forma desordenada, a integração de dados pode levar a erros e inconsistências. Nesse contexto, o *Schema Matching* se apresenta como uma abordagem essencial na integração de dados que consiste na identificação de correspondências entre esquemas de diferentes fontes de dados (Rahm; Bernstein, 2001). Esse processo consiste na identificação e no estabelecimento de correspondências entre os elementos de dois ou mais esquemas de dados diferentes. Essa correspondência permite a integração e o compartilhamento de informações entre sistemas ou fontes de dados heterogêneas, podendo ser aplicado de forma automática, o que leva a uma integração manual mais eficiente e precisa (Bilke; Naumann, 2005).

A contribuição da Inteligência Artificial (IA) para aprimorar o *Schema Matching* é cada vez mais proeminente. Por meio de técnicas de IA, é possível automatizar grande parte do processo de correspondência de esquemas (*Schema Matching*), reduzindo a dependência de intervenção manual intensiva. Isso não apenas acelera a integração de dados, mas também aprimora a precisão, identificando correspondências sutis que poderiam passar despercebidas de outra forma. A IA também tem a capacidade de aprender com as decisões de mapeamentos anteriores, refinando continuamente suas estratégias e adaptações à medida que novos dados e desafios surgem.

No entanto, um mapeamento sistemático sobre a aplicação da IA no *Schema Matching* é fundamental para compreender e sintetizar de maneira completa o estado atual do conhecimento nessa convergência. Dado o constante desenvolvimento das técnicas de IA, tal mapeamento favorecerá uma visão global de abordagens, metodologias e tendências específicas desse domínio, auxiliando na identificação de lacunas de pesquisa, pontos fortes e limitações.

A principal contribuição deste artigo é a apresentação dos resultados de uma abordagem sistemática para mapear as técnicas de IA aplicadas ao *Schema Matching*. No total, 644 estudos foram inicialmente identificados pela *string* de busca, dos quais 68 foram incluídos com base em critérios específicos de seleção e *snowballing*. Notavelmente, as subáreas de *Deep learning* (DL) e Natural Language Processing (NLP) emergiram como as mais amplamente utilizadas na literatura investigada. Além disso, outras técnicas não diretamente relacionadas à IA também foram aplicadas em conjunto com as técnicas de IA, evidenciando a interdisciplinaridade das abordagens utilizadas.

Este artigo está organizado da seguinte forma: inicialmente, são detalhados os métodos adotados no planejamento e na execução do protocolo do Mapeamento Sistemático da Literatura (MSL); em seguida, são apresentados os resultados obtidos ao longo da condução do MSL e a análise correspondente; posteriormente, descrevem-se as ameaças à validade do estudo; e, por fim, apresentam-se as considerações finais e as propostas para trabalhos futuros.

Metodologia

Para desenvolver esta pesquisa, foi utilizado o Mapeamento Sistemático da Literatura (MSL) de acordo com o protocolo proposto por Fabbri et al. (2013) e Petersen, Vakkalanka e Kuzniarz (2015). O protocolo é composto de três etapas principais: Planejamento, Condução e Publicação dos Resultados, apresentado na Figura 1.

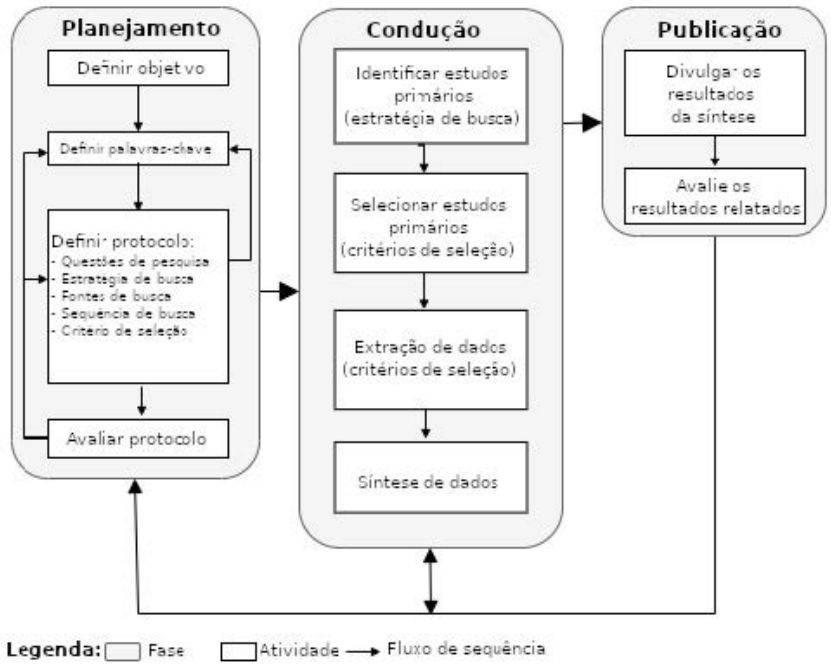


Figura 1 - Fases e atividades do MSL
Fonte: Elaborado pelos(as) autores(as) com base em Fabbri et al. (2013).

Questões de Pesquisa

As questões de pesquisa que expressam os objetivos deste mapeamento foram formuladas seguindo os critérios especificados por PICOC (*Population, Intervention, Comparison, Outcomes, Context*) definido em Budgen e Brereton (2006). Por tratar-se de um mapeamento, apenas PIO foi utilizado. A Tabela 1 exibe os detalhes.

População	*Schema Matching* (SM), Inteligência Artificial (IA), *Machine Learning* (ML), *Deep learning* (DL), Natural Language Processing (NLP)
Intervenção	Métodos/Técnicas/Tecnologias/Ferramentas/Padrões
Resultados	Técnicas ou algoritmos utilizados

Tabela 2 - Critérios do PIO
Fonte: Elaborado pelo(as) autores(as).

No âmbito da IA, foram exploradas três questões de pesquisa que abordam cada uma das suas subáreas: DL, ML e NLP. Além disso, uma quarta questão de pesquisa foi examinada em relação a técnicas que não estão estritamente vinculadas a nenhuma subárea específica.

A Figura 2 mostra um diagrama de Venn da relação entre as subáreas da IA.

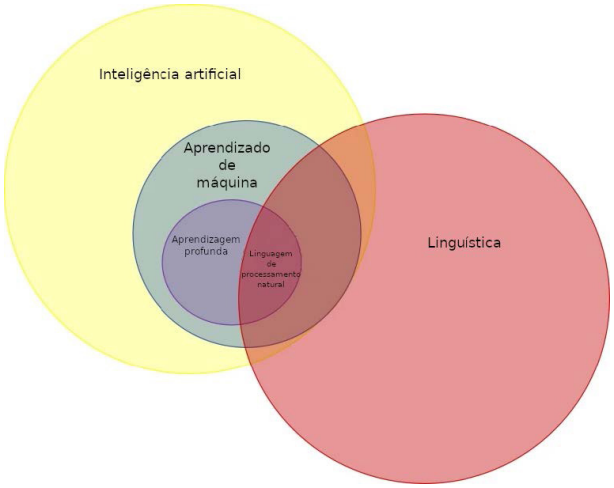


Figura 2 - Diagrama de Venn de subáreas da Inteligência Artificial
Fonte: Elaborado pelo(as) autores(as).

No diagrama da Figura 2 é possível observar que DL está situado dentro do domínio do ML, uma vez que DL é uma abordagem específica que se baseia em redes neurais profundas para o aprendizado de máquina. No entanto, o NLP é representado como uma área separada, pois é um campo especializado que se concentra exclusivamente no processamento e na compreensão da linguagem natural. Embora o NLP utilize técnicas de IA para seu funcionamento, é importante reconhecer sua singularidade devido ao foco específico na linguagem e em desafios linguísticos complexos. A seguir estão as questões de pesquisa (QP) levantadas durante esta fase e sua justificativa.

QP1: Quais técnicas específicas de *Deep learning* são utilizadas na aplicação de *Schema Matching*?

- Procura investigar quais técnicas exclusivamente de *Deep learning* estão sendo aplicadas em *Schema Matching*.

QP2: Quais técnicas específicas[1] de ML têm sido utilizadas na aplicação de *Schema Matching*?

1 Existem técnicas de ML que não são de DL

- Procura investigar quais técnicas exclusivamente de ML estão sendo aplicadas em *Schema Matching*.

  QP3: Quais técnicas específicas[2] de NLP têm sido utilizadas na aplicação de *Schema Matching*?

- Procura investigar quais técnicas exclusivamente de NLP estão sendo aplicadas em *Schema Matching*.

  QP4: Quais técnicas de IA têm sido utilizadas na aplicação de *Schema Matching*?

- Procura investigar quais técnicas exclusivamente de IA que não se enquadram em DL, ML e NLP estão sendo aplicadas em *Schema Matching*.

Identificação dos Estudos

O mapeamento foi conduzido utilizando uma estratégia de busca automática nas bases Scopus, IEEEXplore, ACM e Engineering Village. Além disso, foi aplicada a técnica do *Snowballing* para os estudos selecionados.

STRING DE BUSCA

Fase fundamental para incluir os termos que são pertinentes às questões de pesquisa, ou seja, aqueles relacionados às palavras-chave: inteligência artificial, *deep learning*, *Machine Learning* e *Schema Matching*. Além disso, foi incluído de sinônimos para ampliar a abrangência dos resultados.

A *string* de busca definida para este trabalho foi:

(“*deep learning*”) OR (“*Machine Learning*”) OR (“*Artificial Intelligence*”)) AND ((“*Schema Matching*”) OR (“*ontology matching*”) OR (“*Ontology Alignment*”)

Durante a aplicação da String de busca, houve aprimoramento contínuo, realizando ajustes e refinamentos com base nos resultados obtidos e na análise dos estudos encontrados. Ao longo desse processo de calibração da String, foi possível observar a evolução da String de busca, à medida que se faz o refinamento dos termos-chave, considerando sinônimos, variações linguísticas e acrônimos relevantes para a área de estudo.

Como mostrado a seguir:

Versão 1: (“*deep learning*”) AND ((“*Schema Matching*”) OR (“*ontology matching*”) OR (“entity matching”));

Versão 2: (“*deep learning*”) AND ((“*Schema Matching*”) OR (“*ontology matching*”));

Versão 3: ((“*deep learning*”) OR (“*Deep Neural Network*”)) AND ((“*schema maching*”) OR (“*ontology matching*”));

Versão 4: ((“*deep learning*”) OR (“*Machine Learning*”) OR (“*Artificial Intelligence*”)) AND ((“*Schema Matching*”) OR (“*ontology matching*”) OR (“*Ontology Alignment*”)).

2 Existem técnicas de NLP que não são de ML e DL

Critérios de Seleção

A etapa de definição dos critérios de seleção incluem os critérios de inclusão e exclusão, que são estabelecidos para orientar a seleção dos estudos que serão lidos na íntegra. A seguir, são apresentados os critérios de inclusão (**CI**) e os critérios de exclusão (**CE**) utilizados:

CI1: O estudo apresenta algoritmos ou técnicas de *Deep learning* aplicado a *Schema Matching*/*Ontology matching*/*Ontology Alignment*;

CI2: O estudo apresenta algoritmos ou técnicas de *Machine Learning* aplicado a *Schema Matching*/*Ontology matching*/*Ontology Alignment*;

CI3: O estudo apresenta algoritmos ou técnicas de NLP aplicado a *Schema Matching*/ *Ontology matching*/*Ontology Alignment*;

CI4: O estudo apresenta conjunto de algoritmos ou técnicas IA aplicado a *Schema Matching*/*Ontology matching*/*Ontology Alignment*.

CE1: O estudo não é um estudo primário;

CE2: O estudo não está disponível para acesso gratuito;

CE3: O estudo não está escrito em Inglês ou Português;

CE4: O estudo não foi publicado nos últimos 5 anos;

CE5: O estudo não apresenta algoritmos ou técnicas de Inteligencia Artificial aplicado a *Schema Matching*/*Ontology matching*/*Ontology Alignment*;

CE6: O estudo não é um artigo, ou seja, não foi revisado por partes, sendo uma literatura cinza.

Estudos que atenderam pelo menos um critério de inclusão foram incluídos na seleção inicial e os estudos que atendem a pelo menos um critério de exclusão são excluídos da seleção inicial. Inicialmente, foi aplicado o critério de exclusão de limitar os estudos aos últimos 10 anos. No entanto, percebeu-se que o tema em estudo tem passado por uma rápida evolução, especialmente a partir dos anos de 2018 e 2019, com o surgimento de estudos que aplicam diversas técnicas inovadoras. Diante dessa percepção, foi reduzido o escopo do mapeamento para os últimos 5 anos.

Por meio da extração inicial de 644 estudos e importado na ferramenta[3], foram eliminados os duplicados (151 estudos). Em seguida, foi realizada a leitura de todos os títulos, resumos e até conclusões de cada estudo, a fim de identificar os estudos que realmente atenderam ao tema proposto. Após isso, foi extraída a seleção inicial dos estudos para realizar a leitura completa. Nesta seleção foram extraídos inicialmente 82 estudos. Porém, com a leitura completa percebeu-se que ainda havia estudos que fugiam do tema proposto. Desta forma, os estudos não relevantes foram excluídos, restando ao final 59 estudos para extração.

Extração dos Dados

Durante essa etapa, interrompeu-se o uso da ferramenta, alterando o foco da manipulação dessa extração de dados através de planilhas, com o propósito de obter

3 A ferramenta utilizada foi a Parsif.al (https://parsif.al).

uma perspectiva sobre a evolução das seleções ao longo do protocolo. Essas planilhas foram criadas a partir da plataforma Parsif.al. Durante o processo de extração, todas as questões de pesquisa foram abordadas, e os estudos foram categorizados com base nas respostas para cada uma delas.

2.5. *Snowballing*

Com os estudos identificados, foi aplicado o método *Snowballing*, uma técnica de revisão sistemática que consiste em explorar tanto as referências de estudos selecionados (busca retroativa) quanto os trabalhos que os citaram (busca prospectiva). Esse processo permitiu analisar as referências bibliográficas dos estudos já encontrados e identificar trabalhos que os mencionavam, resultando na descoberta de 26 novos estudos relevantes que inicialmente não haviam sido incluídos. Esses 26 estudos passaram pelo mesmo protocolo de análise, e, ao final, apenas nove foram considerados aptos para o trabalho em questão, totalizando 68 estudos. A aplicação do método de *Snowballing* permitiu não apenas ampliar a base de estudos, mas também garantir uma cobertura mais completa e identificar novas fontes relevantes que enriqueceram a pesquisa. Além disso, a técnica facilitou a descoberta de conexões entre trabalhos e a compreensão da evolução das pesquisas na área.

Resultados

Os estudos relevantes estão organizados em tabelas durante as próximas subseções, onde é possível visualizar a referência completa do estudo e um código de identificação para os estudos acrescido de um valor numérico, que será utilizado como referência.

Quais técnicas ou algoritmos de *Deep learning* têm sido utilizados na aplicação de *Schema Matching*?

Entre as técnicas ou algoritmos da *Deep learning* encontradas destacam-se as Redes Neurais Siameses (SNN) com 23,3%, Redes de Memória de Curto Prazo Longa (LSTM) com 26,7% e Redes Neurais Convolucionais (CNN) com 6,7%. Além disso, o Multi-layer Perceptron (MLP) com 13,3%, Rede Neural Recorrente (RNN) e Gráfico de Redes Convolucionais (GCN) com 10,0%. Outros como a Rede Neural Recursiva (RvNNs), Competitive Learning e Multi-Input com 3,3%. Veja na Figura 3.

As proporções indicam que as LSTM e as SNNs são as técnicas de *Deep learning* mais prevalentes no contexto de *Schema Matching*, seguidas por MLPs, RNNs, GCNs e CNNs. Essas informações sugerem quais técnicas são mais frequentemente empregadas ou consideradas úteis para o *Schema Matching*.

SNN	CNN	LSTM	MLP	RNN	GCN	RvNNs, Competitive Learning e MIMO
E01, E02, E03, E04, E05, E06 e E07	E08 e E09	E10, E11, E12, E13, E14, E15 e E16	E17, E18, E19 e E08	E12, E13 e E16	E09, E20 e E21	E14, E07 e E22

Tabela 1 - Estudos selecionados da questão: Quais técnicas ou algoritmos de *Deep learning* têm sido utilizados na aplicação de *Schema Matching*?
Fonte: Elaborado pelos(as) autores(as).

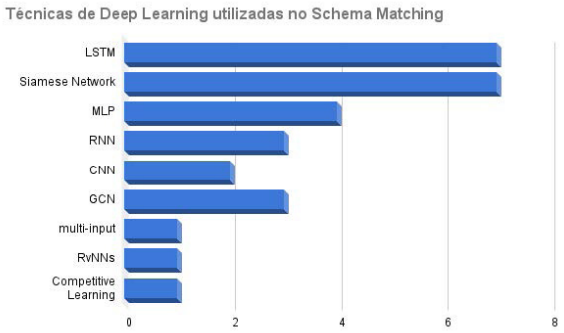


Figura 3 - Técnicas de *Deep learning* que têm sido utilizados no *Schema Matching*
Fonte: Elaborado pelos(as) autores(as).

Identificador	Referência
E01	(Iyer; Agarwal; Kumar, 2020b)
E02	(Iyer; Agarwal; KumaR, 2021)
E03	(Srinivas; Gale; Dolby, 2018)
E04	(Sun; Takeuchi; Yamasaki, 2020)
E05	(Chen *et al.*, 2021)
E06	(Iyer; Agarwal; Kumar, 2020a)
E07	(Xue *et al.*, 2021b)
E08	(Bento; Zouaq; Gagnon, 2020)
E09	(Wang *et al.*, 2022)
E10	(Jiang; XUE, 2020)
E11	(Maji; Rout; Choudhary, 2021)
E12	(Sun; Shen, 2022)
E13	(Mohamed *et al.*, 2022)
E14	(Chakraborty *et al.*, 2021)
E15	(Shraga; GAL, 2022)
E16	(Koutras *et al.*, 2020)
E17	(Xue *et al.*, 2021a)
E18	(Khan; Gubanov, 2020)
E19)	(Shraga; Gal; Roitman, 2020
E20	(Hao *et al.*, 2021)
E21	(Jurisch; Igler, 2019)
E22	(Hulsebos *et al.*, 2019)

Tabela 2 - Referência dos estudos selecionados para a questão: Quais técnicas ou algoritmos de *Deep learning* têm sido utilizados na aplicação de *Schema Matching*?
Fonte: Elaborado pelos(as) autores(as).

Quais técnicas ou algoritmos de *Machine Learning* têm sido utilizados na aplicação de *Schema Matching*?

A Figura 4 apresenta uma variedade de técnicas e algoritmos de ML identificados. Entre as técnicas mais comuns, destacam-se o classificador Naive Bayes, com 20,7%, a árvore de decisão (Decision Tree) com 17,2% e a floresta aleatória (Random Forest), com 13,8%. Além disso, como k-means com 13,8%, JRip, SVM com 3,4%,

6,9% respectivamente. Regressão com 10,3%, Adaboost com 6,9% e K-Nearest com 3,4%. Os resultados desses estudos também evidenciam a eficácia do uso de técnicas e algoritmos de ML em aplicações de *Schema Matching*, o que pode ajudar a melhorar a qualidade e a precisão de sistemas que lidam com dados heterogêneos.

Naive Bayes	Decision Tree	Random Forest	k-means	JRip, SVM e C4.5	Adaboost	K-Nearest	Regression, Gaussian mixture e Algoritmo evolutivo
E23, E24, E25, E26, E27, E28 e E29	E23, E30, E31 e E32	E33, E23, E34 e E35	E36, E28, E37 e E38	E26 e E31	E31 e E23	E23 e E31	E33, E23, E24, E67 e E68

Tabela 3 - Estudos selecionados da questão: Quais técnicas ou algoritmos de *Machine Learning* têm sido utilizados na aplicação de *Schema Matching*?
Fonte: Elaborado pelos(as) autores(as).

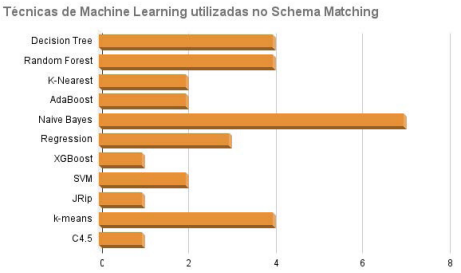


Figura 4 - Técnicas de *Machine Learning* utilizados no *Schema Matching*
Fonte: Elaborado pelos(as) autores(as).

Identificador	Referência
E23	(Lima *et al.*, 2020)
E24	(Bulygin, 2018)
E25	(Xue; Chen; Liu, 2021)
E26	(Laadhar *et al.*, 2019a)
E27	(Schmidts *et al.*, 2019)
E28	(Berlin; Motro, 2002)
E29	(Nikovski *et al.*, 2012)
E30	(Amrouch; Mostefai; Fahad, 2016)
E31	(Nezhadi; Shadgar; Osareh, 2011)
E32	(Rodrigues *et al.*, 2015)
E33	(Bulygin; Stupnikov, 2019)
E34	(Nkisi-Orji *et al.*, 2019)
E35	(Rodrigues; Silva, 2021)
E36	(Jiménez-Ruiz *et al.*, 2018b)
E37	(Belhadi *et al.*, 2023)
E38	(Li; Liu; Zhang, 2005)
E67	(Xue; Chen; Ren, 2019)
E68	(Przyborowski *et al.*, 2021)

Tabela 4 - Referência dos estudos selecionados para a questão: Quais técnicas ou algoritmos de *Machine Learning* têm sido utilizados na aplicação de *Schema Matching*?
Fonte: Elaborado pelos(as) autores(as).

Quais técnicas ou algoritmos de NLP têm sido utilizados na aplicação de *Schema Matching*?

A Figura 5 relata diversas técnicas e algoritmos de Processamento de Linguagem Natural (NLP). Entre as técnicas mais comuns encontradas destaca-se o BERT (Bidirectional Encoder Representations from Transformers), um modelo de linguagem prétreinado baseado na arquitetura Transformers, com 17,6% dos estudos. Além do BERT, outros algoritmos de NLP também foram identificados. Por exemplo, o Word2Vec com 41,2%, enquanto o FastText com 11,8%, o GloVe com 11,8%, TransE e StarSpace com 5,9%. Byte-Pair Encoding (BPE) e abordando também o uso de análise de sentimentos, com 2,9% e também técnicas de embedding sem especificar o algoritmo.

Portanto, o Word2Vec é a técnica mais amplamente utilizada, seguida pelo BERT, FastText e GloVe, com outras técnicas também sendo relevantes, embora com menor frequência de uso.

BERT	Word2Vec	FastText	GloVe	TransE	StarSpace	BPE	Embeddings (Sem especificar)
E11, E39, E40, E41, E42, E43	E44, E45, E46, E47, E34, E48, E42, E49, E50, E21, E51, E16, E52 e E53	E40, E54, E42 e E51	E55, E42, E51 e E52	E21, E56	E36 e E57	E58	E01, E02, E06 e E59

Tabela 5 - Estudos selecionados da questão: Quais técnicas ou algoritmos de NLP têm sido utilizados na aplicação de *Schema Matching*?
Fonte: Elaborado pelos(as) autores(as).

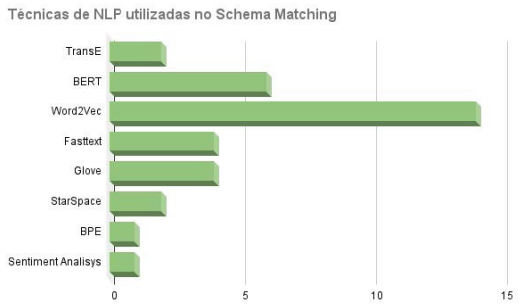


Figura 5 - Técnicas de NLP utilizados no *Schema Matching*
Fonte: Elaborado pelos(as) autores(as).

Identificador	Referência
E44	(Hertling; Portisch; Paulheim, 2020)
E45	(Bulygin, 2018)
E46)	(Teslya; Savosin, 2019
E47	(Li, 2020a)
E34	(Nkisi-Orji *et al.*, 2019)
E48	(Nozaki; Hochin; Nomiya, 2019)
E42	(Pan; Pan; Monti, 2022)
E49	(Li, 2020b)

E50	(Chen *et al.*, 2021)
E21	(Jurisch; Igler, 2019)
E51	(Cappuzzo, 2020)
E16	(Koutras *et al.*, 2020)
E52	(Hättasch *et al.*, 2022)
E53	(ZhanG *et al.*, 2014)
E40	(Yorsh *et al.*, 2022)
E54	(Dhouib; Zucker; Tettamanzi, 2019)
E55	(Ayala *et al.*, 2022)
E56	(Li *et al.*, 2019)
E36	(Jiménez-Ruiz *et al.*, 2018b)
E57	(Jiménez-Ruiz *et al.*, 2018a)
E58	(Zhang *et al.*, 2021)
E01	(Iyer; Agarwal; Kumar, 2020b)
E02	(Iyer; Agarwal; Kumar, 2021)
E06	(Iyer; Agarwal; Kumar, 2020a)
E59	(Li *et al.*, 2021)

Tabela 6 - Referência dos estudos selecionados para a questão: Quais técnicas ou algoritmos de NLP têm sido utilizados na aplicação de *Schema Matching*?
Fonte: Elaborado pelos(as) autores(as).

Quais técnicas ou algoritmos de Inteligência artificial têm sido utilizados na aplicação de *Schema Matching*?

A Figura 6 mostra que também foram encontradosestudos isolados, nos quais utilizaram algoritmos mais generalistas da própria IA. Em alguns desses estudos, não foi especificado qual o algoritmo utilizado, enquanto outros mencionaram apenas a subárea (ML ou DL) sem detalhar o algoritmo específico, com 71,4%, ML, 14,3% DL e o Artificial Bee Colonies com 14,3%.

IA Sem especificar o algoritmo	*Deep learning*	Grasshopper Algorithm	Bee Colonies
E60, E61, E62, E63 e E64	E51	E65	E66

Tabela 7 - Estudos selecionados
Fonte: Elaborado pelos(as) autores(as).

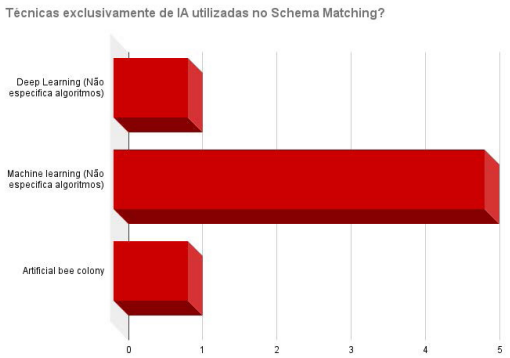


Figura 6 - Técnicas isoladas de IA utilizadas no *Schema Matching*
Fonte: Elaborado pelos(as) autores(as).

Identificador	Referência
E60	(Laadhar *et al.*, 2019b)
E61)	(Aissaoui; Oughdir, 2020
E62	(Doan *et al.*, 2020)
E63	(Mukherjee *et al.*, 2021)
E64	(Shraga, 2022)
E51	(Cappuzzo, 2020)
E65	(Lv, 2022)
E66	(Rangel *et al.*, 2015)

Tabela 8 - Quais técnicas ou algoritmos de Inteligência artificial têm sido utilizados na aplicação de *Schema Matching*?
Fonte: Elaborado pelos(as) autores(as).

DiscussãO

Durante a análise dos estudos, observou-se a importância e o amplo uso de IA aplicada no *Schema Matching* para a integração de dados. A complexidade e o custo de identificar correspondências em uma grande massa de dados pode expandir significativamente o impacto de falhas, tornando o custo e o tempo necessários para resolvê-lo um problema relevante, principalmente quando executada de forma manual.

Além disso, com os dados extraídos do mapeamento, foi constatado a ampla variedade de técnicas de IA aplicadas no *Schema Matching*, sendo uma grande tendência o uso de técnicas de DL e NLP, como na Figura 7. Além disso, também observou-se a presença de abordagens como similaridade e análise de ontologias semânticas junto com técnicas de IA e o uso de mais de uma técnica de IA durante o processo de *Schema Matching*. Essa diversidade de técnicas demonstra que ainda não existe uma única solução ideal para o problema de *Schema Matching*. Pelo contrário, é possível aplicar uma combinação de várias técnicas, adaptando-as ao contexto específico e às necessidades do projeto em questão. A escolha da abordagem mais adequada dependerá das características dos dados, dos requisitos do problema e das metas de qualidade estabelecidas.

Uma das principais lacunas reside na capacidade de compreender semântica e contextos complexos, uma vez que o *Schema Matching* muitas vezes vai além da mera

correspondência de palavras. A escalabilidade e o desempenho também são preocupações, especialmente com o aumento na quantidade de dados e esquemas a serem correspondidos. Além disso, manter a correspondência atualizada e automatizada é uma tarefa desafiadora. Isso inclui a necessidade de aprender com dados limitados e medir o desempenho de maneira objetiva. A diversidade na estrutura e no formato dos esquemas, juntamente com a interoperabilidade em ambientes heterogêneos, é outra área que precisa de atenção. Além disso, sistemas que possam aprender com o feedback dos usuários e melhorar continuamente a correspondência de esquemas são necessários.

No entanto, apesar do amplo uso das subáreas de DL e NLP na pesquisa de *Schema Matching*, essas lacunas servem como indicações claras das áreas onde a pesquisa neste campo ainda tem muito a explorar e aprimorar.

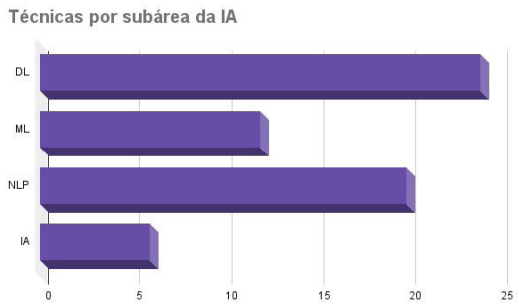


Figura 7 - Técnicas por subárea da IA utilizadas no *Schema Matching*
Fonte: Elaborado pelos(as) autores(as).

Ameaças à validade

Ameaças à validade deste estudo foram identificadas e categorizadasde acordo com Hyman (1982) e Wohlin et al. (2012).

Validade da construção: Pode haver uma possível exclusão de estudos relevantes. Para mitigar esse problema, foram realizadas revisões de cada etapa de condução e extração de dados deste estudo mais de uma vez. A falta da especificidade do tema também foi uma ameaça encontrada. Assim, foram encontrados estudos com foco em técnicas e algoritmos de IA aplicado ao *Schema Matching*, porém sem especificar o algoritmo utilizado. Sendo assim, conseguiu transpor esses resultados identificando pontos aplicados ao *Schema Matching*.

Validade Interna: Podem ter surgido em razão dos métodos de busca escolhidos. Por exemplo, a opção de não incluir algumas bibliotecas digitais pode levar à exclusão de estudos relevantes e ao número relativamente baixo de estudos incluídos. Para mitigar essa ameaça, o protocolo foi previamente avaliado pelos orientadores para identificar possíveis erros.

Validade Externa: Questões externas como a indisponibilidade de estudos foram resolvidas por meio de pesquisa, utilizando o Portal de Periódicos da CAPES. Também utilizou-se a literatura cinza, como Google Acadêmico, Google, Researchgate.

Validade de Conclusão: Possíveis ameaças estão relacionadas ao viés durante a condução e extração de dados, o que pode causar imprecisão na extração de dados, ameaçando a conclusão dos resultados do estudo. Para mitigar essas ameaças, foi apresentado os resultados em uma disciplina de Metogolodia Científica e a um grupo de estudo e pesquisa.

Considerações finais

Neste artigo, foram apresentados os resultados de um mapeamento sistemático sobre a aplicação da Inteligência Artificial (IA) no processo de *Schema Matching*. Foram explorados diversos estudos e pesquisas que utilizaram técnicas de IA em DL, ML e NLP, a fim de identificar o uso de técnicas como Redes Neurais Convolucionais (CNN), Multilayer Perceptron (MLP), dentre outros, para abordar o desafio de *Schema Matching*.

Durante a discussão dos resultados, foi possível observar o amplo uso de IA no contexto do *Schema Matching*. Assim, fica evidente que a IA pode ter um papel crucial na otimização do *Schema Matching*, permitindo a identificação de correspondências entre elementos de esquemas heterogêneos. Essas descobertas enfatizam a necessidade de continuar explorando e aprimorando as técnicas de IA aplicadas ao *Schema Matching*, com o objetivo de compará-las com abordagens não baseadas em IA. Esse enfoque contínuo na evolução das técnicas de IA certamente contribuirá para um avanço significativo no campo do *Schema Matching* e suas aplicações práticas.

Para futuras pesquisas, é importante explorar o desenvolvimento de novos modelos de IA personalizados para abordar os desafios do *Schema Matching*, como o uso de abordagens híbridas que combinam técnicas de IA com métodos tradicionais de *Schema Matching*, que podem oferecer um potencial significativo. A garantia de segurança e privacidade durante o processo de *Schema Matching* para dados críticos também são desafios adicionais que merecem atenção.

Referências

AISSAOUI, O. E.; OUGHDIR, L. A learning style-based ontology matching to enhance learning resources recommendation. *In*: IEEE. *2020 1st international conference on innovative research in applied science, engineering and technology (IRASET)*. [*S. l.*], p. 1-7, 2020.

AMROUCH, S.; MOSTEFAI, S.; FAHAD, M. Decision trees in automatic ontology matching. *International Journal of Metadata, Semantics and Ontologies*, [*s. l.*], v. 11, n. 3, p. 180-190, 2016.

AYALA, D.; HERNÁNDEZ, I.; RUIZ, D.; RAHM, E. Leapme: Learning-based property matching with embeddings. *Data & Knowledge Engineering*, Amsterdã, v. 137, 2022.

BELHADI, A.; DJENOURI, Y.; SRIVASTAVA, G.; LIN, J. C.-W. Fast and accurate framework for ontology matching in web of things. *ACM Transactions on Asian and Low-Resource Language Information Processing*, New York, v. 22, n. 5, p. 1-19, 2023.

BENTO, A.; ZOUAQ, A.; GAGNON, M. Ontology matching using convolutional neural networks. *In*: SYMPOSIUM ON APPLIED COMPUTING, 35., 2020, Brno. *Proceedings* [...]. New York: ACM, 2020.

BERLIN, J.; MOTRO, A. Database schema matching using machine learning with feature selection. *In*: SPACCAPIETRA, S.; MARCH, S. T.; KAMBAYASHI, Y. (ed.). *Conceptual Modeling*: ER 2002. Berlin; Heidelberg: Springer, 2002.

BILKE, A.; NAUMANN, F. Schema matching using duplicates. *In*: INTERNATIONAL CONFERENCE ON DATA ENGINEERING, 21., 2005, Tóquio. *Proceedings* [...]. Washington, DC: IEEE Computer Society, 2005. p. 69-80.

BUDGEN, D.; BRERETON, P. Performing systematic literature reviews in software engineering. *In*: INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING, 28., 2006, Xangai. *Proceedings* [...]. New York: ACM, 2006. p. 1051-1052.

BULYGIN, L. Combining lexical and semantic similarity measures with machine learning approach for ontology and schema matching problem. *Procedia Computer Science*, Amsterdã, v. 128, p. 1370-1379, 2018.

BULYGIN, L.; STUPNIKOV, S. A. Applying of machine learning techniques to combine string-based, language-based and structure-based similarity measures for ontology matching. *In*: IEEE CONFERENCE OF RUSSIAN YOUNG RESEARCHERS IN ELECTRICAL AND ELECTRONIC ENGINEERING (EIConRus), 2019, São Petersburgo. *Proceedings* [...]. Piscataway: IEEE, 2019.

CAPPUZZO, R. *Creating embeddings of heterogeneous relational datasets for data integration tasks*. 2020. Dissertação (Mestrado em Engenharia da Computação) – Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Padova, Pádua, 2020.

CHAKRABORTY, J. *et al*. Onto-connect: Unsupervised ontology alignment with recursive neural network. *In*: INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 2021, Xi'an. *Proceedings* [...]. New York: ACM, 2021. p. 2363-2366.

CHEN, J. *et al*. Augmenting ontology alignment by semantic embedding and distant supervision. *In*: HOSE, K. *et al*. (ed.). *The Semantic Web – ISWC 2021*. Cham: Springer, 2021. p. 124-140. (Lecture Notes in Computer Science, v. 12922).

DHOUIB, M. T.; ZUCKER, C. F.; TETTAMANZI, A. G. An ontology alignment approach combining word embedding and the radius measure. *In*: HITZLER, P. *et al*. (ed.). *Knowledge Engineering and Knowledge Management*. Cham: Springer, 2019. p. 115-130.

DOAN, A. *et al*. Magellan: toward building ecosystems of entity matching solutions. *Communications of the ACM*, New York, v. 63, n. 8, p. 83-91, ago. 2020.

FABBRI, S. C. P. F. *et al*. Externalising tacit knowledge of the systematic review process. *IET Software*, [*s. l.*], v. 7, n. 6, p. 298-307, 2013. DOI: 10.1049/iet-sen.2013.0029.

HAO, J. *et al*. Medto: Medical data to ontology matching using hybrid graph neural networks. *In*: INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 2021, Xi'an. *Proceedings* [...]. New York: ACM, 2021. p. 757-770.

HÄTTASCH, B.; TRUONG-NGOC, M.; SCHMIDT, A.; BINNIG, C. *It's ai match: A two-step approach for schema matching using embeddings*. 2022. arXiv:2203.04366.

HERTLING, S.; PORTISCH, J.; PAULHEIM, H. *Supervised ontology and instance matching with melt*. 2020. arXiv:2009.11102.

HULSEBOS, M. *et al*. Sherlock: A deep learning approach to semantic data type detection. *In*: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY & DATA MINING, 25., 2019, Anchorage. *Proceedings* [...]. New York: ACM, 2019. p. 1514-1524.

HYMAN, R. [Resenha de] Quasi-experimentation: Design and analysis issues for field settings. *Journal of Personality Assessment*, v. 46, n. 1, p. 96-97, 1982.

IYER, V.; AGARWAL, A.; KUMAR, H. *Multifaceted context representation using dual attention for ontology alignment*. arXiv:2010.11721, 2020.

IYER, V.; AGARWAL, A.; KUMAR, H. *Multifaceted context representation using dual attention for ontology alignment*.[1] 2020. arXiv:2010.11721. Disponível em: https://arxiv.org/abs/2010.11721. Acesso em: 15 set. 2025.

IYER, V.; AGARWAL, A.; KUMAR, H. Veealign: multifaceted context representation using dual attention for ontology alignment.[2] *In*: THE WEB CONFERENCE, 2021, Liubliana. *Companion Proceedings* [...]. New York: ACM, 2021. p. 317-321.

JIANG, C.; XUE, X. Matching biomedical ontologies with long short-term memory networks. *In*: IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICINE (BIBM), 2020, Seul. *Proceedings* [...]. Piscataway: IEEE, 2020. p. 1845-1852.

JIMÉNEZ-RUIZ, E. *et al*. *Breaking-down the ontology alignment task with a lexical index and neural embeddings*. 2018. arXiv:1805.12402. Disponível em: https://arxiv.org/abs/1805.12402. Acesso em: 15 set. 2025.

JIMÉNEZ-RUIZ, E. *et al*. We divide, you conquer: from large-scale ontology alignment to manageable sub-tasks with a lexical index and neural embeddings. *In*: INTERNATIONAL WORKSHOP ON ONTOLOGY MATCHING, 13., 2018, Monterey. *Proceedings* [...]. Aachen: CEUR-WS.org, 2018. p. 13-24. (CEUR Workshop Proceedings, v. 2288).

JURISCH, M.; IGLER, B. Graph-convolution-based classification for ontology alignment change prediction. *In*: INTERNATIONAL WORKSHOP ON ONTOLOGY MATCHING, 14., 2019, Auckland. *Proceedings* [...]. Aachen: CEUR-WS.org, 2019. p. 61-72. (CEUR Workshop Proceedings, v. 2536).

KHAN, R.; GUBANOV, M. Weblens: Towards web-scale data integration, training the models. *In*: IEEE INTERNATIONAL CONFERENCE ON DATA ENGINEERING, 36., 2020, Dallas. *Proceedings* [...]. Piscataway: IEEE, 2020. p. 1757-1760.

KOUTRAS, C. *et al*. Rema: Graph embeddings-based relational schema matching.[3] *In*: WORKSHOPS OF THE EDBT/ICDT JOINT CONFERENCE, 2020, Copenhague. *Proceedings* [...]. Aachen: CEUR-WS.org, 2020. p. 219-226. (CEUR Workshop Proceedings, v. 2592).

LAADHAR, A. *et al*. Partitioning and local matching learning of large biomedical ontologies. *In*: INTERNATIONAL CONFERENCE ON EVALUATION OF NOVEL APPROACHES TO SOFTWARE ENGINEERING, 14., 2019, Heraclião. *Proceedings* [...]. Setúbal: SciTePress, 2019. p. 310-317.

LAADHAR, A. *et al*. Pomap++ results for oaei 2019: fully automated machine learning approach for ontology matching. *In*: INTERNATIONAL WORKSHOP ON ONTOLOGY MATCHING, 14., 2019, Auckland. *Proceedings* [...]. Aachen: CEUR-WS.org, 2019. p. 169-174. (CEUR Workshop Proceedings, v. 2536).

LI, G. Deepfca: Matching biomedical ontologies using formal concept analysis embedding techniques. *In*: IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICINE (BIBM), 2020, Seul. *Proceedings* [...]. Piscataway: IEEE, 2020. p. 1829-1836.

LI, G. Improving biomedical ontology matching using domain-specific word embeddings. *In*: IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICINE (BIBM), 2020, Seul. *Proceedings* [...]. Piscataway: IEEE, 2020. p. 1837-1844.

LI, W. *et al*. Multi-view embedding for biomedical ontology matching. *In*: INTERNATIONAL WORKSHOP ON ONTOLOGY MATCHING, 14., 2019, Auckland. *Proceedings* [...]. Aachen: CEUR-WS.org, 2019. p. 13-24. (CEUR Workshop Proceedings, v. 2536).

LI, X. *et al*. Heterogeneous embeddings for relational data integration tasks. *In*: LIN, X.; ZHANG, Y.; ZHANG, W. (ed.). *Database Systems for Advanced Applications*. Cham: Springer, 2021. p. 45-59. (Lecture Notes in Computer Science, v. 12681).

LI, Y.; LIU, D.-B.; ZHANG, W.-M. Schema matching using neural network. *In*: INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND CYBERNETICS, 2005, Guangzhou. *Proceedings* [...]. Piscataway: IEEE, 2005. p. 4467-4472.

LIMA, B. *et al*. Learning reference alignments for ontology matching within and across domains. *In*: INTERNATIONAL WORKSHOP ON ONTOLOGY MATCHING, 15., 2020, Atenas. *Proceedings* [...]. Aachen: CEUR-WS.org, 2020. p. 85-96. (CEUR Workshop Proceedings, v. 2788).

LV, Z. An effective approach for large ontology matching using multi-objective grasshopper algorithm. *Scientific Reports*, London, v. 12, n. 1, p. 1-17, 2022.

MAJI, S.; ROUT, S. S.; CHOUDHARY, S. Dcom: A deep column mapper for semantic data type detection. 2021. arXiv:2106.12871. Disponível em: https://arxiv.org/abs/2106.12871. Acesso em: 15 set. 2025.

MOHAMED, A. *et al*. Schema matching based on deep learning using lstm model. *In*: INTERNATIONAL CONFERENCE ON INNOVATIVE RESEARCH IN APPLIED SCIENCE, ENGINEERING AND TECHNOLOGY, 2., 2022, Erode. *Proceedings* [...]. Piscataway: IEEE, 2022. p. 1-6.

MUKHERJEE, D.; BANDYOPADHYAY, A.; CHOWDHURY, R.; BHATTACHARYA, I. Learning knowledge graph for target-driven schema matching. *In*: INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 2021, Xi'an. *Proceedings* [...]. New York: ACM, 2021. p. 1351-1364.

MUYLAERT, R. Pandemia do novo coronavírus, Parte 6: inteligência artificial (NLP). *Marco Armello*, 19 ago. 2020. Disponível em: https://marcoarmello.wordpress.com/2020/08/19/coronavirus6/. Acesso em: 9 jul. 2023.

NEZHADI, A. H.; SHADGAR, B.; OSAREH, A. Ontology alignment using machine learning techniques. *International Journal of Computer Science & Information Technology*, Chennai, v. 3, n. 2, p. 139-149, 2011.

NIKOVSKI, D.; ESENTHER, A.; YE, X.; SHIBA, M.; TAKAYAMA, S. Bayesian networks for matcher composition in automatic schema matching. *In*: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND INFORMATION RETRIEVAL, 2012, Barcelona. *Proceedings* [...]. Setúbal: SciTePress, 2012. p. 64-73.

NKISI-ORJI, I. *et al*. Ontology alignment based on word embedding and random forest classification. *In*: LAUX, F.; BRUNZEL, M. (ed.). *Declarative AI: Theory, Systems, and Applications*. Cham: Springer, 2019. p. 147-163. (Lecture Notes in Computer Science, v. 11867).

NOZAKI, K.; HOCHIN, T.; NOMIYA, H. Semantic schema matching for string attribute with word vectors. *In*: INTERNATIONAL CONFERENCE ON INFORMATION NETWORKING (ICOIN), 2019, Kuala Lumpur. *Proceedings* [...]. Piscataway: IEEE, 2019. p. 488-493.

PAN, Z.; PAN, G.; MONTI, A. Semantic-similarity-based schema matching for management of building energy data. *Energies*, Basel, v. 15, n. 23, p. 8894, 2022.

PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, [*s. l.*], v. 64, p. 1-18, ago. 2015.

PRZYBOROWSKI, M.; PABIŚ, M.; JANUSZ, A.; ŚLĘZAK, D. *Schema matching using gaussian mixture models with wasserstein distance*. 2021. arXiv:2111.14244. Disponível em: https://arxiv.org/abs/2111.14244. Acesso em: 15 set. 2025.

RAHM, E.; BERNSTEIN, P. A. A survey of approaches to automatic schema matching. *The VLDB Journal*, [*s. l.*], v. 10, n. 4, p. 334-350, dez. 2001.

RANGEL, C. *et al*. An approach for the emerging ontology alignment based on the bees colonies. *In*: INTERNATIONAL CONFERENCE ON COMPUTATIONAL INTELLIGENCE (ICCI), 12., 2015, Madri. *Proceedings* [...]. Atlantis Press, 2015. p. 248-254.

RODRIGUES, D.; SILVA, A. A study on machine learning techniques for the schema matching network problem. *Journal of the Brazilian Computer Society*, Porto Alegre, v. 27, n. 1, p. 1-32, dez. 2021.

RODRIGUES, D.; SILVA, A. da; RODRIGUES, R.; SANTOS, E. dos. Using active learning techniques for improving database schema matching methods. *In*: IEEE SYMPOSIUM ON COMPUTER-BASED MEDICAL SYSTEMS, 28., 2015, São Carlos. *Proceedings* [...]. Piscataway: IEEE, 2015. p. 104-109.

SCHMIDTS, O.; KRAFT, B.; SIEBIGTEROTH, I.; ZÜNDORF, A. Schema matching with frequent changes on semi-structured input files: A machine learning approach on biological product data. *In*: INTERNATIONAL CONFERENCE ON ENTERPRISE INFORMATION SYSTEMS, 21., 2019, Heraclião. *Proceedings* [...]. Setúbal: SciTePress, 2019. v. 1, p. 208-215.

SHRAGA, R.; GAL, A. Humanal: Calibrating human matching beyond a single task. *In*: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 2022, Filadélfia. *Proceedings* [...]. New York: ACM, 2022. p. 1827-1840.

SHRAGA, R.; GAL, A. Powarematch: A quality-aware deep learning approach to improve human schema matching. *ACM Journal of Data and Information Quality*, New York, v. 14, n. 3, p. 1-27, set. 2022.

SHRAGA, R.; GAL, A.; ROITMAN, H. Adnev: Cross-domain schema matching using deep similarity matrix adjustment and evaluation. *Proceedings of the VLDB Endowment*, [*s. l.*], v. 13, n. 9, p. 1401-1415, 2020.

SRINIVAS, K.; GALE, A.; DOLBY, J. *Merging datasets through deep learning*. 2018. arXiv:1809.01604. Disponível em: https://arxiv.org/abs/1809.01604. Acesso em: 15 set. 2025.

SUN, C.; SHEN, D. Towards deep entity resolution via soft schema matching. *Neurocomputing*, v. 471, p. 107-117, fev. 2022.

SUN, J.; TAKEUCHI, S.; YAMASAKI, I. Few-shot ontology alignment model with attribute attentions. *In*: IEEE CONGRESS ON EVOLUTIONARY COMPUTATION (CEC), 2020, Glasgow. *Proceedings* [...]. Piscataway: IEEE, 2020. p. 1-8.

TESLYA, N.; SAVOSIN, S. Matching ontologies with word2vec-based neural network. *In*: ABRAHÃO, S.; ZELENKOV, Y. (ed.). *Software and Compilers for Embedded Systems*. Cham: Springer, 2019. p. 115-125. (Lecture Notes in Computer Science, v. 11789).

WANG, P.; ZOU, S.; LIU, J.; KE, W. Matching biomedical ontologies with gcn-based feature propagation. *Mathematical Biosciences and Engineering*, Springfield, v. 19, n. 8, p. 8479-8504, 2022.

WOHLIN, C. *et al*. *Experimentation in software engineering*. Berlin: Springer Science & Business Media, 2012.

XUE, X.; CHEN, D.; LIU, W. Naive bayesian classifier based semi-supervised learning for matching ontologies. *In*: IEEE INTERNATIONAL CONFERENCE ON COMPUTER SUPPORTED COOPERATIVE WORK IN DESIGN (CSCWD), 24., 2021, Dalian. *Proceedings* [...]. Piscataway: IEEE, 2021. p. 1305-1310.

XUE, X.; CHEN, J.; REN, A. Interactive ontology matching based on evolutionary algorithm. *In*: INTERNATIONAL CONFERENCE ON COMPUTATIONAL INTELLIGENCE AND SECURITY (CIS), 15., 2019, Macau. *Proceedings* [...]. Piscataway: IEEE, 2019. p. 1-5.

XUE, X. *et al*. Artificial neural network based sensor ontology matching technique. *In*: IEEE INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND INFORMATION SYSTEMS (ICAIIS), 2021, Chongqing. *Proceedings* [...]. Piscataway: IEEE, 2021. p. 1-6.

XUE, X. *et al*. Matching sensor ontologies through siamese neural networks without using reference alignment. *PeerJ Computer Science*, San Diego, v. 7, e602, 2021.

YORSH, U.; BEHR, A. S.; KOCKMANN, N.; HOLEŇA, M. *Text-to-ontology mapping via natural language processing models*. 2022. arXiv:2209.04944. Disponível em: https://arxiv.org/abs/2209.04944. Acesso em: 15 set. 2025.

ZHANG, J.; SHIN, B.; CHOI, J. D.; HO, J. C. *Smat: An attention-based deep learning solution to the automation of schema matching*. *In*: INTERNATIONAL CONFERENCE ON ASIAN DIGITAL LIBRARIES, 23., 2021, Virtual Event. *Proceedings* [...]. Cham: Springer, 2021. p. 3-17. (Lecture Notes in Computer Science, v. 13127).

ZHANG, Y. *et al*. Ontology matching with word embeddings. *In*: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 2014, Pequim. *Proceedings* [...]. Beijing: AAAI Press, 2014. p. 1-6.

INSTITUTO FEDERAL
DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
Goiás

tecnia

revista de educação, ciência e tecnologia do IFG

v. 10, Edição Especial 1 | 2025
ISSN: 2526–2130

FAPEG
Fundação de Amparo à Pesquisa do Estado de Goiás

editora ifg

editora@ifg.edu.br
editora.ifg.edu.br