Trabalho de Conclusão de Curso

Uma Pesquisa Sistemática sobre Defeitos em Sistemas de Aprendizagem de Máquina

Sandoval da Silva Almeida Junior

Orientador:

Prof. Dr. Erick de Andrade Barboza

Maceió, Junho de 2022

Sandoval da Silva Almeida Junior

Uma Pesquisa Sistemática sobre Defeitos em Sistemas de Aprendizagem de Máquina

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Engenharia de Computação do Instituto de Computação da Universidade Federal de Alagoas.

Orientador:

Prof. Dr. Erick de Andrade Barboza

Maceió, Junho de 2022

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Helena Cristina Pimentel do Vale – CRB4 –661

A447u Almeida Júnior, Sandoval da Silva.
Uma pesquisa sistemática sobre defeitos em sistemas de aprendizagem de máquina / Sandoval da Silva Almeida Junior. - 2024.
36 f : il.

Orientador: Erick de Andrade Barboza.
Monografia (Trabalho de Conclusão de Curso em Engenharia de Computação) – Universidade Federal de Alagoas, Instituto de Computação. Maceió, 2024.

Bibliografia: f. 32-36.

1. Aprendizagem de máquina. 2. Inteligência artificial. 3. Revisão sistemática da literatura. I. Título.

CDU: 004.8

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Engenharia de Computação do Instituto de Computação da Universidade Federal de Alagoas, aprovada pela comissão examinadora que abaixo assina.

Documento assinado digitalmente
gov.br
ERICK DE ANDRADE BARBOZA
Data: 27/11/2024 22:36:26-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Erick de Andrade Barboza - Orientador
Instituto de Computação
Universidade Federal de Alagoas

Documento assinado digitalmente
gov.br
BALDOINO FONSECA DOS SANTOS NETO
Data: 27/11/2024 17:55:28-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Baldoino Fonseca dos Santos Neto - Examinador
Universidade Federal de Alagoas

Documento assinado digitalmente
gov.br
MARCIO DE MEDEIROS RIBEIRO
Data: 27/11/2024 15:39:53-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Márcio de Medeiros Ribeiro - Examinador
Universidade Federal de Alagoas

Maceió, Junho de 2022

Agradecimentos

Em primeiro lugar, gostaria de expressar minha imensa gratidão ao professor Erick Barboza, que desempenhou um papel crucial na minha trajetória profissional. Foi ele quem abriu as portas do mercado de trabalho para mim, indicando-me como bolsista na Casa da Indústria de Alagoas, o que me proporcionou o início de uma carreira como Engenheiro de Dados. Além disso, sou profundamente grato por todos os ensinamentos que recebi durante a graduação e pelas orientações valiosas ao longo do TCC.

Também gostaria de agradecer aos meus amigos Tayco Murilo e Aldemir Melo, que estiveram comigo em cada etapa da nossa jornada universitária. Nossos sábados e domingos de estudos foram fundamentais para que eu conseguisse a aprovação em todas as disciplinas, e sem as nossas conversas e o apoio mútuo, certamente o caminho teria sido muito mais difícil.

Aos meus pais, agradeço de coração por todo o suporte que me proporcionaram ao longo dessa caminhada. Foram eles que me deram a base e a tranquilidade necessárias para que eu pudesse focar 100% nos meus estudos, sempre me incentivando e acreditando no meu potencial.

Gostaria ainda de fazer um agradecimento especial ao Coldplay e às suas músicas, que me ajudaram a relaxar nos momentos de maior ansiedade. Em muitos momentos, suas canções foram uma verdadeira trilha sonora de calma e inspiração durante essa jornada.

Por último, mas com igual importância, agradeço à minha namorada, Edrei Noemi, por sua incrível paciência, apoio constante, e por estar ao meu lado em todos os momentos. Seu incentivo, conselhos e presença foram essenciais para que eu conseguisse concluir essa etapa com sucesso.

Por fim, quero deixar claro que este diploma não é apenas meu, mas de todos que me ajudaram ao longo dessa caminhada. Sem o apoio, a dedicação e o carinho de cada um de vocês, esta conquista não teria sido possível. Muito obrigado!

“Nobody said it was easy.”

– Chris Martin

Resumo

Nos últimos anos, a integração de sistemas apoiados por aprendizagem de máquina (SAAM) tem se expandido em diversas áreas, impulsionando avanços tecnológicos em setores críticos como saúde, finanças, segurança e transporte. Contudo, à medida que esses sistemas se tornam cada vez mais complexos e difundidos, surgem também desafios significativos relacionados à confiabilidade e robustez das soluções implementadas. Problemas como erros de classificação, falhas de desempenho e vulnerabilidades a ataques adversários podem comprometer não apenas a integridade dos sistemas, mas também a segurança e a confiança nas decisões automatizadas. Nesse cenário, compreender e mitigar esses defeitos é crucial para garantir o avanço seguro e eficaz das tecnologias de aprendizagem de máquina.

Este Trabalho de Conclusão de Curso apresenta uma revisão sistemática da literatura sobre defeitos, erros e falhas que mais comprometem a integridade e desempenho de sistemas apoiados por aprendizagem de máquina (SAAM). A análise foca em identificar e categorizar os principais problemas relatados, bem como os artefatos gerados pela comunidade acadêmica em resposta a esses desafios. O estudo abrange uma ampla gama de artigos publicados nos últimos anos, destacando tendências predominantes na pesquisa de SAAM, como a robustez dos sistemas frente a ataques adversários, a otimização de algoritmos e a gestão de conjuntos de dados.

A partir de uma metodologia de pesquisa e seleção de artigos, problemas são identificados e classificados em subgrupos, revelando que questões como classificação de imagens, detecção de anomalias e defesa contra ataques adversários estão no centro das preocupações atuais. Adicionalmente, são examinados os artefatos gerados pelos estudos, tais como modelos de aprendizagem de máquina, algoritmos de otimização e conjuntos de dados, que são fundamentais para o avanço da pesquisa.

A revisão não apenas mapeia as falhas críticas que podem deteriorar os SAAM, mas também destaca as soluções propostas para mitigar esses riscos, oferecendo uma visão abrangente dos esforços de pesquisa e desenvolvimento na área. A compreensão dessas dinâmicas é essencial para pesquisadores, desenvolvedores e decisores que buscam explorar ou aprimorar a implementação de tecnologias de aprendizado de máquina em sistemas críticos.

Palavras-chave: Aprendizagem de Máquina; Inteligência Artificial; Revisão Sistemática.

Abstract

In recent years, the integration of systems supported by machine learning (SAAM) has expanded across various fields, driving technological advancements in critical sectors such as healthcare, finance, security, and transportation. However, as these systems become increasingly complex and widespread, significant challenges related to the reliability and robustness of implemented solutions also arise. Issues such as classification errors, performance failures, and vulnerabilities to adversarial attacks can compromise not only the integrity of these systems but also the safety and trust in automated decisions. In this context, understanding and mitigating these defects is crucial to ensuring the safe and effective advancement of machine learning technologies.

This Thesis presents a systematic literature review on defects, errors, and failures that most compromise the integrity and performance of machine learning-supported systems (SAAM). The analysis focuses on identifying and categorizing the main reported issues, as well as the artifacts produced by the academic community in response to these challenges. The study covers a wide range of articles published in recent years, highlighting predominant trends in SAAM research, such as system robustness against adversarial attacks, algorithm optimization, and data set management.

Based on a research and article selection methodology, issues are identified and classified into subgroups, revealing that topics like image classification, anomaly detection, and defense against adversarial attacks are central to current concerns. Additionally, the study examines the artifacts generated by these studies, such as machine learning models, optimization algorithms, and data sets, which are essential to advancing research.

The review not only maps the critical failures that can deteriorate SAAM but also highlights proposed solutions to mitigate these risks, providing a comprehensive view of research and development efforts in the field. Understanding these dynamics is essential for researchers, developers, and decision-makers seeking to explore or enhance the implementation of machine learning technologies in critical systems.

Keywords: Machine Learning; Artificial Intelligence; Systematic Review.

Sumário

Lista de Abreviaturas e Siglas

SAAM	Sistemas Apoiados por Aprendizagem de Máquina

Lista de Figuras

1

Introdução

1.1 Justificativa

O crescente uso de Sistemas Apoiados por Aprendizagem de Máquina (SAAM) em áreas como saúde, finanças, segurança e transporte tem transformado processos críticos e impactado diretamente o cotidiano das pessoas. Esses sistemas utilizam algoritmos de aprendizagem de máquina para realizar tarefas complexas, como diagnósticos médicos, previsão de fraudes financeiras, reconhecimento facial em sistemas de segurança e controle autônomo de veículos. Entretanto, a eficácia e a confiabilidade dessas tecnologias ainda enfrentam desafios significativos.

SAAMs são projetados para aprender a partir de grandes quantidades de dados e, com base nisso, tomar decisões ou realizar predições. No entanto, essas decisões podem ser afetadas por uma série de defeitos, como erros de classificação, falhas de desempenho e vulnerabilidades a ataques adversários. Em um exemplo prático, erros de diagnóstico em sistemas de apoio à decisão médica podem comprometer a saúde dos pacientes, resultando em diagnósticos incorretos ou tratamentos inadequados.

Por exemplo, em casos de câncer de mama, sistemas que utilizam algoritmos de aprendizado de máquina para análise de imagens de mamografia podem falhar ao classificar corretamente um tumor como maligno ou benigno. Isso pode levar a diagnósticos equivocados, resultando em tratamentos inadequados, como submeter uma paciente a uma mastectomia desnecessária ou, no extremo oposto, deixar de realizar intervenções necessárias em estágios iniciais da doença. Tais falhas evidenciam a importância de garantir a robustez e a precisão desses sistemas, especialmente em áreas onde erros podem ter impacto direto na vida dos indivíduos.

Já em um contexto de segurança, falhas em sistemas de reconhecimento facial podem gerar identificações equivocadas, o que pode ter sérias consequências para a privacidade e segurança de indivíduos.

A ausência de uma compreensão clara e sistemática sobre as falhas que afetam os SAAMs representa uma lacuna importante na literatura acadêmica e técnica. A literatura existente ainda

não oferece uma abordagem estruturada e abrangente que identifique e classifique as principais vulnerabilidades e defeitos nesses sistemas. Até o momento, não há uma revisão sistemática consolidada que organize o conhecimento sobre os problemas enfrentados pelos SAAMs, o que dificulta o avanço seguro e eficaz dessas tecnologias.

Este trabalho se justifica pela necessidade de conduzir uma pesquisa sistemática para identificar, catalogar e classificar os principais defeitos que impactam a qualidade e o desempenho dos sistemas de aprendizagem de máquina. Ao oferecer uma visão abrangente das falhas recorrentes, este estudo poderá contribuir para o desenvolvimento de soluções mais robustas e seguras. Profissionais e pesquisadores da área de computação, bem como outros setores impactados pela aprendizagem de máquina, poderão se beneficiar de uma base sólida de conhecimento, que servirá para nortear melhorias na concepção, implementação e manutenção desses sistemas.

1.2 Objetivos

1.2.1 Objetivo Geral

O objetivo geral deste trabalho é realizar uma revisão sistemática da literatura e desenvolver um catálogo de tipos de defeitos, erros e falhas que mais prejudicam a qualidade e o desempenho de sistemas apoiados por aprendizagem de máquina (SAAM) com isso o trabalho a seguir busca identificar e classificar os principais problemas encontrados nesses sistemas.

1.2.2 Objetivos Específicos

- Analisar e cassificar defeitos, erros e falhas afetam sistemas apoiados por aprendizagem de máquina (SAAM);
- Avaliar o Impacto dos Problemas Identificados;
- Catalogar os artefatos gerados pelos estudos;

1.3 Resultados esperados

O principal resultado esperado deste trabalho é oferecer aos pesquisadores e profissionais uma visão clara e abrangente das fragilidades que afetam os sistemas apoiados por aprendizagem de máquina (SAAM). Através da catalogação das principais vulnerabilidades relatadas na literatura acadêmica e técnica, busca-se identificar padrões críticos que comprometem a confiabilidade e o desempenho desses sistemas. Espera-se que essa análise não só revele os defeitos e falhas mais recorrentes, mas também sirva como uma base sólida para orientar futuras pesquisas, ajudando

no aprimoramento dos sistemas SAAM. Assim, os desenvolvedores e cientistas poderão implementar soluções mais robustas, minimizando falhas e maximizando a eficiência em ambientes reais.

1.4 Estrutura do Trabalho

O Capítulo 2 será dedicado à apresentação dos conceitos teóricos que fundamentam esta pesquisa, oferecendo um panorama sobre os temas principais, como a revisão sistemática da literatura e a área de Machine Learning.

No Capítulo 3, serão explicados os procedimentos adotados para realizar a revisão sistemática, detalhando as etapas envolvidas e as ferramentas utilizadas para a coleta e organização dos dados da pesquisa.

O Capítulo 4 trará uma análise dos resultados obtidos na revisão, respondendo às questões centrais do estudo com base nas publicações selecionadas. Além disso, será oferecida uma discussão sobre os achados, com foco no estado atual da área e apontamento de possíveis direções para pesquisas futuras.

Finalmente, no Capítulo 5, serão apresentadas as conclusões do trabalho, acompanhadas de recomendações para estudos posteriores.

2

Fundamentação Teórica

2.1 Introdução do capítulo

Este capítulo apresenta a fundamentação teórica necessária para compreender os conceitos e métodos sobre Machine Learning e Revisão Sistemática.

Primeiramente, é abordado os conceitos fundamentais de Machine Learning, os principais tipos de aprendizado, como aprendizado supervisionado, não supervisionado e por reforço.

Em seguida é abordado a metodologia de Revisão Sistemática, uma abordagem rigorosa e estruturada para a análise e síntese de estudos existentes sobre um determinado tema.

Como a proposta deste trabalho é catalogar os principais erros e falhas que mais deterioram sistemas apoiados por aprendizagem de máquina, é necessário ter um conhecimento breve sobre Machine Learning e Revisão Sistemática para compreender as etapas da pesquisa.

2.2 Machine Learning

O objetivo do apredizado de máquina é a construção de programas que melhorem seu desempenho automaticamente com a experiência.(MITCHELL, 1997).

As técnicas de Machine learning são orientadas a dados, isto é, aprendem automaticamente a partir de grandes volumes de dados e quanto maior o quantidade de exemplos para gerar o conhecimento melhor será o aprendizado pois dados mais precisos levam a generalizações mais precisas. São classificados três tipos principais de Aprendizado de Máquina: Supervisionado, Não Supervisionado e por Reforço.

Aprendizado supervisionado: Nesse método, cada exemplo fornecido ao algoritmo de aprendizado vem acompanhado da resposta esperada. Os exemplos são representados por vetores de valores e pelos rótulos das classes correspondentes. O objetivo do algoritmo é criar um classificador capaz de identificar corretamente a classe de novos exemplos que ainda não foram

rotulados. Esse tipo de aprendizagem é frequentemente associado a problemas de classificação e regressão, sendo o método de aprendizado mais comum. Exemplos de algoritmos que utilizam essa abordagem incluem o perceptron multicamadas e as redes neurais profundas.

Aprendizado não supervisionado: Fornece ao algoritmo exemplos não rotulados. Este algoritmo agrupa exemplos com base na semelhança de seus atributos. O algoritmo analisa os exemplos fornecidos e tenta determinar se algum deles pode ser agrupado de alguma forma, formando um agrupamento ou cluster.

Aprendizado por Reforço: Algoritmo não recebe a resposta correta mas recebe um sinal de reforço, de recompensa ou punição. O algoritmo faz uma hipótese baseado nos exemplos e determina se essa hipótese foi boa ou ruim. (LUDERMIR, 2021).

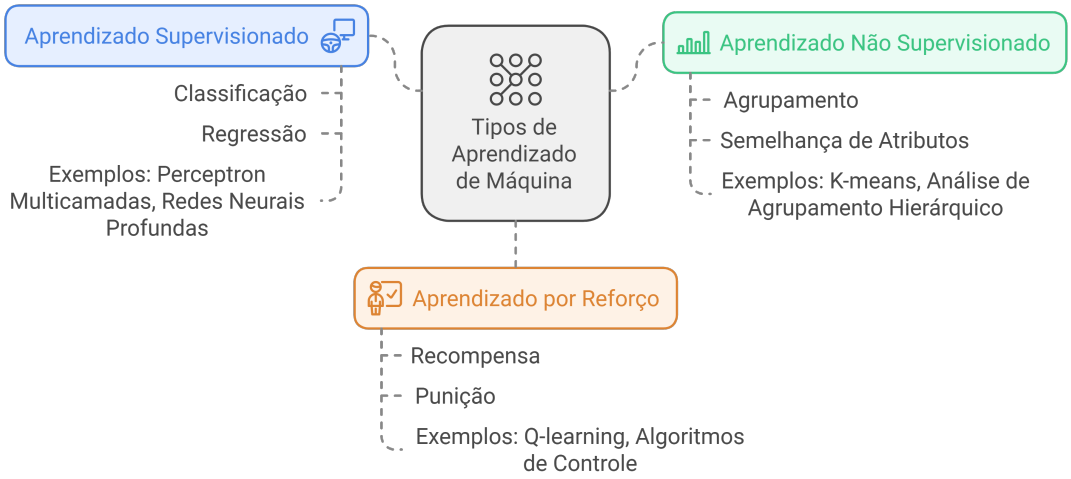

Figura 1: Tipos de Aprendizado de Máquina (Fonte: Elaborada pelo autor).

2.3 Revisão Sistemática

Uma revisão sistemática é uma abordagem rigorosa que visa identificar pesquisas relevantes sobre um tema específico usando métodos de pesquisa explícitos e sistemáticos. Pretende não só situar estes estudos, mas também avaliar a sua qualidade, validade e aplicabilidade no contexto da implementação da mudança.

A revisão sistemática é importante porque oferece uma abordagem estruturada e confiável para reunir e sintetizar evidências, garantindo que as decisões sejam baseadas em informações robustas e bem fundamentadas. (TORRE-UGARTE-GUANILO; TAKAHASHI; BERTOLOZZI, 2011).

A revisão sistemática de acordo com (OKOLI; DUARTE; MATTAR, 2019) pode ser dividida em oito etapas.

1. Identificar o objetivo: esse passo consiste que os revisores identifiquem o propósito da revisão e os objetivos que são pretendidos.

2. Planejar o protocolo e treinar a equipe: Caso a revisão necessite de mais de um revisor, os revisores precisam estar completamente esclarecidos e de acordo sobre o procedimento que seguirão.

3. Aplicar uma seleção prática: também chamada de seleção para inclusão, esta etapa exige que os revisores sejam explícitos sobre quais estudos consideraram para a revisão e quais eliminaram sem maior exame. Para os estudos excluídos, os revisores devem indicar suas razões práticas para não os considerar e justificar.

4. Buscar a bibliografia: é necessário que os revisores sejam os revisores explícitos ao descrever os detalhes da pesquisa bibliográfica.

5. Extrair os dados: após os revisores identificarem todos os estudos que devem ser incluídos, precisam extrair sistematicamente as informações aplicáveis de cada estudo.

6. Avaliar a qualidade dos dados: É necessário que os revisores precisam declarem explicitamente os critérios utilizados para julgar quais artigos serão excluídos por qualidade insuficiente.

7. Sintetizar os estudos: este passo envolve combinar os fatos extraídos dos estudos, usando técnicas quantitativas ou qualitativas apropriadas ou ambas.

8. Escrever a revisão: o processo de uma revisão sistemática de literatura precisa ser descrito com detalhes suficientes de maneira que outros pesquisadores possam, independentemente, reproduzir seus resultados.

2.4 Tipos de Defeitos, Erros ou Falhas que Afetam SAAM

Sistemas Apoiados por Aprendizado de Máquina (SAAM) referem-se a sistemas que utilizam algoritmos de aprendizado de máquina para analisar dados e auxiliar na tomada de decisões. Esses sistemas são aplicados em uma ampla variedade de campos, incluindo saúde, finanças e transporte, onde a análise de grandes volumes de dados em tempo real pode levar a melhorias

significativas em eficiência e eficácia. Nesta seção, serão definidos os conceitos fundamentais relacionados aos tipos de defeitos, erros e falhas que podem afetar os Sistemas Apoiados por Aprendizado de Máquina (SAAM).

Os tipos abordados incluem Ataques Adversários, Dados Desbalanceados, Degradação do Desempenho ao longo do Tempo, Viés de Modelo, Sobreconfiança, Sobreajuste, Subajuste e Vazamento de Dados.

Figura 2: Fluxograma dos Tipos de Defeitos, Erros ou Falhas que Afetam SAAM (Fonte: Elaborada pelo autor).

2.4.1 Ataques Adversários

Um dos principais problemas que afetam os Sistemas Apoiados por Aprendizagem de Máquina (SAAM) são os ataques adversários, que tiram proveito das vulnerabilidades dos modelos de machine learning. Esses ataques envolvem a introdução de pequenas alterações nos dados de entrada, de modo que essas modificações, embora imperceptíveis aos humanos, conseguem confundir o modelo, levando-o a fazer previsões incorretas ou decisões equivocadas.

Segundo (ZHANG; SAKURAI, 2021), ataques adversários manipulam os dados de entrada adicionando ruídos sutis que, apesar de não serem percebidos visualmente ou sensorialmente, provocam resultados inesperados nos sistemas de aprendizado de máquina. Esse tipo de ataque pode gerar falhas críticas em aplicações de alto risco, como veículos autônomos, onde pequenas alterações em sinais de trânsito, imperceptíveis aos olhos humanos, podem levar o modelo a interpretar erroneamente um sinal de "Pare"como um limite de velocidade permitido, resultando em acidentes graves. Em diagnósticos médicos, alterações nos padrões das imagens analisadas podem levar a diagnósticos incorretos, enquanto em sistemas de segurança, modificações quase imperceptíveis podem permitir acessos indevidos.

Esses ataques exploram a sensibilidade dos modelos a pequenas perturbações, mostrando como sistemas aparentemente robustos podem ser enganados de maneiras difíceis de detectar. Dessa forma, a segurança e a confiabilidade dos SAAMs dependem fortemente de mecanismos eficazes de defesa contra essas ameaças.

2.4.2 Dados Desbalanceados

Um dos desafios mais comuns enfrentados por Sistemas Apoiados por Aprendizagem de Máquina (SAAM) são os dados desbalanceados. Esse fenômeno ocorre quando a variável de interesse ou resposta, que se deseja modelar, apresenta uma distribuição desigual entre suas classes, com uma concentração muito maior de exemplos em uma classe do que nas demais. Essa desproporção pode afetar diretamente o desempenho dos algoritmos de machine learning, uma vez que os modelos tendem a se ajustar melhor à classe majoritária, ignorando ou subestimando a classe minoritária, o que compromete a eficácia das predições.

Segundo (CORDEIRO, 2020), uma base de dados é considerada desbalanceada quando há muito mais casos em uma determinada classe em comparação com as outras. Esse problema é recorrente em aplicações como detecção de fraudes, onde há uma quantidade massiva de transações legítimas e um número muito menor de fraudes. Da mesma forma, em diagnósticos médicos, a predominância de casos de pacientes saudáveis nos dados pode dificultar a identificação de doenças raras, que representam a classe minoritária. Isso pode levar os modelos a desconsiderar sinais relevantes de condições menos comuns, impactando negativamente o diagnóstico e o tratamento.

Para mitigar os impactos de dados desbalanceados em SAAMs, são necessárias abordagens específicas, como a reamostragem das classes (aumentando o número de exemplos da classe

minoritária ou diminuindo os da majoritária) e o uso de métricas alternativas, como a F1-Score. Essas estratégias ajudam a equilibrar o aprendizado e garantir que o modelo consiga identificar corretamente tanto as classes majoritárias quanto as minoritárias.

2.4.3 Degradação do Desempenho ao longo do tempo

À medida que os modelos de Sistemas Apoiados por Aprendizado de Máquina (SAAM) avançam em suas aplicações na vida real, manter a qualidade de suas predições ao longo do tempo torna-se um desafio crucial. Isso se deve à natureza dos modelos de aprendizado de máquina, que dependem fortemente dos dados disponíveis no momento de seu treinamento. A degradação do desempenho ocorre à medida que esses dados se tornam desatualizados, um fenômeno conhecido como "envelhecimento da IA".

Segundo o artigo (VELA et al., 2022), esse processo de envelhecimento pode ser descrito como uma degradação complexa e multifacetada da qualidade dos modelos de IA, que ocorre à medida que o tempo passa desde o último ciclo de treinamento. O modelo, que inicialmente foi ajustado com uma determinada distribuição de dados, começa a apresentar erros crescentes conforme surgem novas condições ou mudanças nos padrões dos dados de entrada, fenômeno também conhecido como drift de conceito. Em sistemas de recomendação, por exemplo, alterações nos hábitos dos usuários podem tornar as recomendações baseadas em dados antigos irrelevantes, prejudicando a experiência do usuário e a eficácia do sistema.

Essencialmente, os modelos começam a "envelhecer"quando suas previsões se baseiam em dados históricos que já não representam mais o contexto atual. Sem ciclos regulares de reavaliação e recalibração, o desempenho de um SAAM pode cair drasticamente, levando a decisões errôneas em situações críticas. Dessa forma, a monitorização contínua e o re-treinamento periódico são necessários para garantir a longevidade e a eficácia dos modelos em uso.

2.4.4 Viés de Modelo

O "viés de modelo"se torna uma preocupação crescente à medida que algoritmos de decisão são amplamente adotados em contextos que impactam diretamente a vida das pessoas, como empréstimos, contratações e decisões judiciais. Esses algoritmos, baseados em previsões, têm o potencial de automatizar processos, mas também de perpetuar e amplificar preconceitos existentes, uma vez que são treinados com dados que podem refletir desigualdades sociais e históricas.

Conforme destacado no artigo (PAGANO et al., 2023), a incorporação de preconceitos nos modelos de aprendizado de máquina levanta questões sérias sobre sua justiça e equidade.Por exemplo, em processos de contratação, a predominância histórica de homens em determinadas funções pode influenciar os dados usados no treinamento, resultando em um modelo que favorece esse grupo em detrimento de outros, como mulheres ou minorias. Esse tipo de viés pode reforçar desigualdades já existentes, dificultando o acesso a oportunidades para grupos

sub-representados.

À medida que esses sistemas se tornam parte integrante das operações de governos e organizações, a necessidade de garantir que suas decisões sejam justas e imparciais se torna ainda mais urgente. A utilização de algoritmos em áreas sensíveis expõe a vulnerabilidade a preconceitos que podem afetar negativamente indivíduos ou grupos, resultando em discriminação e injustiça.

Reconhecer e abordar o "viés de modelo"é um desafio complexo, uma vez que a definição de justiça pode variar significativamente entre diferentes contextos culturais e sociais. O artigo enfatiza que as práticas corporativas, a legislação e os compromissos éticos influenciam a percepção do que constitui injustiça. Isso implica que, para mitigar o viés, é necessário um entendimento profundo das implicações sociais dos modelos, bem como a implementação de estratégias que promovam a equidade.

Dessa forma, a luta contra o viés de modelo não é apenas uma questão técnica, mas também uma responsabilidade ética que deve ser assumida por aqueles que desenvolvem e implementam sistemas baseados em aprendizado de máquina.

2.4.5 Sobreconfiança

A sobreconfiança em aprendizado de máquina é um fenômeno que ocorre quando os modelos fazem previsões com um nível de certeza excessivo, mesmo que essas previsões possam ser imprecisas. Essa questão é particularmente crítica em áreas como saúde, segurança e finanças, onde decisões importantes são tomadas com base nas saídas dos modelos. Segundo (TANG et al., 2022), muitos modelos de aprendizado profundo geram probabilidades associadas às suas previsões, mas essas probabilidades frequentemente não refletem adequadamente a verdadeira incerteza, resultando em uma falsa sensação de segurança em suas decisões.

A sobreconfiança pode ser atribuída à forma como os modelos são treinados, que muitas vezes desconsidera a variabilidade dos dados de entrada, levando a uma superestimação da confiança nas previsões.

2.4.6 Sobreajuste e Subajuste

De acordo com o artigo (ALIFERIS; SIMON, 2024), o sobreajuste é caracterizado pela criação de um modelo que se ajusta muito bem aos dados de treinamento, mas não consegue generalizar para novos dados, resultando em um desempenho insatisfatório em situações reais. Esse fenômeno ocorre quando o modelo é excessivamente complexo em relação ao problema em questão.

Por outro lado, o subajuste acontece quando um modelo é muito simples para capturar a complexidade dos dados, levando a um desempenho fraco tanto nos dados de treinamento quanto nos dados de teste. Em suma, um modelo subajustado falha em representar adequada-

mente os padrões presentes nos dados, resultando em um erro de generalização maior do que o do melhor modelo possível que poderia ser ajustado com os dados disponíveis

2.4.7 Vazamento de Dados

De acordo com (MUCCI, 2024), o vazamento de dados em machine learning refere-se à situação em que um modelo utiliza informações durante o treinamento que não estariam disponíveis no momento da previsão. Esse tipo de erro pode fazer com que o modelo pareça eficaz durante a fase de teste, mas, quando colocado em prática, ele tende a falhar, resultando em decisões ruins e em dados interpretados de forma incorreta.

Esse fenômeno pode ocorrer de várias formas, como a inclusão de variáveis que não seriam acessíveis no cenário real ou uma divisão inadequada entre os conjuntos de dados de treinamento e teste. Por exemplo, em um modelo de análise de crédito, a inclusão da variável "status do pagamento"no treinamento pode levar a uma falsa impressão de alta precisão, já que essa informação só estaria disponível após o evento que o modelo deveria prever. Isso resulta em um sistema incapaz de realizar predições corretas no ambiente real, comprometendo sua utilidade prática.

Para mitigar o risco de vazamento de dados, é essencial que os conjuntos sejam separados de maneira rigorosa antes de qualquer processo de pré-processamento, garantindo que as informações utilizadas no treinamento sejam realmente representativas do que o modelo encontrará na prática. Além disso, uma análise criteriosa das variáveis e seu contexto é fundamental para evitar que o modelo dependa de dados irreais ou inapropriados.

3

Metodologia

3.1 Introdução do Capítulo

Este estudo envolveu um exame abrangente da literatura existente, realizado em três fases distintas, planejamento, execução e condução. A fase de planejamento envolveu uma busca na literatura científica, especificamente direcionada a artigos que apresentavam algum tipo de erros ou defeitos que mais deterioram sistemas apoiados por aprendizagem de máquina. A pergunta de pesquisa foi formulada da seguinte forma: "Quais são os principais tipos de erros ou defeitos que afetam a performance de sistemas de aprendizagem de máquina?"

A etapa de execução que seria a segunda parte, foi conduzida utilizando o software StArt (ZAMBONI et al., 2010), que facilita revisões sistemáticas da literatura. Na formulação do protocolo de revisão, o escopo da pesquisa foi delimitado de duas formas, primeiro para incluir apenas publicações especificamente artigos de journal publicados entre 2019 e 2024, garantindo pesquisas mais recentes. Além disso, foram incluídos artigos de conferencia entre 2022 e 2024. Após a identificação das publicações relevantes, uma string de busca e várias palavras-chave para o motor de busca foram definidas, focando nos tipos de erros ou defeitos que mais deterioram sistemas apoiados por aprendizagem de máquina. Esta abordagem garantiu uma coleta abrangente e atualizada de informações, proporcionando uma base sólida para a análise e discussão subsequentes.

A definição da string de busca apresentou um desafio particular, uma vez que os termos "defects", "errors"e "failures"são frequentemente usados de forma intercambiável, mas com nuances específicas dependendo do autor ou contexto. Uma escolha por apenas um desses termos poderia limitar a abrangência da revisão, excluindo estudos relevantes que optassem por outro termo. Por isso, a string foi projetada para capturar todos os três conceitos, permitindo uma análise mais abrangente dos problemas que comprometem a confiabilidade, robustez e tolerância a falhas dos sistemas de aprendizagem de máquina.

A string de busca definida foi ((("machine learning") AND ("defects"OR "errors"OR "failures") AND ("reliability"OR "robustness"OR "fault tolerance"))). Somente publicações em inglês foram consideradas, uma escolha facilitada pela seleção de palavras-chave. As bases de dados de publicações escolhidas foram IEEE, ACM, MDPI, SPRINGER e Sciencedirect.

Na etapa de condução, a estratégia de busca foi aplicada em todas as bases de dados selecionadas, e os resultados foram importados para o software StArt, utilizado para organizar e analisar os dados. Inicialmente, foram identificadas 4188 publicações. Dado o volume elevado de artigos, critérios temporais foram implementados para garantir uma seleção mais precisa e manejável.

Para publicações de periódicos, foram considerados artigos publicados entre 2019 e 2024, enquanto os artigos de conferências foram restringidos ao período de 2022 a 2024. Essa abordagem reduziu o número de artigos, mantendo um total ainda robusto de 4188 publicações para a revisão. Na Figura 3, o gráfico apresenta o quantitativo de publicações oriundas de cada base de dados, bem como, o percentual relativo ao número total concentrado na referida base.

Figura 3: Quantitativo e percentual de publicações conforme base de dados (Fonte: Elaborada pelo autor).

Logo após, foi conduzida uma análise das publicações selecionadas, considerando a leitura dos títulos, resumos e palavras-chave. Durante essa etapa, artigos foram excluídos conforme os critérios de exclusão definidos na Tabela 3.1. Ao término dessa fase, restaram 46 publicações. Em seguida, foram aplicados os critérios de inclusão, juntamente com a leitura completa dos artigos. Nenhum artigo foi descartado nesse momento, mantendo-se o total de 46 publicações. Todo o processo de seleção dos artigos está ilustrado no fluxograma da Figura 2.

Tabela 3.1: Critérios utilizados para inclusão e exclusão de publicações

Critério	**Tipo**
Defeitos, erros e falhas que mais deterioram sistemas apoiados por aprendizagem de máquina	Inclusão
Uso de Aprendizado de Máquina para detecção de falhas (softwares, hardwares, etc.)	Exclusão
O artigo não está em inglês	Exclusão
O artigo não está disponível online	Exclusão
O artigo periódico não é de 2019-2024	Exclusão
O artigo conferência não é de 2022-2024	Exclusão
Não relata defeitos, erros e falhas que mais deterioram sistemas apoiados por aprendizagem de máquina	Exclusão
Artigo duplicado	Exclusão

Abaixo segue a lista contendo os títulos das publicações que foram selecionadas após os critérios.

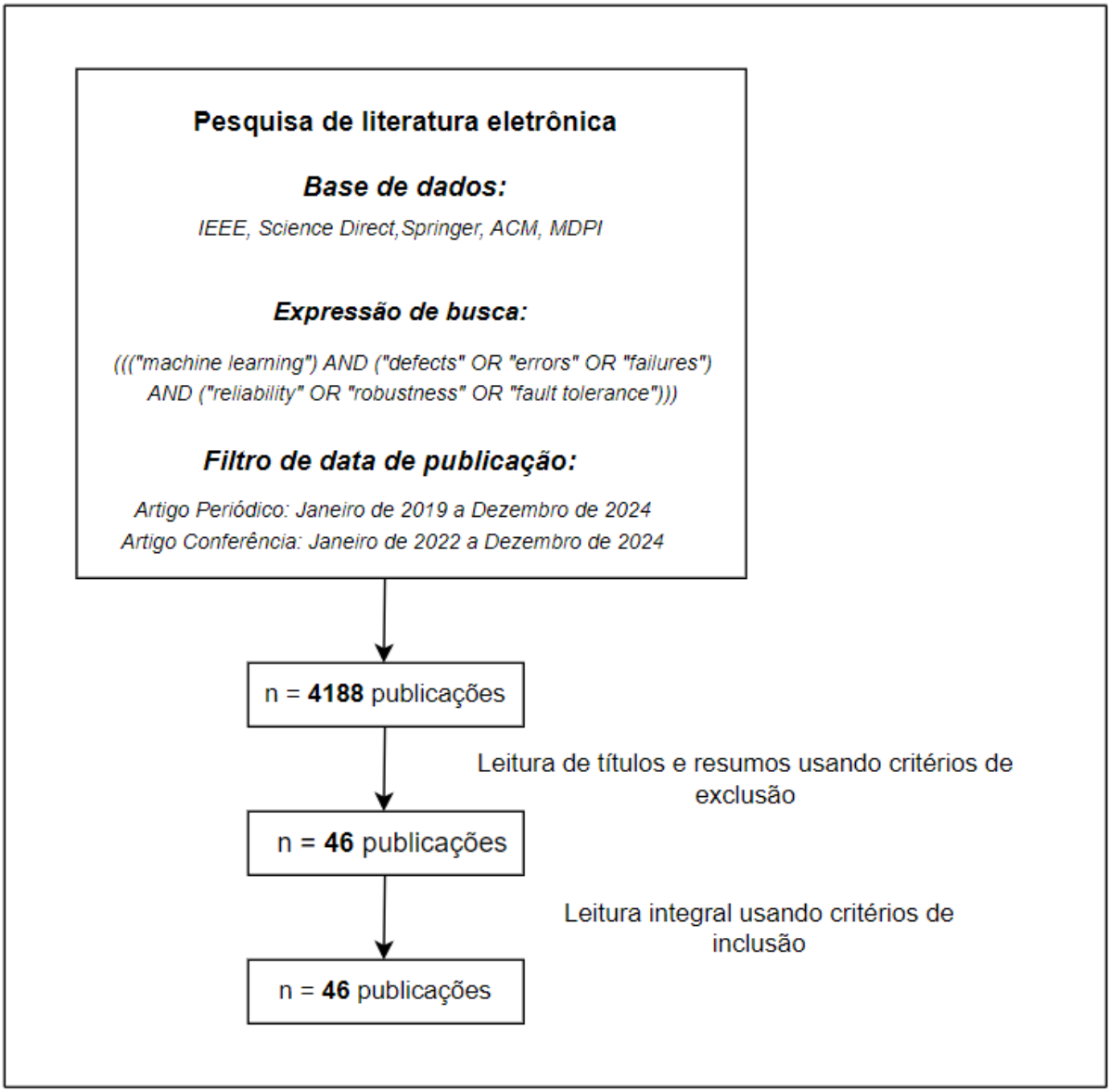

Figura 4: Fluxograma do processo de seleção das publicações (Fonte: Elaborada pelo autor).

- Robustness verification for machine-learning-based power system dynamic security assessment models under adversarial examples: (REN; XU, 2022);
- Responsible and regulatory conform machine learning for medicine: a survey of challenges and solutions: (PETERSEN et al., 2022);
- On Development of Reliable Machine Learning Systems Based on Machine Error Tolerance of Input Images: (HSIEH et al., 2022);
- Vulnerability analysis, robustness verification, and mitigation strategy for machine learning-based power system stability assessment model under adversarial examples: (REN et al., 2021);
- Tolerate Failures of the Visual Camera With Robust Image Classifiers: (ATIF et al., 2023);
- Adversarial attack mitigation strategy for machine learning-based network attack detection model in power system: (HUANG; LI, 2022);
- A more robust model to answer noisy questions in kbqa: (WANG et al., 2023);
- Ghost Loss to Question the Reliability of Training Data: (DELIÈGE; CIOPPA; DROOGENBROECK, 2020);
- Trade-Off Between Robustness and Rewards Adversarial Training for Deep Reinforcement Learning Under Large Perturbations: (HUANG; CHOI; FIGUEROA, 2023);
- Enhancing robustness of on-line learning models on highly noisy data: (ZHAO et al., 2021);
- Attack and defense: Adversarial security of data-driven FDC systems: (ZHUO; YIN; GE, 2022);
- Binary classification under 0 attacks for general noise distribution: (DELGOSHA; HASSANI; PEDARSANI, 2023);
- Adversarial Security Verification of Data-Driven FDC Systems: (ZHUO; GE, 2022);
- Deep face representations for differential morphing attack detection: (SCHERHAG et al., 2020);
- Evaluation of model quantization method on vitis-ai for mitigating adversarial examples: (FUKUDA; YOSHIDA; FUJINO, 2023);
- Model and data-centric machine learning algorithms to address data scarcity for failure identification: (KHAN et al., 2024);

- A new incremental learning for bearing fault diagnosis under noisy conditions using classification and feature level information: (ZHU et al., 2023);
- Distributed optimization under adversarial nodes: (SUNDARAM; GHARESIFARD, 2018);
- A feedback semi-supervised learning with meta-gradient for intrusion detection: (CAI; HAN; LI, 2022);
- Intelligent fault diagnosis method based on full 1-D convolutional generative adversarial network: (GUO et al., 2019);
- Predictive computing of human errors while training machine learning models: (TAO et al., 2023);
- Enhancing the Reliability of Perception Systems using N-version Programming and Rejuvenation: (MENDONÇA; MACHIDA; VÖLP, 2023);
- Comparing the robustness of classical and deep learning techniques for text classification: (TRAN et al., 2022);
- Power, Performance and Reliability Evaluation of Multi-thread Machine Learning Inference Models Executing in Multicore Edge Devices: (ABICH et al., 2023);
- Combining simulation and machine learning for the management of healthcare systems: (RICCIARDI et al., 2022);
- Accounting for Prediction Uncertainty from Machine Learning for Probabilistic Design: (DU, 2023);
- Efficient error-correcting output codes for adversarial learning robustness: (WAN et al., 2022);
- Evaluating the Effect of Common Annotation Faults on Object Detection Techniques: (CHAN et al., 2023);
- Reliability Estimation of ML for Image Perception: A Lightweight Nonlinear Transformation Approach Based on Full Reference Image Quality Metrics: (ZACCHI et al., 2023);
- Handling Imbalanced and Poorly Separated Data: a Multi-Stage Multi-Group Machine Learning Approach: (LEE et al., 2023);
- Simulating Bruise and Defects on Mango images using Image-to-Image Translation Generative Adversarial Networks: (ISMAIL et al., 2022);

- Sensing the Unknowns: A Study on Data-Driven Sensor Fault Modeling and Assessing its Impact on Fault Detection for Enhanced IoT Reliability: (ATTARHA; FÖRSTER, 2024);
- Mitigating Model Poisoning Attacks on Distributed Learning with Heterogeneous Data: (XU et al., 2023);
- Curriculum Defense: An Effective Adversarial Training Method: (YIN; DENG; YAN, 2022);
- Less is more: dimension reduction finds on-manifold adversarial examples in hard-label attacks: (GARCIA et al., 2023);
- A Model-Agnostic approach for learning with noisy labels of arbitrary distributions: (HAO et al., 2022);
- Quality Improvement of Image Datasets using Hashing Techniques: (JOSHI et al., 2023);
- RAIDS: robust autoencoder-based intrusion detection system model against adversarial attacks: (SARIKAYA; KILIÇ; DEMIRCI, 2023);
- Hardening machine learning denial of service (DoS) defences against adversarial attacks in IoT smart home networks: (ANTHI et al., 2021);
- A label-noise robust active learning sample collection method for multi-temporal urban land-cover classification and change analysis: (LI; HUANG; CHANG, 2020);
- Dealing with noise problem in machine learning data-sets: A systematic review: (GUPTA; GUPTA, 2019);
- Evidential classification for defending against adversarial attacks on network traffic: (BEECHEY; LAMBOTHARAN; KYRIAKOPOULOS, 2023);
- Diminishing-feature attack: The adversarial infiltration on visual tracking: (SUTTAPAK; ZHANG; ZHANG, 2022);
- Through the Data Management Lens: Experimental Analysis and Evaluation of Fair Classification: (ISLAM et al., 2022);
- A Distributed Biased Boundary Attack Method in Black-Box Attack: (XIANG et al., 2021);
- Towards Robustifying Image Classifiers against the Perils of Adversarial Attacks on Artificial Intelligence Systems: (ANASTASIOU et al., 2022);

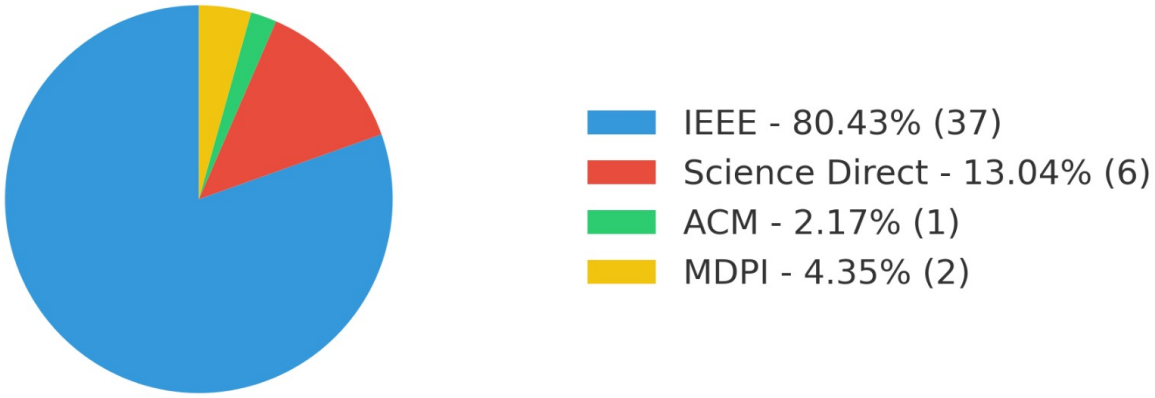

Figura 5: Quantitativo e percentual de publicações conforme base de dados após finalizados os processos de inclusões e exclusões (Fonte: Elaborada pelo autor).

A Figura 5 apresenta um gráfico que, após a conclusão dos processos de inclusão e exclusão, mostra o número de publicações provenientes de cada base de dados, assim como o percentual relativo ao total concentrado em cada uma delas.

Finalmente, foi realizada uma leitura dos 46 artigos no intuito de extrair os seguintes dados:

- Quais são os tipos de defeitos, erros ou falhas identificados no estudo que afetam sistemas apoiados por aprendizagem de máquina (SAAM)?
- Quais artefatos foram gerados pelo estudo?
- Com que tipo de problema o estudo está relacionado?

3.2 Conclusão do Capítulo

Os dados extraídos das publicações foram primeiramente organizados no StArt e depois exportados em formato .csv. Na fase final da revisão, dedicada à elaboração do relatório, foi utilizada a linguagem de programação Python, associada às bibliotecas Pandas, Matplotlib e Seaborn. Essas ferramentas permitiram processar as informações contidas no arquivo e gerar tabelas e gráficos que auxiliam na visualização e interpretação dos resultados apresentados neste trabalho.

4

Resultados

4.1 Análise da Distribuição Anual de Publicações.

Na Figura 6, observa-se a distribuição anual de publicações. Em 2019, início do período analisado, foram registradas apenas 2 publicações, sugerindo que o tema ainda não possuía grande visibilidade. No entanto, em 2020, houve um leve aumento para 4 publicações, sinalizando que a área começava a atrair maior interesse da comunidade científica.

Em 2021, o número de publicações caiu levemente para 3, sugerindo uma possível pausa no ritmo de pesquisas ou a necessidade de maior maturação dos estudos nessa área. Um fator adicional que pode ter contribuído para essa diminuição foi a pandemia de COVID-19, que impactou o ritmo de produção científica em várias áreas. Durante esse período, muitos pesquisadores enfrentaram limitações de recursos, dificuldades logísticas e restrições de acesso a laboratórios, o que pode ter reduzido o número de pesquisas publicadas naquele ano.

No entanto, esse declínio foi temporário, pois em 2022 houve um salto significativo para 12 publicações. Este crescimento pode ser atribuído à expansão das aplicações de SAAM em setores críticos, aumentando a demanda por estudos sobre a confiabilidade desses sistemas.

O ápice do número de publicações ocorreu em 2023, com 21 publicações, o que representa o maior volume de trabalhos registrados até o momento. Esse aumento expressivo pode estar relacionado à crescente urgência em solucionar problemas de confiabilidade em sistemas baseados em aprendizagem de máquina, impulsionada pela rápida adoção dessas tecnologias. Este ano é, até agora, o mais produtivo para a área.

Embora o ano de 2024 ainda esteja em curso, com 5 publicações registradas até o momento, o número final de trabalhos pode crescer, já que mais pesquisas podem ser publicadas ao longo do ano. Mesmo que o volume não ultrapasse o pico de 2023, não se pode interpretar esse dado como uma queda de interesse, mas sim como uma variação natural do ciclo de produção científica.

De forma geral, a análise sugere uma tendência de crescimento substancial na investigação de defeitos e falhas em SAAM, refletindo a maior relevância desse tema à medida que essas tecnologias são integradas em contextos que demandam alta robustez e confiabilidade.

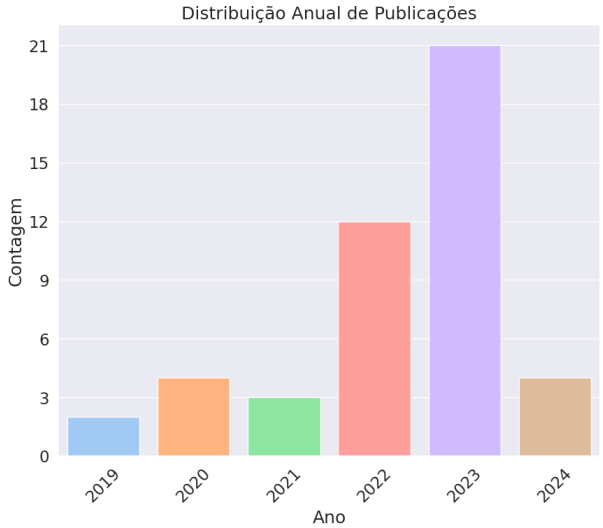

Figura 6: Distribuição Anual de Publicações.(Fonte: Elaborada pelo autor).

4.2 Tipos de defeitos, erros ou falhas identificados no estudo que afetam sistemas apoiados por aprendizagem de máquina (SAAM)

A Figura 7 ilustra os principais tipos de defeitos, erros e falhas que afetam sistemas apoiados por aprendizagem de máquina (SAAM), conforme o número de publicações que abordam cada um desses problemas. Os "Ataques Adversários"aparecem como o problema mais citado, com 20 publicações, indicando que essa vulnerabilidade é amplamente reconhecida como uma das mais críticas para o desempenho e a segurança de sistemas baseados em machine learning. Em seguida, "Dados Desbalanceados"é mencionado em 13 publicações, destacando a importância de dados equilibrados para o treinamento adequado dos modelos. A "Degradação do Desempenho ao Longo do Tempo", com 11 ocorrências, reflete preocupações sobre a capacidade dos modelos de manterem sua eficácia à medida que novas entradas são introduzidas.

Outro aspecto importante é o "Viés do Modelo", citado em 9 publicações, que mostra como a imparcialidade e a representatividade dos dados são questões centrais para a confiabilidade e justiça dos sistemas. "Sobreconfiança", relatada em 3 publicações, aponta para o problema de sistemas que podem gerar previsões incorretas com alto grau de confiança, o que pode ser prejudicial em cenários críticos. Além disso, "Subajuste"e "Sobreajuste", mencionados em 2 e 1 publicações respectivamente, indicam falhas relacionadas ao desempenho inadequado dos modelos, seja pela simplicidade ou complexidade excessiva. Por fim, "Vazamento de Dados"é citado em uma publicação.

Esse panorama revela que os problemas mais frequentemente abordados nos estudos estão relacionados tanto a vulnerabilidades externas, como os ataques adversários, quanto a questões internas de qualidade e representatividade dos dados, como o desbalanceamento e o viés. Conhecer e entender esses problemas é crucial para o desenvolvimento de sistemas de aprendizagem de máquina mais robustos.

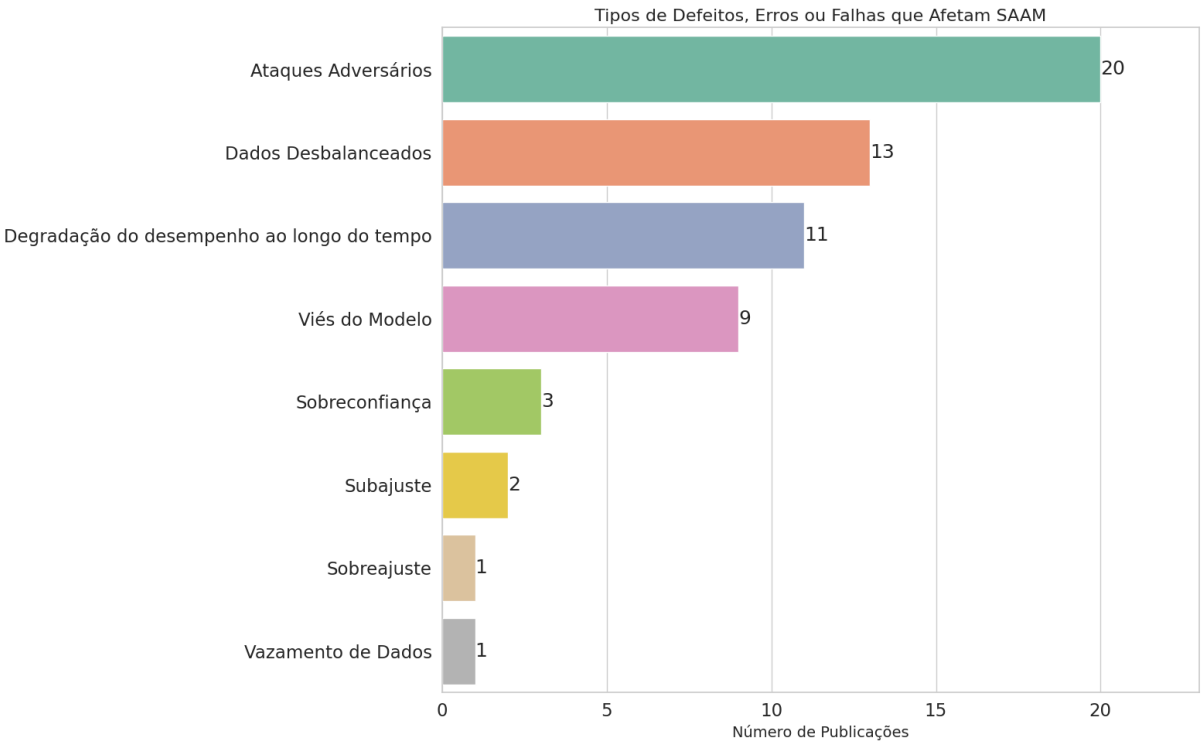

Figura 7: Tipos de Defeitos, Erros ou Falhas que Afetam SAAM (Fonte: Elaborada pelo autor).

4.3 Problema Abordado pelo Estudo

A Tabela 4.1 apresentada oferece uma visão abrangente dos problemas relacionados ao estudo, permitindo uma análise detalhada dos focos de pesquisa. Ela categoriza os tipos de problemas em subgrupos específicos, o que facilita a compreensão das áreas que demandam maior atenção.

A definição dos subgrupos baseou-se em uma análise minuciosa da literatura existente. O objetivo principal foi identificar categorias comuns nos estudos revisados, permitindo uma melhor organização das áreas de pesquisa mais relevantes. Para isso, foram estabelecidos critérios de categorização.

Cada subgrupo foi criado levando em consideração o foco específico dos estudos, abrangendo temas como classificação, detecção, segurança e robustez. Essa abordagem possibilitou o agrupamento de pesquisas com objetivos semelhantes, destacando tendências emergentes na área.

Além disso, a análise levou em conta as técnicas e abordagens utilizadas nos trabalhos revisados, agrupando aqueles que compartilham metodologias similares.

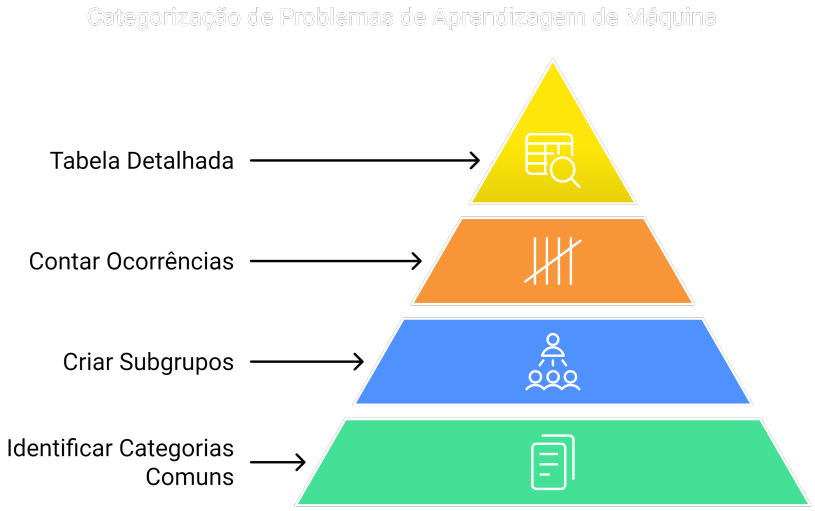

Figura 8: Mapa Conceitual (Fonte: Elaborada pelo autor).

Por exemplo, o subgrupo "Classificação de Imagens"se destaca com um total de 10 estudos, inserindo-se na categoria "Classificação e Detecção". Isso indica uma significativa concentração de esforços na otimização e melhoria da precisão dos algoritmos de classificação de imagens, um campo crucial para aplicações como reconhecimento facial e diagnóstico médico automatizado.

Outro subgrupo notável é "Segurança e Robustez", que abrange pesquisas como "Defesa Contra Ataques Adversários"e "Detecção de Ataques Adversários", totalizando 9 estudos. O foco nessa área enfatiza a crescente preocupação com a segurança e a robustez dos sistemas de aprendizado de máquina frente a ataques e interferências, um aspecto essencial para garantir a confiabilidade em aplicações críticas, como as de sistemas financeiros e de defesa.

O subgrupo "Classificação e Detecção"também possui 7 estudos identificados, evidenciando a importância da detecção de padrões atípicos que podem indicar erros, fraudes ou falhas. Essa habilidade de identificar anomalias é fundamental para a manutenção da confiabilidade e operação em diversos setores, especialmente no contexto industrial e de serviços.

Outros subgrupos, como "Dados e Ruído"e "Predição e Falhas", ressaltam a necessidade de desenvolver métodos eficazes para lidar com dados imperfeitos e prever falhas antes que resultem em perdas significativas. Essas áreas demandam abordagens que minimizem o impacto de ruídos nos dados e garantam a antecipação de problemas em sistemas críticos.

Dessa forma, a tabela proporciona uma visão clara das tendências e lacunas na pesquisa sobre Sistemas Apoiados por Aprendizado de Máquina (SAAM), permitindo a identificação de áreas que podem se beneficiar de investigações adicionais e inovações tecnológicas.

Tabela 4.1: Relação de categorias e número de ocorrências por tipo de problema estudado

Subgrupo	**Tipo de problema que o estudo está relacionado**	**N° de Ocorrências**
Análise de Desempenho de Algoritmos	Análise de Desempenho de Algoritmos de Classificação	1
	Desempenho e Eficiência de Modelos de Aprendizado de Máquina	1
	Quantificação de Incerteza em Modelos de Aprendizado de Máquina	1
Classificação e Detecção	Classificação Binária com Perturbação Adversária	1
	Classificação de Imagens	10
	Classificação de Texto	1
	Detecção de Anomalias	7
	Detecção de Intrusões e Ataques	1
	Rastreamento Visual	1
Dados e Ruído	Gestão de Dados Desbalanceados	1
	Identificação de Ruídos em Dados	1
	Treinamento de Modelos Robustos ao Ruído	1
Predição e Falhas	Detecção de Falhas	1
	Predição de Falhas	5
Segurança e Robustez	Avaliação de Segurança Dinâmica de Sistemas de Energia	1
	Defesa Contra Ataques Adversários	3
	Detecção de Ataques Adversários	4
	Robustez e Confiabilidade de Sistemas de Machine Learning em Ambientes Críticos	1
Controle e Robótica	Controle Robótico e Robustez contra Perturbações	1
Saúde e Simulação	Simulação de Processos de Saúde	1
Outros Temas	Monitoramento de Mudanças de Uso e Cobertura da Terra (LULC)	1
	Resposta a Perguntas Baseada em Conhecimento	1

4.4 Artefatos gerados pelos pelo estudo

A Tabela 4.2 apresenta uma compilação dos artefatos gerados pelos estudos, abordando a diversidade e a quantidade de contribuições na literatura. Com "Modelos de Aprendizagem de Máquina"liderando o número de publicações, a Tabela reflete a intensa atividade de pesquisa e o desenvolvimento contínuo nesta área, com 9 estudos.

Os "Conjuntos de Dados", com 7 publicações, enfatizam a importância da disponibilidade de dados adequados e específicos para o treinamento e teste de algoritmos de aprendizado de máquina, crucial para o avanço da pesquisa e aplicação prática dessas tecnologias.

Observa-se também a presença de "Algoritmo de Otimização"e "Framework", cada um com 2 publicações, indicando esforços para aprimorar a eficiência e a aplicabilidade dos sistemas de aprendizado de máquina em diferentes contextos.

Além desses, a Tabela lista uma variedade de outros artefatos, cada um com uma única publicação, destacando a ampla gama de pesquisas focadas em diferentes aspectos dos sistemas de aprendizado de máquina, desde "Métodos de Ataque e Defesa"até "Modelos de Regressão". Essa diversidade mostra a profundidade e o escopo das abordagens que os pesquisadores estão explorando para resolver problemas complexos e melhorar a eficácia desses sistemas.

Cada entrada na Tabela não apenas representa uma contribuição única ao campo, mas também serve como um recurso potencial para futuras investigações, oferecendo bases para a construção e teste de novas teorias e aplicações no mundo real. Esta Tabela serve como um guia para entender onde estão concentrados os esforços e as inovações no campo de sistemas apoiados por aprendizado de máquina.

Tabela 4.2: Artefatos gerados pelo estudo e número de publicações

Artefato	**N° de Publicações**
Modelos de Aprendizagem de Máquina	9
Conjuntos de Dados	7
Algoritmo de Otimização	2
Framework	2
Método de Classificação com Truncamento	1
Modelos de Redes Neurais	1
Método de Aprendizado Robusto ao Ruído	1
Método de Classificação Evidencial	1
Algoritmo de Decomposição de Gradientes (DeSGD)	1
Método de Teste de Tolerância a Erros	1
Modelo de Treinamento	1
Método de Verificação de Robustez Adversarial	1
Métodos de Ataque e Defesa	1
Resultados de Experimentos	1
Revisão Sistemática da Literatura	1
Sistemas de Diagnóstico	1
Modelo para Processamento de Imagens	1
Modelo de Detecção de Ataques	1
Modelo de Regressão	1
Algoritmo de Detecção de Configurações de Ruído	1
Imagens Sintéticas	1
Gerador de Ruído Adversário	1
Framework Distribuído para Ataques Adversários	1
Dados e resultados de experimentos que analisam a robustez dos modelos a diferentes perturbações	1
Dados de ataques adversários gerados por GANs	1
Análise de Abordagens de Classificação Justa	1
Amostras de Ataques Adversários	1
Algoritmos de Detecção de Anomalias	1
Algoritmos de Defesa	1
Índice Empírico	1

4.5 Conclusão do capítulo

Os resultados apresentados ao longo deste capítulo fornecem uma visão abrangente sobre o cenário atual da pesquisa em sistemas apoiados por aprendizado de máquina (SAAM). A análise da distribuição anual de publicações revela uma tendência de crescimento, principalmente a partir de 2022, com um pico significativo em 2023. Este aumento pode ser atribuído à crescente relevância dos SAAM em diversos setores, como saúde, finanças e segurança, onde a confiabilidade e robustez desses sistemas são críticas. A rápida adoção de tecnologias de machine learning em contextos que demandam alta precisão impulsiona o interesse da comunidade científica, refletido no volume crescente de publicações.

Os tipos de defeitos identificados nos estudos demonstram que, embora haja uma ampla variedade de problemas, alguns se destacam pela sua criticidade, como os ataques adversários e o desbalanceamento de dados. Isso sugere que as questões de segurança e qualidade dos dados são pontos de maior preocupação na comunidade, uma vez que essas vulnerabilidades podem comprometer o desempenho e a confiança nos sistemas. A alta incidência de pesquisas voltadas para a defesa contra ataques adversários, por exemplo, reflete a urgência em proteger SAAM contra interferências externas, enquanto o foco em dados desbalanceados ressalta a importância de garantir a equidade e eficácia dos modelos de aprendizado.

Além disso, a categorização dos problemas abordados pelos estudos revela uma concentração de esforços em áreas como a classificação de imagens, detecção de anomalias e defesa contra ataques adversários. A predominância dessas áreas indica uma resposta direta às demandas práticas e desafios enfrentados por sistemas SAAM em ambientes críticos. Esse foco, no entanto, aponta para possíveis lacunas em outras áreas, como predição de falhas e robustez em cenários menos explorados, sugerindo que essas áreas podem se beneficiar de maior atenção em futuros estudos.

Em síntese, os resultados discutidos destacam tanto os avanços significativos quanto os desafios persistentes na pesquisa de SAAM. A evolução do número de publicações, o foco em vulnerabilidades críticas e a variedade de artefatos desenvolvidos demonstram o progresso contínuo da área. No entanto, há oportunidades para explorar questões menos abordadas, como métodos de detecção de falhas e a aplicação em novos contextos, que podem oferecer valiosas contribuições para a robustez e confiabilidade dos sistemas no futuro.

5

Conclusão

Este trabalho apresentou uma revisão sistemática com o objetivo de identificar os tipos de defeitos, erros e falhas que mais afetam os sistemas apoiados por aprendizagem de máquina (SAAM). A partir da análise de artigos selecionados, foi possível responder às perguntas de pesquisa propostas e fornecer um panorama abrangente sobre os principais problemas que deterioram esses sistemas, bem como os artefatos gerados e o contexto dos problemas abordados pelos estudos.

Durante o desenvolvimento, foram enfrentados alguns desafios. O principal deles foi o grande volume de artigos retornados pela string de busca utilizada. A filtragem desses artigos na etapa de "leitura de títulos e resumos usando critérios de exclusão"foi uma tarefa extremamente trabalhosa e demorada, exigindo um rigoroso processo de seleção para garantir a relevância dos estudos incluídos na análise. Essa etapa consumiu mais tempo do que o previsto, mas foi fundamental para assegurar que os resultados obtidos refletissem com precisão os principais desafios da área.

Apesar dessas dificuldades, o trabalho conseguiu apresentar uma visão abrangente sobre os defeitos em SAAM e contribuir para o entendimento de como mitigar esses problemas no futuro.

A primeira pergunta de pesquisa, que buscava identificar os tipos de defeitos, erros ou falhas que afetam SAAM, revelou que as vulnerabilidades mais recorrentes incluem ataques adversários, dados desbalanceados e a degradação do desempenho ao longo do tempo. Esses problemas são críticos para a confiabilidade dos sistemas, uma vez que podem comprometer tanto a segurança quanto a precisão dos modelos. Os ataques adversários, em particular, destacaram-se como o tipo de falha mais citado, evidenciando a crescente preocupação com a segurança em sistemas que utilizam machine learning, especialmente em aplicações de alto risco.

Em relação à segunda pergunta, que investigava os artefatos gerados pelos estudos, observou-se uma diversidade significativa nas contribuições. Modelos de aprendizado de máquina e conjuntos de dados foram os artefatos mais frequentemente gerados, refletindo o esforço contínuo da comunidade científica em desenvolver ferramentas e recursos que permitam testar

e melhorar a robustez dos sistemas. Além disso, a criação de frameworks e algoritmos de otimização também foi recorrente, mostrando que os estudos buscam não apenas identificar falhas, mas também fornecer soluções práticas para mitigar seus efeitos.

Por fim, ao responder à terceira pergunta de pesquisa, foi possível categorizar os estudos de acordo com o tipo de problema com o qual estão relacionados. A maioria dos estudos analisados está vinculada a problemas de classificação e detecção, especialmente em áreas como a classificação de imagens e a detecção de anomalias. Outro grupo relevante de estudos focou em segurança e robustez, com destaque para as pesquisas que abordam a defesa contra ataques adversários. Essa concentração sugere que os principais desafios enfrentados por SAAM estão ligados à garantia de segurança e precisão, em especial em aplicações críticas, onde a integridade dos modelos é essencial.

Em conclusão, esta revisão sistemática revelou uma preocupação crescente com a segurança e a qualidade dos dados em sistemas de aprendizagem de máquina, bem como a necessidade de desenvolver soluções mais robustas para mitigar os efeitos de falhas e defeitos. A pesquisa demonstrou que, embora já existam muitos avanços nessa área, ainda há desafios a serem superados, especialmente em relação à adaptação dos modelos a contextos dinâmicos e à prevenção de falhas antes que causem danos significativos. Dessa forma, o campo de SAAM continua sendo uma área de pesquisa ativa e em rápida evolução, com oportunidades para novas investigações que possam contribuir para a criação de sistemas mais confiáveis e eficientes.

5.1 Trabalhos Futuros

A presente revisão sistemática forneceu uma visão abrangente sobre os principais desafios enfrentados pelos sistemas apoiados por aprendizagem de máquina (SAAM). No entanto, diversos caminhos ainda podem ser explorados em pesquisas futuras para melhorar o entendimento e desenvolver soluções mais robustas para lidar com esses problemas.

1. **Explorar o Impacto de Dados Desbalanceados e Viés em Diferentes Contextos:** Embora o desbalanceamento de dados e o viés do modelo tenham sido amplamente identificados como problemas críticos, seria interessante realizar estudos mais específicos em diferentes áreas de aplicação, como saúde, segurança e finanças. Pesquisas futuras poderiam analisar como esses defeitos impactam cada setor. Dessa forma, soluções para mitigar esses problemas poderiam ser desenvolvidas para contextos específicos.

2. **Defesas Contra Ataques Adversários:** Dado que os ataques adversários foram identificados como a vulnerabilidade mais recorrente nos SAAMs, pesquisas futuras poderiam focar no desenvolvimento de novos mecanismos de defesa para melhorar a segurança dos modelos. Uma análise mais profunda sobre como esses ataques evoluem em diferentes aplicações também seria relevante para fortalecer a proteção em áreas críticas, como veículos autônomos e diagnósticos médicos.

3. **Desenvolvimento de Métodos de Avaliação de Robustez e Confiabilidade:** Como os SAAMs estão se tornando cada vez mais integrados a setores de alta criticidade, futuras pesquisas poderiam focar no desenvolvimento de métricas e metodologias que permitam avaliar a robustez e a confiabilidade desses sistemas. Tais métodos ajudariam a antecipar falhas e prevenir problemas em fases anteriores ao uso em produção, evitando erros que possam causar grandes danos.

Referências Bibliográficas

ABICH, G. et al. Power, performance and reliability evaluation of multi-thread machine learning inference models executing in multicore edge devices. In: IEEE. *2023 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*. [S.l.], 2023. p. 1–6.

ALIFERIS, C.; SIMON, G. Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and ai. In: ____. [S.l.: s.n.], 2024. p. 477–524. ISBN 978-3-031-39354-9.

ANASTASIOU, T. et al. Towards robustifying image classifiers against the perils of adversarial attacks on artificial intelligence systems. *Sensors*, v. 22, n. 18, 2022. ISSN 1424-8220. Disponível em: <https://www.mdpi.com/1424-8220/22/18/6905>.

ANTHI, E. et al. Hardening machine learning denial of service (dos) defences against adversarial attacks in iot smart home networks. *computers & security*, Elsevier, v. 108, p. 102352, 2021.

ATIF, M. et al. Tolerate failures of the visual camera with robust image classifiers. *IEEE Access*, IEEE, v. 11, p. 5132–5143, 2023.

ATTARHA, S.; FÖRSTER, A. Sensing the unknowns: A study on data-driven sensor fault modeling and assessing its impact on fault detection for enhanced iot reliability. In: IEEE. *2024 19th Wireless On-Demand Network Systems and Services Conference (WONS)*. [S.l.], 2024. p. 33–40.

BEECHEY, M.; LAMBOTHARAN, S.; KYRIAKOPOULOS, K. G. Evidential classification for defending against adversarial attacks on network traffic. *Information Fusion*, v. 92, p. 115–126, 2023. ISSN 1566-2535. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1566253522002342>.

CAI, S.; HAN, D.; LI, D. A feedback semi-supervised learning with meta-gradient for intrusion detection. *IEEE Systems Journal*, IEEE, v. 17, n. 1, p. 1158–1169, 2022.

CHAN, A. et al. Evaluating the effect of common annotation faults on object detection techniques. In: IEEE. *2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE)*. [S.l.], 2023. p. 474–485.

CORDEIRO, T. V. B. *Predição de default de empresas: técnicas de machine learning em dados desbalanceados*. 2020. Accessed: 13-Oct-2024. Disponível em: <https://hdl.handle.net/10438/29873>.

DELGOSHA, P.; HASSANI, H.; PEDARSANI, R. Binary classification under 0 attacks for general noise distribution. *IEEE Transactions on Information Theory*, IEEE, 2023.

DELIÈGE, A.; CIOPPA, A.; DROOGENBROECK, M. V. Ghost loss to question the reliability of training data. *IEEE Access*, IEEE, v. 8, p. 44774–44782, 2020.

DU, X. Accounting for prediction uncertainty from machine learning for probabilistic design. In: IEEE. *2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*. [S.l.], 2023. p. 1–6.

FUKUDA, Y.; YOSHIDA, K.; FUJINO, T. Evaluation of model quantization method on vitis-ai for mitigating adversarial examples. *IEEE Access*, IEEE, 2023.

GARCIA, W. et al. Less is more: dimension reduction finds on-manifold adversarial examples in hard-label attacks. In: IEEE. *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. [S.l.], 2023. p. 254–270.

GUO, Q. et al. Intelligent fault diagnosis method based on full 1-d convolutional generative adversarial network. *IEEE Transactions on Industrial Informatics*, IEEE, v. 16, n. 3, p. 2044–2053, 2019.

GUPTA, S.; GUPTA, A. Dealing with noise problem in machine learning data-sets: A systematic review. *Procedia Computer Science*, Elsevier, v. 161, p. 466–474, 2019.

HAO, S. et al. A model-agnostic approach for learning with noisy labels of arbitrary distributions. In: IEEE. *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. [S.l.], 2022. p. 1219–1231.

HSIEH, T.-Y. et al. On development of reliable machine learning systems based on machine error tolerance of input images. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, IEEE, v. 42, n. 4, p. 1323–1335, 2022.

HUANG, J.; CHOI, H. J.; FIGUEROA, N. Trade-off between robustness and rewards adversarial training for deep reinforcement learning under large perturbations. *IEEE Robotics and Automation Letters*, IEEE, 2023.

HUANG, R.; LI, Y. Adversarial attack mitigation strategy for machine learning-based network attack detection model in power system. *IEEE Transactions on Smart Grid*, IEEE, v. 14, n. 3, p. 2367–2376, 2022.

ISLAM, M. T. et al. Through the Data Management Lens: Experimental Analysis and Evaluation of Fair Classification. In: *SIGMOD '22: International Conference on Management of Data*. ACM, 2022. p. 232–246. Disponível em: <https://doi.org/10.1145/3514221.3517841>.

ISMAIL, M. H. bin et al. Simulating bruise and defects on mango images using image-to-image translation generative adversarial networks. In: IEEE. *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*. [S.l.], 2022. p. 110–114.

JOSHI, A. et al. Quality improvement of image datasets using hashing techniques. In: IEEE. *2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*. [S.l.], 2023. p. 18–23.

KHAN, L. Z. et al. Model and data-centric machine learning algorithms to address data scarcity for failure identification. *Journal of Optical Communications and Networking*, Optica Publishing Group, v. 16, n. 3, p. 369–381, 2024.

LEE, E. K. et al. Handling imbalanced and poorly separated data: a multi-stage multi-group machine learning approach. In: IEEE. *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. [S.l.], 2023. p. 4336–4343.

LI, J.; HUANG, X.; CHANG, X. A label-noise robust active learning sample collection method for multi-temporal urban land-cover classification and change analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, Elsevier, v. 163, p. 1–17, 2020.

LUDERMIR, T. B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. *Estudos Avançados*, Instituto de Estudos Avançados da Universidade de São Paulo, v. 35, n. 101, p. 85–94, Jan 2021. ISSN 0103-4014. Disponível em: <https://doi.org/10.1590/s0103-4014.2021.35101.007>.

MENDONÇA, J.; MACHIDA, F.; VÖLP, M. Enhancing the reliability of perception systems using n-version programming and rejuvenation. In: IEEE. *2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*. [S.l.], 2023. p. 149–156.

MITCHELL, T. M. *Machine Learning*. 1. ed. McGraw-Hill, 1997. (McGraw-Hill series in computer science). ISBN 9780070428072,0070428077. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=f3aa83fb7adab9c8675871a717db6231>.

MUCCI, T. *What is data leakage in machine learning?* 2024. Accessed: 2024-10-13. Disponível em: <https://www.ibm.com/think/topics/data-leakage-machine-learning>.

OKOLI, C.; DUARTE, T. p. W. A.; MATTAR, R. t. e. i. Guia para realizar uma revisão sistemática de literatura. *EaD em Foco*, v. 9, n. 1, abr. 2019. Disponível em: <https://eademfoco.cecierj.edu.br/index.php/Revista/article/view/748>.

PAGANO, T. P. et al. Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, v. 7, n. 1, 2023. ISSN 2504-2289. Disponível em: <https://www.mdpi.com/2504-2289/7/1/15>.

PETERSEN, E. et al. Responsible and regulatory conform machine learning for medicine: a survey of challenges and solutions. *IEEE Access*, IEEE, v. 10, p. 58375–58418, 2022.

REN, C. et al. Vulnerability analysis, robustness verification, and mitigation strategy for machine learning-based power system stability assessment model under adversarial examples. *IEEE Transactions on Smart Grid*, IEEE, v. 13, n. 2, p. 1622–1632, 2021.

REN, C.; XU, Y. Robustness verification for machine-learning-based power system dynamic security assessment models under adversarial examples. *IEEE Transactions on Control of Network Systems*, IEEE, v. 9, n. 4, p. 1645–1654, 2022.

RICCIARDI, C. et al. Combining simulation and machine learning for the management of healthcare systems. In: IEEE. *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*. [S.l.], 2022. p. 335–339.

SARIKAYA, A.; KILIÇ, B. G.; DEMIRCI, M. Raids: robust autoencoder-based intrusion detection system model against adversarial attacks. *Computers & Security*, Elsevier, v. 135, p. 103483, 2023.

SCHERHAG, U. et al. Deep face representations for differential morphing attack detection. *IEEE transactions on information forensics and security*, IEEE, v. 15, p. 3625–3639, 2020.

SUNDARAM, S.; GHARESIFARD, B. Distributed optimization under adversarial nodes. *IEEE Transactions on Automatic Control*, IEEE, v. 64, n. 3, p. 1063–1076, 2018.

SUTTAPAK, W.; ZHANG, J.; ZHANG, L. Diminishing-feature attack: The adversarial infiltration on visual tracking. *Neurocomputing*, v. 509, p. 21–33, 2022. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231222010633>.

TANG, Q. et al. Alleviating overconfident failure predictions via masking predictive logits in semantic segmentation. In: *Artificial Neural Networks and Machine Learning – ICANN 2022: 31st International Conference on Artificial Neural Networks, Bristol, UK, September 6–9, 2022, Proceedings, Part II*. Berlin, Heidelberg: Springer-Verlag, 2022. p. 1–13. ISBN 978-3-031-15930-5. Disponível em: <https://doi.org/10.1007/978-3-031-15931-2_1>.

TAO, H. et al. Predictive computing of human errors while training machine learning models. In: IEEE. *2023 International Conference on Intelligent Computing and Next Generation Networks (ICNGN)*. [S.l.], 2023. p. 1–6.

TORRE-UGARTE-GUANILO, M. C. De-la; TAKAHASHI, R. F.; BERTOLOZZI, M. R. Revisão sistemática: noções gerais. *Revista da Escola de Enfermagem da USP*, Universidade de São Paulo, Escola de Enfermagem, v. 45, n. 5, p. 1260–1266, Oct 2011. ISSN 0080-6234. Disponível em: <https://doi.org/10.1590/S0080-62342011000500033>.

TRAN, Q. et al. Comparing the robustness of classical and deep learning techniques for text classification. In: IEEE. *2022 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2022. p. 1–10.

VELA, D. et al. Temporal quality degradation in ai models. *Scientific Reports*, v. 12, 07 2022.

WAN, L. et al. Efficient error-correcting output codes for adversarial learning robustness. In: IEEE. *ICC 2022-IEEE International Conference on Communications*. [S.l.], 2022. p. 2345–2350.

WANG, Z. et al. A more robust model to answer noisy questions in kbqa. *IEEE Access*, IEEE, v. 11, p. 22756–22766, 2023.

XIANG, F. et al. A distributed biased boundary attack method in black-box attack. *Applied Sciences*, v. 11, n. 21, 2021. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/11/21/10479>.

XU, J. et al. Mitigating model poisoning attacks on distributed learning with heterogeneous data. In: IEEE. *2023 International Conference on Machine Learning and Applications (ICMLA)*. [S.l.], 2023. p. 738–743.

YIN, H.; DENG, X.; YAN, J. Curriculum defense: An effective adversarial training method. In: IEEE. *2022 41st Chinese Control Conference (CCC)*. [S.l.], 2022. p. 7399–7406.

ZACCHI, J.-V. et al. Reliability estimation of ml for image perception: A lightweight nonlinear transformation approach based on full reference image quality metrics. In: IEEE. *2023 IEEE 16th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*. [S.l.], 2023. p. 186–193.

ZAMBONI, A. et al. Start uma ferramenta computacional de apoio à revisão sistemática. In: UFBA. *Proc.: Congresso Brasileiro de Software (CBSoft'10), Salvador, Brazil*. [S.l.], 2010. p. 91–96.

ZHANG, H.; SAKURAI, K. Conditional generative adversarial network-based image denoising for defending against adversarial attack. *IEEE Access*, v. 9, p. 169031–169043, 2021.

ZHAO, Z. et al. Enhancing robustness of on-line learning models on highly noisy data. *IEEE Transactions on Dependable and Secure Computing*, IEEE, v. 18, n. 5, p. 2177–2192, 2021.

ZHU, J. et al. A new incremental learning for bearing fault diagnosis under noisy conditions using classification and feature level information. *IEEE Transactions on Instrumentation and Measurement*, IEEE, 2023.

ZHUO, Y.; GE, Z. Adversarial security verification of data-driven fdc systems. *IEEE Transactions on Reliability*, IEEE, v. 72, n. 4, p. 1580–1593, 2022.

ZHUO, Y.; YIN, Z.; GE, Z. Attack and defense: Adversarial security of data-driven fdc systems. *IEEE Transactions on Industrial Informatics*, IEEE, v. 19, n. 1, p. 5–19, 2022.