UNIVERSIDADE FEDERAL DO AMAPÁ
PRÓ-REITORIA DE ENSINO E GRADUAÇÃO
DEPARTAMENTO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
COLEGIADO DO CURSO DE CIÊNCIA DA COMPUTAÇÃO

MARCO ANTONIO DA SILVA NEVES FILHO

MAPEAMENTO SISTEMÁTICO DE TÉCNICAS EMERGENTES DE PRÉ-PROCESSAMENTO PARA MITIGAÇÃO DE VIÉS EM DADOS DE TREINAMENTO

MACAPÁ

2025

MARCO ANTONIO DA SILVA NEVES FILHO

MAPEAMENTO SISTEMÁTICO DE TÉCNICAS EMERGENTES DE PRÉ-PROCESSAMENTO PARA MITIGAÇÃO DE VIÉS EM DADOS DE TREINAMENTO

Trabalho de Conclusão de Curso submetido à Banca Examinadora do Curso de Ciência da Computação do Departamento de Ciências Exatas e Tecnológicas da Universidade Federal do Amapá, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Julio Cezar Costa Furtado.

MACAPÁ

2025

Dados Internacionais de Catalogação na Publicação (CIP)
Biblioteca Central/UNIFAP-Macapá-AP
Elaborado por Cristina Fernandes – CRB-2 / 1569

N518m Neves Filho, Marco Antonio da Silva.
Mapeamento sistemático de técnicas emergentes de pré-processamento para mitigação de viés em dados de treinamento / Marco Antonio da Silva Neves Filho. - Macapá, 2025.
1 recurso eletrônico.
62 f.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal do Amapá, Coordenação do Ciência da Computação, Macapá, 2025.
Orientador: Prof. Dr. Julio Cezar Costa Furtado.
Coorientador: .

Modo de acesso: World Wide Web.
Formato de arquivo: Portable Document Format (PDF).

1. Viés. 2. Aprendizado de máquina. 3. Pré-processamento. I. Furtado, Julio Cezar Costa, orientador. II. Universidade Federal do Amapá. III. Título.

CDD 23. ed. – 004

NEVES FILHO, Marco Antonio da Silva. **Mapeamento sistemático de técnicas emergentes de pré-processamento para mitigação de viés em dados de treinamento**. Orientador: Prof. Dr. Julio Cezar Costa Furtado. 2025. 62 f. Trabalho de Conclusão de Curso (Graduação) - Ciência da Computação. Universidade Federal do Amapá, Macapá, 2025.

UNIVERSIDADE FEDERAL DO AMAPÁ
DEPARTAMENTO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
COORDENAÇÃO DO CURSO DE CIÊNCIA DA COMPUTAÇÃO

ATA DE DEFESA DE TCC

Realizou-se no dia 18 de julho de 2025, às 19h00 a defesa do projeto de TCC intitulado: **"MAPEAMENTO SISTEMÁTICO DE TÉCNICAS EMERGENTES DE PRÉ-PROCESSAMENTO PARA MITIGAÇÃO DE VIÉS EM DADOS DE TREINAMENTO"**, do discente Marco Antonio da Silva Neves Filho, matrícula 2022008410. A Banca Examinadora foi composta pelo Prof. Dr. JULIO CEZAR COSTA FURTADO, presidente da banca e orientador; Prof. Me. THIAGO PINHEIRO DO NASCIMENTO e Prof. Me. MARCO ANTÔNIO LEAL DA SILVA, examinadores. Concluída a defesa, foram realizadas as arguições e comentários. Em seguida, procedeu-se o julgamento pelos membros da Banca Examinadora, tendo o projeto sido APROVADO com nota 9,25.

E, para constar, eu, Prof. Dr. JULIO CEZAR COSTA FURTADO, orientador e presidente da Banca Examinadora, lavrei a presente ata que, após lida e achada conforme, foi assinada por mim e demais membros da Banca Examinadora.

gov.br Documento assinado digitalmente
JULIO CEZAR COSTA FURTADO
Data: 22/07/2025 19:29:30-0300
Verifique em https://validar.iti.gov.br

Macapá, 18 de julho de 2025.

Prof. Dr. JULIO CEZAR COSTA FURTADO
Orientador do TCC

gov.br Documento assinado digitalmente
THIAGO PINHEIRO DO NASCIMENTO
Data: 23/07/2025 12:45:16-0300
Verifique em https://validar.iti.gov.br

Prof. Me. THIAGO PINHEIRO DO NASCIMENTO
Examinador (UNIFAP)

gov.br Documento assinado digitalmente
MARCO ANTONIO LEAL DA SILVA
Data: 23/07/2025 16:48:26-0300
Verifique em https://validar.iti.gov.br

Prof. Me. MARCO ANTÔNIO LEAL DA SILVA
Examinador (UNIFAP)

AGRADECIMENTOS

Sou eternamente grato pela minha família, e a ela dedico este trabalho. Sem vocês, isso não seria possível.

Agradeço à minha fiel esposa e companheira Maria Melissa, que segurou as pontas enquanto eu não pude estar presente em nossas responsabilidades compartilhadas. Obrigado, meu bem, por ter segurado firmemente minha mão nessa jornada.

À minha maior heroína, e muito sábia, Jacimary Cascaes Santos. Mãe, não sei o que eu seria sem você, e é um enorme prazer poder compartilhar essa vitória contigo. Obrigado por todas as vezes que esteve presente quando precisamos, mesmo quando de última hora e sem prévio aviso. Por cuidar de problemas antes mesmo que chegassem aos nossos ouvidos. Por auxiliar na criação de meu filho. E muitíssimo obrigado por sempre me apoiar, mesmo quando incerta sobre os desafios futuros. Você merece isso tanto quanto eu.

À minha estimada sogra, Dora Figueiredo, por toda a base que nos proporcionou e pela fé que sempre depositou em nosso potencial. Obrigado por tudo. Serei sempre grato quando você se fez presente em tempos de necessidade.

Ao meu maior amor, meu filho Arthur. Obrigado por todas as vezes que seu sorriso me salvou. Você sempre será tudo de mais importante pra mim. A vida ao seu lado é incrível, e desejo ser sempre uma grande inspiração pra você.

Agradeço também aos meus professores, e pelo aprendizado que me proporcionaram durante a duração do curso. Ao meu orientador, Julio Furtado, agradeço especialmente, pela paciência e pela pressa quando essa também foi necessária. Obrigado pelos ensinamentos, atenção e compreensão que me proporcionou durante essa jornada.

Obrigado, também, aos meus amigos, pela paciência, ajuda e conforto.

Por fim, agradeço a todos que acreditaram no objetivo deste estudo e que auxiliaram para a sua conclusão.

RESUMO

O presente trabalho apresenta um mapeamento sistemático da literatura que identifica técnicas emergentes de mitigação de viés em dados de treinamento para aprendizado de máquina, aplicadas durante o pré-processamento, entre o período de 2020 e 2024. Dada a crescente popularidade do campo de estudo de inteligência artificial e de diferentes áreas dependerem cada vez mais desses modelos, é fundamental o controle e monitoramento dos vieses presentes em meio a estes. A pesquisa teve como base metodológica os critérios estabelecidos por Kitchenham e Charters (2007), com *string* de busca refinada e definição de critérios de inclusão/exclusão aplicados sobre as bases *IEEE Xplore* e *ACM Digital Library*. 85 estudos primários foram analisados, resultando em 244 técnicas identificadas, classificadas em sete categorias principais, sendo estas: balanceamento de dados, transformação de *labels*, modificação de atributos sensíveis, reponderação, geradores de dados justos, censura de *score/feature* e normalização e transformações de *feature*. Os resultados apontam para uma crescente adoção de abordagens híbridas e relevantes ao contexto, além da utilização de múltiplas métricas para a avaliação do desempenho e equidade. O estudo conclui que o pré-processamento permanecerá como uma etapa crítica para a construção de sistemas mais justos, e que a vigilância contínua das técnicas emergentes é essencial para o avanço ético e técnico da área.

Palavras-chave: Mitigação de Viés, Viés em Aprendizado de Máquina, Aprendizado de Máquina Justo, Pré-processamento de Dados, Equidade Algorítmica, Técnicas Emergentes, Mapeamento Sistemático.

ABSTRACT

This study presents a systematic mapping of the literature aimed at identifying emerging bias mitigation techniques in training data for machine learning, specifically those applied during the preprocessing stage, within the period from 2020 to 2024. Given the growing popularity of artificial intelligence as a field of study, and the increasing dependence of various domains on these models, the control and monitoring of biases inherent in them become essential. The research adopted the methodological framework established by Kitchenham and Charters (2007), employing a refined search string and well-defined inclusion and exclusion criteria applied to the IEEE Xplore and ACM Digital Library databases. A total of 85 primary studies were analyzed, resulting in the identification of 244 techniques, which were categorized into seven main groups: data balancing, label transformation, sensitive attribute modification, reweighting, fair data generators, score/feature censoring and feature normalization and transformation. The findings indicate a rising adoption of hybrid and context-relevant approaches, as well as the use of multiple metrics to assess both performance and fairness. The study concludes that preprocessing will remain a critical step in building fairer systems, and that the continuous monitoring of emerging techniques is crucial for the ethical and technical advancement of the field.

Keywords: Bias Mitigation, Bias in Machine Learning, Fair Machine Learning, Data Preprocessing, Algorithmic Fairness, Emerging Techniques, Systematic Mapping.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACM	Association for Computing Machinery
ADASYN	Adaptive Synthetic Sampling
ANN	Artificial Neural Networks
ANOVA	Analysis of Variance
ASMO	Adaptive Scattering-Based Minority Class Oversampling
AUROCC	Area Under the Receiver Operating Characteristic Curve
BAR	Biased Action Recognition
BIBM	IEEE International Conference on Bioinformatics and Biomedicine
BigData	IEEE International Conference on Big Data
CBR	Cluster-Based Resampling
CCUS	Cluster Centroid Undersampling
CE	Critério de Exclusão
CFM	Class-aware Feature MixUp
CGANs	Improved Conditional Generative Adversarial Networks
CGDM	Cross-Domain Gradient Discrepancy Minimization
CI	Critério de Inclusão
CMW-Net	Class-aware Meta-Weight-Net
covRew	Coverage-Based Rewriting
CSALI	Cyclic Style ALI
CVF	Computer Vision Foundation
CVPR	IEEE/CVF Conference on Computer Vision and Pattern Recognition
CVPRW	Conference on Computer Vision and Paterna Recognition Workshops
DeFT	Demographic Fairness Transformer
DEM	Digital Elevation Maps
ENN	Edited Nearest Neighbour
EOD	Equal Opportunity Difference
GAN	Generative Adversarial Network
GKDPCA	Gaussian Kernel Density Peak Clustering Algorithm
IEEE	Institute of Electrical and Electronics Engineers
IJCNN	International Joint Conference on Neural Networks
IMITATE	Identify and Mitigate Selection Bias

ISMSIT	International Symposium on Multidisciplinary Studies and Innovative Technologies
KDE	Kernel Density Estimation
K-S	Kolmogorov-Smirnov
LFR	Learning Fair Representations
LID	Local Intrinsic Dimension
LOF	Local Outlier Factor
LRP	Layer-wise Relevance Propagation
MCD-GAN	Maximum Classifier Discrepancy GAN
MCO	Minority Centroid Oversampling
ML	Machine Learning
MMD	Maximum Mean Discrepancy
MNIST	Modified National Institute of Standards and Technology
MSL	Mapeamento Sistemática da Literatura
NICO	Natural Interactive Conversation
NMUS	Near Miss Undersampling
PB	Progressively-balanced
PCA	Principal Component Analysis
PICO	Population, Intervention, Comparison e Outcomes.
QP	Questão da Primária
RFC	Recalibrated Feature Compensation
RFC	Request for Comments
SAUS	Simulated annealing based undersampling
SCM	Structural Causal Model
SMOTE	Synthetic Minority Over-sampling Technique
SSCI	IEEE Symposium Series on Computational Intelligence
STEM	SMOTE-ENN + Mixup
WACV	IEEE/CVF Winter Conference on Applications of Computer Vision

SUMÁRIO

1 INTRODUÇÃO

Além de uma ferramenta técnica, Machine Learning (ML) tornou-se, atualmente, um catalizador de transformações econômicas (ATHEY, 2019), sociais e culturais (FEHER e ZELENKAUSKAITE, 2020). Ao otimizar decisões, personalizar experiências, antecipar comportamentos e acelerar resultados, ele amplia a capacidade humana de lidar com dados e complexidade.

Estudar ML vai além de buscar relevância profissional ou autonomia intelectual. Trata-se de compreender uma tecnologia que já está moldando o presente e que será ainda mais central no futuro da sociedade. E, embora ML traga inúmeros benefícios, também envolve sérios riscos — especialmente quando mal implementado. Por isso, avaliar com olhar crítico ML não é apenas uma questão técnica: é uma forma de se preparar para participar ativamente das transformações sociais, econômicas e éticas do nosso tempo (SIAM e BHATTACHARJEE, 2024).

O pré-processamento de dados é uma das etapas essenciais no processo de desenvolvimento de modelos de ML, pois algoritmos aprendem diretamente com os dados que recebem — e se esses dados estiverem inconsistentes, incompletos ou desbalanceados, o desempenho do modelo será prejudicado. Técnicas como normalização, remoção de ruídos, tratamento de valores ausentes, codificação de variáveis categóricas e seleção de atributos relevantes são fundamentais para garantir que o modelo aprenda de forma confiável (SURESH e GUTTAG, 2019). Logo, o desenvolvimento de um modelo começa na preparação adequada dos dados, antes mesmo do treinamento.

O estudo de ML evolui aceleradamente, com muitos estudos publicados nos últimos anos e com diversas tendências ganhando destaque. Essas tendências refletem uma busca contínua por maior eficiência, acessibilidade e responsabilidade em sistemas inteligentes.

1.1 Justificativa

Num cenário de constante transformação, é fundamental avaliar e acompanhar as técnicas emergentes em ML. Abordagens novas surgem para superar limitações de métodos tradicionais e, ao estudar essas inovações, profissionais da área tornam-se capazes de identificar soluções mais adequadas para problemas específicos, adotando práticas que estejam alinhadas com avanços técnicos e éticos mais recentes. Ignorar tais tendências é correr o risco de

estagnação, e mesmo que bem validados, modelos antigos eventualmente tornam-se obsoletos ou ineficientes.

Avaliar essas técnicas também é uma forma de acompanhar o que está sendo validado na indústria e na academia. Publicações científicas, benchmarks e desafios de competição revelam os métodos que apresentam melhores resultados em problemas reais. Ademais, analisando o desempenho de técnicas emergentes em diferentes domínios — como saúde, finanças ou sistemas embarcados —, é possível antecipar soluções com grande potencial de impacto. Portanto, acompanhar ativamente essa evolução orienta decisões mais fundamentadas na escolha de algoritmos e arquiteturas.

Em suma, para compreender a responsabilidade em se tratar bem os dados, acompanhar as transformações do campo e analisar criticamente eventuais inovações, é imprescindível o estudo de ML. Assim, estuda-se continuamente abordagens e tendências emergentes, o que permite melhorar a performance dos sistemas e alinhar seu uso aos princípios de transparência, equidade e eficiência. Logo, o aprendizado de ML é essencial não apenas pelo contexto técnico, mas também pelo seu papel estratégico, ético e social — sendo uma ferramenta fundamental para quem deseja atuar com responsabilidade e prudência na construção e no desenvolvimento tecnológico da sociedade.

1.2 Objetivos

1.2.1 Geral

Identificar, categorizar e analisar técnicas emergentes de pré-processamento voltadas à mitigação de viés em dados de treinamento, no contexto de aprendizado de máquina, entre os anos de 2020 e 2024, com o intuito de compreender as abordagens mais recentes, suas aplicações e impactos sobre a equidade e o desempenho dos modelos.

1.2.2 Específicos

I. Elaborar um protocolo de mapeamento sistemático com base em diretrizes consolidadas na literatura, definindo a questão de pesquisa, critérios de inclusão e exclusão, e estratégias de busca.

II. Selecionar os artigos relevantes publicados entre 2020 e 2024 que abordem técnicas de pré-processamento para mitigação de viés em dados de treinamento, utilizando filtros automáticos e revisão manual.
III. Extrair e organizar metadados dos estudos selecionados e identificar o tipo de técnica aplicada para a mitigação de viés.
IV. Classificar as técnicas encontradas em categorias com base em suas características e objetivos.
V. Analisar criticamente as tendências observadas e destacar possíveis direções para pesquisas futuras.

1.3 Metodologia

Este trabalho segue os princípios de um Mapeamento Sistemático da Literatura (MSL), conforme orientações de Kitchenham e Charters (2007). O processo inicia com a elaboração de um protocolo de pesquisa contendo a definição da questão central, critérios de inclusão e exclusão, e a estratégia de busca. Foram utilizadas as bases de dados I*EEE Xplore* e *ACM Digital Library*, com *strings* otimizadas para recuperar estudos publicados entre 2020 e 2024. Após a aplicação dos filtros automáticos e da revisão manual de títulos, resumos e textos completos, os estudos selecionados passaram por um processo de extração de dados estruturado. As informações obtidas englobam os métodos utilizados, métricas de avaliação de equidade e desempenho, bem como o contexto de aplicação das técnicas. Estas foram, então, extraídas e organizadas em categorias, e as técnicas mais recorrentes foram analisadas.

O capítulo 4 detalha melhor a metodologia utilizada.

1.4 Estrutura do Trabalho

Além deste capítulo introdutório, este trabalho irá apresentar: o Capítulo 2 apresenta a fundamentação teórica que sustenta esta pesquisa. No Capítulo 3, são discutidos trabalhos relacionados, com uma breve exposição de seus principais resultados. Os o Capítulos 4 e 5 detalham o protocolo adotado para o MSL, e o Capítulo 6 relata os resultados obtidos até a etapa de seleção dos artigos. No Capítulo 7 encontra-se o principal foco da pesquisa, onde é abordada a questão primária. Por fim, o Capítulo 8 trata das considerações finais deste estudo.

2 FUNDAMENTAÇÃO TEÓRICA

Aqui apresentaremos definições importantes para o devido entendimento da pesquisa realizada.

2.1 Machine Learning

Machine Learning, ou aprendizado de máquina, é uma subárea da Inteligência Artificial que busca desenvolver algoritmos capazes de aprender padrões a partir de dados. Em vez de seguir instruções programadas diretamente por humanos, modelos de ML ajustam seus próprios parâmetros com base nas informações que recebem, o que os permite realizarem tarefas como classificação, previsão, agrupamento ou recomendação de forma automatizada (ALPAYDIN, 2016). Tal abordagem tem sido aplicada em diversos contextos, como reconhecimento de voz, diagnósticos médicos, sistemas de recomendação e análise de dados financeiros.

De forma geral, o processo envolve a coleta e preparação de dados, a escolha de um modelo apropriado, o treinamento do modelo com os dados disponíveis e, enfim, a avaliação de seu desempenho. Existem diferentes tipos de aprendizado, sendo os mais comuns o supervisionado, o não-supervisionado e o por reforço. O aprendizado supervisionado utiliza dados rotulados para ensinar o modelo, enquanto o não-supervisionado identifica padrões em dados sem rótulos. Já o aprendizado por reforço é baseado em recompensas e penalidades, sendo mais utilizado em contextos interativos, como jogos ou robótica (ALPAYDIN, 2016).

Apesar do grande potencial, o uso de Machine Learning também levanta desafios importantes, como o risco de viés nos dados, interpretações dos modelos e necessidade de grandes volumes de informação para alcançar bons resultados. Compreender os fundamentos e suas limitações, portanto, é essencial tanto para o uso responsável quanto para o avanço da área. O estudo contínuo de novas técnicas e aplicações permite que pesquisadores e profissionais contribuam com soluções mais eficientes, éticas e alinhadas às necessidades reais da sociedade.

2.2 Pré-processamento

Uma etapa fundamental no desenvolvimento de modelos de aprendizado de máquina, o pré-processamento de dados é responsável por preparar os dados brutos para que sejam utilizados mais eficientemente pelos algoritmos. Este processo busca, de forma geral, limpar dados, (i.e., remoção de valores ausentes ou inconsistentes), normalizar variáveis numéricas, codificar variáveis categóricas e reduzir dimensionalidade. O objetivo principal, portanto, é

garantir que os dados estejam num formato apropriado, livre de ruídos e distorções que possam comprometer o desempenho do modelo (GÉRON, 2019).

Além de melhorar a performance, o pré-processamento pode ter impacto direto na equidade e generalização dos modelos. Desequilíbrios na distribuição de classes ou presença de atributos sensíveis podem introduzir vieses que afetam a imparcialidade das previsões. Balanceamento de classes e remoção de atributos sensíveis são exemplos de estratégias comuns nessa fase, e tais técnicas e afins são essenciais para garantir modelos mais robustos e confiáveis (KELLEHER; NAMEE; D'ARCY, 2015).

2.3 Comparando técnicas de mitigação

Em geral, a comparação entre técnicas de despolarização (ou mitigação de viés) depende de critérios quantitativos e qualitativos, que buscam balancear equidade, eficiência, precisão. Para a comparação de técnicas de pré-processamento, a análise envolve simultaneamente métricas de justiça e desempenho preditivo, dado que a alteração nos dados pode impactar de forma significativa a imparcialidade e a acurácia do modelo final. Sendo assim, os modelos devem não apenas reduzir a disparidade dentre os grupos, mas também preservar performance. O estudo de Friedler *et al.* (2019) evidencia os *trade-offs* resultantes de métodos aplicados em diferentes métricas e datasets padronizados, quando estes buscam alcançar equidade sem comprometer a utilidade.

Barocas, Hardt e Narayanan (2019) detalham como não existe uma única métrica de justiça que seja absolutamente adequada para todos os contextos, o que torna essencial a utilização de diferentes avaliações. Ademais, os autores também enfatizam que justiça não deve ser reduzida apenas a fórmulas matemáticas, e que as escolhas de métrica, técnica e intervenção devem estar alinhadas com as aplicações reais.

3 TRABALHOS RELACIONADOS

A pesquisa de Wang e Liu (2023) faz um levantamento abrangente sobre métodos de despolarização (de-biasing) em redes neurais, especificamente aquelas treinadas com dados de imagem.

Apresentaram uma definição formal do problema de mitigação de viés e discutiram tópicos relevantes. A meta é treinar um modelo em que a saída dependa apenas do atributo alvo e não dos atributos enviesados.

Utilizam conjuntos de dados comumente usados em estudos sobre mitigação de viés, como MNIST (onde se introduz viés associando classes a cores específicas), conjuntos de dados de reconhecimento facial IMDB, CelebA, UTKFace (onde vieses de idade, gênero e tom de pele são introduzidos ou já são inerentes), e conjuntos de dados do mundo real como Biased Action Recognition (Bar) e NICO.

Wang e Liu (2023) destacam que, embora as redes neurais demonstrem um desempenho notável em várias tarefas de aprendizado de máquina, elas são frequentemente afetadas por vieses de representação. Esses vieses podem levar a previsões injustas ou incorretas, especialmente quando aplicadas a novos dados. A pesquisa detalha os esforços significativos feitos nos últimos anos para melhorar a eficácia da despolarização, particularmente na visão computacional. O objetivo comum de todos os métodos de despolarização é reduzir os impactos de características ruidosas e irrelevantes nos dados de treinamento e melhorar a precisão e a equidade do modelo. Wang e Liu (2023) concluem que a redução de viés é uma direção de pesquisa importante no aprendizado profundo, com o objetivo de construir modelos mais justos e robustos.

Hort *et al.* (2022) conduziram uma revisão abrangente de métodos de mitigação de viés para classificadores de aprendizado de máquina, analisando 341 publicações. Eles categorizaram as abordagens em pré-processamento, em-processamento e pós-processamento, e discutiram as métricas e conjuntos de dados comumente utilizados na avaliação dessas técnicas.

Hort *et al.* (2022) observam que a avaliação dos métodos de mitigação de viés é dificultada pela falta de padronização nos conjuntos de dados e nas métricas utilizadas. A pesquisa destaca a necessidade de benchmarks mais consistentes para permitir comparações mais eficazes entre diferentes abordagens.

Os autores sugerem que pesquisadores e profissionais considerem cuidadosamente o contexto de aplicação ao escolher métodos de mitigação de viés, levando em conta as características dos dados e os objetivos específicos de equidade. Além disso, enfatizam a importância de desenvolver e utilizar benchmarks padronizados para avaliar a eficácia das diferentes técnicas de mitigação.

Este estudo serve como um recurso valioso para entender o panorama atual das técnicas de mitigação de viés em aprendizado de máquina e orienta futuras pesquisas e aplicações práticas na área.

4 O MAPEAMENTO SISTEMÁTICO DA LITERATURA

A pesquisa realizada neste trabalho objetiva identificar tendências dentre os principais métodos e técnicas aplicados para despolarização de dados de treinamento utilizados para aprendizado de máquina. Para tanto, optou-se pela elaboração de um Mapeamento Sistemática da Literatura.

O MSL é um tipo de estudo científico que busca identificar, selecionar e sintetizar as evidências relevantes já publicadas para descrever o panorama geral de um campo. Utiliza uma metodologia padronizada, reprodutível e transparente (KITCHENHAM e CHARTERS, 2007).

De acordo com Kitchenham e Charters (2007), o MSL cobre três fases: planejamento, condução e relatório dos resultados. Estes devem ser bem descritos, mas a análise não considera o impacto das questões da pesquisa nos procedimentos de revisão ou mesmo especifica em detalhes as ferramentas necessárias para a meta-análise.

Durante o planejamento, é realizada a identificação da necessidade da revisão, a definição das questões de pesquisa e o desenvolvimento do protocolo de revisão. Durante a condução, a pesquisa tem fontes de pesquisa, *string* de busca e critérios de seleção definidos. Também é onde é documentada a extração dos dados. Por fim, o relatório dos resultados apresenta e discute o objetivo principal do estudo: as tendências emergentes que foram identificadas.

5 PLANEJAMENTO DO TRABALHO

É necessário avaliar técnicas emergentes de mitigação em viés em dados de treinamento pois, além de ajudar a identificar caminhos promissores, também é mapeado como o campo tem evoluído em resposta aos desafios contemporâneos.

Nos estudos recentes, nota-se foco crescente no desenvolvimento de métodos mais adaptáveis e que não dependam tanto de suposições sobre os dados. Ademais, é evidente a tendência de hibridização, dado que muitos trabalhos combinam múltiplas estratégias, como documentado no estudo amplamente citado de Hort *et al.* (2022).

5.1 Questões da pesquisa

As questões da pesquisa guiam o Mapeamento Sistemático da Literatura, e especificá-las é a parte mais importante do estudo. O principal objetivo deste trabalho busca responder a seguinte questão primária (QP):

- QP: Quais as técnicas emergentes para mitigação de viés em dados de treinamento de Aprendizado de Máquina?

A QP foi estruturada de acordo com os critérios PICO definidos por Sackett *et al.* (2000): *Population* (População), *Intervention* (Intervenção), *Comparison* (Comparação) e *Outcomes* (Resultados). No entanto, optou-se pela exclusão da Comparação, dado que não é relevante para este mapeamento.

Este estudo objetiva identificar e relacionar tendências — i.e. as técnicas utilizadas nos últimos anos — na mitigação de viés em dados de treinamento de aprendizado de máquina. A Tabela 1 demonstra como a QP se alinha com os critérios PIO.

TABELA 1 – QP DE ACORDO COM PIO

População (P)	Dados de treinamento de aprendizado de máquina
Intervenção (I)	Técnicas de mitigação de viés
Resultado (O)	Abordagens recentes identificadas na literatura

Fonte: elaborado pelo autor (2025).

Também foram elaboradas questões secundárias (QS1, QS2 e QS3) relacionadas à QP, com objetivo de guiarem a análise dos estudos primários e auxiliar a identificação de características importantes para a síntese da QP.

- QS1: Os estudos analisados são de iniciativa acadêmica ou industrial?
- QS2: Que métricas são utilizadas para avaliar a eficácia das técnicas apresentadas?
- QS3: Os estudos realizados utilizam de medidas além de técnicas pré-processamento para a mitigação de viés?

5.2 Protocolo para o mapeamento

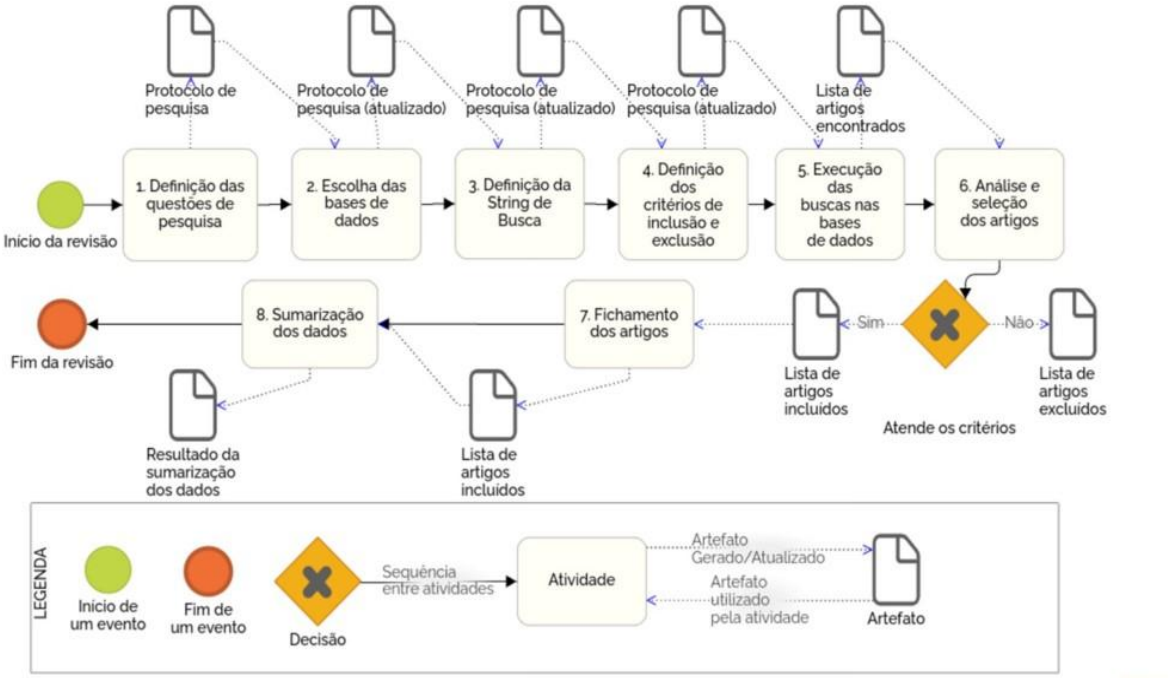
Para garantir a qualidade e reprodutibilidade do estudo, primeiramente definem-se as etapas a serem seguidas para a seleção e análise dos resultados do trabalho. A Figura 1 esquematiza o processo a ser seguido.

Inicialmente, são definidas as questões de pesquisa e as bases de dados a serem consultadas. Em seguida, a *string* de busca é elaborada e refinada, utilizando palavras-chave extraídas a partir da QP. Os critérios de seleção são então elaborados e é feita a busca nas bases

de dados. O resultado é analisado e pré-selecionado. Cada estudo é, enfim, avaliado e selecionado de acordo com os critérios estabelecidos. A etapa de condução finaliza com a análise e sumarização do resultado final da busca, e é aqui que as Questões Secundárias são abordadas.

A etapa seguinte cobre o objetivo principal do mapeamento. Portanto, a Questão Primária é respondida em detalhes e, com isto, busca-se disponibilizar os resultados obtidos para servirem como referência para trabalhos futuros.

FIGURA 1 – ELABORAÇÃO DO MAPEAMENTO SISTEMÁTICO DA LITERATURA



Fonte: elaboração pelo autor (2025).

5.2.1 Fontes de pesquisa

As fontes de dados a serem consultadas devem atender aos seguintes critérios:

- Oferecer suporte para consultas por meio digital, online e automatizado;
- Viabilizar o acesso integral às publicações através do Portal de Periódicos da CAPES, do domínio da Universidade Federal do Amapá ou a partir do Google e/ou Google Scholar;
- Deter de publicações que estejam em português, espanhol ou inglês;
- Não deve acarretar ônus financeiro para os pesquisadores.

Dadas as condições, as fontes de dados selecionadas foram da IEEE Xplore e da ACM Digital Library. Além de atenderem aos critérios, estas bases de dados dispõem de grandes acervos, com alta qualidade e confiabilidade, e são referência para as áreas de tecnologia, engenharia e computação.

5.2.2 *String* de busca

Extraindo termos e palavras-chave da Questão Primária e dos critérios PIO definidos na Tabela 1, a *string* de busca é elaborada. Esta é apresentada na Tabela 2 de forma geral, dado que as bases de dados selecionadas usam de diferentes motores de busca e requerem ajustes na *string* para a adaptação da consulta.

TABELA 2 – *STRING* DE BUSCA

(bias OR fairness OR discrimination) AND (dataset* OR "data set*" OR "training data*") AND (adjust* OR mitigat* OR remov* OR correct* OR optimizat* OR reduc* OR minimiz*) AND ("machine learning" OR "deep learning" OR "supervised learning") AND ("case study" OR experiment OR empirical OR "real-world" OR implement* OR practical OR propos* OR novel) NOT (interview OR survey OR "position paper")

Fonte: elaboração pelo autor (2025).

Os termos inclusivos adicionados na quinta cláusula e os termos exclusivos adicionados na última cláusula têm como objetivo filtrar publicações que não apresentem ou implementem as técnicas de mitigação.

Devido o estudo buscar a identificação de tendências recentes, o período estabelecido foi entre janeiro de 2020 e dezembro de 2024, compreendendo um intervalo de 5 anos.

5.2.3 Critérios de seleção

Os critérios de seleção são divididos em Critério de Inclusão (CI) e Critério de Exclusão (CE), descritos na Tabela 3. Estes são estabelecidos pelos pesquisadores de forma a garantir

que a seleção possa classificar os estudos adequadamente (KITCHENHAM e CHARTES, 2007). Aqui, há um cuidado para definir os critérios de forma que os estudos sejam acessíveis sem comprometer o alcance da análise.

TABELA 3 – DESCRIÇÃO DOS CRITÉRIOS DE SELEÇÃO

ID	Critérios de Inclusão (CI) e de Exclusão (CE)
CI1	Estudos que apresentem, primária ou secundariamente, técnicas de mitigação de viés em dados de treinamento para *Machine Learning* (ML).
CE1	Artigos que não estejam disponíveis livremente para consulta ou download (em versão completa) nas fontes de pesquisa ou por meio de busca manual (para artigos que não sejam fornecidos na íntegra) realizada nas ferramentas de busca Google (http://www.google.com.br/) e/ou Google Scholar (http://scholar.google.com.br/).
CE2	Artigos não relacionados aos objetivos da pesquisa.
CE3	Artigos repetidos (em mais de uma fonte de busca) tiveram apenas sua primeira ocorrência considerada.
CE4	Estudos enquadrados como resumos, keynote speeches, cursos, tutoriais e afins.
CE5	Artigos que não mencionam as palavras-chave da pesquisa no título, resumo ou nas palavras-chave do artigo.
CE6	Excluir se o estudo não estiver apresentado em uma das linguagens aceitas (Inglês e Português).

Fonte: elaboração pelo autor (2025).

5.2.4 Procedimentos de seleção

A seleção dos estudos inicialmente foi realizada através da análise de títulos, resumos e palavras-chave. Caso estes dados não sejam suficientes para determinar a inclusão ou exclusão do mapeamento, a cópia íntegra do estudo foi analisada.

Após esta pré-seleção, os critérios de seleção CI e CE são aplicados. Os estudos restantes são a seleção final para o mapeamento sistemático.

5.2.5 Extração dos dados

Os metadados extraídos de cada trabalho foram:

- Tipo de origem (i.e. conferência, periódico, simpósio ou workshop);

- Ano de publicação;
- Instituições responsáveis pelo estudo;
- Países das instituições responsáveis;
- *Publisher* (i.e. entidade responsável pela publicação);
- Origem;
- Autores;
- Título, resumo e palavras-chave.

5.2.6 Síntese dos dados extraídos

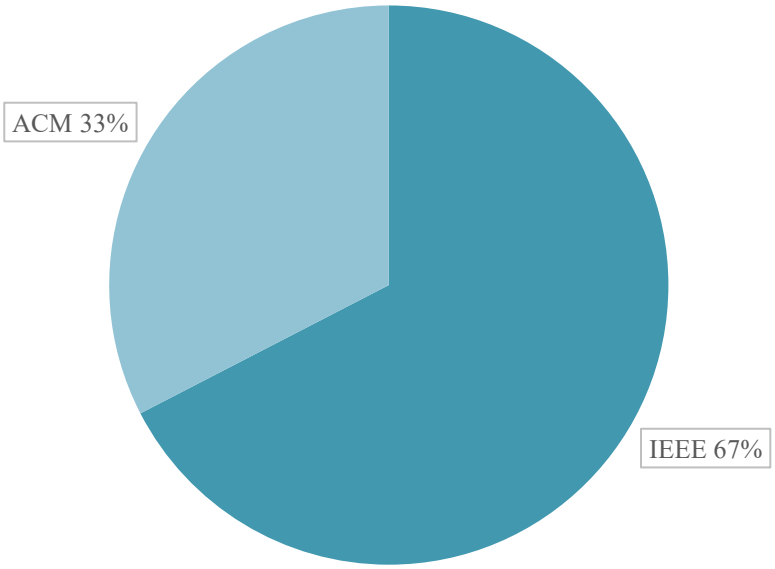
Os estudos são classificados de acordo com os metadados extraídos. Os dados serão tabelados ou graficamente representados de acordo com:

- Número de publicações por ano;
- Número de publicações por país envolvido;
- Número de publicações por tipo de origem do estudo;
- Origem com maior número publicações, por tipo de origem;
- Entidades com maior número de publicações, por tipo de entidade (i.e. instituição, autor ou país).

6 RESULTADOS DOS ESTUDOS PRIMÁRIOS

Com a string de busca anteriormente definida, realizou-se a busca nas bases de dados IEEE Xplore e ACM Digital Library, das quais resultaram em 844 e 408 trabalhos, respectivamente. Os resultados somados enumeram um total de 1252 trabalhos retornados, conforme a Figura 2.

FIGURA 2 – PROPORÇÃO DE ESTUDOS POR BASE DE DADOS



Fonte: elaboração pelo autor (2025).

Durante a pré-seleção, analisando títulos, resumos e palavras-chave (e nos casos aplicáveis, da cópia íntegra do estudo), 169 estudos foram selecionados. Em seguida, com a classificando dos estudos de acordo com os CE e CI definidos, 86 artigos restaram, os quais foram selecionados para análise e extração de dados.

A Tabela 4 revela que o CE2 foi o responsável por 68 eliminações: o maior quantitativo dentre as exclusões. Em contraste, os critérios CE5 e CE6 não eliminam nenhum. O baixo quantitativo de exclusões no CE1 (3) sugere que este estudo obteve sucesso em isolar um bom quantitativo de estudos para análise, através das bases de dados, *string* de busca e critérios de seleção definidos.

TABELA 4 – RESULTADO DA SELEÇÃO DE ESTUDOS POR CRITÉRIO

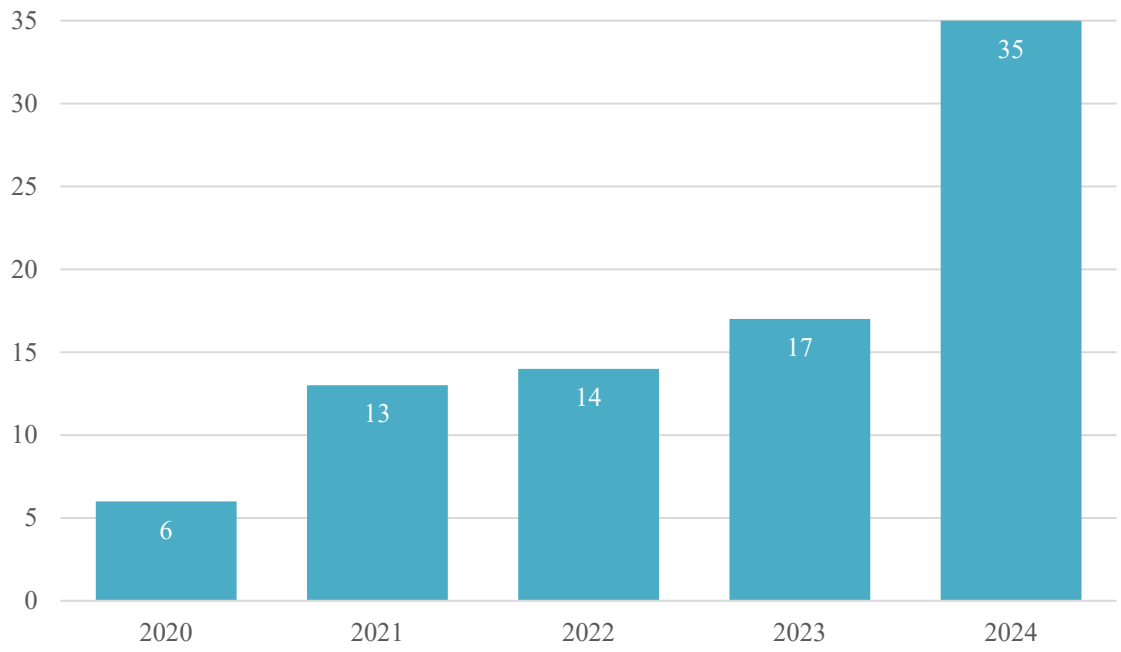
	IEEE	**ACM**	**Total**
Resultado da busca	844	408	1252
Pré-seleção	128	41	169
CE6	0	0	0
CE5	0	0	0
CE4	1	0	1

CE3	2	10	12
CE2	59	9	68
CE1	0	3	3
CI1	67	19	85

Fonte: elaboração pelo autor (2025).

Os estudos são inicialmente classificados de acordo com o ano de publicação, representado na Figura 3. A disposição demonstra um aumento gradual de número de estudos acerca do tema da pesquisa durante o período analisado. Os dados sugerem uma tendência que reforça os objetivos deste mapeamento.

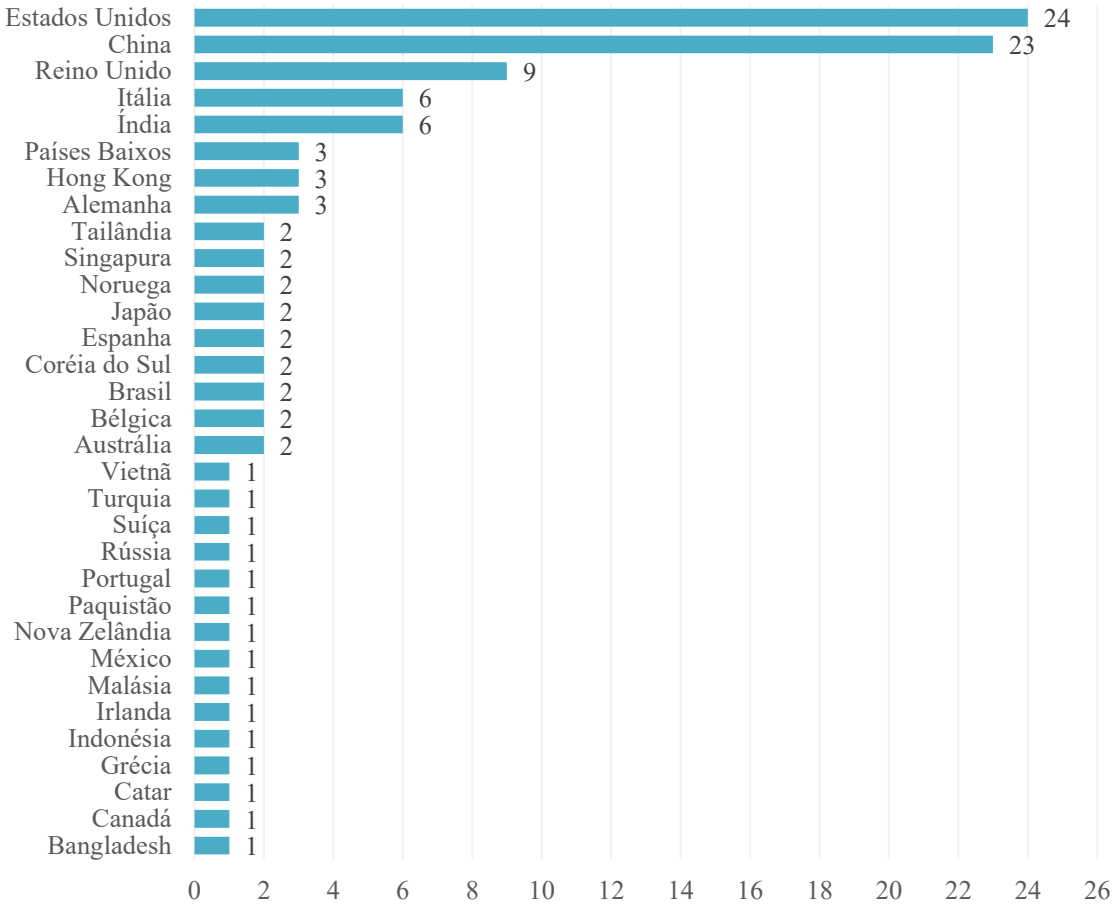
FIGURA 3 – PUBLICAÇÕES POR ANO



Fonte: elaboração pelo autor (2025).

Em seguida, classificamos os trabalhos de acordo com o país. Importante ressaltar que alguns estudos envolvem múltiplos países, o que foi levado em conta na extração dos metadados e na quantificação de publicações por país. A Figura 4 representa essa distribuição, de maior para menor, com Estados Unidos, China e Reino Unido com o maior quantitativo.

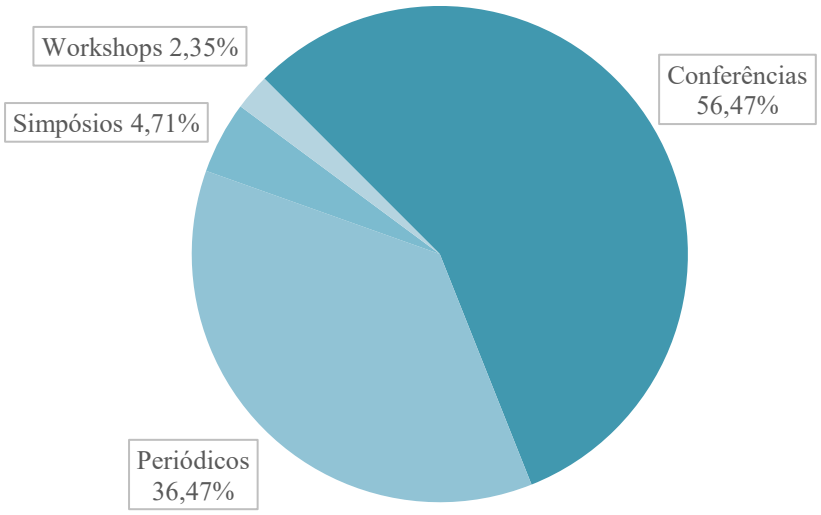
FIGURA 4 – PUBLICAÇÕES POR PAÍS



Fonte: elaboração pelo autor (2025).

A Figura 5 apresenta a proporção de publicações por tipo de origem. A grande maioria das publicações foram originadas de conferências (56,47%) e periódicos (36,47%), enquanto apenas 6 foram oriundas de simpósios (4,71%) e workshops (2,35%).

FIGURA 5 – PROPORÇÃO DE PUBLICAÇÕES POR TIPO DE ORIGEM



Fonte: elaboração pelo autor (2025).

Foram contabilizados 48 estudos publicados em conferências, 31 em periódicos, 4 em simpósios e 2 em workshops. A Tabela 5 registra as origens com maior número de publicações, divididas entre tipos.

TABELA 5 – ORIGENS COM MAIOR NÚMERO DE PUBLICAÇÕES, POR TIPO

Nome	Publicações
Conferências	
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)	6
IEEE International Conference on Big Data (BigData)	3
IEEE International Conference on Bioinformatics and Biomedicine (BIBM)	2
IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)	2
IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)	2
International Joint Conference on Neural Networks (IJCNN)	2
Simpósios	
IEEE Intelligent Vehicles Symposium (IV)	1
IEEE International Geoscience and Remote Sensing Symposium	1
IEEE Symposium Series on Computational Intelligence (SSCI)	1
International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)	1
Periódicos	
ACM Transactions on Knowledge Discovery from Data	2
Artificial Intelligence in Medicine	2
Expert Systems with Applications	2
IEEE Access	2
IEEE Transactions on Pattern Analysis and Machine Intelligence	2
Information Sciences	2
Neurocomputing	2
Workshops	
IEEE/ACM International Workshop on Equitable Data & Technology (FairWare)	1
IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)	1

Fonte: elaboração pelo autor (2025).

A Tabela 6 apresenta as entidades com maior quantidade de publicações, separadas por instituições, autores e países. Da mesma forma que múltiplos países foram identificados para determinados estudos, conforme descrito anteriormente, alguns estudos também foram elaborados por múltiplas instituições. Aqui, isso é levado em conta novamente.

Dentre as três instituições, temos a University of Genoa, da Itália, com 4 publicações, e a San Jose State Universty e a New York University, ambas dos Estados Unidos, com 4 e 2 publicações, respectivamente.

Os autores com mais publicações são Danilo Franco, Luca Oneto e Nicolò Navarin, todos com 2 publicações e oriundos da Itália. Nas duas publicações selecionadas, estes pesquisadores trabalharam juntos. Oneto e Franco estão com a University of Genoa, e apenas Navarin com a University of Padua.

A Itália também está presente entre os países com maior número de publicações, com 6 ocorrências, junto da Índia, com o mesmo número. Os Estados Unidos e a China lideram o gráfico, com 25 e 23 publicações, respectivamente. O Brasil foi identificado apenas em 2 ocorrências dentre os trabalhos selecionados.

TABELA 6 – ENTIDADES COM MAIOR NÚMERO DE PUBLICAÇÕES, POR TIPO

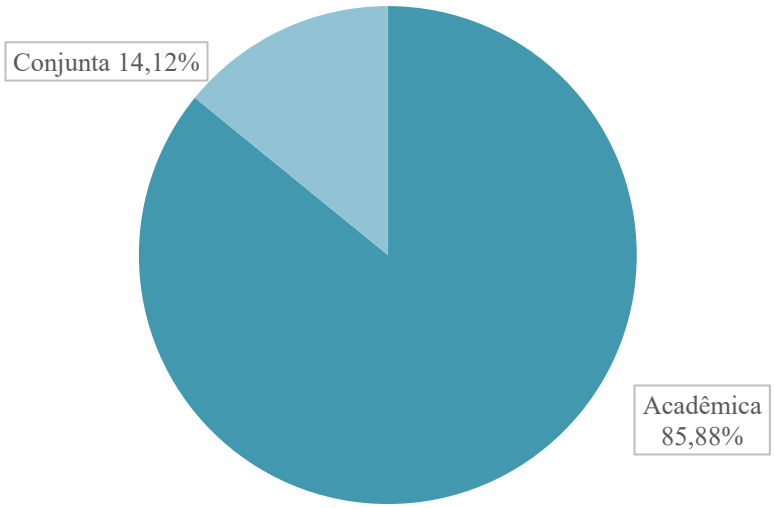
Nome	Publicações
Instituições	
San Jose State University	4
University of Genoa	4
New York University	2
Autores	
Danilo Franco	2
Luca Oneto	2
Nicolò Navarin	2
Países	
Estados Unidos	24
China	23
Reino Unido	9
Índia	6
Itália	6

Fonte: elaboração pelo autor (2025).

- **QS1: Os estudos analisados são de iniciativa acadêmica ou industrial?**

Nenhuma das publicações selecionadas são de iniciativa exclusivamente industrial, embora 73 delas sejam de iniciativa exclusivamente acadêmica. Os outros 12 estudos são de iniciativa conjunta, isto é, resultam da colaboração entre instituições acadêmicas e industriais. A Figura 6 demonstra a proporção entre as quantidades.

FIGURA 6 – PROPORÇÃO DE INICIATIVA

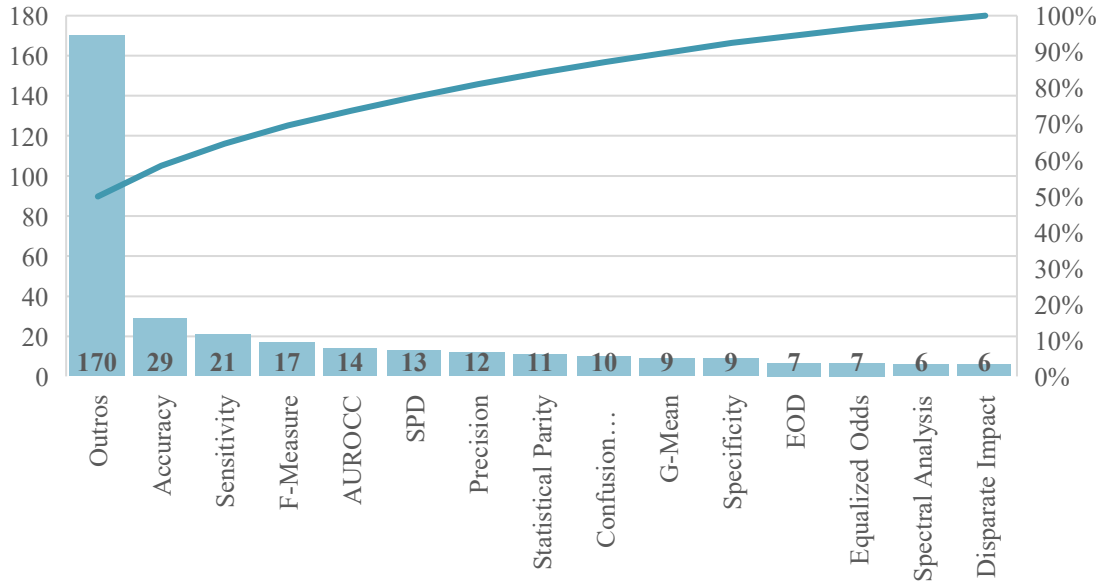


Fonte: elaboração pelo autor (2025).

- **QS2: Que métricas são utilizadas para avaliar a eficácia das técnicas apresentadas?**

Dezenas de métricas foram extraídas dos trabalhos analisados, de diversas categorias, como acurácia, precisão, classificação, erros, correlação, equidade, performance e eficiência. De 341 ocorrências distribuídas em um total de 153 diferentes métricas identificadas, 171 são atribuídas a apenas 14 métricas, o que representa aproximadamente metade do total de ocorrências (49,56%). A Figura 7 relaciona as métricas mais recorrentes, sendo estas: *Accuracy* (29), *Sensitivity* (21), *F-Measures* (17), *Area Under the Receiver Operating Characteristic Curve – AUROCC* (14), *Statistical Parity Difference – SPD* (13), *Precision* (12), *Statistical Parity* (11), *Confusion Matrices* (10), *Geometric Mean – G-Mean* (9), *Specificity* (9), *Equal Opportunity Difference – EOD* (7), *Equalized Odds* (7), *Disparate Impact* (6) e *Spectral Analysis* (6).

FIGURA 7 – PRINCIPAIS MÉTRICAS IDENTIFICADAS

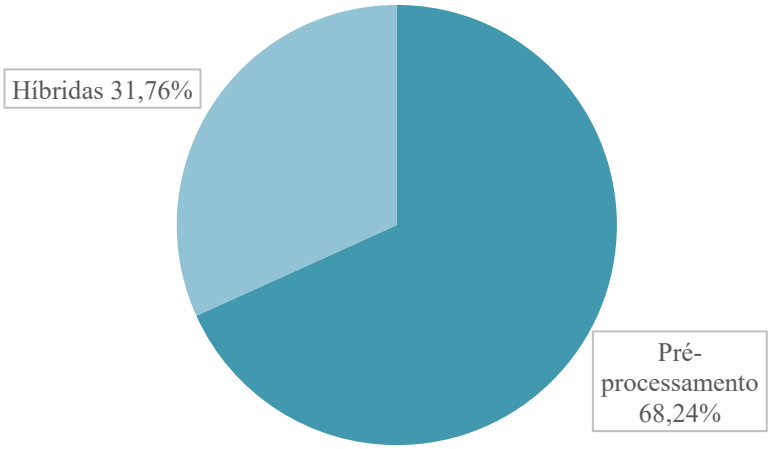


Fonte: elaboração pelo autor (2025).

- **QS3: Os estudos realizados utilizam de medidas além de técnicas pré-processamento para a mitigação de viés?**

Dentre os 85 estudos selecionados, 27 apresentam abordagens híbridas ao realizarem a mitigação de viés através do conjunto de abordagens além da etapa de pré-processamento, sejam estas de em-processamento ou pós-processamento. A Figura 8 representa essa proporção.

FIGURA 8 – PROPORÇÃO DE ABORDAGENS HÍBRIDAS E PURAMENTE PRÉ-PROCESSAMENTO



Fonte: elaboração pelo autor (2025).

7 ANÁLISE DOS RESULTADOS QUANTO À QUESTÃO PRIMÁRIA

Neste capítulo, abordamos a Questão Primária da pesquisa. Para tanto, foram extraídas de cada estudo as técnicas de mitigação de viés utilizadas ou introduzidas durante a fase de pré-processamento, e descartadas as técnicas identificadas em fases adiantes.

O Apêndice A lista todas as técnicas registradas, agrupadas em sete categorias. Estas serão discutidas adiante no capítulo. Para cada entrada, há também os códigos de referência dos estudos que aplicaram. O Apêndice B representa uma tabela com o título de cada estudo, com devida referência.

O total de técnicas identificadas foi 244. Dado o grande quantitativo listado e o objetivo do presente estudo ser a identificação de tendências emergentes, optou-se pela descrição apenas das técnicas mais utilizadas. Isto é, das que foram aplicadas em pelo menos dois estudos diferentes. A ordenação foi realizada prioritariamente de acordo com o número de ocorrências e secundariamente por ordem alfabética.

7.1 Balanceamento de dados

Ajustam a distribuição das classes (ou grupos sensíveis, tipo gênero ou raça) para que fiquem mais equilibradas.

7.1.1 *Random Oversampling*

É uma técnica de reamostragem que replica aleatoriamente exemplos da classe minoritária até que haja equilíbrio entre as classes no conjunto de dados. É um método simples e amplamente utilizado como *baseline* em cenários de desbalanceamento, tanto em classificação geral quanto em contextos sensíveis como gênero ou raça. Apesar de melhorar o aprendizado da classe minoritária durante o treinamento, pode aumentar o risco de *overfitting*, já que os exemplos adicionados são cópias exatas dos existentes (CHAWLA *et al.*, 2002).

Identificada nos estudos: E1, E8, E34, E57, E77 e E78.

7.1.2 *Class Imbalance Rebalancing*

Uma técnica de pré-processamento que busca corrigir distribuições desiguais entre classes ou grupos sensíveis no conjunto de treinamento. Isso é feito por meio de estratégias como *oversampling* (replicação ou geração de exemplos da classe minoritária), *undersampling* (remoção de exemplos da classe majoritária) ou amostragem balanceada com base em atributos protegidos. O objetivo é garantir representação justa das classes e reduzir o viés aprendido pelo modelo, evitando que ele favoreça padrões da classe dominante (BUDA *et al.*, 2018).

Identificada nos estudos: E12, E25, E46, E56 e E65.

7.1.3 *ADASYN (Adaptive Synthetic Sampling)*

Uma técnica de *oversampling* em nível de dados que gera amostras sintéticas para a classe minoritária com base na dificuldade de aprendizado de cada instância. Diferente de métodos como o SMOTE, que distribuem as amostras sintéticas uniformemente, o ADASYN concentra a geração de novas amostras nas regiões onde a minoria é mais difícil de classificar, ou seja, onde há maior sobreposição com a classe majoritária. Isso ajuda a reduzir o viés do modelo em relação à maioria e melhora a representação da classe minoritária, promovendo maior equilíbrio e potencialmente melhorando a equidade do modelo (HE *et al.,* 2008).

Identificada nos estudos: E13, E33 e E57.

7.1.4 *Balanced Training Data*

Busca mitigar vieses e *overfitting* promovendo representatividade equitativa entre grupos no conjunto de treino. Isso pode envolver remover exemplos de grupos super-representados (*undersampling*) ou ajustar proporções de forma explícita (e.g., 50:50), garantindo que o modelo não favoreça características específicas de uma classe ou grupo demográfico. A ideia central é que, ao treinar com dados balanceados, o modelo tende a aprender padrões mais gerais e justos, evitando viés sistemático (ZHAO *et al.*, 2017).

Identificada nos estudos: E6, E39 e E61.

7.1.5 *Hybrid Resampling*

Definido como a combinação de técnicas de *oversampling* e *undersampling* para lidar com o desequilíbrio de classes. O objetivo é aproveitar as vantagens de ambas: aumentar a representatividade das classes minoritárias sem inflar demais o conjunto de dados, e reduzir o

excesso de exemplos da classe majoritária sem perda drástica de informação (NEWAZ e HAQ, 2022).

Identificada nos estudos: E5, E23 e E68.

7.1.6 *Minority Class Augmentation*

Data augmentation focada em classes minoritárias é uma técnica de pré-processamento usada para reduzir o desequilíbrio de classes em *datasets*, especialmente quando há poucas amostras disponíveis para determinadas categorias ("*tail classes*"). A ideia central é aplicar transformações (como rotação, espelhamento, adição de ruído ou métodos como *Mixup*) exclusivamente ou prioritariamente nas classes com menor representação, de forma a ampliar sua presença no conjunto de treino sem gerar sobreposição indesejada com classes majoritárias. Isso melhora a capacidade do modelo de generalizar para essas classes raras e reduz o viés em favor das classes dominantes (ZHANG *et al.*, 2018).

Identificada nos estudos: E9, E17 e E58.

7.1.7 *Neighbourhood-based Undersampling*

Um *undersampling* que consiste em remover aleatoriamente amostras da classe majoritária, considerando a proximidade entre exemplos para minimizar a perda de informações relevantes. Essa abordagem busca balancear o conjunto de dados ao reduzir a predominância da classe majoritária, evitando, na medida do possível, a exclusão de padrões importantes ao levar em conta a vizinhança das amostras removidas (KUBAT; MATWIN, 1997).

Identificada nos estudos: E8, E15 e E69.

7.1.8 *SMOTE-ENN (SMOTE + Edited Nearest Neighbour)*

Sendo uma técnica híbrida e de reamostragem, o SMOTE-ENN combina o *oversampling* do SMOTE com o *undersampling* do Edited Nearest Neighbour (ENN). O SMOTE gera exemplos sintéticos da classe minoritária por interpolação entre vizinhos, enquanto o ENN remove amostras da classe majoritária (e possivelmente sintéticas) que são mal classificadas por seus vizinhos mais próximos. Essa combinação mitiga tanto o desbalanceamento entre classes quanto o ruído interno, limpando regiões de sobreposição e melhorando a separabilidade entre as classes (BATISTA *et al.*, 2004).

Identificada nos estudos: E57, E75 e E78.

7.1.9 *Statistical Parity Correction*

No pré-processamento, essas técnicas visam reduzir disparidades nos resultados entre grupos privilegiados e não privilegiados, ajustando os dados antes do treinamento. Uma técnica comum é o *Disparate Impact Remover*, que modifica os valores dos atributos de entrada de forma a tornar menos distinguíveis os grupos sensíveis (por exemplo, gênero ou raça), ao mesmo tempo em que tenta preservar a utilidade preditiva. O objetivo é garantir que a taxa de resultados favoráveis seja mais equilibrada entre os grupos, reduzindo o viés sistêmico do modelo (FELDMAN *et al.*, 2015).

Identificada nos estudos: E25, E51 e E73.

7.1.10 *Missing Data Handling*

Refere-se ao processo de tratar dados incompletos para evitar vieses que possam surgir devido à ausência de informações em variáveis-chave. Uma abordagem comum é excluir completamente os registros com dados faltantes, prevenindo que valores ausentes distorçam a distribuição ou influenciem negativamente o treinamento do modelo. Contudo, essa exclusão pode reduzir o tamanho do conjunto de dados, e alternativas como imputação podem ser consideradas dependendo do contexto (LITTLE; RUBIN, 2019).

Identificada nos estudos: E22 e E33.

7.1.11 *Undersampling*

Técnicas que reduzem o número de exemplos da classe majoritária, especialmente em regiões onde essa classe está excessivamente representada, para equilibrar a distribuição das classes. Métodos comuns usam critérios estatísticos, como limiares de z-score, para identificar áreas ou grupos com predominância da classe majoritária e aplicam remoção aleatória de amostras, buscando manter a diversidade dos dados enquanto mitigam o viés do modelo em favor da classe dominante (LIU *et al.*, 2008).

Identificada nos estudos: E52 e E72.

7.2 Transformação de Labels

Mudam os rótulos (targets) para correção de viés histórico. Se decisões passadas discriminaram, rotular novamente pode resultar num julgamento mais justo.

7.2.1 *Label Adjustment*

Consiste em modificar os rótulos de amostras para corrigir disparidades entre grupos protegidos, buscando equilibrar a taxa de resultados positivos entre eles. O método pode envolver trocar rótulos negativos de membros de grupos desfavorecidos por positivos (*positive correction*) e vice-versa para grupos favorecidos (*negative correction*), conforme regras que garantam um limiar mínimo de paridade. Essa abordagem ajusta diretamente a distribuição das classes sem alterar o tamanho do conjunto, mas pode introduzir distorções nos dados originais e apresenta desafios relacionados a discriminação por proxy (KAMIRAN; CALDERS, 2012).

Identificada nos estudos: E10, E41, E46 e E68.

7.2.2 *Soft Label Smoothing*

É usada para lidar com ruído nos rótulos (*label noise*), substituindo rótulos binários ou categóricos rígidos (*hard labels*) por distribuições probabilísticas (*soft labels*). Em vez de tratar um exemplo como pertencente 100% a uma classe, atribui-se uma distribuição de probabilidade sobre as classes, o que permite ao modelo aprender com incertezas e evita o sobreajuste a rótulos incorretos. Estratégias como o uso de *pseudo-labels* ou combinação com previsões do próprio modelo ajudam a refinar os rótulos ao longo do treinamento, reduzindo o viés causado por erros de anotação (REED *et al.*, 2015).

Identificada nos estudos: E14 e E21.

7.3 Modificação de Atributos Sensíveis

Remove ou modifica atributos sensíveis (como sexo ou raça) para reduzir influência direta ou indireta no modelo.

7.3.1 *Fair Representation Learning*

Técnica de pré-processamento que busca transformar os dados de entrada em uma nova representação latente que seja independente de atributos sensíveis, como gênero ou raça, ao mesmo tempo em que preserva a utilidade preditiva dos dados. Essa abordagem geralmente envolve a otimização de múltiplos objetivos: minimizar a perda de informação, garantir acurácia preditiva e satisfazer critérios de justiça, como *statistical parity* ou *equal opportunity*. Métodos como o *Learning Fair Representations (LFR)* utilizam estratégias probabilísticas ou treinamento adversarial para garantir que o novo espaço latente não revele informações discriminatórias, promovendo previsões mais justas (ZEMEL *et al.*, 2013).

Identificada nos estudos: E10, E25, E38, E45, E51, E59, E60, E61 e E73.

7.3.2 *Sensitive Attribute Suppression*

Busca remover atributos sensíveis (como raça, gênero) — ou atributos altamente correlacionados com eles — do conjunto de dados antes do treinamento do modelo. O objetivo é evitar discriminação direta ou indireta, impedindo que o modelo tenha acesso explícito ou implícito a essas informações. Embora essa abordagem reduza o risco de viés direto, ela não garante imparcialidade, pois o modelo ainda pode inferir atributos sensíveis a partir de variáveis correlacionadas remanescentes (DWORK *et al.*, 2012).

Identificada nos estudos: E46, E62 e E73.

7.3.3 *Adversarial Debiasing*

Utiliza uma estrutura adversarial para remover informações sobre atributos sensíveis (como gênero ou raça) das representações latentes aprendidas. O método envolve dois componentes principais: um codificador, que aprende a representar os dados, e um adversário, que tenta prever o atributo sensível a partir da representação. O codificador é treinado para minimizar a perda de tarefa principal (ex.: classificação) e maximizar a perda do adversário, forçando as representações a se tornarem invariantes aos atributos sensíveis e, assim, promovendo maior equidade (EDWARDS; STORKEY, 2016).

Identificada nos estudos: E54 e E59.

7.4 Reponderação

Atribui pesos diferentes aos exemplos, favorecendo os subgrupos menos representados ou prejudicados. Não altera os dados, mas manipula a importância deles.

7.4.1 *Instance Reweighting*

Estratégia usada para mitigar vieses em modelos de aprendizado de máquina por meio da atribuição de pesos diferenciados às instâncias de treino. A ideia central é que exemplos de grupos sub-representados ou frequentemente mal classificados recebam maior influência durante o treinamento, enquanto instâncias de grupos majoritários ou bem representados tenham peso reduzido. Essa técnica pode ser aplicada tanto na função de perda quanto diretamente nos dados de entrada, seja por métodos estáticos (com pesos pré-definidos) ou dinâmicos (ajustados ao longo do treinamento, via meta-learning ou frameworks como AIF360). O objetivo é equilibrar a representação e reduzir disparidades nos resultados entre grupos protegidos e não protegidos (KAMIRAN; CALDERS, 2012).

Identificada nos estudos: E3, E14, E19, E21, E25, E39, E44, E45, E46, E58 e E73.

7.5 Geradores de Dados Justos

Fabrica dados sintéticos balanceados, com menos viés. Os dados gerados devem promover equidade, sem distorcer a realidade.

7.5.1 *Synthetic Minority Oversampling Technique (SMOTE)*

SMOTE (Synthetic Minority Over-sampling Technique) é uma técnica de *oversampling* que gera exemplos sintéticos da classe minoritária por meio da interpolação entre instâncias existentes e seus vizinhos mais próximos. Diferente do *Random Oversampling*, que apenas duplica exemplos, o SMOTE cria novas amostras artificiais, o que reduz o risco de *overfitting*. É amplamente usado para lidar com desequilíbrio de classes em tarefas de classificação, aumentando a representatividade da minoria sem perder diversidade no conjunto de dados (CHAWLA *et al.*, 2002).

Identificada nos estudos: E8, E12, E34, E48, E50, E57, E75 e E78.

7.5.2 *Generative Data Augmentation*

Utiliza modelos generativos, como GANs ou difusão, para criar amostras artificiais de grupos sub-representados, com o objetivo de balancear a distribuição do dataset em relação a atributos protegidos (ex.: raça, idade, tonalidade de pele). As identidades geradas podem variar em pose, iluminação e expressão para preservar diversidade intra-classe, enquanto a manipulação controlada de atributos garante a representatividade de minorias. Essa abordagem corrige desequilíbrios sem remover dados originais, mas requer cuidado para evitar artefatos e *outliers* irreais (ANTONIOU *et al.*, 2017).

Identificada nos estudos: E10, E32, E56 e E76.

7.6 Censura de Score/Feature

Oculta ou remove variáveis ou scores que contenham viés implícito. Funciona como um filtro que auxilia o modelo a evitar decisões enviesadas. Censura de Score/Feature foi a única categoria em que nenhuma técnica foi aplicada mais de uma vez.

7.7 Normalização e Transformações de Feature

Devem evitar que diferenças numéricas injustificadas entre grupos influenciem o modelo.

7.7.1 *Fairness-driven Binning*

Binning é uma técnica de pré-processamento que transforma atributos contínuos em categorias discretas, agrupando valores em intervalos definidos. Essa abordagem é usada para reduzir vieses associados a escalas numéricas contínuas, como em atributos sensíveis (ex: idade), que podem afetar desproporcionalmente decisões do modelo. Além de facilitar a análise estatística e emparelhamento entre grupos, o *binning* também pode ajudar na estabilidade do modelo ao lidar com ruído em dados tabulares (DWORK *et al.*, 2012).

Identificada nos estudos: E16, E22 e E46.

7.7.2 *Principal Component Analysis (PCA)*

É uma técnica de redução de dimensionalidade que transforma dados originais em um novo espaço de variáveis ortogonais (componentes principais), ordenadas por variância. Ao

reter apenas os componentes mais informativos (por exemplo, 80% da variância), o PCA elimina redundância e ruído, o que ajuda a mitigar viés causado por atributos irrelevantes ou altamente correlacionados em espaços de alta dimensionalidade. Isso torna os modelos mais eficientes e menos propensos a superajuste em padrões espúrios (JOLLIFFE; CADIMA, 2016).

Identificada nos estudos: E1, E3 e E77.

7.7.3 *Data Normalization*

Visa padronizar a escala dos atributos numéricos para evitar que variáveis com magnitudes maiores dominem o processo de aprendizado. Isso é feito com métodos como *z-score* (zero média e desvio padrão unitário) ou *min-max scaling*, garantindo que todas as features tenham influência proporcional na função de perda. A normalização contribui para a redução de vieses estruturais ligados à escala dos dados e melhora a convergência dos algoritmos de treinamento (JAIN *et al.*, 2005).

Identificada nos estudos: E13 e E40.

8 CONSIDERAÇÕES FINAIS

O trabalho desenvolvido buscou, através de um MSL, a identificação das técnicas emergentes de pré-processamento aplicadas à mitigação de viés em dados de treinamento, com foco no período de 2020 a 2024, e também identificar tendências. Para tanto, elaborou-se um protocolo com base em diretrizes estabelecidas para MSL, com devidos critérios de inclusão e exclusão, além de uma estratégia de busca estruturada em repositórios científicos relevantes. Os estudos selecionados foram analisados, com extração e categorização das técnicas identificadas.

Foram identificadas 244 técnicas distintas, distribuídas em sete categorias principais, definidas com base nas características de implementação. Observou-se um aumento no uso de múltiplas abordagens, sensíveis ao contexto e combinadas com múltiplas métricas de avaliação de justiça e desempenho. Isso evidencia a evolução das soluções propostas na literatura recente, com maior sofisticação e preocupação ética.

Os dados sintetizados por este mapeamento sistemático podem auxiliar pesquisadores e profissionais da área a entender o atual estado da pesquisa científica sobre mitigação de viés em pré-processamento, o que fornece base para decisões metodológicas mais embasadas. Ao reunir e categorizar técnicas recentes durante uma fase específica do treinamento, o estudo oferece um panorama valioso para orientar a aplicação de abordagens existentes, assim como o desenvolvimento de novas soluções mais eficazes e justas em sistemas de aprendizado de máquina.

Sugere-se, para trabalhos futuros, a ampliação da pesquisa para outras bases científicas importantes ou o foco em fases seguintes do processo de treinamento, ainda no contexto de busca e sintetização das informações. Outro caminho interessante seria a realização de estudos empíricos que avaliassem e comparassem o desempenho e a efetividade das técnicas mapeadas em diferentes domínios de aplicação. Também seriam relevantes o estudo e desenvolvimento de abordagens automatizadas de seleção de técnicas de mitigação, com base nas individualidades da área de aplicação, das características dos dados e das métricas de equidade desejadas.

REFERÊNCIAS

ALPAYDIN, Ethem. **Aprendizado de máquina**. 3. ed. Porto Alegre: Bookman, 2016.

ANTONIOU, Antreas; STORKEY, Amos; EDWARDS, Harrison. **Data augmentation generative adversarial networks**. 2017. Disponível em: <https://doi.org/10.48550/arXiv.1711.04340>. Acesso em: 4 jul. 2025.

ATHEY, Susan. **The impact of machine learning on economics**. In: AGRAWAL, Ajay; GANS, Joshua; GOLDFARB, Avi (org.). The economics of artificial intelligence: an agenda. Chicago: University of Chicago Press, 2019. p. 507-547.

BAROCAS, Solon; HARDT, Moritz; NARAYANAN, Arvind. **Fairness and machine learning**: limitations and opportunities. Cambridge: The MIT Press, 2019. 340 p. Disponível em: <https://fairmlbook.org/>. Acesso em: 4 jul. 2025.

BATISTA, Gustavo E. A. P.; PRATI, Ronaldo C.; MONARD, Maria Carolina. A study of the behavior of several methods for balancing machine learning training data. **ACM SIGKDD Explorations Newsletter**, v. 6, n. 1, p. 20-29, 2004. DOI: https://doi.org/10.1145/1007730.1007735.

BUDA, Mateusz; MAKI, Atsuto; MAZUROWSKI, Maciej A. **A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks**, v. 106, p. 249-259, 2018. DOI: https://doi.org/10.1016/j.neunet.2018.07.011.

CALIVÁ, Francesco; RIBEIRO, Fabio Sousa De; MYLONAKIS, Antonios; DEMAZIÈRE, Christophe; VINAI, Paolo; LEONTIDIS, Georgios. A deep learning approach to anomaly detection in nuclear reactors. In: INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS – IJCNN, 2018, Rio de Janeiro, Brasil. **Proceedings**... Los Alamitos: IEEE, 2018. DOI: https://doi.org/10.1109/IJCNN.2018.8489130.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321-357, 2002. DOI: https://doi.org/10.1613/jair.953.

DWORK, Cynthia; HARDT, Moritz; PITASSI, Toniann; REINGOLD, Omer; ZEMEL, Richard. Fairness through awareness. In: INNOVATIONS IN THEORETICAL COMPUTER SCIENCE CONFERENCE, 3., 2012, Cambridge, MA. **Proceedings**... New York: ACM, 2012. p. 214-226. DOI: https://doi.org/10.1145/2090236.2090255.

EDWARDS, Harrison; STORKEY, Amos. **Censoring representations with an adversary**. 2016. Disponível em: <https://doi.org/10.48550/arXiv.1511.05897>. Acesso em: 4 jul. 2025.

FEHER, Katalin; ZELENKAUSKAITE, Asta. AI in society and culture: decision making and values. In: CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS EXTENDED ABSTRACTS – CHI 2020, 2020, Honolulu. **Proceedings**... New York: ACM, 2020. DOI: https://doi.org/10.48550/arXiv.2005.02777.

FELDMAN, Michael; FRIEDLER, Sorelle A.; MOELLER, John; SCHEIDEGGER, Carlos; VENKATASUBRAMANIAN, Suresh. Certifying and removing disparate impact. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 21., 2015, Sydney. **Proceedings**... New York: ACM, 2015. p. 259-268. DOI: https://doi.org/10.1145/2783258.2783311.

FRIEDLER, S. A.; SCHEIDEGGER, C.; VENKATASUBRAMANIAN, S. A comparative study of fairness-enhancing interventions in machine learning. In: CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY – FAT*, 2019. **Proceedings**... New York: ACM, 2019. p. 329–338.

GÉRON, Aurélien. **Mãos à obra**: aprendizado de máquina com scikit-learn, keras e tensorflow. 2. ed. Rio de Janeiro: Alta Books, 2019.

HE, Haibo; BAI, Yang; GARCIA, E. A.; LI, Shutao. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: IEEE INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IEEE WORLD CONGRESS ON COMPUTATIONAL INTELLIGENCE), 2008, Hong Kong. **Proceedings**... Piscataway: IEEE, 2008. p. 1322-1328. DOI: https://doi.org/10.1109/IJCNN.2008.4633969.

HORT, Max; CHEN, Zhenpeng; ZHANG, Jie M.; HARMAN, Mark; SARRO, Federica. **Bias mitigation for machine learning classifiers**: a comprehensive survey. 2022. Disponível em: <https://arxiv.org/abs/2207.07068>. Acesso em: 4 jul. 2025.

JAIN, A. K.; DUIN, R. P. W.; MAO, Jianchang. Statistical pattern recognition: a review**. IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 22, n. 1, p. 4-37, jan. 2005. DOI: https://doi.org/10.1109/34.824819.

JOLLIFFE, Ian T.; CADIMA, Jorge. Principal component analysis: a review and recent developments. **Philosophical Transactions of the Royal Society A**: Mathematical, Physical and Engineering Sciences, v. 374, art. 20150202, 13 abr. 2016. DOI: https://doi.org/10.1098/rsta.2015.0202.

KAMIRAN, Faisal; CALDERS, Toon. Data preprocessing techniques for classification without discrimination. **Knowledge and Information Systems,** v. 33, p. 1-33, 2012. DOI: https://doi.org/10.1007/s10115-011-0463-8.

KELLEHER, John D.; NAMEE, Brian Mac; D'ARCY, Aoife. **Fundamentals of machine learning for predictive data analytics**: algorithms, worked examples, and case studies. Cambridge: The Mit Press, 2015.

KITCHENHAM, Barbara; CHARTERS, Stuart M. Guidelines for performing systematic literature reviews in software engineering. Version 2.3. **EBSE Technical Report EBSE-2007-01**. Keele; Durham: Software Engineering Group, School of Computer Science and Mathematics, Keele University; Department of Computer Science, University of Durham, 9 jul. 2007. Disponível em: <https://www.researchgate.net/publication/302924724>. Acesso em: 4 jul. 2025.

KUBAT, Miroslav; MATWIN, Stan. Addressing the curse of imbalanced training sets: one-sided selection. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 14.,

1997, Nashville. **Proceedings...** San Francisco: Morgan Kaufmann, 1997. p. 179-186. Disponível em: <https://www.researchgate.net/publication/2624358_Addressing_the_Curse_of_Imbalanced_Training_Sets_One-Sided_Selection>. Acesso em: 4 jul. 2025.

LITTLE, Roderick J. A.; RUBIN, Donald B. **Statistical analysis with missing data**. 3. ed. Hoboken: Wiley, 2019. DOI: https://doi.org/10.1002/9781119482260.

LIU, Xu-Ying; WU, Jianxin; ZHOU, Zhi-Hua. Exploratory undersampling for class-imbalance learning. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, v. 39, n. 2, p. 539-550, abr. 2008. DOI: https://doi.org/10.1109/TSMCB.2008.2007853.

NEWAZ, Asif; HAQ, Farhan Shahriyar. **A novel hybrid sampling framework for imbalanced learning**. 2022. Disponível em: <https://doi.org/10.48550/arXiv.2208.09619>. Acesso em: 4 jul. 2025.

PEREZ, Luis; WANG, Jason. **The effectiveness of data augmentation in image classification using deep learning**. 2017. Disponível em: <https://doi.org/10.48550/arXiv.1712.04621>. Acesso em: 4 jul. 2025.

REED, Scott; LEE, Honglak; ANGUELOV, Dragomir; SZEGEDY, Christian; ERHAN, Dumitru; RABINOVICH, Andrew. **Training deep neural networks on noisy labels with bootstrapping**. 2015. Disponível em: <https://doi.org/10.48550/arXiv.1412.6596>. Acesso em: 4 jul. 2025.

SACKETT, David L.; STRAUS, Sharon E.; RICHARDSON, W. Scott; ROSENBERG, William; HAYNES, R. Brian. **Evidence-based medicine**: how to practice and teach EBM. 2. ed. London: Churchill Livingstone, 2000. p. 24.

SIAM, Md Kamrul Hossain; BHATTACHARJEE, Manidipa; MAHMUD, Shakik; SARKAR, Md. Saem; RANA, Md. Masud. **The impact of machine learning on society**: an analysis of current trends and future implications. 2024. Disponível em: <https://arxiv.org/abs/2404.10204>. Acesso em: 23 maio 2025.

SURESH, Harini; GUTTAG, John V. A framework for understanding sources of harm throughout the machine learning life cycle. **Communications of the ACM**, v. 64, n. 8, p. 62-71, Aug. 2019.

WANG, Yixin; LIU, Han. De-biasing methods in neural networks: a survey. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND CYBERNETICS (ICMLC), 2023, Shenzhen. **Proceedings...** Los Alamitos: IEEE, 2023. ISBN 979-8-3503-0378-0. DOI: https://doi.org/10.1109/ICMLC58545.2023.10327985.

ZEMEL, Rich; WU, Yu; SWERSKY, Kevin; PITASSI, Toni; DWORK, Cynthia. Learning fair representations. In: DASGUPTA, Sanjoy; MCALLESTER, David (eds.). Proceedings of the 30th International Conference on Machine Learning, ICML 2013, 2013**. Proceedings of Machine Learning Research**, v. 28, n. 3, p. 325-333. PMLR, 2013. Disponível em: <https://proceedings.mlr.press/v28/zemel13.html>. Acesso em: 4 jul. 2025.

ZHAO, Jieyu; WANG, Tianlu; YATSKAR, Mark; ORDONEZ, Vicente; CHANG, Kai-Wei.

Men also like shopping: reducing gender bias amplification using corpus-level constraints. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING – EMNLP, 2017, Copenhagen. **Proceedings...** Stroudsburg: Association for Computational Linguistics, 2017. p. 2979-2989. DOI: https://doi.org/10.18653/v1/D17-1323.

APÊNDICE A – TÉCNICAS POR ESTUDO, POR CATEGORIA

Técnica	Estudos
Balanceamento de Dados	
ADASYN (Adaptive Synthetic Sampling)	E13, E33, E57
Balanced Sampling for Fairness	E40
Balanced Sampling for Initial Labeled Data	E48
Balanced Training Data	E6, E39, E61
Biased Edge Dropout	E42
Binary vs. Multi-class Categorization	E12
Chunking of Signal Data	E7
Class Imbalance Rebalancing	E12, E25, E46, E56, E65
Class Overlap-Based Undersampling	E69
Client Data Rebalancing via Mediators	E72
Cluster Centroid Undersampling (CCUS)	E8
Cluster-Based Over-Sampling	E64
Cluster-Based Resampling (CBR)	E2
Cluster-Based Under-Sampling	E64
Cluster-Based Under-Sampling (K-Means Clustering)	E4
Common Nearest Neighbors Search	E69
Comparison with Random Oversampling (ROT)	E50
Cost-sensitive Metaheuristics	E5
Coverage-Based Rewriting (covRew)	E67
Cross-Validation	E70
Data Filtering and Selection	E39
Data Matching Algorithm	E22
Data-Centric Strategy (Filtering External Data)	E29
Dataset Balancing	E36
Dataset Construction (ALIDA)	E65
Dataset Diversity Analysis	E39
Dataset Re-sampling	E66

Dynamic Instance Selection with Penalty for Skipped Instances	E5
Entropy-based Sample Selection	E34
Entropy-guided Oversampling	E34
Exclusion of Multi-Fault and Corrupted Data	E11
Face Detection Filtering	E85
Fair Dataset Construction (Harvard-FairVLMed)	E43
Fairlet Construction (MakeFairlets Algorithm)	E41
Filtering and Selection of Isolated Damage Cases	E80
Graded Boundary-Sensitive Majority Class Undersampling (GBMU)	E23
Grouped Data Splitting Strategies	E7
Hybrid Approach (ANN-based Resampling)	E8
Hybrid Approach (FOFO-BS-I)	E50
Hybrid Resampling	E5, E23, E68
Inclusion of All MRI Slices (Normal and Tumorous)	E6
K-means Clustering for Scene Separation	E17
Local Outlier Factor (LOF) for Outlier Removal	E84
Minority Class Augmentation	E9, E17, E58
Missing Data Handling	E22, E33
Modified Tomek-Link Undersampling	E69
Near Miss Undersampling (NMUS)	E8
Neighbourhood-based Undersampling	E8, E15, E69
Preferential Sampling	E68
Progressively-balanced (PB) Re-sampling	E65
Random Oversampling	E1, E8, E34, E57, E77, E78
Recursive Matching for Outliers	E22
Recursive Search	E69
Resampling for Attribute Bias Mitigation	E71
SMOTE-ENN (SMOTE + Edited Nearest Neighbour)	E57, E75, E78
Standard Deviation-Based Segment Sampling	E4
Statistical Parity Correction	E25, E51, E73

STEM (SMOTE-ENN + Mixup)	E75
Stratified Sampling Within Clusters	E4
Test-Train Distribution Separation	E39
Threshold-Based Random Sampling for Small Clusters	E4
Train-Test Split with Stratified Sampling	E70
Two-Step Oversampling (FOFO Framework)	E50
Undersampling	E52, E72
Transformação de Labels	
Adaptive Rounding	E55
Auxiliary Demographic Loss for Training	E27
Cross-Domain Gradient Discrepancy Minimization (CGDM)	E20
Fairness-Aware Rewriting	E67
Flexible Threshold Adjustment for Pseudo-labeling	E49
Gender-Neutral Training	E36
Label Adjustment	E10, E41, E46, E68
Label Definition Standardization	E40
Label Flipping (iFlipper)	E55
Labeling Both Tumor and Brain Parts	E6
Local Intrinsic Dimension (LID)-Based Client Selection	E79
Manual Annotation of Artifacts	E26
Meta-data Generation for Clean Labels	E21
Noise Label Identification and Correction	E79
Noisy Label Training	E60
Pseudo-Subgroup Labeling	E29
Reverse Greedy Optimization	E55
Self-Labeling and Confidence-Based Selection	E58
Self-Supervised Learning with Pseudo-Labels	E20
Semantic Similarity Constraints for Text Perturbation	E16
Situation Testing for Fair Labeling	E48
Soft Label Smoothing	E14, E21

Strategy-Agnostic Pseudo-Label Integration	E37
Total Error Constraint	E55
Uniformity Loss Optimization	E60
Unsupervised Clustering for Pseudo-Labeling	E37
Variable Simplification	E22
Modificação de Atributos Sensíveis	
Adversarial Debiasing	E54, E59
Binary Differential Privacy for Categorical Attributes	E28
Counterfactual Reasoning for Video Bias Removal	E71
Defining Intersectional Groups	E44
De-Identification of Sensitive Information	E43
Differential Privacy Enforcement	E28
Distance Correlation Minimization (Dist-fair)	E61
Exclusion of Confounding Cases	E22
Fair Representation Learning	E10, E25, E38, E45, E51, E59, E60, E61, E73
Feature Selection Based on Common Features	E70
Feature Selection Using ANOVA F-value and Correlation Analysis	E62
Filtering Sensitive Attributes	E40
Handling Proxy Attributes	E40
Protected Attribute Identification	E16
Selection of Relevant Confounding Variables	E22
Sensitive Attribute Suppression	E46, E62, E73
Unsupervised Identification of Pseudo-Attributes via Clustering	E82
Reponderação	
Adversarial Reweighting	E19
Causal Graph-Based Reweighting	E19
Class-aware Meta-Weight-Net (CMW-Net)	E21
Cluster-wise Reweighting Scheme	E82
Exponential Moving Average for Weight Updates	E82
Identifying Privileged and Unprivileged Groups	E44

Instance Reweighting	E3, E14, E19, E21, E25, E39, E44, E45, E46, E58, E73
Intelligent Batch Sampling (Weighted Selection Based on Latent Features)	E15
K-Means Clustering for Grouping	E3
Probabilistic Demographic Weights	E27
Regularization Constraint	E19
Trade-off Parameter (β)	E47
Geradores de Dados Justos	
Adaptive Scattering-Based Minority Class Oversampling (ASMO)	E23
Attribute-SCM Learning	E35
Auto-encoders for Fairness	E59
Bias Detection via Generative Modeling	E31
Class-aware Feature MixUp (CFM) Augmentation	E49
Collaborative Adversarial Training	E30
Conditional Generator with Random Labels	E30
Confidence-Based Filtering	E84
Constraint-Based Optimization (Disc-Less)	E67
Controlled Image Manipulation (ContraCLIP+DiffAE)	E56
Controlled Latent Space Mutation	E85
CycleGAN for Style Transfer	E20
Data Augmentation at Word Level (using nlpaug)	E78
Data Augmentation via Subgroup Mixing	E24
Dataset Augmentation	E66
Distributionally Generative Augmentation	E31
Diversity Loss Functions	E32
Evolutionary Search Algorithm	E85
Fair Data Integration (Schema Mapping Adjustments)	E46
Fairness-aware Subgroup Mixup	E24
Feature Embedding and Fusion	E30
Fisher Information-Based Optimization	E14

GAN (Generative Adversarial Network) Integration with ADASYN	E13
Gaussian Kernel Density Peak Clustering Algorithm (GKDPCA)	E83
Generative Data Augmentation	E10, E32, E56, E76
Human Evaluation	E35
IMITATE (Identify and Mitigate Selection Bias)	E84
Implicit Differentiation for Hyperparameter Optimization	E14
Improved Conditional Generative Adversarial Networks (CGANs)	E83
Induced Dataset Creation	E3
Initial Pool for Fairness Estimation	E40
Intra-Class Variation Generation	E85
Joint Optimization Framework	E76
Kolmogorov-Smirnov (K-S) Test for Distribution Comparison	E33
Label-Guided Data Generation (DEM - Generation Data Augmentation)	E9
Latent Factor Construction	E62
LLM-Based Data Augmentation (GPT-3.5 Turbo)	E78
MCD-GAN (Maximum Classifier Discrepancy GAN)	E20
Minority Centroid Oversampling (MCO)	E2
MixFeat (Proposed Data Augmentation Method)	E1
Multi-Generator Architecture	E32
Multi-Head Encoder with Soft Attention Fusion	E27
Oversampling (Synthetic Data Generation)	E52
Parallel Generation for Different Demographics	E85
PatchMix Data Augmentation	E74
Per-Point Augmentation Hyperparameters	E14
Pseudo De-biased Test Sets	E18
Pseudo-labeling for Synthetic Data	E56
Random Rejection Sampling with Demographic Constraints	E85
Secondary Pre-Disaster Image Integration	E80
Self-Attention Mechanism in GANs	E32
Structural Causal Model (SCM) Integration	E35

Synthetic Data Augmentation via Poisson Blending	E80
Synthetic Minority Oversampling Technique (SMOTE)	E8, E12, E34, E48, E50, E57, E75, E78
Trap Sets	E26
Unsupervised Scene Feature Extraction using GAN	E17
Use of Auxiliary Network Information	E18
Wasserstein Distance with Gradient Penalty	E83
Weight Sharing Scheme	E30
Weighted Least Squares Optimization for Gaussian Fitting	E84
Z-score-based Data Augmentation	E72
Censura de Score/Feature	
Counterfactual Reasoning for Question Bias Removal	E71
Explainability Heatmap Extraction (LRP-based)	E37
Fairness Metric Based on Correlation Reduction	E47
Model-Agnostic Data Attribution	E66
Object-Level vs. Pixel-Level Scoring	E80
Region Masking	E66
Role Assignment Based on Data Quality	E79
Targeted Noise Addition	E66
Normalização e Transformações de Feature	
Adaptive Distribution Recalibration (within RFC)	E53
Autoencoder for Dimensionality Reduction	E63
Bias Field Modeling and Correction	E11
Categorical Variable Encoding	E33
Causal Disentanglement	E18
Channel-Wise Attention Mechanism	E27
Clinical Note Summarization	E43
Clip Function for Valid Perturbations	E16
ComBat Harmonization	E20
Cyclic Style ALI (CSALI) for Improved Reconstruction	E35
Data Normalization	E13, E40

Demographic Fairness Transformer (DeFT)	E27
Dependency-Based Filtering	E67
Discretization for Candidate Selection Conditions	E67
Fair Attribute Preprocessing	E42
Fairness-aware PCA (FairDR)	E47
Fairness-driven Binning	E16, E22, E46
Feature Extraction for Class-Imbalance Handling	E63
Feature Extraction for Traditional Models	E7
Feature Selection (High-Level vs. Low-Level Features)	E1
Graph Regularization (Graph-fair)	E61
Hellinger Distance as Similarity Metric	E7
Hierarchical Representation Learning	E60
Idempotent Encoder for Explicit Bias Elimination	E54
Interquartile Range (IQR) for Outlier Detection	E50
Kernel Density Estimation (KDE) for Small Datasets	E84
Local Pre-Piecewise Fitting Function	E11
Log-Transformation and Mean Subtraction	E77
Manifold Learning (Dimensionality Reduction)	E64
Masking of Non-Brain Voxels	E77
Maximum Mean Discrepancy (MMD)	E18
Microaggregation (Microaggregate Algorithm)	E41
Normalized-Background Dataset	E26
One-Hot Encoding for Categorical Features	E13
Optimized Data Term for Sensitivity Enhancement	E11
Optimized Pre-processing	E46
Patch-wise Supervised Contrastive Loss	E74
Percentile Transformation	E81
Principal Component Analysis (PCA)	E1, E3, E77
Recalibrated Feature Compensation (RFC)	E53
Reconstruction Loss for Information Preservation	E54

Selection of Non-Sensitive Attributes with Low Correlation to Sensitive Attributes	E62
Similarity Matrix Construction	E55
Spatial Normalization	E77
Statistical Feature Selection	E57
Text Preprocessing (Indirect Bias Mitigation)	E57
Thresholded L1 Loss for Appearance Preservation	E76
Variational Denoised Term	E11
Word Embedding for Text Data	E16

APÊNDICE B – ESTUDOS SELECIONADOS

ID	Título	DOI
E1	“It’s not Fair!” – Fairness for a Small Dataset of Multi-modal Dyadic Mental Well-being Coaching	10.1109/ACII59096.2023.10388118
E2	A clustering-based resampling technique with cluster structure analysis for software defect detection in imbalanced datasets	10.1016/J.INS.2024.120724
E3	A dual algorithmic approach to deal with multiclass imbalanced classification problems	10.1016/J.BDR.2024.100484
E4	A K-means Clustering Based Under-Sampling Method for Imbalanced Dataset Classification	10.1109/ICOIN59985.2024.10572133
E5	A Novel Multi-class Classification Architecture Combining Population-based Sampling and Multi-expert Classifier for Imbalanced Data	10.1109/SMC52423.2021.9659252
E6	A Novel Pipeline on Medical Object Detection for Bias Reduction: Preliminary Study for Brain MRI	10.1109/INISTA52262.2021.9548623
E7	A Proposal to Mitigate Similarity Bias for the Paderborn Bearing Data Set	10.1109/IECON48115.2021.9589342
E8	An ANN-Based Resampling Approach for Handling Imbalance and Overlapped Data	10.1109/ETNCC63262.2024.10767538
E9	An Effective and Head-Friendly Long-Tailed Object Detection Network with Improving Data Enhancement and Classifier	10.1109/ICCC59590.2023.10507396
E10	An information theoretic approach to reducing algorithmic bias for machine learning	10.1016/J.NEUCOM.2021.09.081
E11	An optimized denoised bias correction model with local pre-fitting function for weak boundary image segmentation	10.1016/J.SIGPRO.2024.109448
E12	Analyzing and Addressing Data-driven	10.1109/ICCECE51049.2023.10085470

	Fairness Issues in Machine Learning Models used for Societal Problems	
E13	Application of TPE-Random Forest in Microwave Oven Fault Diagnosis Based on GAN and ADASYN Algorithm	10.1109/ISMSIT63511.2024.10757240
E14	AutoDO: Robust AutoAugment for Biased Data with Label Noise via Scalable Probabilistic Implicit Differentiation	10.1109/CVPR46437.2021.01633
E15	Automated Bias Reduction in Deep Learning Based Melanoma Diagnosis using a Semi-Supervised Algorithm	10.1109/BIBM52615.2021.9669772
E16	Automatic Fairness Testing of Neural Classifiers Through Adversarial Sampling	10.1109/TSE.2021.3101478
E17	Balance Scene Learning Mechanism for Offshore and Inshore Ship Detection in SAR Images	10.1109/LGRS.2020.3033988
E18	Causal Disentanglement for Implicit Recommendations with Network Information	10.1145/3582435
E19	Causal Fairness-Guided Dataset Reweighting using Neural Networks	10.1109/BigData59044.2023.10386836
E20	CGDM-GAN: An Adversarial Network Approach with Self-supervised Learning for Site Effect Removal	10.1109/EMBC53108.2024.10782176
E21	CMW-Net: Learning a Class-Aware Sample Weighting Mapping for Robust Deep Learning	10.1109/TPAMI.2023.3271451
E22	Construction of a confounder-free clinical MRI dataset in the Mass General Brigham system for classification of Alzheimer's disease	10.1016/J.ARTMED.2022.102309
E23	Data Augmentation by Hybrid Data Resampling: Towards Enhanced Performance in Automatic Cardiac Arrhythmia Detection	10.1109/CALCON63337.2024.10914252
E24	Data Augmentation via Subgroup Mixup for Improving Fairness	10.1109/ICASSP48485.2024.10446564

E25	Data vs. Model Machine Learning Fairness Testing: An Empirical Study	10.1145/3643786.3648022
E26	Debiasing Skin Lesion Datasets and Models? Not So Fast	10.1109/CVPRW50498.2020.00378
E27	Demographic Fairness Transformer for Bias Mitigation in Face Recognition	10.1109/IJCB62174.2024.10744457
E28	Differential Privacy for Fair Deep Learning Models	10.1109/SysCon48628.2021.9591252
E29	Discover and Mitigate Multiple Biased Subgroups in Image Classifiers	10.1109/CVPR52733.2024.01037
E30	Distribution Bias Aware Collaborative Generative Adversarial Network for Imbalanced Deep Learning in Industrial IoT	10.1109/TII.2022.3170149
E31	Distributionally Generative Augmentation for Fair Facial Attribute Classification	10.1109/CVPR52733.2024.02151
E32	DM-GAN: A Data Augmentation-Based Approach for Imbalanced Medical Image Classification	10.1109/BIBM62325.2024.10821792
E33	Enhancing Students' Academic Performance Classifier using ADASYN and MLP	10.1109/SCOReD64708.2024.10872712
E34	E-SMOTE: Entropy Based Minority Oversampling for Heart Failure and AIDS Clinical Trails Analysis	10.1109/COMPSAC61105.2024.00291
E35	Evaluating and Mitigating Bias in Image Classifiers: A Causal Perspective Using Counterfactuals	10.1109/WACV51458.2022.00393
E36	Evaluating Gender-Neutral Training Data for Automated Image Captioning	10.1109/BigData52589.2021.9671774
E37	ExMap: Leveraging Explainability Heatmaps for Unsupervised Group Robustness to Spurious Correlations	10.1109/CVPR52733.2024.01142
E38	Explainability and Fairness in Machine Learning: Improve Fair End-to-end Lending for Kiva	10.1109/SSCI47803.2020.9308371
E39	Exploring Biases and Prejudice of Facial	10.1109/IJCNN52387.2021.9534287

	Synthesis via Semantic Latent Space	
E40	Fair active learning	10.1016/J.ESWA.2022.116981
E41	Fair and Private Data Preprocessing through Microaggregation	10.1145/3617377
E42	Fair graph representation learning: Empowering NIFTY via Biased Edge Dropout and Fair Attribute Preprocessing	10.1016/J.NEUCOM.2023.126948
E43	FairCLIP: Harnessing Fairness in Vision-Language Learning	10.1109/CVPR52733.2024.01168
E44	FairDETOCS: An approach to detect and connect unfair models	10.1109/ICECCME62383.2024.10796175
E45	FairGridSearch: A Framework to Compare Fairness-Enhancing Models	10.1109/WI-IAT59888.2023.00064
E46	Fairness-aware Data Integration	10.1145/3519419
E47	Fairness-Aware Dimensionality Reduction	10.23919/EUSIPCO58844.2023.10289717
E48	Fair-SSL: Building fair ML Software with less data	10.1145/3524491.3527305
E49	FGBC: Flexible graph-based balanced classifier for class-imbalanced semi-supervised learning	10.1016/J.PATCOG.2023.109793
E50	FOFO: Fused Oversampling Framework by addressing Outliers	10.1109/ESCI50559.2021.9397056
E51	GenEthos: A Synthetic Data Generation System With Bias Detection And Mitigation	10.1109/IC3SIS54991.2022.9885653
E52	G-SOMO: An oversampling approach based on self-organized maps and geometric SMOTE	10.1016/J.ESWA.2021.115230
E53	Hierarchical Bias Mitigation for Semi-Supervised Medical Image Classification	10.1109/TMI.2023.3247440
E54	Idempotence-Constrained Representation Learning: Balancing Sensitive Information Elimination and Feature Robustness	10.1109/BigData62323.2024.10825539

E55	iFlipper: Label Flipping for Individual Fairness	10.1145/3588688
E56	Improving Fairness using Vision-Language Driven Image Augmentation	10.1109/WACV57701.2024.00463
E57	Improving Sentiment Analysis Performance on Imbalanced Dataset Using Data Resampling and Statistical Feature Selection	10.1109/InCIT63192.2024.10810634
E58	LCSL: Long-Tailed Classification via Self-Labeling	10.1109/TCSVT.2024.3421942
E59	Learn and Visually Explain Deep Fair Models: an Application to Face Recognition	10.1109/IJCNN52387.2021.9533659
E60	Learning Fair Representations through Uniformly Distributed Sensitive Attributes	10.1109/SaTML54575.2023.00014
E61	Learning Fair Representations via Distance Correlation Minimization	10.1109/TNNLS.2022.3187165
E62	Leverage Generative Adversarial Network for Enhancing Fairness and Utility	10.1109/NaNA60121.2023.00090
E63	Leveraging Autoencoder and Focal Loss for Imbalanced Data Classification	10.1109/ITME56794.2022.00110
E64	Manifold cluster-based evolutionary ensemble imbalance learning	10.1016/J.CIE.2021.107523
E65	Mitigating Bias of Deep Neural Networks for Trustworthy Traffic Perception in Autonomous Systems	10.1109/IV55156.2024.10588805
E66	Mitigating Bias Using Model-Agnostic Data Attribution	10.1109/CVPRW63382.2024.00028
E67	Mitigating Representation Bias in Data Transformations: A Constraint-based Optimization Approach	10.1109/BigData59044.2023.10386181
E68	Mitigating Subgroup Unfairness in Machine Learning Classifiers: A Data-Driven Approach	10.1109/ICDE60146.2024.00171
E69	Neighbourhood-based undersampling approach for handling imbalanced and	10.1016/J.INS.2019.08.062

	overlapped data	
E70	Prediction of diabetes using cost sensitive learning and oversampling techniques on Bangladeshi and Indian female patients	10.1109/ICITR51448.2020.9310892
E71	Removing Bias of Video Question Answering by Causal Theory	10.1109/ICIEA61579.2024.10665208
E72	Self-Balancing Federated Learning With Global Imbalanced Data in Mobile Systems	10.1109/TPDS.2020.3009406
E73	Simulated annealing based undersampling (SAUS): a hybrid multi-objective optimization method to tackle class imbalance	10.1007/S10489-021-02369-4
E74	Source-Free Domain Adaptation With Domain Generalized Pretraining for Face Anti-Spoofing	10.1109/TPAMI.2024.3370721
E75	STEM Rebalance: A Novel Approach for Tackling Imbalanced Datasets using SMOTE, Edited Nearest Neighbour, and Mixup	10.1109/ICCP60212.2023.10398660
E76	Style Transfer with Bio-realistic Appearance Manipulation for Skin-tone Inclusive rPPG	10.1109/ICCP54855.2022.9887649
E77	Subspace corrected relevance learning with application in neuroimaging	10.1016/J.ARTMED.2024.102786
E78	The Applicability of LLMs in Generating Textual Samples for Analysis of Imbalanced Datasets	10.1109/ACCESS.2024.3463400
E79	TrustBCFL: Mitigating Data Bias in IoT Through Blockchain-Enabled Federated Learning	10.1109/JIOT.2024.3379363
E80	Uncovering Bias in Building Damage Assessment from Satellite Imagery	10.1109/IGARSS53475.2024.10642347
E81	Understanding and mitigating multi-sided exposure bias in recommender systems	10.1145/3566100.3566103
E82	Unsupervised Learning of Debiased Representations with Pseudo-Attributes	10.1109/CVPR52688.2022.01624

E83	Using Improved Conditional Generative Adversarial Networks to Detect Social Bots on Twitter	10.1109/ACCESS.2020.2975630
E84	Your Best Guess When You Know Nothing: Identification and Mitigation of Selection Bias	10.1109/ICDM50108.2020.00115
E85	Zero-Shot Racially Balanced Dataset Generation using an Existing Biased StyleGAN2	10.1109/IJCB57857.2023.10449028