Uma Revisão Sistemática sobre a Relação de Impacto da Qualidade de Dados na Justiça Algorítmica para Classificação de Imagens

Maristela de Freitas Riquelme[1], Lucas Freire de Lima[1], Luiz Fernando F. P. de Lima[2], Danielle Rousy Dias Ricarte[1]

[1]Centro de Informática – Universidade Federal da Paraíba (UFPB) - João Pessoa, PB, Brasil

[2]Centro de Estudos e Sistemas Avançados do Recife (CESAR) - Recife, PE, Brasil

maristela.riquelme@academico.ufpb.br, lucas.freire@estudantes.ufpb.br, lffpl@cesar.org.br, danielle@ci.ufpb.br

Abstract. *As medical image classification systems become more widespread, the debate surrounding their fairness and impartiality intensifies. Seeking to understand how this issue is being discussed, a systematic review was conducted on the impact of data quality on biases in machine learning systems for medical image classification. After analyzing the articles, methods were identified to ensure the quality of the datasets. It is concluded that the quality of the dataset impacts the performance of the models, potentially leading to incorrect or imprecise clinical diagnoses.*

Resumo. *À medida que os sistemas de classificação de imagens médicas são difundidos, intensifica-se o debate acerca da imparcialidade e justiça destes. Buscando compreender a forma como a temática vem sendo debatida, realizou-se uma revisão sistemática sobre o impacto da qualidade de dados na ocorrência de vieses em sistemas de aprendizado de máquina na classificação de imagens médicas. Após a análise dos artigos, foram identificados métodos para assegurar a qualidade dos conjuntos de dados. Conclui-se, que a qualidade do conjunto de dados impacta no desempenho dos modelos, podendo ocasionar em diagnósticos clínicos incorretos ou imprecisos.*

1. Introdução

À medida que o uso de sistemas de inteligência artificial para a classificação de imagens se populariza, a necessidade de sistemas imparciais e justos se torna ainda mais crucial [El-Sappagh *et al* 2023]. No entanto, esse ainda é um grande desafio, principalmente devido à variação nos contextos de aplicação que está diretamente relacionado à qualidade dos conjuntos de dados.

Por exemplo, no contexto médico, na classificação de imagens dermatoscópicas, as doenças comuns como nevo melanocítico geralmente têm conjuntos de dados robustos, enquanto as doenças raras como dermatofibroma são mais difíceis de encontrar representações adequadas [Lei *et al* 2023]. Isso resulta em classificações menos precisas e robustas, comprometendo sua eficácia em ambientes clínicos.

Garantir a qualidade dos dados em aprendizagem de máquina é fundamental, pois em inúmeros casos, esses modelos estão propensos a vieses, a depender da base de treinamento utilizada, o que impacta diretamente a precisão na relação entre os atributos dos dados e os resultados obtidos [Yang *et al* 2023].

Conforme destacado por Arora (2023), se os dados usados para treinar um algoritmo forem tendenciosos em relação a grupos demográficos específicos, é provável que o algoritmo tenha um desempenho inferior quando aplicado a esses grupos no mundo real. A ausência de diversidade, permite que os modelos tenham um bom desempenho em casos comuns, mas enfrentarão dificuldades com casos incomuns ou sub-representados. Isso evidencia como a falta de qualidade dos dados limita a capacidade do sistema de lidar com cenários diversos e representativos da realidade.

Com isso, é importante compreender o que a literatura atual define sobre o conceito de conjunto de dados de qualidade, seu impacto no treinamento e desenvolvimento de sistemas justos no contexto médico, bem como os desafios éticos associados. Dessa forma, este trabalho apresenta uma revisão sistemática com o objetivo de mapear os estudos que abordam os aspectos mencionados anteriormente e obter um melhor entendimento acerca do conhecimento atual sobre o tema.

Além desta Seção introdutória, este artigo divide-se em mais quatro Seções. A Seção 2 descreve a metodologia adotada na condução da pesquisa. A Seção 3 expõe os resultados obtidos e, por fim, a Seção 4 aborda a conclusão com base no que foi desenvolvido.

2. Metodologia

Nesta Seção, apresentamos a metodologia da revisão sistemática. Segundo Kitchenham (2004), esse método consiste em um planejamento no qual é definido um protocolo de revisão que especifica a questão da pesquisa a ser abordada, os métodos utilizados para realizá-la e, posteriormente, o desenvolvimento desse protocolo, concluindo com a apresentação dos resultados.

2.1. Questões da Pesquisa

Com os artigos considerados relevantes para a pesquisa, buscamos responder às seguintes questões: (**QP01**) Qual o contexto (domínio) de aplicação está sendo analisado? (**QP02**) Como a qualidade do conjunto de dados impacta no desenvolvimento e na precisão de sistemas justos de aprendizagem de máquina? (**QP03**) Como a qualidade dos dados pode ser assegurada em conjuntos de dados para evitar vieses nos modelos de aprendizagem de máquina? (**QP04**) Quais são os desafios éticos associados à qualidade de conjunto de dados em projetos de aprendizado de máquina?

2.2. Estratégia de Busca

Para o nosso estudo, foram definidas *strings* de busca utilizadas para pesquisar artigos nas bases de dados Springer, ScienceDirect, ACM e IEEE, utilizando como critério temporal os últimos cinco anos (2018 - 2023).

Com base na *string* (“data quality” and “fairness” and “image classification”), que melhor se adequava aos nossos objetivos, realizou-se a análise e seleção de artigos, inicialmente selecionando-os com a leitura do título e resumo e, posteriormente, seguindo os critérios de exclusão e inclusão previamente definidos.

2.3. Critérios de Seleção

Os critérios de inclusão foram definidos como: o artigo aborda a qualidade dos dados em aprendizado de máquina; discute questões de justiça, ética ou vieses; relaciona ética com qualidade de dados; aponta como a qualidade de dados influencia na ética em contexto clínico. Já os critérios de exclusão: não aborda a qualidade de *datasets* em aprendizado de máquina para o desenvolvimento de sistemas justos; o acesso ao arquivo completo não está disponível; publicação não científica e artigos duplicados.

3. Resultados

A seleção de artigos foi dividida em etapas. Na etapa 1, que consiste na consulta aos engenhos de busca com a *string* ("data quality" and "fairness" and "image classification"), considerando o período entre os anos de 2018 a 2023, na ACM, ScienceDirect, IEEE e Springer, foram obtidos 162, 180, 3 e 1.426 resultados, respectivamente. Entretanto, dos 1.426 resultados apresentados na base Springer, apenas os 1.000 primeiros foram apresentados, devido a uma restrição da própria plataforma. Na etapa 2, após a seleção com base na leitura do título e resumo, foram coletados 1, 2, 3 e 2 artigos nas bases ACM, IEEE, ScienceDirect e Springer, por essa ordem.

Dos 8 artigos pré-selecionados, 2 foram excluídos: 1 da base de dados ScienceDirect e 1 da Springer, após a leitura da introdução dos mesmos, por não estarem alinhados ao objetivo da pesquisa. Todavia, em relação à questão QP01, observamos que 2 desses artigos não pertencem ao contexto médico, não indicando claramente a área de atuação. A Tabela 1 apresenta um resumo dos achados da nossa pesquisa, identificando os trabalhos analisados e o contexto no qual estão inseridos (QP01).

Tabela 1. Trabalhos analisados e o contexto no qual estão inseridos

Identificador	Artigo	QP01
[A1]	Trustworthy Machine Learning for Health Care: Scalable Data Valuation with the Shapley Value Pandl, K. D. et al. (2021)	Contexto médico
[A2]	Evaluating and Mitigating Bias in Image Classifiers: A Causal Perspective Using Counterfactuals Dash, S., Vineeth, N. B. and Sharma, A. (2022)	Não especifica
[A3]	Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods Band, S. S. et al. (2023)	Contexto médico
[A4]	Trustworthy artificial intelligence in Alzheimer's disease: state of the art, opportunities, and challenges Pandl, K. D. et al. (2021)	Contexto médico
[A5]	Face Recognition Fairness Assessment based on Data Augmentation: An Empirical Study Tian, F. et al. (2022)	Não especifica

[A6]	Category-aware feature attribution for Self-Optimizing medical image classification Lei, J. et al. (2023)	Contexto médico

3.1. (QP02) Como a qualidade do conjunto de dados impacta no desenvolvimento e na precisão de sistemas justos de aprendizagem de máquina?

Em relação ao impacto da qualidade de dados no desenvolvimento de sistemas de aprendizagem de máquina justos, o artigo [A1] aponta que a confiabilidade, o bom desempenho e a precisão são fatores que impactam no desenvolvimento e na precisão dos sistemas justos. Já em [A2], os autores apresentam argumentos que indicam que quando não há qualidade de dados, isso pode ser um impeditivo para que as imagens geradas sejam utilizadas em aplicações em que haja avaliação de imparcialidade, impossibilitando o estabelecimento de um padrão de relações. Em [A3] os autores apontam que dados desbalanceados podem levar a modelos tendenciosos em direção a classe majoritária, acarretando em má avaliação dos modelos implementados, causando resultados imprecisos ou não confiáveis, e em último caso, afetando a qualidade do diagnóstico e tratamento. O trabalho [A4] indica que a falta de qualidade nos dados pode ocasionar em vieses sociais, e, assim como em [A3], levar à falta de credibilidade no sistema e diagnósticos incorretos. No artigo [A5], o preconceito presente na formação do modelo pode ser amplificado. E em [A6], não demonstrando resultados suficientemente precisos e robustos para uso clínico, comprometendo a classificação das imagens.

3.2. (QP03) Como a qualidade dos dados pode ser assegurada em conjuntos de dados para evitar vieses nos modelos de aprendizado de máquina?

No contexto da garantia da qualidade de dados para evitar vieses em sistemas de aprendizado de máquina, o artigo [A1] destaca que identificar e corrigir instâncias mal rotuladas nos conjuntos de dados, aplicar métodos de valoração de dados, detectar e tratar vieses, são formas de assegurar essa qualidade. Em [A2], os autores afirmam que as explicações das previsões de um classificador e a avaliação de sua imparcialidade, são medidas relevantes. Já em [A3], sugere-se o uso do conhecimento conceitual e a fusão de informações de forma abrangente e integrativa. No artigo [A4], destaca-se a importância da transparência e rastreabilidade, de modo que todos os dados, processos e algoritmos devam ser documentados, além da remoção de vieses na fase inicial da coleta de dados. Em [A5], compreendendo quais propriedades dos dados influenciam as decisões dos modelos e aumentando os dados. E no trabalho [A6], por meio da promoção da aprendizagem de características por categoria.

3.3. (QP04) Quais são os desafios éticos associados à qualidade de conjunto de dados em projetos de aprendizado de máquina?

Referente à qualidade de dados em projetos de aprendizado de máquina, o trabalho [A1] revela que proteger a privacidade das informações dos pacientes, representar as minorias além de previsões erradas ou imprecisas, constituem os principais desafios éticos. No artigo [A2], o autor expõe que o principal desafio ético é a marginalização demográfica de grupos. Em [A3], são apontadas questões como a falta de transparência em certos

algoritmos, as preocupações com a privacidade dos dados usados para treinar modelos de IA, bem como questões de segurança e responsabilidade que podem surgir quando a IA é usada em ambientes clínicos. Já em [A4], a preocupação recai sobre a integridade dos dados e, assim como em [A1], a proteção à privacidade dos pacientes. Em [A5], o desafio está em combater o viés humano e a discriminação em algumas aplicações, o que pode gerar desigualdades no processo de tomada de decisão. Por fim, em [A6], enfrenta-se a limitação imposta pelas amostras difíceis inevitáveis e pelas amostras de classes minoritárias, ao mesmo tempo, é difícil entender o mecanismo de previsão e depurar previsões incorretas.

3.4. Discussão

Após a análise dos artigos, tornou-se evidente que a qualidade dos dados impacta diretamente na confiabilidade, no desempenho e na precisão dos sistemas. Além disso, dados inconsistentes podem acarretar em vieses sociais, discriminação racial e ética, amplificando o preconceito. Na tentativa de mitigar esses vieses, a qualidade dos dados pode ser assegurada através da identificação e correção das instâncias mal rotuladas, aplicação de métodos de valoração de dados, bem como garantir a transparência, rastreabilidade e documentar os processos e algoritmos. Ademais, os desafios éticos associados à qualidade do conjunto de dados de aprendizagem de máquina estão principalmente relacionados a questões de privacidade, integridade dos dados e limitação das amostras.

Realizada a análise, identificamos que o contexto médico é a área que mais faz uso da classificação de imagens. Entretanto, os sistemas de aprendizado de máquina muitas vezes não apresentam resultados precisos e robustos para uso clínico podendo comprometer a qualidade do diagnóstico e tratamento.

Neste sentido, observamos que para que haja avanços nessa área, é essencial buscar maneiras de identificar e corrigir vieses, garantir a transparência nos processos e aprimorar as práticas de coleta e integridade dos dados, por exemplo, utilizando avaliação de equidade, análise de viés estatístico ou geração de dados sintéticos. Além de que buscar por soluções que garantam sistemas mais precisos e imparciais deve ser vista como prioridade para gerar resultados clínicos precisos.

4. Conclusão

Este trabalho teve como objetivo realizar uma revisão sistemática para entender o estado da arte no que diz respeito ao impacto da qualidade de dados para alcançar resultados justos em tarefas de classificação de imagens. Após a análise dos resultados, identificamos fatores que contribuem para a presença de vieses na classificação de imagens na área médica. E assim, compreendemos melhor a importância de um conjunto de dados de qualidade e como isso afeta diretamente o desempenho dos modelos de aprendizado de máquina.

Agora, com a conclusão da fase atual da pesquisa, planejamos dar continuidade ao trabalho desenvolvendo experimentos avaliativos. Nosso propósito inclui realizar uma análise técnica dos conjuntos de dados balanceados em conjunto com técnicas de aprendizado de máquina. Isso nos permitirá identificar características e técnicas

essenciais para uma representação mais precisa da população de determinada região nos conjuntos de dados, reduzindo assim, potenciais fontes de discriminação.

Dessa forma, buscamos contribuir significativamente para o avanço da área desenvolvendo abordagens mais eficazes para lidar com os vieses na classificação de imagens médicas. O trabalho faz parte de um projeto de pesquisa, em andamento, que visa avaliar a influência de conjunto de dados com maior diversidade na construção de aplicações de aprendizado de máquina justos considerando o contexto médico de análise de imagens de caixa-torácica.

5. Referências

Arora, A. *et al*. (2023) "The value of standards for health datasets in artificial intelligence-based applications", Em: Nature Medicine, 29, 2929-2938, https://doi.org/10.1038/s41591-023-02608-w

Band, S. S. *et al*. (2023) "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods", Em: Informatics in Medicine Unlocked, 40, 101286, https://doi.org/10.1016/j.imu.2023.101286

Dash, S., Vineeth, N. B. and Sharma, A. (2022) "Evaluating and Mitigating Bias in Image Classifiers: A Causal Perspective Using Counterfactuals", Em: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, p. 3879-3888, doi: 10.1109/WACV51458.2022.00393.

El-Sappagh, S. *et al.* (2023) "Trustworthy artificial intelligence in Alzheimer's disease: state of the art, opportunities, and challenges", Em: Artificial Intelligence Review, 56, p. 11149 – 11296, https://doi.org/10.1007/s10462-023-10415-5

Kitchenham, B. (2004) "Procedures for Performing Systematic Reviews", https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf

Lei, J. *et al*. (2023) "Category-aware feature attribution for Self-Optimizing medical image classification", Em: Displays, 77, 102397, https://doi.org/10.1016/j.displa.2023.102397

Pandl, K. D. *et al*. (2021) "Trustworthy machine learning for health care: scalable data valuation with the shapley value", Em: CHIL '21: Proceedings of the Conference on Health, Inference, and Learning, p. 47 - 57, https://doi.org/10.1145/3450439.3451861

Tian, F. *et al*. (2022) "Face Recognition Fairness Assessment based on Data Augmentation: An Empirical Study", Em: 2022 IEEE 22nd International Conference on Software Quality, Reliability, and Security Companion (QRS-C), Guangzhou, China, p. 315-318, doi: 10.1109/QRS-C57518.2022.00053.

Yang, J. *et al*. (2023) "Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning", Em: Nature Machine Intelligence, 5, 884-894, https://doi.org/10.1038/s42256-023-00697-3.