Framework para avaliação de fairness em modelos de classificação aplicados à saúde

Letícia Lopes Mendes da Silva, Diedre Santos do Carmo, Letícia Rittner
Faculdade de Engenharia Elétrica e de Computação, Universidade de Campinas (UNICAMP)
Campinas, SP, Brasil

***Resumo*—Modelos de inteligência artificial (IA) têm sido amplamente utilizados em aplicações médicas, com destaque para tarefas de classificação de imagens. No entanto, há uma crescente preocupação sobre o viés algorítmico, especialmente em cenários em que atributos sensíveis como sexo ou idade impactam a performance do modelo. Neste trabalho, propomos um *framework* visual e interativo para avaliação de *fairness* em diferentes modelos classificadores, a partir de diferentes métricas de justiça algorítmica. Como estudo de caso, foi escolhida a tarefa de detecção automática de achados clínicos em radiografias de tórax, utilizando três conjuntos de dados públicos e variantes da arquitetura EfficientNet. Os resultados mostram que, mesmo em modelos com alto desempenho segundo as métricas tradicionais de classificação, as métricas de *fairness* revelam disparidades entre diferentes grupos de pacientes, reforçando a importância de abordagens específicas para avaliação de equidade. O *framework* proposto se mostra útil para auxiliar na análise crítica de modelos em contextos médicos.**

***Abstract*—Artificial intelligence (AI) models have been widely used in medical applications, particularly in tasks such as image classification. However, growing concerns have emerged regarding algorithmic bias, especially in scenarios where sensitive attributes like gender or age influence model performance. In this work, we propose an interactive visual framework for evaluating fairness across different classification models, incorporating a variety of algorithmic fairness metrics. As a case study, we have chosen the automatic detection of clinical findings in chest X-rays, using three public datasets and variants of the EfficientNet architecture. The results show that even in models with high performance according to traditional classification metrics, the fairness metrics reveal significant disparities between certain groups of patients, reinforcing the importance of specific approaches for equity assessment. The proposed framework proves to be useful for assisting in the critical analysis of models in medical contexts.**

I. Introdução

A incorporação de técnicas de inteligência artificial (IA) tem transformado significativamente a área médica, por exemplo, ao permitir a classificação automatizada de exames clínicos, com o objetivo de apoiar o diagnóstico, a triagem de pacientes e a tomada de decisões em ambientes hospitalares. Modelos classificadores baseados em aprendizado profundo têm alcançado desempenho elevado em diferentes aplicações, aproximando-se ou até superando, em certos casos, o desempenho de especialistas humanos [1].

Apesar dos avanços recentes, cresce a preocupação quanto ao uso ético e justo de sistemas de inteligência artificial na área da saúde. Diversos estudos apontam que as predições geradas por esses modelos podem apresentar variações sistemáticas de desempenho entre diferentes subgrupos populacionais, especialmente em relação a atributos sensíveis como sexo, idade e etnia [2], [3]. Tais disparidades em predições de modelos baseados em aprendizado de máquina muitas vezes têm origem em bases de dados de treinamento desbalanceadas [4], o que pode levar a decisões clínicas imprecisas ou até prejudiciais.

Esse cenário tem impulsionado o desenvolvimento de abordagens voltadas à avaliação de *fairness* (justiça) e mitigação de vieses em modelos de IA aplicados à medicina. No entanto, a avaliação prática de justiça em modelos de classificação ainda é limitada, em grande parte pela dificuldade de interpretação das métricas e pela falta de ferramentas que facilitem a análise, especialmente em casos envolvendo múltiplas classes e grupos simultaneamente. Alguns *frameworks*, como AI Fairness 360 [5], Aequitas [6] e Fairlearn [7], oferecem suporte a métricas de *fairness* utilizadas na literatura, mas foram originalmente desenvolvidas para tarefas de classificação binária, dificultando sua aplicação em cenários multirrótulo e multiclasse, como é o caso de muitas aplicações médicas.

Neste trabalho, é proposto um *framework* visual e interativo para análise de *fairness* em tarefas de classificação médica. A ferramenta permite a visualização das principais métricas de identificação de viés propostas na literatura [2], permitindo a comparação entre diferentes modelos, classes de predição e atributos sensíveis. O objetivo é facilitar a avaliação prática de *fairness*, promovendo uma análise mais acessível e interpretável, especialmente em contextos clínicos que envolvem múltiplos achados ou atributos.

II. Métricas para Identificação de Viés

Ao investigar a presença de viés, é comum que estudos analisem o desempenho do modelo de forma desagregada, calculando métricas como taxa de predições positivas (PR), taxa de verdadeiros positivos (TPR) ou de falsos positivos (FPR) separadamente para diferentes subgrupos populacionais [3].

Com o avanço das discussões sobre ética algorítmica, ferramentas como o AI Fairness 360 [5] foram desenvolvidas para sistematizar e implementar métricas voltadas à identificação de desigualdades em tarefas de classificação. Essas métricas requerem a definição de termos importantes ao explorar *fairness*: rótulo favorável, como, por exemplo, a presença ou ausência de determinada patologia; atributos sensíveis, como sexo ou idade; e grupos privilegiados, comumente associados a populações historicamente favorecidas ou grupos majoritários em certos contextos. Neste trabalho, foram utilizadas quatro

métricas amplamente referenciadas na literatura e implementadas no *framework* AI Fairness 360 (Tab. I).

Tabela I
PRINCIPAIS MÉTRICAS DE FAIRNESS UTILIZADAS NO FRAMEWORK.

Métrica	**Cálculo para Identificação de Viés**	**Intervalo de *Fairness***
Disparate Impact (DI)	Razão entre taxas de predições positivas (PR).	$(0.8, 1.2)$
Statistical Parity Difference (SPD)	Diferença entre as taxas de predições positivas (PR).	$(-0.1, 0.1)$
Equal Opportunity Difference (EOD)	Diferença entre as taxas de verdadeiros positivos (TPR).	$(-0.1, 0.1)$
Average Odds Difference (AOD)	Média da diferença entre TPR e FPR.	$(-0.1, 0.1)$

É importante notar que as métricas DI e SPD são calculadas apenas com base nas predições do modelo, independentemente da acurácia dos resultados. Por esse motivo, são consideradas métricas de **paridade estatística**, úteis para identificar disparidades no acesso ao resultado favorável entre grupos. Já as métricas EOD e AOD incorporam também os rótulos verdadeiros, avaliando a equidade nas taxas de acerto e erro do modelo entre os grupos. Estas últimas são, portanto, mais indicadas quando se busca garantir igualdade de oportunidade ou de erro, relevantes em contextos sensíveis como a saúde.

Cada métrica captura diferentes aspectos do problema de justiça algorítmica e possui intervalos comumente aceitos para indicar o que é considerado justo [2]. A utilização combinada dessas métricas fornece uma avaliação mais abrangente, mitigando conclusões enviesadas que poderiam surgir da análise isolada de apenas uma delas.

III. O FRAMEWORK PROPOSTO

Este trabalho propõe um *framework* de avaliação de justiça algorítmica que consiste em uma expansão do AI Fairness 360 para modelos de classificação multiclasse e *multi-label*, comumente encontrados em aplicações de inteligência artificial na área da saúde. O sistema visa facilitar a análise de métricas de justiça preditiva por classe (ou *label*), atributo e modelo, permitindo comparações entre múltiplos experimentos e a geração de visualizações interpretáveis.

A implementação foi realizada em *Python*, com interface construída em *Streamlit*[1], promovendo uma interação dinâmica com os parâmetros de análise (Fig. 1). Os dados de entrada são fornecidos em arquivos no formato CSV, sendo necessário ao menos um arquivo contendo os atributos e os rótulos verdadeiros do conjunto de dados, além de um ou mais arquivos com as predições de modelos, em formato binário ou probabilístico, para cada *label*.

O fluxo de análise se inicia pela seleção das classes preditivas de interesse, o que pode ser relevante em cenários *multi-label* com muitos rótulos, nos quais, por exemplo, algumas patologias podem possuir maior impacto clínico. Em seguida, define-se o atributo sensível a ser avaliado (como sexo ou

[1]https://streamlit.io/

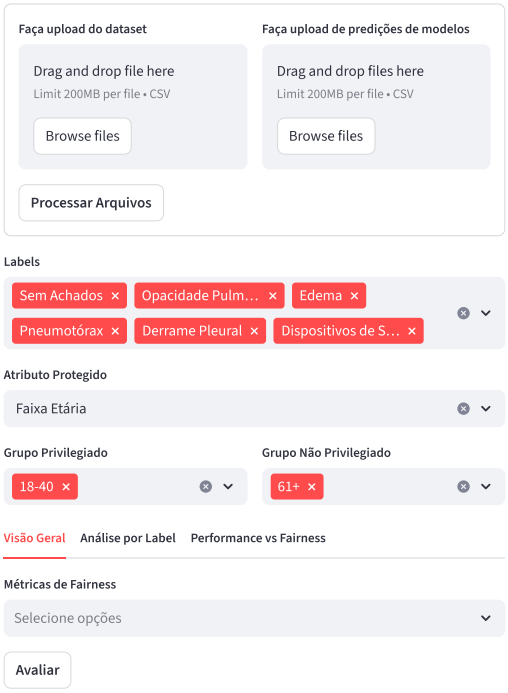


Figura 1. Interface do *framework* proposto. Na parte superior se encontram os botões para seleção dos arquivos de entrada. Logo abaixo, o usuário escolhe as classes (*labels*) a serem consideradas, o atributo protegido, o grupo privilegiado e não privilegiado. Na parte inferior da interface é possível escolher as métricas de *fairness* e a forma de visualização.

faixa etária), bem como os grupos que serão considerados privilegiados e não privilegiados. Cabe destacar que os grupos de referência não implicam necessariamente maior benefício em todas as classes preditivas: a direção do viés pode variar conforme a classe e a distribuição dos dados [4]. Na etapa seguinte, selecionam-se as métricas a serem avaliadas.

O *framework* utiliza a biblioteca `aif360` para o cálculo das métricas de *fairness*, instanciando, para cada classe, objetos do tipo `StandardDataset`, a partir dos quais derivam-se as classes `BinaryLabelDatasetMetric` e `ClassificationMetric`. Além disso, métricas tradicionais de desempenho, como F1-score, são obtidas por meio da biblioteca `scikit-learn`.

A saída da ferramenta é composta por visualizações gráficas das métricas selecionadas, organizadas em três possíveis modos de apresentação:

- **Visão geral:** Matriz que relaciona modelos e classes, exibindo os valores das métricas de justiça para cada combinação. A coloração indica se o valor está dentro do intervalo considerado justo (*fair range*). Essa visualização permite uma análise rápida e compacta sobre a presença de viés entre modelos e categorias preditivas.
- **Análise por classe:** Gráficos individuais para cada classe, no formato de barras horizontais, com os valores das métricas para cada modelo e o *fair range* destacado. Essa visualização permite comparações mais focalizadas entre modelos em cada classe do problema.

- **Performance vs Fairness:** Para cada *label*, são exibidos gráficos de dispersão relacionando métricas tradicionais de desempenho e métricas para avaliação de *fairness*. Essa visualização também exibe a faixa considerada justa, facilitando a identificação de possíveis *trade-offs* que podem existir entre esses dois aspectos.

As figuras são geradas a partir da biblioteca `matplotlib`, permitindo alto grau de personalização gráfica e a inclusão de elementos auxiliares, como faixas de referência e legendas explicativas que facilitam a interpretação das métricas.

A estrutura modular da ferramenta permite sua aplicação em diversos contextos, bastando ajustar os arquivos de entrada e os parâmetros de avaliação conforme o cenário desejado. Além disso, o código foi desenvolvido com flexibilidade, possibilitando a adaptação da entrada para integração com sistemas em tempo real, como bancos de dados clínicos, pipelines automatizados ou interfaces de apoio a diagnóstico.

Dessa forma, o *framework* apresenta-se como uma solução mais completa e flexível para a análise de justiça algorítmica em modelos de classificação, com ênfase especial em tarefas multiclasse ou *multi-label* frequentemente encontradas em cenários clínicos.

IV. Estudo de Caso

Para demonstrar a aplicabilidade do *framework* proposto, foi realizado um estudo de caso envolvendo o treinamento e avaliação de modelos de classificação multirrótulo em radiografias de tórax.

A. Conjunto de dados

Os dados utilizados são provenientes de 3 bases públicas: CheXpert [8], MIMIC-CXR [9] e BRAX [10]. São mais de 600.000 imagens de raio-X do tórax, com vistas frontais e laterais, rotuladas com a presença ou ausência de 14 achados clínicos relevantes: aumento do cardiomediastino; cardiomegalia; lesão pulmonar; opacidade pulmonar; edema; consolidação; pneumonia; atelectasia; pneumotórax; derrame pleural; outro pleural; fratura; dispositivos de suporte; e ausência de achados. Tais classes podem estar correlacionadas e frequentemente podem coexistir, característica de um problema multirrótulo. Além disso, as imagens são acompanhadas de metadados, incluindo informações dos pacientes, como sexo e idade.

B. Modelos

A arquitetura escolhida foi a EfficientNet, por seu equilíbrio entre desempenho computacional e acurácia em tarefas de classificação de imagens, inclusive no contexto médico [11]. Foram avaliadas diversas variantes da EfficientNet, incluindo as versões da série original (B0, B4 e B7) [12] e diferentes tamanhos da EfficientNet-V2 [13]. Essas variantes diferem principalmente em profundidade, largura e resolução de entrada, permitindo investigar como a complexidade arquitetural influencia tanto o desempenho quanto as métricas de *fairness*.

C. Treinamento e Avaliação

Foi realizado o ajuste fino (*fine-tuning*) dos modelos selecionados, previamente treinados no conjunto ImageNet. As redes foram adaptadas para a tarefa de classificação multirrótulo, visando detectar os 14 achados clínicos e foram avaliadas quanto ao desempenho por métricas tradicionais: *F1-score*, *Precision*, *Recall* e curvas ROC.

Para a avaliação de *fairness*, foram considerados apenas os conjuntos CheXpert e BRAX, já que o MIMIC-CXR não disponibiliza metadados de pacientes. Foram definidos como grupo privilegiado pacientes com idade entre 18 e 40 anos, e como grupo não privilegiado pacientes idosos (com mais de 60 anos), refletindo abordagens adotadas na literatura [3], [14]. As inferências realizadas no conjunto de teste foram processadas pelo *framework*, permitindo a geração de visualizações interpretáveis das métricas de *fairness*.

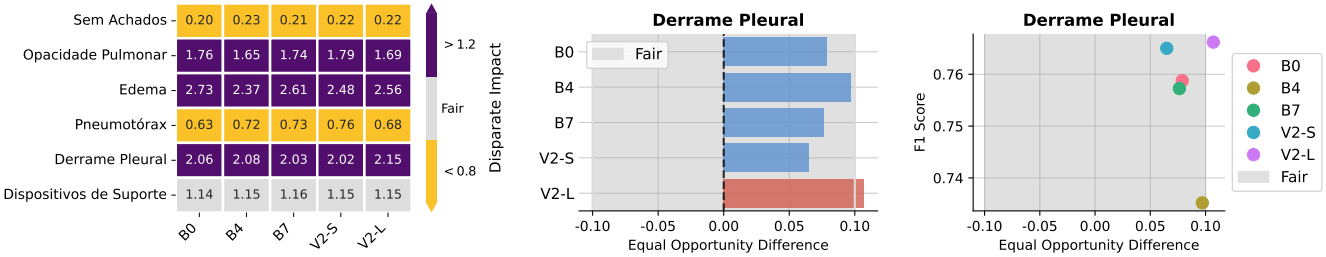
V. Resultados e Discussões

Considerando as métricas tradicionais de desempenho, como F1-score micro e macro, os modelos obtiveram desempenhos semelhantes. Dentre os modelos, foram selecionados aqueles que apresentaram o melhor *trade-off* entre desempenho e tamanho, com a EfficientNet-V2-L alcançando uma performance levemente melhor em relação às demais.

A avaliação da equidade preditiva foi realizada com o auxílio do *framework*, considerando especificamente as métricas DI e EOD. Para simplificar a visualização, foram consideradas somente as 6 melhores classes preditivas: *Sem Achados*, *Opacidade Pulmonar*, *Edema*, *Pneumotórax*, *Derrame Pleural* e *Dispositivos de Suporte*. Além disso, a idade dos pacientes foi considerada como atributo sensível.

Visualizando a métrica DI para cada modelo em cada classe (Fig. 2a), observa-se que há certa similaridade de comportamento entre os modelos EfficientNet, seguindo a tendência observada nas métricas tradicionais de desempenho. Além disso, é possível identificar claramente a presença de viés na maioria das classes (cores mais fortes), que estão fora do intervalo de *fairness*. Achados clínicos como *Opacidade Pulmonar*, *Derrame Pleural* e *Edema* apresentam grandes disparidades nas taxas de predições positivas (PR), favorecendo pacientes acima de 60 anos. Por outro lado, a classe *Sem Achados* mostra um viés oposto: o valor de PR é maior para pacientes com idade entre 18 e 40 anos. É importante destacar que este viés pode refletir a realidade do problema, ou seja, pacientes mais jovens tendem a ser mais saudáveis estatisticamente, e não necessariamente indicam um modelo injusto.

Para uma investigação mais focada, o *framework* permite comparar de forma visual diferentes modelos em relação a uma única classe, como exemplificado para *Derrame Pleural* (Fig. 2b). Embora a visão geral (Fig. 2a) tenha indicado um forte viés segundo a métrica DI, a métrica EOD revela que somente a EfficientNet-V2-L não se comporta de maneira justa. Isto é, há uma maior taxa de verdadeiros positivos (TPR) para pacientes idosos, indicando que o modelo possui menor taxa de acertos na tarefa de identificar a presença de *Derrame Pleural* em indivíduos mais jovens.



(a) (b) (c)

Figura 2. Resultados gerados a partir do *framework* proposto, considerando a idade como atributo sensível. Foram avaliadas cinco variantes da arquitetura EfficientNet (B0, B4, B7, V2-S e V2-L). Valores abaixo e acima do intervalo de *fairness* indicam, respectivamente, viés contra pacientes com idade acima de 60 anos e viés contra pacientes com idade entre 18 e 40 anos. (a) Matriz com os valores da métrica *Disparate Impact* (DI) para todas as classes e modelos avaliados; (b) Valores da métrica *Equal Opportunity Difference* (EOD) para a classe *Derrame Pleural*; (c) Relação entre F1-score e EOD para a classe *Derrame Pleural*.

Este resultado sugere que, ao implementar tal modelo em um ambiente hospitalar, esse grupo está sujeito a mais falhas clínicas, dado que é mais provável que condições graves não sejam diagnosticadas corretamente. A partir dessa análise, nota-se que a divergência entre métricas DI e EOD ressalta a importância de análises complementares para uma avaliação mais abrangente de equidade.

Diante da existência de viés, é importante avaliar qual modelo seria mais adequado para aplicação em ambiente clínico, considerando tanto o desempenho geral quanto as métricas de *fairness* avaliadas. A visualização do F1-score em conjunto com a métrica EOD (Fig. 2c) indica que a EfficientNet-V2-S, além de apresentar a menor disparidade na TPR, mantém um bom desempenho preditivo, destacando-se como uma alternativa mais equilibrada entre acurácia e equidade. Tal resultado reforça que mesmo redes neurais com alta qualidade preditiva (EfficientNet-V2-L) não garantem, por si só, justiça nos resultados, evidenciando a importância de considerar métricas de *fairness* na avaliação de modelos destinados a contextos clínicos sensíveis.

VI. Conclusão

A crescente adoção de modelos de IA em aplicações médicas exige não apenas alta performance, mas também garantias de equidade entre diferentes grupos demográficos. Neste trabalho, foi proposto um *framework* visual e interativo para avaliação de *fairness* em tarefas de classificação supervisionada, aplicado ao caso da detecção de achados em radiografias de tórax. Os resultados evidenciaram vieses relevantes entre grupos etários a partir de múltiplas métricas especializadas, reforçando a importância do uso combinado de diferentes indicadores de justiça algorítmica. Nesse sentido, o *framework* desenvolvido fornece uma interface interativa e de fácil uso para uma avaliação mais robusta e crítica da equidade em cenários médicos. Como trabalho futuro, pretende-se incluir um mecanismo de recomendação das métricas de *fairness* mais adequadas, de acordo com o tipo de problema e os atributos sensíveis, dada a relevância dessa escolha para a interpretação clínica.

Agradecimentos

Os autores agradecem à empresa NeuralMind pelo apoio financeiro e pela infraestrutura computacional disponibilizada para a realização deste trabalho.

Referências

[1] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nat. Med.*, vol. 28, no. 1, pp. 31–38, 2022.

[2] F. Dehghani, N. Malik, J. Lin, S. Bayat, and M. Bento, "Fairness in healthcare: Assessing data bias and algorithmic fairness," in *Proc. 20th Int. Symp. Med. Inf. Process. Anal. (SIPAIM)*, 2024, pp. 1–6.

[3] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi, "CheXclusion: Fairness gaps in deep chest x-ray classifiers," in *Proc. Biocomputing 2021: Pac. Symp.*, 2020, pp. 232–243.

[4] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *P. Natl. Acad. Sci. U.S.A.*, vol. 117, no. 23, pp. 12 592–12 594, 2020.

[5] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan *et al.*, "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," 2018, arXiv:1810.01943.

[6] P. Saleiro, B. Kuester, A. Stevens, A. Anisfeld, L. Hinkson, J. London *et al.*, "Aequitas: A bias and fairness audit toolkit," 2018, arXiv:1811.05577.

[7] H. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, and M. Madaio, "Fairlearn: Assessing and improving fairness of AI systems," *J. Mach. Learn. Res.*, vol. 24, no. 257, pp. 1–8, 2023.

[8] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," 2019, arXiv:1901.07031.

[9] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. ying Deng *et al.*, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, no. 1, p. 317, 2019.

[10] E. P. Reis, J. P. Q. de Paiva, M. C. B. da Silva, G. A. S. Ribeiro, V. F. Paiva, L. Bulgarelli *et al.*, "BRAX, Brazilian labeled chest x-ray dataset," *Sci. Data*, vol. 9, no. 1, p. 487, 2022.

[11] F. Zulfiqar, U. I. Bajwa, and Y. Mehmood, "Multi-class classification of brain tumor types from MR images using EfficientNets," *Biomed. Signal Process. Control*, vol. 84, p. 104777, 2023.

[12] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, vol. 97, Jun 2019, pp. 6105–6114.

[13] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, vol. 139, Jul 2021, pp. 10 096–10 106.

[14] Y. Yang, Y. Liu, X. Liu, A. Gulhane, D. Mastrodicasa, W. Wu *et al.*, "Demographic bias of expert-level vision-language foundation models in medical imaging," *Sci. Adv.*, vol. 11, no. 13, p. eadq0305, 2025.