Journal of Health Informatics

J Health Inform. 2026;18:e1853
DOI: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1853

ARTIGO COMPLETO
CBIS 2026

Ponderação dinâmica, diretriz médica ou frequência? O impacto destas estratégias de custos em triagem torácica

Dynamic weighting, clinical guideline or frequency? The impact of these cost strategies on chest screening

Gustavo Pedreira[1] ✉, Henrique Fernandes[1], Marcelo Zanchetta do Nascimento[1]

ORCID: 0009-0005-9523-4583 · 0000-0002-7078-9620 · 0000-0003-3537-0178

[1] Universidade Federal de Uberlândia, Uberlândia (MG), Brasil.

FICHA DO ARTIGO

TIPO
Artigo completo · CBIS 2026

IDENTIFICADOR
e1853

DOI
10.59681/2175-4411.v18.iEspecial.2026.1853

RECEBIDO
17/05/2026

APROVADO
20/07/2026

PROCESSO DE AVALIAÇÃO
Double Blind Review

EDITOR RESPONSÁVEL
Juliano de Souza Gaspar

ACESSO
Aberto, CC BY 4.0

AUTOR CORRESPONDENTE
Gustavo Pedreira
gustavo.pedreira@ufu.br

LICENÇA

Este é um artigo publicado em acesso aberto sob a licença Creative Commons CC BY 4.0, que permite uso, distribuição e reprodução em qualquer meio, desde que o trabalho original seja corretamente citado.

CC BY 4.0

RESUMO

Objetivo: Investigar o impacto de estratégias de ponderação de custo na triagem multirrótulo de radiografias torácicas, com ênfase em patologias de cauda longa. **Método:** Avaliaram-se nove cenários de custo (estatísticos, clínicos, híbridos e dinâmicos) aplicados à DenseNet-121 com Weighted Asymmetric Loss e Pairwise Ranking Loss em uma amostra controlada do NIH ChestX-ray14 (23.635 imagens, 6 classes), com validação em cinco sementes ($N = 5$). **Resultados:** A estratégia estatística Inverso da Frequência (InvFreq) foi a mais eficaz, alcançando AUC macro de 0,8863 ± 0,0016, F1 macro de 0,6115 ± 0,0145 e F1 de 0,5473 ± 0,0812 para Hérnia. O impacto do custo foi assimétrico: a cauda extrema (Hérnia, 0,57%) demonstrou alta sensibilidade, enquanto a classe intermediária (Pneumotórax, 13,62%) manteve desempenho estável. **Conclusão:** A estratégia de custo é crítica para a segurança clínica na triagem de achados raros, sendo a abordagem estatística a mais eficaz.

Descritores: Radiografia torácica; Aprendizado de máquina; Diagnóstico por computador.

ABSTRACT

Objective: To investigate the impact of cost-weighting strategies on multi-label chest X-ray screening, with an emphasis on long-tail pathologies. **Methods:** Nine cost scenarios (statistical, clinical, hybrid, and dynamic) applied to a DenseNet-121 with Weighted Asymmetric Loss and Pairwise Ranking Loss were evaluated on a controlled sample of the NIH ChestX-ray14 dataset (23,635 images, 6 classes), with validation across five seeds (N=5). **Results:** The statistical Inverse Frequency (InvFreq) strategy was the most effective, achieving a macro AUC of 0.8863 ± 0.0016, a macro F1 of 0.6115 ± 0.0145, and an F1 of 0.5473 ± 0.0812 for Hernia. The impact of the cost function was asymmetric: the extreme long-tail class (Hernia, 0.57%) showed high sensitivity, whereas the intermediate class (Pneumothorax, 13.62%) maintained stable performance. **Conclusion:** The cost strategy is critical for clinical safety in screening rare findings, with the statistical approach being the most effective.

Keywords: Radiography, Thoracic; Machine Learning; Diagnosis, Computer-Assisted.

CBIS 2026 XXI CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE

Artigo apresentado no XXI Congresso Brasileiro de Informática em Saúde (CBIS 2026), realizado de 23 a 25 de setembro de 2026, em Brasília (DF), e publicado na seção CBIS 2026 do Journal of Health Informatics.

COMO CITAR ESTE ARTIGO: Pedreira G, Fernandes H, do Nascimento MZ. Ponderação dinâmica, diretriz médica ou frequência? O impacto destas estratégias de custos em triagem torácica. J Health Inform 2026;18:e1853. doi: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1853

1 INTRODUÇÃO

A radiografia de tórax (RXT) é um dos exames de imagem mais realizados no mundo e permanece como referência primária na propedêutica clínica há décadas [1]. Sua ampla adoção e sustentabilidade nos sistemas de saúde globais devem-se à rápida aquisição de imagens, ao baixo custo operacional e à reduzida exposição do paciente à radiação ionizante [2]. Diariamente, dezenas de milhões de exames radiográficos são utilizados para o rastreio inicial, diagnóstico e manejo terapêutico de diversas patologias intratorácicas, abrangendo um espectro complexo de anormalidades pulmonares, cardíacas, ósseas e pleurais [1,3,4].

A RXT tem ganhado notório destaque em estudos de rastreio oportunístico apoiados por inteligência artificial (IA) [1,5-7]. Avanços recentes em aprendizado profundo (AP) demonstraram a viabilidade de utilizar RXT de rotina para extrair biomarcadores fenotípicos não diretamente observáveis ao olho humano, possibilitando predições que extrapolam o sistema respiratório. A literatura recente relata a estimativa de risco de mortalidade cardiovascular [1,5], a detecção incidental e precoce de diabetes tipo 2 [6] e a avaliação de viabilidade e custo-efetividade para o rastreio de osteoporose em larga escala [7]. Essa emergente linha de pesquisa amplia substancialmente o valor clínico deste exame e a mitigação de vieses em distribuições de cauda longa se faz essencial [8,9].

Apesar de sua consolidação tecnológica, a interpretação humana da RXT permanece intrinsecamente complexa e suscetível a falhas perceptivas e cognitivas [3]. A compressão geométrica de estruturas torácicas tridimensionais em uma projeção bidimensional, englobando o feixe vascular, o arcabouço ósseo costal e os tecidos moles, produz inevitáveis sobreposições anatômicas. Tais artefatos visuais dificultam a identificação de alterações patológicas sutis [2,9]. Consequentemente, a acurácia da avaliação radiológica depende de forma desproporcional do tempo de experiência clínica e da heurística visual prévia do profissional, conferindo ao processo analítico um inegável grau de subjetividade [1,2]. Essa dependência humana resulta em considerável variabilidade interobservador e intraobservador [3,10]. Corroborando essa vulnerabilidade, análises retrospectivas, incluindo a avaliação multicêntrica de mais de 1,6 milhão de exames, demonstram uma taxa persistente de discrepâncias interpretativas que, aliada aos dados históricos, revela que radiologistas podem omitir até 30% dos achados positivos (falsos-negativos) em leituras de rotina [3,4].

Este desafio diagnóstico é severamente agravado pelo volume atual da demanda institucional [11,12]. Consolidada como modalidade de triagem de primeira linha nas emergências e enfermarias, a quantidade de imagens de RXT geradas diariamente cresce em ritmo exponencialmente superior à capacidade de formação de novos especialistas, exacerbando o já crítico cenário de escassez global de radiologistas [11,13]. Tal descompasso demográfico e ocupacional submete as equipes médicas a cargas intensas e ininterruptas de trabalho, impondo uma pressão sistêmica por laudos rápidos [3, 13]. Inevitavelmente, a fadiga visual e a sobrecarga cognitiva associadas a esse ecossistema de alto rendimento reduzem a acurácia diagnóstica, elevam a incidência de erros por omissão de achados incidentais e comprometem os desfechos clínicos e a segurança do paciente [3,4].

A transição da validação teórica dos sistemas de diagnóstico auxiliado por computador (CAD) para a sua implementação prática ganha contornos críticos ao analisar a integração da IA em grandes sistemas públicos de saúde, como o Sistema Único de Saúde (SUS) no Brasil e o *National Health Service* (NHS) no Reino Unido [12,14,15]. Em redes públicas de saúde, o uso de algoritmos automatizados de leitura de RXT tem se mostrado uma intervenção estratégica para otimizar fluxos de triagem, acelerar o diagnóstico em linhas de cuidado de alta urgência, como no rastreamento de câncer de pulmão ou na busca ativa de tuberculose em populações vulneráveis, e mitigar graves assimetrias regionais de acesso a especialistas [10,14,15]. Contudo, a implantação em escala nesses ecossistemas enfrenta gargalos operacionais e institucionais severos, que vão desde baixos índices de digitalização e interoperabilidade de sistemas até complexidades de governança, conformidade regulatória e gestão de mudança com as equipes de saúde [12,14]. Além disso, em cenários de rastreio populacional, inclusive os rastreios oportunísticos, e triagem em massa, como em estabelecimentos prisionais ou unidades periféricas, a eficácia do CAD depende criticamente da capacidade do modelo de manter alta especificidade e sensibilidade sem gerar um volume inviável de falsos-positivos ou negligenciar comorbidades raras [10,15].

Diante deste panorama, o desenvolvimento de sistemas automatizados de diagnóstico auxiliado por computador tornou-se uma intervenção tecnológica prioritária para mitigar a sobrecarga médica e racionalizar o fluxo de triagem [10,11]. Contudo, a formulação de modelos preditivos robustos e generalizáveis para RXT enfrenta um obstáculo estatístico limitante: a assimetria patológica de cauda longa combinada à intrincada presença de múltiplos rótulos diagnósticos concorrentes em uma única imagem [9,16-18]. Conforme evidenciado por Pedreira *et al.* na concepção da arquitetura FRCS-Net [19], abordagens baseadas predominantemente na minimização da entropia cruzada binária (BCE, do inglês *binary cross-entropy*) tendem a negligenciar a topologia da cauda longa. Esse viés de otimização favorece o aprendizado e a convergência rápida para anomalias altamente prevalentes, mas reduz drasticamente a sensibilidade

algorítmica para doenças raras, fenômeno estatístico que invalida a segurança clínica do modelo em ambientes reais de rastreio [16,17,19].

Em uma recente revisão sistemática por Gomes *et al.*, que avaliou a classificação de pneumonias por métodos de IA, reportaram-se acurácias diagnósticas teóricas na ordem de 99,7% [20]. O estado da arte, entretanto, exige cautela na interpretação dessas métricas, pois esses níveis de desempenho quase sempre se restringem a conjuntos de dados exaustivamente curados, de escopo patológico binário ou artificialmente balanceados [8, 20]. No ecossistema clínico real, a prevalência de patologias torácicas obedece a uma curva de potência altamente assimétrica. Nesses ambientes, a vasta maioria das anomalias ocorre de forma esporádica e com poucas amostras representativas, induzindo as redes neurais convolucionais convencionais a concentrarem sua capacidade representacional de ativação em classes majoritárias (como cardiomegalia e derrames pleurais), em detrimento da identificação de achados críticos minoritários, como hérnias ou nódulos sutis [9,16,18].

Historicamente, o enfrentamento desse desbalanceamento tem recorrido a estratégias centradas puramente na reestruturação dos dados, notadamente abordagens de reamostragem baseadas em *oversampling* (duplicação de instâncias minoritárias) ou *undersampling* (supressão de instâncias majoritárias) [8]. Entretanto, evidências empíricas sólidas indicam que, em contextos médicos multirrótulo, a manipulação direta do *dataset* pode descaracterizar as complexas dependências de coocorrência anatômico-patológica, introduzindo ruídos semânticos severos que degradam a generalização do modelo [8,9,16]. Por conseguinte, a fronteira da pesquisa tem convergido para o paradigma de aprendizado sensível ao custo [8,10]. Essa família de algoritmos preserva a integridade e a distribuição original das imagens, atuando exclusivamente na reestruturação matemática da função de perda. O método penaliza o modelo de forma assimétrica e mais severa pelos erros cometidos nas classes minoritárias durante o cálculo do gradiente, contornando os riscos sintéticos da reamostragem [8,10,16].

Atualmente, a aplicação de funções de custo é dominada por estratégias de pesos estáticos. Nesse escopo, uma parcela substancial da literatura emprega o inverso das frequências de classe para escalonar a entropia cruzada, impondo um rigor matemático maior às doenças menos prevalentes [8,10,16]. Em contraponto a essa visão, estudos recentes advogam pela incorporação de conhecimentos clínicos especializados (*medical consensus*) diretamente no cômputo da perda [10,19]. O modelo FRCS-Net propôs uma função híbrida otimizada em dois estágios, calibrada por métricas de severidade médica padronizadas, equilibrando a detecção de patologias de alto risco sistêmico mediante restrições de ranqueamento pareado [19]. Os testes desse modelo evidenciaram ganhos de F1-*score* para anomalias raras sem deterioração estatística da especificidade global [19].

Apesar do avanço promovido por arquiteturas correlatas orientadas pelo custo radiológico [10,19], o campo ainda colide com uma restrição técnica sistêmica: a rigidez algorítmica. Consensos médicos validados previamente, ou mesmo frequências globais estáticas, são hiperparâmetros fixos, permanecendo inertes durante todas as épocas de treinamento, independentemente de como a rede evolui [8,16]. Em marcante contraste, as modernas abordagens de ponderação dinâmica recalibram os tensores de custo de forma contínua e autônoma, pautando-se exclusivamente pela magnitude dos erros de predição observados nas iterações imediatamente anteriores [13, 16,17,21]. Esse mecanismo de retroalimentação estatística refina dinamicamente o vetor de otimização em direção às amostras e classes intrinsecamente mais difíceis de mapear, sem depender de uma classificação prévia de pertencerem à cauda longa ou ao espectro majoritário da base [16,17].

Diante da consolidação dessas duas vertentes, permanece sem resposta metodológica conclusiva qual matriz de ponderação entrega sistemas CAD mais responsivos, seguros e operacionalmente viáveis para a prática clínica diária. Neste trabalho, propõe-se uma análise comparativa rigorosa em que estratégias puramente dinâmicas, balizadas pela oscilação temporal do erro de predição [13,16,17,21], são confrontadas com modelos de pesos estáticos ancorados em criticidade clínica ou frequência estatística. O escopo central desta pesquisa é investigar essa dicotomia estrutural e determinar, de maneira empírica, se a fluidez do ajuste estatístico dinâmico possui superioridade preditiva frente à estabilidade dos consensos médicos cristalizados na proteção radiológica das classes minoritárias.

2 MÉTODOS

O estudo utiliza um subconjunto curado da base pública NIH ChestX-ray14, totalizando 23.635 radiografias após a filtragem de comorbidades, com foco em cinco patologias de interesse (hérnia, enfisema, massa, efusão e pneumotórax). Essa seleção buscou isolar o fenômeno de cauda longa e garantir que os gradientes de erro fossem provenientes exclusivamente das classes monitoradas. O problema foi formulado como classificação *multirrótulo* com seis categorias. Observou-se severo desbalanceamento, caracterizando uma distribuição de cauda longa: a classe majoritária, correspondente à cabeça da distribuição, foi *No Finding* (53,84%); as classes intermediárias incluíram efusão (20,92%), pneumotórax (13,62%), massa (12,13%) e enfisema (6,27%); e a classe de cauda extrema foi hérnia (0,57%). O fator de

desbalanceamento entre a classe majoritária e a minoritária foi de aproximadamente 94:1.

Para garantir a integridade da avaliação clínica, a divisão dos dados (70% treino, 10% validação e 20% teste) foi realizada estritamente no nível do paciente, evitando vazamento de radiografias seriadas do mesmo indivíduo entre os conjuntos, o que comprometeria a validade da avaliação, especialmente nas classes raras.

A formulação da perda constitui a variável independente central do estudo. O modelo de referência (*baseline*) adotado foi a BCE não ponderada. Para o tratamento da cauda longa, foi utilizada a arquitetura de custo FRCSLoss [19], composta por duas frentes operando em estágios.

A frente base é a *Weighted Asymmetric Loss* (WASL), que estende a ASL com pesos vetoriais (w_c). Para priorizar instâncias de difícil classificação e mitigar o domínio de exemplos negativos triviais, foram definidos os parâmetros de decaimento $\gamma neg = 2$ e $\gamma pos = 0$, com inclusão de um fator de *clipping* (0,05) para estabilidade numérica.

Os pesos (w_c) compõem os nove cenários avaliados: (1) frequência inversa normalizada no intervalo [1,5] (InvFreq); (2) criticidade baseada no American College of Radiology (ACR); (3) Híbrido 1 (0,7 · InvFreq + 0,3 · ACR); (4) Híbrido 2 (média geométrica √(InvFreq · ACR) renormalizada em [1,5]); (5) Híbrido 3 (0,3 · InvFreq + 0,7 · ACR); (6) pesos originais de Pedreira *et al.*; e (7-8) abordagens dinâmicas (Curriculum Learning e Loss-Aware Dynamic Weights). A Tabela 1 apresenta a matriz de pesos.

No segundo estágio, acopla-se a *Pairwise Ranking Loss*, que penaliza instâncias em que a probabilidade atribuída a uma classe negativa supera a de uma classe positiva. Nos modelos dinâmicos, o vetor de pesos atingido ao final do primeiro estágio é congelado e mantido durante a aplicação do segundo estágio com *ranking*. Essa etapa foi calibrada por função *softplus*, com margem $\tau = 0,7$ e taxa de modulação $\lambda = 1,0$.

Adotou-se a arquitetura *DenseNet-121* pré-treinada na *ImageNet*, modificada com um cabeçalho linear para C = 6 classes. O pré-processamento incluiu redimensionamento das imagens para 512 × 512 pixels e equalização adaptativa de histograma limitada por contraste (CLAHE, do inglês *Contrast Limited Adaptive Histogram Equalization*; p = 0,8) para realce do contraste radiológico. O aumento de dados contemplou transformações afins espacialmente conservadoras, como rotações (±15°), translações (10%) e espelhamento horizontal.

O treinamento utilizou o otimizador *AdamW*, associado à técnica de acúmulo de gradientes (tamanho de lote físico de 16 e lote efetivo de 64) e à *automatic mixed precision* (AMP). A taxa de aprendizado foi ajustada dinamicamente por *ReduceLROnPlateau*.

A acurácia global foi preterida por métricas robustas ao viés de prevalência (AUC, F1-*score*, sensibilidade/*recall* e sensibilidade a 95% de especificidade). A determinação do limiar ótimo de decisão (*threshold*) para cada classe foi obtida buscando a maximização do F1-*score* exclusivamente no conjunto de validação. Esses limiares foram congelados e aplicados diretamente ao conjunto de teste independente, eliminando qualquer risco de *data leakage*.

Para a avaliação estatística rigorosa, foi aplicado um funil de seleção a partir do experimento inicial (*seed* = 42): fixou-se a variante "Com *Ranking*" e eliminaram-se estratégias com F1-*score* para Hérnia inferior ao *baseline* BCE (0,0571). Os 5 cenários finalistas (InvFreq, Híbrido 1, Híbrido 2, *Curriculum Learning* e *Loss-Aware Dynamic Weights*) foram submetidos a um protocolo de robustez com 5 sementes determinísticas independentes, avaliando-se a média e o desvio-padrão das métricas no conjunto de teste. O modelo de Pedreira *et al.* [19] foi excluído da etapa de múltiplas sementes por possuir pesos definidos de forma empírica (*ad hoc*), não se enquadrando nas metodologias comparadas.

Tabela 1. Pesos dos cenários de avaliação (wc) por classe patológica.

Cenário	Hérnia	Enfisema	Massa	Efusão	Pneumotórax	No Finding
0 – BCE (*Baseline*)	1,000	1,000	1,000	1,000	1,000	1,000
1 – InvFreq (Estatístico)	5,000	1,325	1,147	1,067	1,126	1,000
2 – Criticidade (Clínico)	1,000	1,500	3,000	2,500	5,000	1,000
3 – Híbrido 1 (Estat.+)	3,800	1,377	1,703	1,497	2,288	1,000
4 – Híbrido 2 (Neutro)	4,601	2,193	3,491	2,845	5,000	1,000
5 – Híbrido 3 (Clínico+)	2,200	1,447	2,444	2,070	3,838	1,000
6 – Pedreira *et al.* [19]	3,000	2,250	1,500	1,500	2,250	1,000
7 – Curriculum Learning	Dinâmico	Dinâmico	Dinâmico	Dinâmico	Dinâmico	Dinâmico
8 – Loss-Aware Dynamic Weights	Dinâmico	Dinâmico	Dinâmico	Dinâmico	Dinâmico	Dinâmico

Fonte: Elaborado pelos autores (2026).

Tabela 2. Desempenho por classe e macro sob restrição ativa de Ranking Loss.

Cenário	AUC Macro	F1 Macro	AUC Hérnia	F1 Hérnia	AUC Pneumo	F1 Pneumo
BCE (*Baseline*)	0,8783	0,5351	0,9039	0,0571	0,8741	0,5877
InvFreq (Estatístico)	0,8816	0,5858	0,9322	0,3636	0,8675	0,5873
Criticidade (Clínico)	0,8723	0,5272	0,8882	0,0303	0,8680	0,5768
Híbrido 1 (Estat.+)	0,8824	0,5383	0,9480	0,1053	0,8711	0,5866
Híbrido 2 (Neutro)	0,8804	0,5837	0,9310	0,3478	0,8680	0,5618
Híbrido 3 (Clínico+)	0,8828	0,5175	0,9304	0,0000	0,8753	0,5779
Pedreira *et al.* [19]	0,8847	0,6044	0,9528	0,4348	0,8661	0,6047
Curriculum Learning	0,8778	0,5403	0,9599	0,1818	0,8584	0,5901
Loss-Aware Dyn. Wts.	0,8826	0,5406	0,9444	0,1176	0,8668	0,5975

Fonte: Elaborado pelos autores (2026).

3 RESULTADOS

A avaliação do desempenho preditivo indica que a incorporação da penalização por *ranking* produziu ganhos em relação às arquiteturas treinadas apenas com perdas pontuais, padrão observado em todos os cenários de função de custo analisados. Tal mecanismo esteve associado à redução da supressão das classes minoritárias em múltiplos cenários, resultando em incrementos nas métricas avaliadas independentemente da estratégia de ponderação adotada na camada base.

A introdução da penalização por *ranking* no modelo *baseline* (BCE) elevou o *recall* macro de 0,5495 para 0,5726 e o F1-*score* macro de 0,5221 para 0,5351. Tendência análoga foi observada nas formulações estáticas sensíveis ao custo. O modelo InvFreq (estatístico) atingiu F1-*score* macro de 0,5653 sem *ranking* e evoluiu para 0,5858 com a inclusão da perda de ranqueamento. A estratégia baseada na criticidade clínica (ACR) apresentou elevação de AUC macro (0,8574 para 0,8723) e F1-*score* macro (0,5131 para 0,5272). O modelo Híbrido 2 (Neutro), concebido para equilibrar as vertentes clínica e estatística, registrou ganho de F1-*score* (0,5484 para 0,5837). Os pesos originais de Pedreira *et al.* [19] alcançaram a maior AUC macro (0,8847) e o maior F1-*score* macro (0,6044) sob restrição de *ranking* no experimento inicial. Nas abordagens dinâmicas, o congelamento dos pesos ao final do primeiro estágio seguido do acoplamento do *ranking* manteve o ganho competitivo, com F1-*score* macro de 0,5403 (*Curriculum Learning*) e 0,5406 (*Loss-Aware Dynamic Weights*). Pode-se afirmar que os 9 modelos com *ranking* superaram ou empataram em desempenho, pelas métricas utilizadas, com seus pares sem *ranking*.

Ao analisar os nove cenários operando sob restrição ativa de *ranking loss* (Tabela 2), evidencia-se que o impacto das funções de custo variou sensivelmente conforme a frequência amostral da patologia. O modelo InvFreq (Estatístico) alcançou F1-*score* de 0,3636 para Hérnia (cauda extrema, 0,57%) e 0,5873 para Pneumotórax. A estratégia Pedreira *et al.* [19] destacou-se com o maior F1-*score* para Hérnia (0,4348) e F1-*score* de 0,6047 para Pneumotórax. Em contrapartida, a estratégia puramente clínica (Criticidade) obteve F1-*score* reduzido para Hérnia (0,0303), situando-se abaixo do BCE (0,0571), enquanto o Híbrido 3 obteve F1-*score* nulo (0,0000). Dentre as técnicas dinâmicas, a abordagem *Curriculum Learning* obteve AUC de 0,9599 e F1-*score* de 0,1818 para Hérnia, enquanto a *Loss-Aware* obteve 0,9444 de AUC e 0,1176 de F1-*score*.

Para achados de prevalência intermediária, como o Pneumotórax (13,62%), a oscilação do desempenho entre as diferentes funções de custo revelou-se consideravelmente menor. A AUC situou-se na faixa entre 0,8584 e 0,8753, enquanto o F1-*score* variou de 0,5618 a 0,6047. Esse comportamento indica maior estabilidade preditiva das redes neurais em patologias com representação amostral intermediária, reduzindo a sensibilidade do modelo às alterações na matriz de custos, o que era esperado.

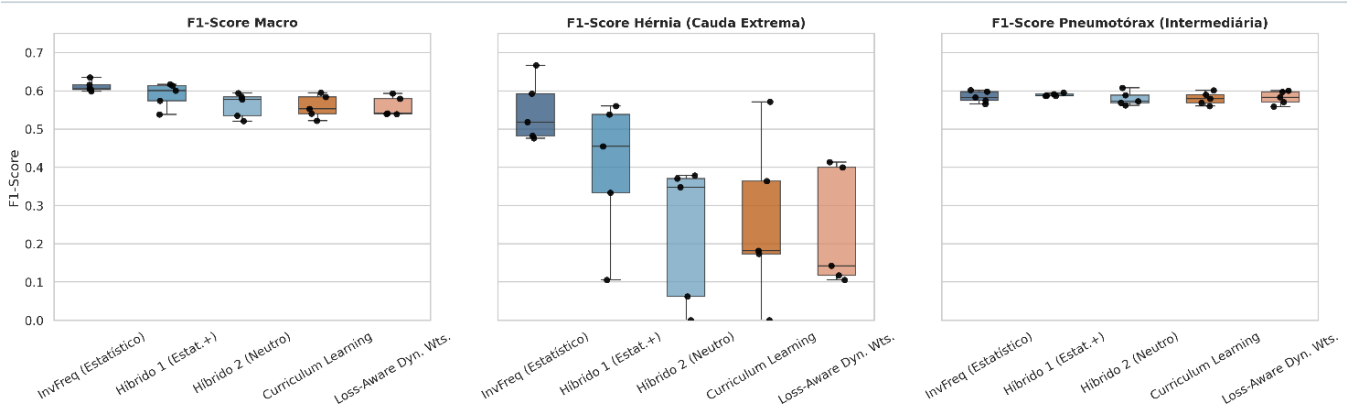
A avaliação de robustez estatística conduzida em 5 sementes determinísticas independentes para os cenários finalistas (Tabela 3 e Figura 1) revelou superioridade da estratégia InvFreq (Estatístico). Esse modelo obteve a maior média de F1-*score* macro (0,6115 ± 0,0145) e a maior média de F1-*score* para Hérnia (0,5473 ± 0,0812), demonstrando também a menor dispersão e maior estabilidade na cauda extrema. O modelo Híbrido 1 obteve o segundo melhor F1-*score* para Hérnia (0,3983 ± 0,1864) e o maior F1 para Pneumotórax (0,5895 ± 0,0039).

Tabela 3. Síntese estatística da análise com múltiplas sementes.

Estratégia de Custo	AUC Macro	F1 Macro	F1 Hérnia (Cauda Extrema)	F1 Pneumotórax (Intermediária)
InvFreq (Estatístico)	0,8863 ± 0,0016	0,6115 ± 0,0145	0,5473 ± 0,0812	0,5845 ± 0,0153
Híbrido 1 (Estat.+)	0,8845 ± 0,0026	0,5887 ± 0,0328	0,3983 ± 0,1864	0,5895 ± 0,0039
Loss-Aware Dyn. Wts.	0,8863 ± 0,0027	0,5582 ± 0,0261	0,2359 ± 0,1567	0,5822 ± 0,0176
Híbrido 2 (Neutro)	0,8835 ± 0,0053	0,5621 ± 0,0320	0,2318 ± 0,1848	0,5798 ± 0,0184
Curriculum Learning	0,8820 ± 0,0045	0,5589 ± 0,0304	0,2582 ± 0,2173	0,5799 ± 0,0163

Fonte: Elaborado pelos autores (2026).

Figura 1. Variabilidade do F1-score em cinco sementes determinísticas independentes.



Fonte: Elaborado pelos autores (2026).

Em contrapartida, o modelo Híbrido 2 apresentou acentuada queda de desempenho na cauda extrema quando avaliado sob múltiplas sementes (F1-*score* médio para Hérnia de 0,2318 ± 0,1848), acompanhado de alta volatilidade estocástica (variação de 0,0000 a 0,4000). As abordagens dinâmicas demonstraram desempenho intermediário para Hérnia (*Curriculum*: 0,2582 ± 0,2173; *Loss-Aware*: 0,2359 ± 0,1567), confirmando que a recalibração dinâmica não superou a simplicidade da ponderação por frequência estatística inversa em cenários de extrema escassez amostral.

4 DISCUSSÃO

Os resultados revelam que a estratégia de custo é determinante em triagem multirrótulo em RXT, mas seu impacto é modulado pela posição da patologia na distribuição de frequências. O efeito mostrou-se assimétrico: enquanto classes de cauda extrema (<1%, como hérnia) demonstraram sensibilidade crítica ao esquema de ponderação, classes de frequência intermediária (pneumotórax, 13,62%) mantiveram-se estáveis entre os cenários.

O contraste entre estratégias puras evidencia essa dinâmica. O cenário Criticidade (Clínico), fundamentado nas diretrizes do ACR, atribui peso máximo ao pneumotórax (wc = 5), atingindo F1-*score* de 0,5768. Contudo, reduziu o sinal de gradiente da hérnia a um F1-*score* de 0,0303, inferior à BCE não ponderada (0,0571). Esse fenômeno evidencia um *trade-off* estrutural: priorizar patologias intermediárias de alta criticidade via penalização estática pontual pode penalizar classes de cauda extrema e agravar o viés de distribuição.

Em marcante contraponto, a avaliação estatística rigorosa em cinco sementes (N = 5) estabeleceu a estratégia InvFreq (Estatístico) como a abordagem mais robusta e eficaz do estudo (F1 macro médio de 0,6115 ± 0,0145 e F1 Hérnia de 0,5473 ± 0,0812). Esse achado desafia a premissa de que esquemas complexos de ponderação clínica ou adaptações dinâmicas sejam indispensáveis para o aprendizado em cauda longa: o escalonamento baseado puramente na frequência empírica aparenta ser suficiente para fornecer um sinal de gradiente proporcional à escassez amostral, minimizando a supressão das classes minoritárias sem induzir instabilidade. A formulação Híbrida 2 (Neutro), embora teoricamente promissora, demonstrou acentuada volatilidade na cauda extrema sob múltiplas inicializações (F1 Hérnia médio de 0,2318 ± 0,1848), indicando que a calibração geométrica estática entre frequências e pesos de severidade clínica pode amplificar o ruído estocástico quando aplicada a classes de altíssima escassez amostral.

A *Pairwise Ranking Loss* demonstrou efeito complementar consistente, alinhando-se aos achados da arquitetura FRCS-Net [19]. A introdução de restrições de ordenamento relativo no segundo estágio atuou como um regularizador topológico eficaz, elevando o *recall* e o F1-*score* em praticamente todas as matrizes de custo avaliadas.

As abordagens dinâmicas não superaram os melhores cenários estáticos em F1 macro. A estratégia *Loss-Aware* obteve elevada AUC individual para Hérnia (0,9683 ± 0,0142), indicando alta capacidade discriminativa no espaço ROC. Entretanto, seu F1-score médio para Hérnia permaneceu modesto (0,2359 ± 0,1567). Essa dissociação revela que a recalibração dinâmica baseada no erro de época ajusta com eficiência os *logits* para a ordenação das amostras, mas pode descalibrar a distribuição de probabilidades no ponto de operação (limiar de decisão congelado), reduzindo a precisão prática do modelo binarizado. Para o pneumotórax, as técnicas dinâmicas geraram resultados equivalentes às estáticas (F1 de 0,5799 a 0,5822), reiterando que o ajuste dinâmico não confere vantagem em classes com amostragem suficiente.

5 CONCLUSÕES

Este estudo demonstrou que a estratégia de custo adotada durante o treinamento é uma variável crítica para a segurança clínica de sistemas CAD multirrótulo em radiografias torácicas, com impacto assimétrico entre as classes patológicas. Classes de cauda extrema, como hérnia (0,57%), mostraram alta sensibilidade ao esquema de ponderação, enquanto classes de frequência intermediária, como pneumotórax (13,62%), mantiveram desempenho estável. A estratégia InvFreq (Estatístico) emergiu como a formulação de melhor desempenho e maior estabilidade sistêmica, alcançando AUC macro de 0,8863 ± 0,0016, F1-score macro de 0,6115 ± 0,0145 e F1-score de 0,5473 ± 0,0812 para hérnia no protocolo de cinco sementes (N = 5). A avaliação estocástica expandida desconstruiu a pretensa superioridade de abordagens mais complexas, revelando que penalizar a perda com base na frequência empírica inversa supera tanto a calibração híbrida com diretrizes clínicas (que apresentou alta volatilidade na cauda extrema) quanto a adaptação dinâmica contínua. A incorporação da *Pairwise Ranking Loss* no segundo estágio revelou-se um mecanismo de regularização consistentemente benéfico em todos os cenários, corroborando e ampliando os achados da arquitetura FRCS-Net [19].

O estudo apresenta limitações que condicionam a interpretação dos resultados. A avaliação foi restrita a um subconjunto curado de seis classes, a um único *backbone* (*DenseNet-121*) e a um número reduzido de instâncias positivas no conjunto de teste para a cauda extrema. Essa delimitação, necessária para isolar o fenômeno de cauda longa com rigor experimental, limita a generalização direta para o espectro completo de patologias torácicas e para arquiteturas alternativas.

Como desdobramentos prioritários, recomendam-se: (i) a replicação do protocolo no conjunto completo do NIH ChestX-ray14 (112.120 imagens, 14 classes), permitindo avaliar o comportamento das estratégias de custo em um espectro de cauda longa ampliado; e (ii) a validação externa em bases independentes (*CheXpert, MIMIC-CXR e PadChest*), aferindo a generalização a diferentes populações e protocolos de aquisição.

Contribuição dos autores. Os autores contribuíram igualmente em todas as etapas de construção do artigo: concepção e delineamento do estudo; coleta, análise e interpretação dos dados; redação e revisão crítica do manuscrito; e aprovação da versão final.

Financiamento. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Disponibilidade de dados e materiais. Os autores não declararam a disponibilidade dos dados e materiais da pesquisa.

Conflitos de interesse. Os autores não declararam conflitos de interesse.

Declaração de uso de IA. Os autores declaram o uso da ferramenta de inteligência artificial Gemini durante a etapa de preparação e redação final deste manuscrito. A ferramenta foi utilizada exclusivamente como apoio à redação e à revisão textual, com o objetivo de aprimorar clareza, correção gramatical e organização do texto acadêmico.

REFERÊNCIAS

1. Ueda D, Matsumoto T, Ehara S, Yamamoto A, Walston SL, Ito A, et al. Artificial intelligence-based model to classify cardiac functions from chest radiographs: a multi-institutional, retrospective model development and validation study. Lancet Digit Health. 2023;5(8):e525-e533. doi: https://doi.org/10.1016/S2589-7500(23)00107-3
2. Wang Y, Bao C, Wang Z, Shi Y, Yang J. CR-MSNet: a dual-branch multi-scale attention network for multi-label chest X-ray classification. Sci Rep. 2026;16(1):14585. doi: https://doi.org/10.1038/s41598-026-44591-5
3. Bruno MA, Walker EA, Abujudeh HH. Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction. Radiographics. 2015;35(6):1668-76. doi: https://doi.org/10.1148/rg.2015150023
4. Lamoureux C, Hanna TN, Sprecher D, Weber S, Callaway E. Radiologist errors by modality, anatomic region, and pathology for 1.6 million exams: what we have learned. Emerg Radiol. 2021;28(6):1135-41. doi: https://doi.org/10.1007/s10140-021-01959-6

5. Weiss J, Raghu VK, Paruchuri K, Zinzuwadia A, Natarajan P, Aerts HJWL, et al. Deep Learning to Estimate Cardiovascular Risk From Chest Radiographs : A Risk Prediction Study. Ann Intern Med. 2024;177(4):409-17. doi: https://doi.org/10.7326/M23-1898
6. Pyrros A, Borstelmann SM, Mantravadi R, Zaiman Z, Thomas K, Price B, et al. Opportunistic detection of type 2 diabetes using deep learning from frontal chest radiographs. Nat Commun. 2023;14(1):4039. doi: https://doi.org/10.1038/s41467-023-39631-x
7. Reginster JY, Schmidmaier R, Alokail M, Hiligsmann M. Cost-effectiveness of opportunistic osteoporosis screening using chest radiographs with deep learning in Germany. Aging Clin Exp Res. 2025;37(1):149. doi: https://doi.org/10.1007/s40520-025-03048-x
8. Zhou D, Gao S, Huang X. Strategies for Class-Imbalanced Learning in Multi-Sensor Medical Imaging. Sensors (Basel). 2026;26(6):1998. doi: https://doi.org/10.3390/s26061998
9. Nguyen-Mau TH, Huynh TL, Le TD, Nguyen HD, Tran MT. Advanced Augmentation and Ensemble Approaches for Classifying Long-Tailed Multi-Label Chest X-Rays. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). [s.l.]: IEEE; 2023. p. 2721-30. doi: https://doi.org/10.1109/iccvw60793.2023.00288
10. Ranjan DN, Balkhande DB, Tembhare L, More A, Mahajan S, Razakova R, et al. Addressing diagnostic resource imbalance in pulmonary tuberculosis detection from chest radiographs through cost-aware learning. Indian J Tuberc. 2025;72 Suppl 3:S69-S76. doi: https://doi.org/10.1016/j.ijtb.2025.10.011
11. Jin Y, Lu H, Zhu W, Huo W. Deep learning based classification of multi-label chest X-ray images via dual-weighted metric loss. Comput Biol Med. 2023;157:106683. doi: https://doi.org/10.1016/j.compbiomed.2023.106683
12. Farias LGM, Siqueira ACC, Ribas FM, Dos Santos KFA. Inteligência artificial na radiologia do SUS: panorama atual, barreiras e estratégias de integração. BioSCIENCE. 2025;83(S2):e00014. doi: https://doi.org/10.55684/2025.83.s2.e00014
13. Liu Z, Cheng Y, Tamura S. Multi-Label Local to Global Learning: A Novel Learning Paradigm for Chest X-Ray Abnormality Classification. IEEE J Biomed Health Inform. 2023;27(9):4409-20. doi: https://doi.org/10.1109/JBHI.2023.3281466
14. Shelmerdine SC, Togher D, Rickaby S, Dean G. Artificial intelligence (AI) implementation within the National Health Service (NHS): the South West London AI Working Group experience. Clin Radiol. 2024;79(9):665-72. doi: https://doi.org/10.1016/j.crad.2024.05.018
15. Soares TR, Oliveira RD, Liu YE, Santos ADS, Santos PCPD, Monte LRS, et al. Evaluation of chest X-ray with automated interpretation algorithms for mass tuberculosis screening in prisons: a cross-sectional study. Lancet Reg Health Am. 2023;17:100388. doi: https://doi.org/10.1016/j.lana.2022.100388
16. Park W, Park I, Kim S, Ryu J. Robust Asymmetric Loss for Multi-Label Long-Tailed Learning. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). [s.l.]: IEEE; 2023. p. 2703-12. doi: https://doi.org/10.1109/iccvw60793.2023.00286
17. Anh DM, Hung PD. Multi-label Long-Tailed Disease Recognition on Chest X-ray Images. In: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland; 2025. p. 231-40. doi: https://doi.org/10.1007/978-3-031-81596-6_21
18. Lin YC, Chen YS. Weighted Stratification in Multi-label Contrastive Learning for Long-Tailed Medical Image Classification. In: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland; 2026. p. 677-87. doi: https://doi.org/10.1007/978-3-032-05169-1_65
19. Pedreira G, Fernandes H, Zanchetta M. FRCS-Net: Superando a Cauda Longa em Radiografias de Tórax Via Aprendizado em Dois Estágios e Ranking Sensível ao Custo. In: Anais do XXVI Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2026). [s.l.]: Sociedade Brasileira de Computação - SBC; 2026. p. 37-48. doi: https://doi.org/10.5753/sbcas.2026.20305
20. Gomes GM, Da Silva KAL, Soares F, De Fonseca AU, Fernandes D. Uma visão sobre a classificação de pneumonia viral e bacteriana por radiografias de tórax. J Health Inform. 2024;16(Especial). doi: https://doi.org/10.59681/2175-4411.v16.iespecial.2024.1341
21. Hanif MS, Bilal M, Alsaggaf AH, Al-Saggaf UM. Enhancing Multi-Label Chest X-Ray Classification Using an Improved Ranking Loss. Bioengineering (Basel). 2025;12(6):593. doi: https://doi.org/10.3390/bioengineering12060593

NOTAS EDITORIAIS

O Journal of Health Informatics é uma publicação científica mantida pela Sociedade Brasileira de Informática em Saúde, associação científica sem fins lucrativos dedicada ao desenvolvimento da Informática em Saúde e da Saúde Digital no Brasil. O processo editorial é conduzido com independência científica, sem interferência institucional, comercial ou de patrocinadores, e segue as boas práticas internacionais de ética em publicação científica.

Este artigo foi avaliado por pares, em sistema duplo-cego, por especialistas da área. A equipe editorial usou ferramentas de inteligência artificial, com supervisão humana, na triagem inicial, na conferência do texto, das referências e dos números e, quando necessário, no redesenho de figuras, quadros e tabelas, aprovado pelos autores na prova. As ferramentas e as salvaguardas estão na Política de uso de inteligência artificial da revista. As decisões editoriais foram tomadas por pessoas.

SBIS Sociedade Brasileira de Informática em Saúde