FRCS-Net: Superando a Cauda Longa em Radiografias de Tórax Via Aprendizado em Dois Estágios e Ranking Sensível ao Custo

Gustavo Pedreira[1], **Henrique Fernandes**[1], **Marcelo Zanchetta**[1]

[1]Faculty of Computing – Federal University of Uberlândia (UFU)
Postgraduate Program in Computer Science (PPGCO)
Uberlândia – MG – Brazil

{gustavo.pedreira, henrique.fernandes, marcelo.nascimento}@ufu.br

Abstract. *Chest X-ray screening is fundamental, but severe class imbalance in public datasets hampers the detection of rare pathologies. Cross-entropy-based approaches ignore the long tail, while resampling techniques introduce noise in multi-label scenarios. This work proposes the Focal-Ranking Cost-Sensitive (FRCS) Loss, a hybrid loss function designed to prioritize clinical severity. We employed a DenseNet-121 architecture in a two-stage training protocol (stabilization via asymmetric loss and ranking refinement), validated on a curated subset of the NIH ChestX-ray14 dataset. Experiments demonstrate that the method outperforms the baseline, raising the F1-Score of the hernia class (the rarest) from* 0.00 *to* 0.63*, without significant degradation in global specificity. The approach confirms that imposing pairwise ranking constraints is superior to traditional classification for ensuring clinical safety in long-tailed diagnoses.*

Resumo. *A radiografia de tórax é fundamental para triagem, mas o desbalanceamento severo de classes em bases de dados públicas prejudica a detecção de patologias raras. Abordagens baseadas em entropia cruzada ignoram a cauda longa, enquanto técnicas de reamostragem introduzem ruído em cenários multirrótulo. Este trabalho propõe a Focal-Ranking Cost-Sensitive (FRCS) Loss, uma função de perda híbrida desenhada para priorizar a severidade clínica. Utilizou-se uma arquitetura DenseNet-121 em protocolo de treinamento de dois estágios (estabilização via perda assimétrica e refinamento de ordenação), validado em um subconjunto curado da base NIH ChestX-ray14. Os experimentos demonstram que o método supera o modelo base, elevando o F1-Score da classe hérnia (a mais rara) de* $0{,}00$ *para* $0{,}63$*, sem degradação significativa na especificidade global. A abordagem confirma que a imposição de restrições de ordenação por pares é superior à classificação tradicional para garantir a segurança clínica em diagnósticos de cauda longa.*

1. Introdução

A radiografia de tórax ou Raio-X de Tórax (RXT) é o exame de imagem mais solicitado e realizado em escala global, sendo fundamental na prática clínica por sua ampla disponibilidade, custo-benefício e capacidade de representar diversas condições cardiopulmonares [Gefter et al. 2023, Gefter and Hatabu 2023]. Estima-se bilhões de procedimentos anuais deste exame [Tanno et al. 2025, Prinster et al. 2024]. No Brasil, a relevância

é evidenciada pela alta incidência do câncer de pulmão que, segundo dados do INCA [Instituto Nacional de Câncer 2022], registra estimativas de 18.020 novos casos em homens e 14.540 em mulheres para o triênio 2023-2025, configurando-se como a principal causa de morte por câncer entre o público masculino.

Contudo, a interpretação desses exames por radiologistas é uma tarefa complexa, sujeita a vieses cognitivos, como a satisfação de busca, e à alta variabilidade interobservador, fatores que contribuem diretamente para a ocorrência de erros diagnósticos [Lamoureux et al. 2021, Prinster et al. 2024]. Estudos como o de Lamoureux demonstram taxas de discrepância chegando a 30% em exames anormais. Estima-se que 90% dos erros diagnósticos de câncer de pulmão ocorram especificamente em RXT e tais falhas são predominantemente perceptivas, causadas pela complexidade anatômica e por "zonas cegas", como regiões retrocardíacas e hilos, que obscurecem lesões de baixa conspicuidade [Ciello et al. 2017, Gefter et al. 2023]. Adicionalmente, a consistência diagnóstica é prejudicada por inconsistências intraobservador motivadas por fadiga, onde o foco em uma patologia óbvia leva à negligência de achados secundários e raros [Ciello et al. 2017, Gefter et al. 2023, Prinster et al. 2024].

Esse cenário desafiador tem impulsionado o desenvolvimento de sistemas de auxílio à detecção e ao diagnóstico por computador (CAD/CADx) baseados em aprendizado profundo (AP) [Gonçalves et al. 2024], visando mitigar a carga de trabalho e reduzir a taxa de falhas na detecção de patologias [Tanno et al. 2025]. Apesar do progresso recente, a aplicação prática desses sistemas enfrenta um obstáculo crítico: a natureza desbalanceada e multirrótulo dos dados clínicos, onde uma pequena parcela de doenças comuns domina os conjuntos de treinamento, enquanto patologias raras, porém clinicamente severas, residem na "cauda longa" da distribuição [Park et al. 2023, Hanif et al. 2025]. O fenômeno de cauda longa consiste em um desbalanceamento severo onde poucas classes (cabeça) dominam os dados, enquanto a escassa representatividade da maioria das patologias (cauda) degrada a convergência do modelo para achados raros [Nguyen-Mau et al. 2023].

A problematização central deste trabalho reside no fato de que modelos treinados com funções de perda convencionais, como a entropia cruzada binária (ECB), tendem a exibir um viés acentuado em favor das classes majoritárias. Em cenários multirrótulo, esse fenômeno é exacerbado pela coocorrência de doenças, onde a presença de múltiplas etiquetas em uma única imagem dilui o gradiente de representação das patologias minoritárias [Liu et al. 2023]. Clinicamente, essa falha algorítmica traduz-se em taxas inaceitáveis de falsos negativos (FN) para condições como hérnia, pneumotórax ou nódulos, cuja detecção tardia compromete o prognóstico do paciente [Ranjan et al. 2025, Tyagi et al. 2022].

Diante deste cenário, o presente trabalho propõe superar o desbalanceamento de classes através de uma abordagem fundamentada na eficiência algorítmica e no aprendizado em estágios. Para atingir esse objetivo foi proposto a *Focal-Ranking Cost-Sensitive* (FRCS) *Loss*, uma formulação que integra perdas de *ranking* com pesos de custo, reduzindo FN em classes de cauda longa sem a necessidade de reamostragem ruidosa. O protocolo de treinamento foi estruturado em dois estágios *curriculum learning* que estabiliza a extração de características com uma *Weighted Asymmetric Loss* (WASL) e, posteriormente, refina a fronteira de decisão com o componente de ranking. A avaliação

desta formulação foi centrada na segurança clínica e separabilidade e não somente em métricas globais, analisando o F1-Score e o Recall nas classes minoritárias para garantir o equilíbrio entre sensibilidade e precisão. Adicionalmente, apresenta-se uma análise qualitativa das densidades de probabilidade (histogramas de predição) no intuito de avaliar se a FRCS Loss desloca efetivamente a distribuição de scores das patologias raras para regiões de maior confiança.

2. Trabalhos relacionados

Segundo [Park et al. 2023] e [Hanif et al. 2025], a distribuição de patologias em bases radiológicas é severamente desigual, onde achados frequentes ofuscam doenças raras localizadas na "cauda" da distribuição. Além da disparidade quantitativa, a natureza multirrótulo impõe desafios de coocorrência, nos quais a presença de múltiplas doenças em uma mesma imagem dilui o gradiente de representação das classes minoritárias [Lin and Chen 2026]. Historicamente, o desbalanceamento de classes em RXT tem sido abordado via técnicas de reamostragem ou estratégias de aumento de dados [Tsaniya et al. 2023, Marques and Machado 2023]. Contudo, a literatura recente aponta limitações severas desta estratégia em cenários multirrótulo. Conforme argumentado por [Park et al. 2023] e por [Nguyen-Mau et al. 2023], métodos de *oversampling* são inerentemente ambíguos quando múltiplas patologias coexistem na mesma imagem; a manipulação artificial de uma classe da "cauda" impacta inevitavelmente a distribuição das classes "cabeça", distorcendo a probabilidade conjunta. Com isso, as principais contribuições para esse problema vêm sendo tratada nos trabalhos de [Tsaniya et al. 2023, Ridnik et al. 2021, Park et al. 2023, Hanif et al. 2025, Tyagi et al. 2022, Ranjan et al. 2025, Kim 2023, Li et al. 2024, Hu et al. 2024, Sujay et al. 2024, Liu et al. 2023, Lai et al. 2024, Chen et al. 2023, Vimala et al. 2024].

No trabalho apresentado por [Tsaniya et al. 2023] os autores demonstram empiricamente que, embora abordagens como o MLSMOTE possam elevar o *recall* de patologias minoritárias, elas frequentemente degradam a precisão global ao introduzir ruído anatômico e artefatos de borda, o que tem direcionado o estado da arte para soluções focadas na otimização algorítmica. Os trabalhos de [Tsaniya et al. 2023] e [Nguyen-Mau et al. 2023] argumentam que técnicas de reamostragem são ambíguas em cenários multirrótulo, pois a coexistência de etiquetas de diferentes frequências em uma única instância torna a sobreamostragem conflitante.

Consequentemente, a evolução das funções de perda tem buscado desacoplar a dificuldade do exemplo da frequência da classe. A *Asymmetric Loss* (ASL), formalizada por [Ridnik et al. 2021] e aplicada em RXT por [Park et al. 2023], representa um avanço significativo sobre a *Focal Loss* tradicional ao permitir o controle independente dos gradientes positivos e negativos, mitigando a supressão causada pela vasta quantidade de negativos fáceis.

[Ranjan et al. 2025] e [Vimala et al. 2024] defendem o aprendizado sensível ao custo (*cost-sensitive*), modulando a penalidade do erro conforme a raridade da patologia. Avançando neste paradigma, [Hanif et al. 2025] propuseram a *Focal ZLPR*, que introduz noções de *ranking* para priorizar o aprendizado de patologias raras. Entretanto, abordagens puramente estatísticas muitas vezes falham em capturar a urgência médica. [Ranjan et al. 2025] e [Tyagi et al. 2022] exploraram o *cost-sensitive learning* através de

pesos de frequência inversa, estratégia que, embora válida, carece de uma integração direta com a otimização da margem de separação entre classes críticas e benignas.

No âmbito arquitetural, observa-se uma dicotomia entre complexidade e eficiência. Modelos híbridos massivos, como o CheXFusion [Kim 2023] e integrações CNN-Transformer [Li et al. 2024, Hu et al. 2024], atingem o desempenho do estado da arte através da fusão de múltiplas vistas e mecanismos de atenção cruzada. Contudo, estas arquiteturas impõem uma infraestrutura computacional onerosa. Em contrapartida, trabalhos recentes como o de [Sujay et al. 2024] sugerem que *backbones* eficientes (como DenseNets), quando acoplados a estratégias de treinamento robustas, podem oferecer um balanço superior entre custo computacional e acurácia clínica. Por fim, consolidam-se estratégias de aprendizado progressivo para organizar o espaço latente. [Liu et al. 2023] inovam com o aprendizado por currículo (ML-LGL), demonstrando que o escalonamento da dificuldade estabiliza a convergência de funções de perda complexas. [Chen et al. 2023] utilizam aprendizado métrico para compactar representações, enquanto [Lai et al. 2024] abordam a robustez a ruídos de anotação.

3. Metodologia

A presente metodologia descreve o arcabouço experimental delineado para validar a eficácia da *Focal-Ranking Cost-Sensitive Loss* (FRCS) em cenários de desbalanceamento severo. A estratégia fundamenta-se na premissa de que a intervenção na função de perda deve ser validada em um ambiente controlado, livre de ruídos de anotação externos, mas preservando a distribuição de cauda longa natural das patologias.

3.1. Construção do *dataset* e processamento de imagem

Embora o *dataset* NIH ChestX-ray14 compreenda originalmente 112.120 imagens e 14 patologias, este estudo adotou um protocolo de curadoria para isolar o fenômeno do desbalanceamento. Foi construído um subconjunto focado composto por 5 patologias de interesse — hérnia, enfisema, massa, efusão e pneumotórax — além da classe de controle (sem achados). Este subconjunto foi formado com 23.635 imagens, conforme detalhamento na tabela 1. Esta seleção foi estratégica para incluir tanto classes da "cabeça" (efusão, $\sim 11\%$) quanto da "cauda extrema" (hérnia, $< 0{,}5\%$), desafiando a capacidade de ordenação do modelo.

O processo de reamostragem seguiu três etapas críticas de controle de qualidade. A primeira etapa realizou a filtragem de imagens que continham comorbidades fora do escopo do estudo (ex: cardiomegalia associada a efusão), garantindo que os gradientes de erro fossem provenientes exclusivamente das classes monitoradas. Segundo, optou-se por manter a proporção original de "sem achados", aproximadamente 54%, no subconjunto, preservando a prevalência natural das doenças raras. Isso assegura que o modelo seja testado em um cenário realista de triagem, onde a maioria dos exames é normal. Por fim, para evitar a possibilidade de vazamento de dados, a divisão dos conjuntos de treino (70%), validação (10%) e teste (20%) foi realizada estritamente por ID de Paciente. Isso impede que radiografias seriadas do mesmo indivíduo apareçam simultaneamente no treino e no teste, o que inflaria artificialmente as métricas de performance [Zhang et al. 2025].

O *pipeline* de pré-processamento adotou uma estratégia de redimensionamento seguida de recorte central para 512×512 pixels. Esta abordagem garante a uniformidade dos dados de entrada sem comprometer a proporção anatômica original.

Para mitigar o sobreajuste e aumentar a invariância do modelo a fatores de aquisição, aplicou-se um protocolo de aumento de dados estocástico em tempo de execução. As transformações incluíram: CLAHE (*Contrast Limited Adaptive Histogram Equalization*) com *clip limit* 2,0 para realce de estruturas anatômicas; rotação aleatória ($\pm 15^{\circ}$); deslocamentos horizontais e verticais de até 10%; e espelhamento horizontal. A distribuição final das classes no subconjunto curado, juntamente com seus respectivos pesos de severidade (ω_c), é detalhada na Tabela 1. Ressalta-se que a estrutura de cauda longa foi preservada para validar a eficácia da função de perda proposta, com a classe minoritária (hérnia) apresentando uma frequência aproximadamente 50 vezes menor que a classe majoritária positiva (efusão).

A partição dos dados seguiu a proporção 70/10/20 (treinamento: 16.544; validação: 2.364; e teste: 4.727 imagens), respeitando o isolamento por ID de paciente. Ressalta-se que as técnicas de aumento de dados foram aplicadas exclusivamente durante a fase de treinamento, mantendo os conjuntos de validação e teste inalterados para assegurar a fidelidade da avaliação clínica.

Tabela 1. Distribuição de classes e pesos de severidade no dataset curado.

Classe	**Qtd. Imagens**	**Frequência (%)**	**Peso (w_c)**
Hérnia	134	0,57	3,00
Enfisema	1.481	6,27	2,25
Massa	2.868	12,13	1,50
Efusão	4.945	20,92	1,50
Pneumotórax	3.218	13,62	2,25
Sem Achados	12.725	53,84	1,00
Total	**23.635**	**100,00**	–

3.2. Arquitetura do modelo e estratégia de treinamento

A arquitetura fundamenta-se na DenseNet-121, pré-treinada na *ImageNet*, selecionada por sua robustez no reuso de características e eficiência de parâmetros, vitais para evitar *overfitting* em patologias com poucas amostras. A camada de classificação original foi substituída por uma projeção linear dimensionada para as 6 classes alvo.

O protocolo de otimização inova ao adotar uma abordagem de Aprendizado em Dois Estágios, ilustrada na Figura 1. No Estágio 1 (Estabilização), a rede é treinada com a WASL partindo de uma taxa de aprendizado base de 10^{-4}. Diferente de abordagens estáticas, utiliza-se um escalonador dinâmico que decai a taxa ao detectar estagnação na métrica de validação, priorizando a convergência robusta da representação latente. No Estágio 2 (Refinamento de *Ranking*), realiza-se o ajuste fino com uma taxa inicial conservadora (10^{-5}), introduzindo a FRCS Loss completa. Neste passo, a restrição *pairwise* penaliza geometricamente a inversão de prioridades entre classes severas e benignas, forçando o reordenamento na cauda da distribuição enquanto o escalonador garante a otimização precisa dos parâmetros da rede.

3.3. Função de perda e algoritmos

A FRCS *Loss* foi empregada no estágio de refinamento. Esta função híbrida supera a otimização tradicional ao somar a robustez da classificação assimétrica com uma restrição geométrica de ordenação por pares. Essa função é expressa por:

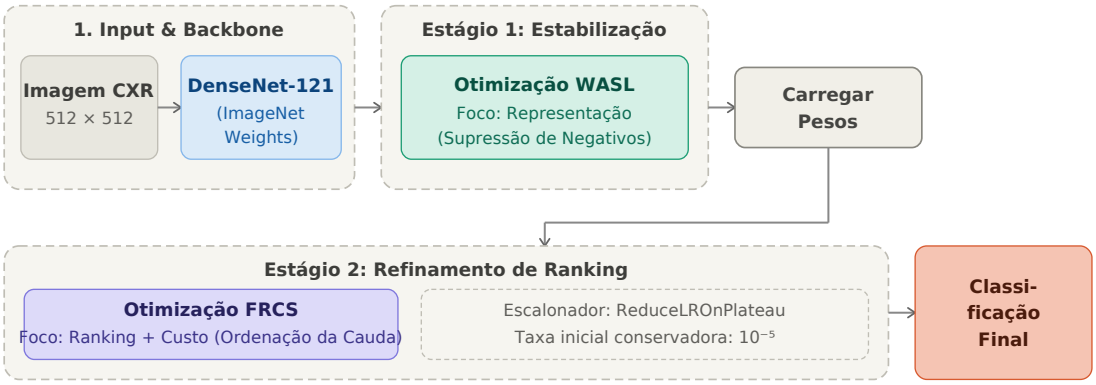

Figura 1. Esquema do protocolo com escalonamento dinâmico.

$$L_{FRCS} = L_{WASL}(W_c) + \lambda \cdot \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \text{softplus}\left(\frac{z_j - z_i}{\tau}\right) \tag{1}$$

Onde:

- L_{FRCS}: Função de perda global.
- L_{WASL}: Componente de perda assimétrica ponderada.
- W_c: Matriz ou vetor de pesos de custo.
- λ: Hiperparâmetro escalar de regularização que pondera a contribuição do termo de ranking na perda total.
- $|\mathcal{P}|$: Cardinalidade do conjunto $\mathcal{P}$, utilizada como fator de normalização para garantir a invariância da perda em relação ao tamanho do lote (*batch size*).
- $\mathcal{P}$: Conjunto de pares de índices (i, j) válidos amostrados, onde i denota uma classe positiva e j uma classe negativa.
- softplus$(\cdot)$: Função de ativação suave e convexa.
- z_j: *Logit* (pontuação bruta pré-ativação) predito pela rede para a classe negativa j.
- z_i: *Logit* (pontuação bruta pré-ativação) predito pela rede para a classe positiva i.
- τ: Parâmetro de temperatura (*temperature scaling*).

O componente assimétrico ponderado (L_{WASL}) é uma extensão da *Asymmetric Loss* (ASL) [Ridnik et al. 2021] que integra pesos de severidade estáticos (w_c) para modular a magnitude do gradiente. Ela permite o tratamento independente de amostras positivas e negativas, suprimindo o sinal de negativos fáceis enquanto preserva a contribuição de patologias raras através de parâmetros de focalização distintos (γ_- e γ_+) [Park et al. 2023]. Fixou-se $\gamma_- = 2$ para suprimir o gradiente proveniente dos numerosos "negativos fáceis" (fundo da imagem), enquanto $\gamma_+ = 0$ preserva o sinal integral das patologias raras, impedindo que o modelo ignore classes da cauda longa. Para o componente de sensibilidade ao custo (W_c), em substituição a pesos dinâmicos instáveis, integra-se um vetor de severidade estático W_c [Zhang et al. 2025]. Classes críticas como hérnia e pneumotórax recebem multiplicadores elevados ($3{,}0\times$ e $2{,}25\times$), penalizando a omissão diagnóstica diretamente na magnitude do gradiente da L_{WASL}. Os pesos de severidade (w_c) foram estabelecidos seguindo uma abordagem heurística sensível ao custo,

que pondera dois fatores: (i) a criticidade clínica definida pelos critérios de adequação do American College of Radiology (ACR), onde patologias com maior risco de mortalidade recebem maior penalização; e (ii) a frequência relativa na amostra, aplicando o inverso da frequência para compensar a escassez de dados da cauda longa. Os valores finais resultantes desta composição estão detalhados na Tabela 1. Por fim, o componente de *ranking* ($L_{Ranking}$) buscou otimizar a separabilidade através de comparações *pairwise*. A função penaliza pares onde o *logit* de uma classe negativa (z_j) supera o de uma positiva (z_i). A função de ativação *Softplus* suaviza a margem, e a temperatura τ calibra a sensibilidade a diferenças sutis de pontuação.

O algoritmo de otimização selecionado foi o AdamW, devido à sua eficiência no desacoplamento do decaimento de peso (*weight decay*). O controle da taxa de aprendizado utiliza a estratégia *ReduceLROnPlateau*, que monitora a AUC de validação e reduz a taxa (fator $0{,}1$) mediante estagnação, garantindo um ajuste fino preciso nos mínimos locais da superfície de erro. Esses valores foram definidos empiricamente.

3.4. Protocolo de treinamento e avaliação

O protocolo experimental foi delineado como um estudo comparativo entre paradigmas (modelo base treinado com ECB *versus* a abordagem proposta em dois estágios: estabilização com WASL e refinamento com FRCS). Com o intuito de mitigar o viés de inicialização de pesos e variações estocásticas na amostragem de mini-lotes, todos os experimentos foram executados de forma independente em cinco rodadas (*runs*), utilizando sementes aleatórias distintas (42, 123, 7, 2024 e 99). Ressalta-se que a partição dos dados (treino, validação e teste) foi fixada através de uma semente global estática para assegurar que todos os modelos fossem treinados e avaliados exatamente sob o mesmo subconjunto de pacientes, isolando assim exclusivamente o efeito da otimização algorítmica.

O ambiente computacional baseou-se no *framework* PyTorch, operando em uma GPU NVIDIA A100 (40 GB). Para maximizar a eficiência de memória sem comprometer a estabilidade dos gradientes, adotou-se o treinamento com precisão mista automática (AMP - *Automatic Mixed Precision*) aliado à técnica de acúmulo de gradientes (*gradient accumulation*). Especificamente, utilizou-se um tamanho de lote físico de 16 imagens com 4 passos de acúmulo, resultando em um lote efetivo de 64 instâncias. A otimização em todas as fases foi conduzida pelo algoritmo AdamW, acompanhado pelo escalonador *ReduceLROnPlateau* para o ajuste fino dinâmico da taxa de aprendizado.

A avaliação de desempenho das abordagens transcende métricas globais de acurácia, focando em três dimensões críticas para cenários de cauda longa: separabilidade global, sensibilidade na cauda e qualidade de *ranking*. Consequentemente, os resultados passam a ser reportados em termos de Média $\pm$ Desvio Padrão ao longo das cinco execuções para as seguintes métricas: AUC-ROC *Macro* (utilizada como métrica primária para monitoramento de convergência), *F1-Score*, *Recall* (Sensibilidade) específico por classe e a Sensibilidade no ponto operacional de 95% de Especificidade (Sens@95%Spec) — uma exigência clínica vital para minimizar falsos alarmes em sistemas de triagem autônoma. Adicionalmente, manteve-se a análise qualitativa das distribuições de probabilidade predita por meio de estimativa de densidade de *kernel* (KDE Histograms), a fim de constatar visualmente a ampliação da margem de separabilidade e a mitigação da indecisão do modelo estruturada pela FRCS Loss.

4. Resultados e discussão

Os experimentos foram conduzidos comparando-se o modelo base (DenseNet-121 treinada com BCE) com os dois estágios do método proposto: a estabilização via WASL e o refinamento final via FRCS. Para assegurar a validade estatística das conclusões frente a variações estocásticas de inicialização, todos os resultados reportados refletem a Média $\pm$ Desvio Padrão consolidados a partir de cinco execuções (*runs*) independentes no conjunto de teste.

A Tabela 2 sumariza as métricas obtidas. Observa-se que, embora a AUC *Macro* global apresente uma estabilidade estatística entre as abordagens ($0{,}8800 \pm 0{,}0098$ no Baseline *vs.* $0{,}8834 \pm 0{,}0081$ na FRCS), o impacto clínico do método proposto revela-se de forma acentuada nas métricas de decisão (F1-Score e Recall).

Tabela 2. Comparação de desempenho consolidado (5 *runs*) por patologia. Valores representam Média $\pm$ Desvio Padrão. O melhor desempenho em F1-Score para as classes minoritárias está destacado.

Patologia	AUC (Base)	AUC (FRCS)	F1 (Base)	F1 (WASL)	F1 (FRCS)
Hérnia ($< 0{,}5\%$)	$0{,}9229 \pm 0{,}0277$	$\mathbf{0{,}9444 \pm 0{,}0161}$	$0{,}0000 \pm 0{,}0000$	$0{,}6136 \pm 0{,}1097$	$\mathbf{0{,}6332 \pm 0{,}1400}$
Enfisema ($\sim 6\%$)	$0{,}9246 \pm 0{,}0021$	$0{,}9162 \pm 0{,}0097$	$0{,}5029 \pm 0{,}0293$	$0{,}5800 \pm 0{,}0134$	$\mathbf{0{,}5827 \pm 0{,}0136}$
Massa ($\sim 12\%$)	$0{,}8281 \pm 0{,}0045$	$\mathbf{0{,}8325 \pm 0{,}0034}$	$0{,}4991 \pm 0{,}0394$	$0{,}5349 \pm 0{,}0152$	$\mathbf{0{,}5496 \pm 0{,}0058}$
Pneumotórax ($\sim 13\%$)	$0{,}8588 \pm 0{,}0105$	$\mathbf{0{,}8688 \pm 0{,}0059}$	$0{,}4442 \pm 0{,}0853$	$0{,}5442 \pm 0{,}0370$	$\mathbf{0{,}5784 \pm 0{,}0195}$
Efusão ($\sim 21\%$)	$\mathbf{0{,}8726 \pm 0{,}0080}$	$0{,}8707 \pm 0{,}0076$	$0{,}6325 \pm 0{,}0291$	$0{,}6458 \pm 0{,}0051$	$\mathbf{0{,}6472 \pm 0{,}0082}$
Sem Achados ($\sim 54\%$)	$\mathbf{0{,}8729 \pm 0{,}0063}$	$0{,}8676 \pm 0{,}0057$	$\mathbf{0{,}8156 \pm 0{,}0069}$	$0{,}8048 \pm 0{,}0053$	$0{,}7999 \pm 0{,}0081$
Média (Macro)	$0{,}8800 \pm 0{,}0098$	$\mathbf{0{,}8834 \pm 0{,}0081}$	$0{,}4824 \pm 0{,}0317$	$0{,}6206 \pm 0{,}0309$	$\mathbf{0{,}6318 \pm 0{,}0325}$

O modelo *Baseline* demonstrou o viés clássico do AP em dados médicos não tratados: maximizou a performance na classe majoritária (*No Finding*, F1 $0{,}8156$), mas sofreu um colapso na classe mais rara (Hérnia), falhando em predizer verdadeiros positivos, o que resultou em um F1-Score nulo ($0{,}0000 \pm 0{,}0000$). A introdução do protocolo de dois estágios reverteu drasticamente este cenário. Apenas com a fase de estabilização (WASL), o F1-Score da Hérnia atingiu $0{,}6136$, saltando para $0{,}6332 \pm 0{,}1400$ com o refinamento FRCS. Esse ganho excepcional foi alcançado com uma penalização marginal estatisticamente insignificante na classe de controle ($0{,}8156 \rightarrow 0{,}7999$), comprovando que a imposição de restrições de *ranking pairwise* induz o modelo a respeitar a severidade clínica independentemente da escassez estatística.

Para aplicações de triagem clínica autônoma, a métrica de Sensibilidade no limite de 95% de Especificidade (*Sens@95%Spec*) é mandatória, pois traduz a capacidade do CADx de reter doentes sem gerar falsos alarmes em grande quantidade. Nesta métrica, o modelo FRCS demonstrou ser superior. Na classe Hérnia, a sensibilidade operacional subiu de $0{,}6960 \pm 0{,}1043$ no *Baseline* para $0{,}8320 \pm 0{,}0438$ (FRCS). Isso significa que o modelo proposto é capaz de detectar mais de 83% dos casos da patologia mais extrema da cauda longa gerando apenas 5% de falsos positivos sistêmicos.

A eficácia do componente de otimização estrutural da FRCS é visualmente corroborada pela análise de densidade de probabilidade (KDE), ilustrada na Figura 2. No modelo *Baseline* (linha pontilhada cinza), a distribuição dos *scores* preditos para patologias minoritárias apresentava supressão severa, concentrando as probabilidades em zonas de baixa confiança ($p < 0{,}2$), resultando em falsos negativos contínuos. A estabilização via WASL (linha tracejada azul) inicia o tratamento do gradiente e a FRCS (linha sólida vermelha) que consolida o deslocamento e a ancoragem das curvas de densidade para

a direita ($p > 0{,}8$). O modelo passa a gerar um espaço onde a decisão diagnóstica para doenças críticas ocorre em um regime de maior confiança, mitigando a indecisão na cauda da distribuição.

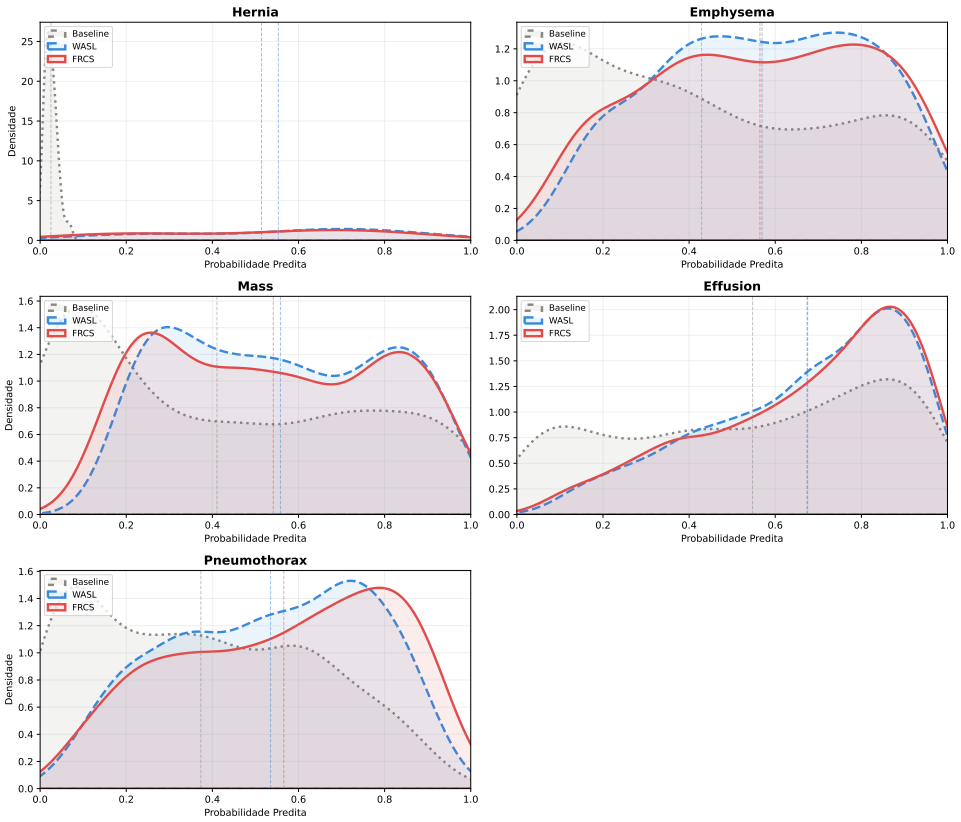

Figura 2. Comparação das densidades de probabilidade (KDE) entre Baseline e FRCS.

5. Conclusão

Este trabalho apresentou uma abordagem algorítmica robusta para mitigar o desbalanceamento de cauda longa em RXT, rejeitando soluções baseadas em manipulação de dados em favor de uma otimização consciente do custo clínico. Os resultados experimentais confirmam que a arquitetura proposta, culminando na aplicação da *FRCS Loss*, é superior ao treinamento convencional com ECB. A principal evidência reside no ganho expressivo de sensibilidade para a classe Hérnia (a mais rara do conjunto), onde o *F1-Score* atingiu 0,63 em relação aos valores nulos do *baseline*, demonstrando que a imposição de restrições de *ranking pairwise* obriga o modelo a respeitar a severidade da patologia, independentemente de sua frequência estatística.

Entretanto, o estudo apresenta limitações que devem ser consideradas. Primeiramente, a validação restringiu-se a um subconjunto curado de 5 patologias e uma classe de controle, em vez das 14 classes originais do *dataset* NIH. Embora essa redução tenha sido metodologicamente estratégica para isolar o fenômeno da cauda longa livre de ruídos de anotação excessivos, a escalabilidade da *FRCS Loss* para o espectro completo de doenças torácicas, onde a coocorrência de múltiplas anomalias raras é frequente, carece de verificação empírica. Em segundo lugar, a Matriz de Severidade Clínica utilizada

foi estática, baseada em consenso pré-definido, o que não captura a subjetividade ou a incerteza inerente a casos limítrofes.

Como direções para trabalhos futuros, sugere-se a expansão do protocolo experimental para validação cruzada em bases de dados externas, a fim de testar a robustez do *ranking* diante de variações de equipamento e protocolos de aquisição. Adicionalmente, planeja-se integrar o cálculo de incerteza (*uncertainty estimation*) ao processo de inferência do modelo, permitindo o tratamento de rótulos ambíguos e conferindo maior confiabilidade clínica às predições. Por fim, vislumbra-se a oportunidade de aplicar o algoritmo em bases de dados que, além das patologias torácicas clássicas, contemplem o rastreio oportunístico (*opportunistic screening*), possibilitando a identificação de condições como baixa densidade óssea e doenças cardiovasculares, ampliando assim o impacto diagnóstico da ferramenta proposta.

Referências

Chen, K., Lei, W., Zhao, S., et al. (2023). PCCT: Progressive Class-Center Triplet Loss for Imbalanced Medical Image Classification. *IEEE Journal of Biomedical and Health Informatics*, 27(4):2026–2036. DOI: 10.1109/JBHI.2023.3240136.

Ciello, A. D., Franchi, P., Contegiacomo, A., et al. (2017). Missed lung cancer: when, where, and why? *Diagnostic and Interventional Radiology*, 23(2):118–126. DOI: 10.5152/dir.2016.16187.

Gefter, W. B. and Hatabu, H. (2023). Reducing Errors Resulting From Commonly Missed Chest Radiography Findings. *Chest*, 163(3):634–649. DOI: 10.1016/j.chest.2022.12.003.

Gefter, W. B., Post, B. A., and Hatabu, H. (2023). Commonly Missed Findings on Chest Radiographs. *Chest*, 163(3):650–661. DOI: 10.1016/j.chest.2022.10.039.

Gonçalves, J. V. L., de Souza, D. V., dos Santos, C. I. A. T., do Nascimento, C. E. T., da Cruz, L. B., Junior, D. A. D., and Diniz, J. O. B. (2024). D.IAgnóstica: Ferramenta CADx para diagnóstico de doenças pulmonares em imagens radiológicas. In *Anais do XXIV Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, Porto Alegre, RS, Brasil. SBC.

Hanif, M. S., Bilal, M., Alsaggaf, A. H., et al. (2025). Enhancing Multi-Label Chest X-Ray Classification Using an Improved Ranking Loss. *Bioengineering*, 12(6):593. DOI: 10.3390/bioengineering12060593.

Hu, Z., Hongyu, C., Mei, W., and Wang, Y. (2024). Fetransnet: An Enhanced Lung Disease Classification Approach Combining Efficientnet and Transformer with Adaptive Focal Loss. DOI: 10.2139/ssrn.4884895.

Instituto Nacional de Câncer (2022). Estimativa 2023: incidência de câncer no brasil. Disponível em: https://www.gov.br/inca/pt-br/assuntos/cancer/tipos/pulmao. Acesso em: 18 fev. 2026.

Kim, D. (2023). CheXFusion: Effective Fusion of Multi-View Features using Transformers for Long-Tailed Chest X-Ray Classification. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2694–2702, Paris, France. IEEE. DOI: 10.1109/ICCVW60793.2023.00285.

Lai, H., Yao, Q., He, Z., et al. (2024). Long-Tailed Multi-Label Classification with Noisy Label of Thoracic Diseases from Chest X-Ray. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, Athens, Greece. IEEE. DOI: 10.1109/ISBI56570.2024.10635361.

Lamoureux, C., Hanna, T. N., Sprecher, D., et al. (2021). Radiologist errors by modality, anatomic region, and pathology for 1.6 million exams: what we have learned. *Emergency Radiology*, 28(6):1135–1141. DOI: 10.1007/s10140-021-01959-6.

Li, X., Xu, X., Liu, Y., and Zhao, X. (2024). CheX-DS: Improving Chest X-ray Image Classification with Ensemble Learning Based on DenseNet and Swin Transformer. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 5295–5301, Lisbon, Portugal. IEEE. DOI: 10.1109/BIBM62325.2024.10822262.

Lin, Y.-C. and Chen, Y.-S. (2026). Weighted Stratification in Multi-label Contrastive Learning for Long-Tailed Medical Image Classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pages 677–687, Cham. Springer Nature Switzerland. DOI: 10.1007/978-3-032-05169-1.

Liu, Z., Cheng, Y., and Tamura, S. (2023). Multi-Label Local to Global Learning: A Novel Learning Paradigm for Chest X-Ray Abnormality Classification. *IEEE Journal of Biomedical and Health Informatics*, 27(9):4409–4420. DOI: 10.1109/JBHI.2023.3281466.

Marques, A. G. M. and Machado, A. M. C. (2023). Convolutional neural networks with approximation of Shapley values for the classification and interpretation of pneumonia in X-ray images. In *Anais do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, Porto Alegre, RS, Brasil. SBC.

Nguyen-Mau, T.-H., Huynh, T.-L., Le, T.-D., et al. (2023). 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)*, pages 2721–2730, Paris, France. IEEE. DOI: 10.1109/ICCVW60793.2023.00288.

Park, W., Park, I., Kim, S., et al. (2023). Robust Asymmetric Loss for Multi-Label Long-Tailed Learning. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2703–2712, Paris, France. IEEE. DOI: 10.1109/ICCVW60793.2023.00286.

Prinster, D., Mahmood, A., Saria, S., et al. (2024). Care to Explain? AI Explanation Types Differentially Impact Chest Radiograph Diagnostic Performance and Physician Trust in AI. *Radiology*, 313(2):e233261. DOI: 10.1148/radiol.233261.

Ranjan, N., Balkhande, B., Tembhare, L., et al. (2025). Addressing diagnostic resource imbalance in pulmonary tuberculosis detection from chest radiographs through cost-aware learning. *Indian Journal of Tuberculosis*, 72:S69–S76. DOI: 10.1016/j.ijtb.2025.10.011.

Ridnik, T., Ben-Baruch, E., Zamir, N., et al. (2021). Asymmetric Loss For Multi-Label Classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 82–91, Montreal, QC, Canada. IEEE. DOI: 10.1109/ICCV48922.2021.00015.

Sujay, J. K., Surakshith, D. T., Uday, T. Y., et al. (2024). Hybrid Approach for Handling Class Imbalance on Medical Data. In *2024 International Conference on Data Science and Network Security (ICDSNS)*, pages 1–6, Tiptur, India. IEEE. DOI: 10.1109/ICDSNS62112.2024.10691062.

Tanno, R., Barrett, D. G. T., Sellergren, A., et al. (2025). Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 31(2):599–608. DOI: 10.1038/s41591-024-03302-1.

Tsaniya, H., Fatichah, C., and Suciati, N. (2023). Comparison of sampling methods for handling imbalance data in deep learning-based predictions of chest X-ray abnormality tags. In *2023 the 7th International Conference on Medical and Health Informatics (ICMHI)*, pages 6–10, Kyoto, Japan. ACM. DOI: 10.1145/3608298.3608300.

Tyagi, M., Roy, S., and Bansal, V. (2022). Custom Weighted Balanced Loss function for Covid 19 Detection from an Imbalanced CXR Dataset. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2707–2713, Montreal, QC, Canada. IEEE. DOI: 10.1109/ICPR56361.2022.9956580.

Vimala, A. G., Legapriyadharshini, N., Dhanalakshmi, R., et al. (2024). Cost Sensitive Learning using Chest X-Ray with CNN for Covid-19 Detection with Lung Diseases which Lead to Class Imbalance. In *2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN)*, pages 489–495, Dhulikhel, Nepal. IEEE. DOI: 10.1109/ICIPCN63822.2024.00086.

Zhang, M., Hu, X., Gu, L., et al. (2025). A New Benchmark: Clinical Uncertainty and Severity Aware Labeled Chest X-Ray Images With Multi-Relationship Graph Learning. *IEEE Transactions on Medical Imaging*, 44(1):338–347. DOI: 10.1109/TMI.2024.3441494.