Classificação de Pneumonia em Radiografias Torácicas com Transfer Learning: Análise Clínica Orientada à Sensibilidade e Interpretabilidade

Maria Luísa Brandão de Luna Barros, UFPE (luisa.brandao@ufpe.br).

I. INTRODUÇÃO

A pneumonia é uma das principais causas de mortalidade infantil global, impactando severamente países de baixa renda onde a escassez de radiologistas limita o diagnóstico. Embora a radiografia torácica seja o instrumento primordial para identificar infiltrados, sua interpretação é subjetiva e restrita pela distribuição geográfica de especialistas, tornando os sistemas de visão computacional ferramentas de alto impacto para democratizar o acesso a diagnósticos precisos.

Modelos baseados em **Redes Neurais Convolucionais** (CNNs) alcançam desempenho competitivo ao de especialistas por aprenderem representações visuais hierárquicas diretamente dos dados, eliminando a necessidade de extração manual de características. A técnica de *transfer learning*, utilizando modelos pré-treinados no *ImageNet*, viabiliza o uso dessas arquiteturas em conjuntos de dados médicos moderados, acelerando a convergência e mitigando o risco de sobreajuste.

O modelo *Baseline* adotado é uma **ResNet18** com *transfer learning*, augmentation leve e sem ponderação de classes. A **ResNet18** foi escolhida por três razões complementares: sua arquitetura de conexões residuais (*skip connections*) resolve o problema do gradiente desvanecente, permitindo treinamento estável em redes mais profundas; seu histórico em tarefas de classificação de imagens médicas é amplamente consolidado na literatura; e sua eficiência computacional — menor número de parâmetros em comparação a arquiteturas mais profundas — é favorável para datasets de tamanho moderado, onde modelos mais complexos tendem a sobreajustar. O *Baseline* estabelece o ponto de referência sobre o qual cada hipótese experimental introduz uma única modificação controlada, garantindo que diferenças de desempenho sejam atribuíveis ao fator investigado. A partir dessa fundamentação, foram formuladas três hipóteses experimentais:

Hipótese 1 – Arquitetura (DenseNet121): a **ResNet18** opera por adição residual — cada bloco aprende um resíduo sobre sua entrada. A **DenseNet121**, por sua vez, conecta cada camada a todas as anteriores por concatenação, reutilizando features em múltiplos níveis de abstração simultaneamente. Dado que a pneumonia se manifesta de forma heterogênea — consolidações de variada extensão, infiltrados uni ou bilaterais, opacidades em diferentes lobos —, a hipótese é que essa propagação densa de representações em diferentes escalas confira à **DenseNet121** maior sensibilidade a padrões sutis, superando o *Baseline* em métricas clínicas.

Hipótese 2 – Data Augmentation Intenso: modelos treinados em datasets de tamanho moderado são susceptíveis a sobreajuste às particularidades do conjunto de treino. Estratégias de *augmentation* mais intensas — rotação ±15°, transformações afins e jitter de brilho e contraste — no *Baseline* introduzem variabilidade artificial durante o treinamento, forçando o modelo a aprender representações mais invariantes. A hipótese é que essa regularização implícita melhore a capacidade de generalização para casos não vistos, reduzindo a lacuna entre desempenho de treino e validação.

Hipótese 3 – Ponderação de Classes: o dataset apresenta desbalanceamento entre as classes, com predominância de casos positivos. Em presença de desbalanceamento, o minimizador da função de perda padrão tende a favorecer a classe majoritária, resultando em Recall artificialmente reduzido para a classe Pneumonia. A aplicação de *class weighting* — atribuindo maior penalidade a erros na classe minoritária — no *Baseline* reequilibra esse viés durante o treinamento, com o objetivo de aumentar a sensibilidade diagnóstica e reduzir Falsos Negativos. No contexto clínico, uma pneumonia não detectada representa risco direto à vida do paciente, impondo um custo incomparavelmente maior que um alarme falso — o que torna esta hipótese a de maior relevância clínica do estudo.

Diferentemente de abordagens puramente orientadas a desempenho numérico, este trabalho busca não apenas maximizar a métrica **ROC-AUC**, mas também avaliar criticamente o comportamento dos modelos por meio de métricas clínicas (Recall, matriz de confusão) e técnicas de interpretabilidade baseadas em **Grad-CAM**. Essa análise integrada permite investigar se o modelo fundamenta suas decisões em regiões anatômicas plausíveis, mitigando o risco de soluções do tipo “caixa-preta”.

II. MATERIAIS E MÉTODOS

A. Base de Dados

O conjunto de dados utilizado consiste em radiografias torácicas previamente rotuladas em duas classes: Normal e Pneumonia. A base de teste possui 624 imagens. O desbalanceamento da base de treinamento marcado pelo predomínio de classes de Pneumonia pode ser observado na Tabela 1.

Tabela 1
DESBALANCEAMENTO ENTRE CLASSES

	Dataset de Treino	Split de Treino	Split de Validação
Pneumonia	3883	3011	872
Normal	1349	1082	267

Para garantir reprodutibilidade experimental e comparabilidade entre as hipóteses avaliadas, o particionamento entre treino e validação foi realizado uma única vez e salvo em arquivos CSV fixos, mantendo-se constante ao longo de todos os experimentos. Essa estratégia evita variações decorrentes de diferentes divisões de dados e assegura que as comparações entre modelos sejam metodologicamente justas. O desbalanceamento observado motivou a avaliação de estratégias de ponderação de classes, investigadas como uma das hipóteses experimentais deste trabalho.

B. *Estratégia Experimental e Pré-processamento*

Todos os experimentos compartilham uma base controlada comum: mesmo conjunto de dados, mesmos splits de treino e validação (congelados em CSV com seed 42), mesmos hiperparâmetros base (lr=1×10^{-4}, batch size = 32, 10 épocas, otimizador Adam, *scheduler ReduceLROnPlateau*) e pesos inicializados com pré-treinamento no *ImageNet*.

O pré-processamento é padronizado para todos os experimentos: *resize* para 224×224 pixels seguido de *CenterCrop* de 90% (~ 201×201 px), com posterior normalização pela média e desvio padrão do ImageNet (μ = [0.485, 0.456, 0.406], σ = [0.229, 0.224, 0.225]). O *CenterCrop* foi motivado pela análise de variância pixel-a-pixel da EDA, que identificou maior concentração de ruído e artefatos de aquisição nas regiões periféricas das imagens. O augmentation do *Baseline* — e de **H1** e **H3** — consiste em *RandomHorizontalFlip* (p=0.5) e *RandomRotation* (±5°), transformações conservadoras que introduzem variabilidade sem distorcer estruturas anatômicas relevantes. Na hipótese **H2**, empregou-se *augmentation* mais intenso (rotações ±15°, transformações afins e maior variação fotométrica), aumentando a diversidade geométrica do conjunto O modelo é salvo pela melhor época segundo o ROC-AUC de validação, e o histórico completo de métricas por época (AUC, F1, Recall, Precision, matriz de confusão, curva ROC) é persistido em `.pkl` para análise posterior.

C. *Modelos*

A escolha por *Transfer Learning* fundamentou-se na limitação do volume amostral e na necessidade de estabilidade de treinamento em um domínio médico com alta variabilidade intra-classe. Ao inicializar as redes com pesos pré-treinados no *ImageNe*t, aproveitam-se representações visuais de baixo e médio nível já consolidadas (bordas, texturas e padrões estruturais), reduzindo o risco de sobreajuste e acelerando a convergência durante o fine-tuning.

A **ResNet18** foi selecionada como arquitetura base por empregar conexões residuais, nas quais cada bloco aprende uma função residual $y = F(x) + x$, facilitando o fluxo de gradiente e mitigando degradação em profundidade. Sua configuração relativamente compacta favorece estabilidade e bom desempenho em cenários com dados moderados.

Já a **DenseNet121** utiliza conexões densas entre camadas, permitindo que cada bloco receba como entrada as ativações de todas as camadas anteriores. Essa estratégia promove forte reutilização de características e maior eficiência paramétrica, potencialmente capturando padrões texturais mais sutis. A comparação entre ambas permitiu avaliar o impacto de diferentes mecanismos de propagação de informação na tarefa de classificação radiológica.

III. RESULTADOS E DISCUSSÕES

A. *Desempenho Geral - equivalência em AUC*

Os quatro experimentos convergiram para valores de ROC-AUC equivalentes e elevados, tanto na validação interna (0.9989–0.9992) quanto no *leaderboard* privado da competição (0.9907–0.9954), demonstrando que todas as configurações testadas produzem classificadores com alta capacidade discriminativa global. Essa equivalência revela a limitação do ROC-AUC como critério de seleção: por agregar o desempenho ao longo de todos os limiares possíveis, a métrica obscurece diferenças críticas no ponto de operação clínico. A análise em *threshold* fixo (0.5), apresentada nas seções seguintes, expõe diferenças substanciais em Recall, Specificity e número de Falsos Negativos — evidenciando que a escolha do modelo clinicamente mais relevante não pode ser fundamentada no ROC-AUC isoladamente.

Tabela 2
DESEMPENHO INTERNO DO AUC

Experimento	AUC
Baseline (ResNet18)	0.9990
H1 — DenseNet121	0.9989
H2 — Strong Aug + Baseline	0.9990
H3 — Class Weight + Baseline	0.9992

B. *Análise por Hipótese - FNs, Recall e Specificity*

A análise em threshold fixo revela um padrão claro e progressivo nos Falsos Negativos. A H1 não confirmou a hipótese, a **DenseNet121** aumentou os FNs e reduziu o Recall para 0.9782, sugerindo que sua maior complexidade é desfavorável para um dataset de tamanho moderado, onde a **ResNet18** converge com maior estabilidade e mais rapidamente. A **H2** produziu o pior resultado clínico, com 24 FNs e Recall de 0.9725. O que indica que o augmentation intenso, ao aplicar transformações geométricas agressivas, distorceu padrões patológicos como

consolidações e infiltrados — estruturas sensíveis a rotações e deformações afins —, degradando justamente o sinal que o modelo precisava aprender. A **H3** foi a única hipótese confirmada: ao introduzir penalização assimétrica na função de perda, o class weighting produziu Recall de 0.9943, F1 de 0.9931 e apenas 5 FNs — redução de 69% em relação ao *Baseline* —, sendo o único experimento explicitamente projetado para o custo assimétrico inerente ao problema clínico.

Tabela 3
DESEMPENHO INTERNO DE MÉTRICAS CLÍNICAS

	Best Epoch	Recall	Sensitivity	Specificity	FN
Baseline	5	0.9817	0.9817	0.9850	16
H1	10	0.9782	0.9782	0.9888	19
H2	10	0.9725	0.9725	0.9925	24
H3	6	0.9943	0.9943	0.9738	5

C. Tradeoff Clínico - Sensitivity vs Specificity

O ganho de Sensitivity do H3 (0.9943 vs 0.9817 do *Baseline*) ocorre à custa de redução de Specificity (0.9738 vs 0.9850), conforme apresentado na Tabela 3. Esse *tradeoff* deve ser interpretado no contexto clínico: em triagem diagnóstica, um Falso Negativo, paciente com pneumonia dispensado sem tratamento, representa risco direto à vida, enquanto um Falso Positivo implica apenas investigação complementar. O custo assimétrico entre esses erros justifica a priorização de Sensitivity. Observa-se que o H3 apresenta a maior Sensitivity entre os experimentos, enquanto Baseline, **H1** e **H2** mantêm maior Specificity à custa de menor Recall e maior número de Falsos Negativos, perfil menos adequado ao objetivo clínico do sistema.

IV. Interpretabilidade

A análise de interpretabilidade foi conduzida por meio do Grad-CAM (*Gradient-weighted Class Activation Mapping*), técnica que projeta os gradientes da classe alvo sobre os mapas de ativação da última camada convolucional, gerando um mapa de calor que indica quais regiões da imagem mais influenciaram a decisão do modelo. A probabilidade de saída do softmax (p) quantifica a confiança do modelo na classe Pneumonia: valores próximos de 1.0 indicam alta confiança na predição positiva; valores próximos de 0.0, alta confiança na predição negativa. Nos erros, o valor de p revela se o modelo falhou com convicção ou com incerteza — distinção clinicamente relevante.

Verdadeiros Positivos (TP): em ambos os modelos, os mapas de calor nos acertos concentram-se nas regiões pulmonares centrais e inferiores, com ativação sobre áreas de maior opacidade — padrão anatomicamente consistente com consolidações e infiltrados. O *Baseline* apresenta ativações mais compactas e localizadas; o **H3** exibe cobertura mais difusa e bilateral, coerente com a maior sensibilidade a manifestações heterogêneas da doença. Ambos classificam com p=1.00, indicando alta confiança nos acertos.

Falsos Negativos (FN) — padrão de erro divergente: este é o achado mais relevante da análise. Nos FNs do *Baseline* (p=0.44, 0.03, 0.44), os mapas ativam estruturas periféricas — costelas, clavícula, bordas torácicas — sem foco nas regiões pulmonares. O modelo não detectou o sinal patológico e errou com probabilidade próxima ao limiar de decisão (p≈0.44) ou com falsa certeza negativa (p=0.03), sem sinalizar dúvida. Nos FNs do **H3** (p=0.17, 0.15, 0.07), o comportamento é qualitativamente distinto: o modelo erra com baixa confiança, sinalizando incerteza calibrada. Em um sistema de triagem clínica, essa distinção é operacionalmente valiosa — casos com p baixo podem ser automaticamente encaminhados para revisão por especialista, transformando o erro em um alerta em vez de uma dispensa silenciosa.

Comparação direta — mesmos casos, modelos opostos: a figura de comparação apresenta três casos de pneumonia classificados erroneamente pelo *Baseline* (p=0.36–0.44) e corretamente pelo **H3** (p=0.83–0.89). Os mapas divergem de forma consistente: o *Baseline* ativa regiões não-pulmonares com foco compacto e incorreto; o **H3** ativa os campos pulmonares com padrão difuso bilateral, identificando consolidações que o *Baseline* ignorou. Esse resultado demonstra que o class weighting não apenas ajusta o limiar de decisão — ele altera a representação interna aprendida, tornando o modelo mais sensível a padrões de menor intensidade e maior heterogeneidade espacial.

Verdadeiros Negativos (TN): em ambos os modelos, os mapas nos casos normais ativam o mediastino e a coluna vertebral — estruturas proeminentes em pulmões saudáveis, onde a ausência de opacidade pulmonar é o próprio sinal discriminativo. As probabilidades são p=0.00 em todos os TNs analisados, confirmando que a redução de Specificity do **H3** está concentrada em casos limítrofes, não em normais claramente saudáveis.

A análise Grad-CAM confirma que os ganhos quantitativos do **H3** correspondem a um padrão de atenção anatomicamente mais coerente e a uma gestão de incerteza superiror — dois critérios essenciais para qualquer sistema de suporte ao diagnóstico clinicamente confiável.

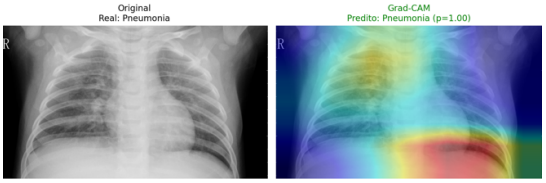

Figura 1. TP do *Baseline* com p = 1.00.

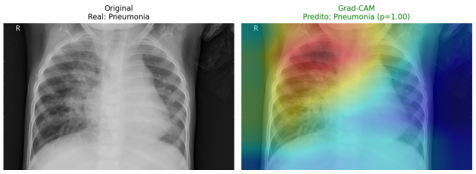

Figura 2. TP do CW com p = 1.00.

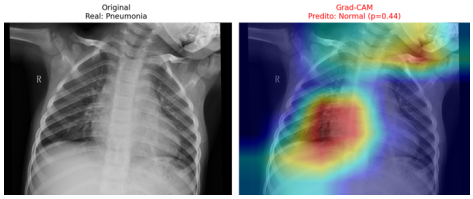

Figura 3. FN do *Baseline* com p = 0.44.

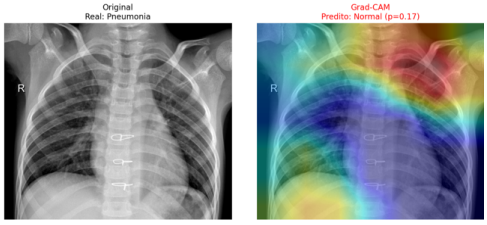

Figura 4. FN do CW com p = 0.17.

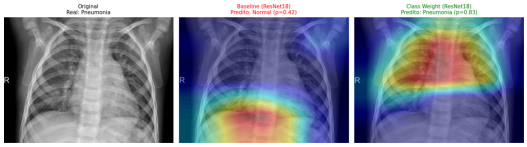

Figura 5. Comparação de modelos com *Baseline* (p = 0.42) e

CW (p = 0.83).

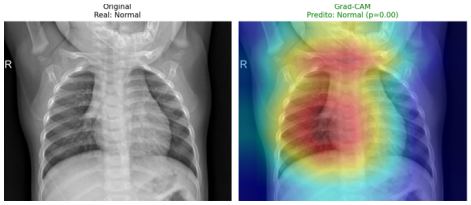

Figura 6. TN do *Baseline* com p = 0.00.

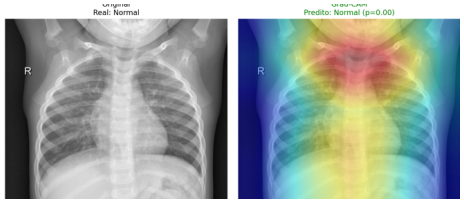

Figura 7. TN do CW com p = 0.00.

V. CONCLUSÃO

Este trabalho avaliou quatro experimentos controlados de classificação de pneumonia via transfer learning (ResNet18 e **DenseNet121**). Todos apresentaram ROC-AUC elevado e equivalente (0.9989–0.9992), confirmando a viabilidade da abordagem; entretanto, a análise em threshold fixo revelou diferenças clinicamente relevantes.

A **H1** (**DenseNet121**) não foi confirmada, elevando os Falsos Negativos de 16 para 19, indicando que maior complexidade não favorece datasets moderados. A **H2** (augmentation intenso) reduziu o Recall para 0.9725 e aumentou os FNs para 24, sugerindo que transformações geométricas agressivas distorcem padrões radiológicos sutis. A **H3** foi confirmada: o class weighting reduziu os FNs de 16 para 5 (−69%), elevando o Recall para 0.9943 e ajustando o modelo ao custo clínico assimétrico do problema.

A análise por Grad-CAM sustenta os achados quantitativos: o **H3** mantém ativações anatomicamente plausíveis e, quando erra, o faz com baixa confiança ($p < 0.20$), enquanto o *Baseline* concentra parte dos erros próximos ao limiar de decisão. Assim, embora o ROC-AUC seja robusto para comparação global, é insuficiente isoladamente em contexto diagnóstico. Recomenda-se a ResNet18 com class weighting para suporte à triagem de pneumonia, por combinar alta sensibilidade, coerência interpretativa e incerteza calibrada.

VI. LIMITAÇÕES E TRABALHOS FUTUROS

Como limitações, destacam-se: treinamento restrito a 10 épocas sem busca sistemática de hiperparâmetros; threshold fixo em 0.5 sem otimização; definição proporcional do class weighting sem ajuste fino; e análise de interpretabilidade predominantemente qualitativa.

Como extensões, propõem-se: otimização do threshold por curva ROC; validação cruzada para ajuste do peso das classes; investigação de alternativas como focal loss; expansão para classificação multiclasse; e validação em datasets externos para avaliar generalização e robustez a domain shift.

VII. PRINCIPAIS REFERÊNCIAS

- KUNDU, R. et al. Pneumonia detection in chest x-ray images using an ensemble of deep learning models. PLoS ONE, v. 16, n. 9, p. e0256630, 2021.
- SALEHI, M. et al. Automated detection of pneumonia cases using deep transfer learning with paediatric chest X-ray images. The British Journal of Radiology, v. 94, n. 1121, p. 20201263, 1 maio 2021.