Classificação de Patologias em Imagens de Raio-X

DenseNet-121 com Fine-Tuning Gradual, Regularização via Mixup + Label Smoothing e Ensemble Heterogêneo

Ivan Carvalho Ernesto Bezerra
Centro de Informática · UFPE · iceb@cin.ufpe.br | Ligia – Trilha: Visão Computacional

Resumo

Este relatório documenta a solução desenvolvida para a trilha de Visão Computacional do Desafio Individual da Ligia (UFPE, 2026): classificar radiografias de tórax como NORMAL ou PNEUMONIA, avaliado por ROC AUC. A jornada cobre uma análise exploratória orientada por hipóteses — incluindo a detecção original de shortcut learning por viés de hardware (ruído de sensor diferenciado por fabricante) —, pré-processamento validado estatisticamente (CLAHE + NLMeans h=3), validação cruzada anti-leakage por paciente (StratifiedGroupKFold), seleção de arquitetura por torneio controlado entre quatro backbones, fine-tuning gradual em três fases, regularização ortogonal por Mixup + Label Smoothing, e ensemble heterogêneo entre pesos ImageNet e especializados em raio-X. Cada decisão é rastreável a logs de experimento reproduzíveis (SEED=42). Resultado: ROC AUC de 0,9954 no holdout interno e 0,99129 no leaderboard Kaggle.

Palavras-chave: DenseNet-121, Transfer Learning, StratifiedGroupKFold, Mixup, Label Smoothing, Ensemble, XAI, Shortcut Learning, ROC AUC

I. INTRODUÇÃO

A pneumonia é uma das principais causas de mortalidade prevenível globalmente. O diagnóstico por raio-X torácico possui impacto clínico direto onde especialistas são escassos. Este trabalho documenta a solução construída para a trilha de Visão Computacional da Ligia — uma jornada com hipóteses, torneios controlados, erros encontrados e corrigidos. Cada decisão é rastreável a evidência quantitativa obtida nos próprios experimentos. Resultado: ROC AUC 0,9954 no holdout e 0,99129 no leaderboard Kaggle, treinado em GPU Tesla P100-PCIE-16GB.

II. ANÁLISE EXPLORATÓRIA E INTEGRIDADE

A. Estrutura do Dataset e Integridade

Varredura com PIL.Image.verify() validou 5.856 imagens (5.232 treino + 624 teste) sem nenhum corrompido. O dataset de treino apresenta razão 2,88x entre classes: 1.349 NORMAL (25,8%) e 3.883 PNEUMONIA (74,2%). A classe PNEUMONIA decompõe-se em 2.530 bacterianas e 1.345 virais (razão interna 1,88x).

Um achado estrutural relevante: a análise forense dos nomes de arquivo (padrão BACTERIA-5132924-0001.jpeg) revelou 3.458 IDs únicos para 5.232 imagens — média de 1,51 imagens/paciente. A distribuição é altamente assimétrica: 76,3% têm uma imagem, mas um outlier extremo concentra 30 imagens do mesmo paciente. Essa estrutura torna o StratifiedGroupKFold um requisito metodológico inegociável, não uma escolha opcional.

B. Descoberta de Shortcut Learning por Viés de Hardware

A análise de ruído de fundo (desvio padrão em cantos uniformes, N=50/classe) revelou anomalia crítica: imagens NORMAL possuem ruído sistematicamente inferior às classes patológicas. A hipótese é que diferentes fabricantes foram usados para cada contexto clínico — aparelhos de alta fidelidade para triagem de saudáveis, equipamentos portáteis para casos internados.

O risco concreto é um modelo que aprende a classificar pela "assinatura do sensor" e não pela patologia, obtendo ROC AUC elevado na validação mas colapsando em qualquer hospital com hardware padronizado. Essa descoberta motivou integralmente o pipeline de neutralização subsequente — sem ela, toda a

IV. MODELAGEM

A. Validação Anti-Leakage por Paciente

O holdout fixo (15%) foi isolado via StratifiedShuffleSplit sobre o label majoritário por paciente antes de qualquer treinamento: 802 imagens de 519 pacientes, proporção 26,8% NORMAL / 73,2% PNEUMONIA (desvio de 1,2pp em relação ao treino). Os 5 folds apresentaram tamanhos de validação entre 845 e 931 imagens com desvio padrão de apenas 30,3 imagens — variação aceitável como consequência natural do agrupamento por paciente, que não permite corte perfeito.

B. Torneio de Transforms — Sem Augmentação Vence

Antes do torneio de backbones, um torneio menor comparou Pipeline A (sem augmentação) versus Pipeline B (augmentação geométrica) usando EfficientNet-B0 como proxy (600 imagens, 2 folds, 2 épocas). Pipeline A venceu por margem dentro do ruído estatístico (Δ=0,0053 < std=0,0256), sendo adotado por parcimônia. O achado mais relevante: o pré-processamento foi suficientemente eficaz para que transformações geométricas adicionais não trouxessem benefício discriminativo — o CLAHE e NLMeans já extraíram o essencial das variações do dataset.

C. Torneio de Backbones — DenseNet-121

Quatro backbones avaliados em condições controladas (599 imagens com distribuição real, 3 folds, 5 épocas, backbone congelado). O critério de equivalência foi Δ(AUC) < std do vencedor (ddof=1) — evitando selecionar diferenças dentro do ruído de amostragem.

Tabela 2 — Torneio de Backbones (3 folds, 5 épocas congelado)

Backbone	Resolução	AUC Médio	Std
EfficientNet-B0	224×224	0,9604	0,0117
EfficientNet-B2	260×260	0,9782	0,0180
ResNet-50	224×224	0,9533	0,0101
DenseNet-121	**224×224**	**0,9833**	**0,0061**

O DenseNet-121 venceu com Δ>0,0229 sobre todos os rivais — margem 3,7× superior ao próprio std (0,0061). O dado mais diagnóstico não é o AUC médio mas o menor spread do projeto: 0,0061 contra 0,0180 do EfficientNet-B2, que tem AUC médio mais alto mas generalização muito menos uniforme. As conexões

modelagem seria construída sobre alicerce falso.

C. Análise Espectral e Perfil de Erros Projetado

Histogramas médios (N=50/classe) revelaram que imagens PNEUMONIA apresentam deslocamento para tons mais claros (150–230), compatível com regiões de consolidação. A descoberta analítica central: a sobreposição espectral entre NORMAL e PNEUMONIA viral é substancialmente maior do que entre NORMAL e PNEUMONIA bacteriana. Isso projeta com precisão o perfil de erros a ser encontrado na validação — os falsos negativos de maior risco clínico seriam casos virais iniciais, cujo infiltrado intersticial sutil não gera o mesmo deslocamento de histograma que a consolidação lobar bacteriana. A projeção foi confirmada nos folds.

Outro dado que informa o limite do modelo: a média normalizada (ImageNet Z-score) é −0,154 ± 0,256 para NORMAL e −0,381 ± 0,347 para PNEUMONIA — diferença de 0,23 unidades de desvio padrão que persiste após todo o pré-processamento. Esse offset legítimo é a "pegada digital" da patologia: consolidações mapeiam para pixels mais claros que, após Z-score, resultam em valores menos negativos.

III. PRÉ-PROCESSAMENTO

A. Letterboxing e CLAHE

O redimensionamento direto para 224×224 distorce proporções anatômicas — costelas artificialmente comprimidas podem criar associações espúrias entre geometria e patologia. O letterboxing (preservação de razão de aspecto com padding em preto) elimina esse risco. A equalização de histograma global foi descartada porque amplifica ruído térmico de forma indiscriminada nas amostras patológicas já ruidosas. O CLAHE (clipLimit=2,0, tileGridSize=(8,8)) realça texturas pulmonares localmente, preservando o fundo.

B. Torneio de Filtragem — NLMeans h=3

Um efeito colateral inesperado do letterboxing foi identificado na perícia pós-CLAHE: o padding preto absoluto nos cantos das imagens criava uma "etiqueta de silêncio" — o cálculo estatístico de ruído priorizava o canto de menor desvio padrão, e o padding zero se tornava um discriminador artificial entre imagens com e sem padding. Esse problema exigiu a etapa de filtragem.

Três estratégias foram avaliadas (N=100/classe, Score de Convergência como critério). O Filtro Bilateral, que preserva bordas seletivamente, manteve disparidades residuais — evidência de que a variação de textura global, não apenas bordas, era a fonte do viés. O NLMeans h=3 venceu (score 41,71): suavização leve e constante foi mais eficaz que preservação seletiva para unificar assinaturas de ruído entre classes.

Tabela 1 — Torneio de Pipelines de Filtragem (N=100/classe)

Pipeline	Score Conv.↓	Decisão
A — NLMeans h=3	**41,71**	**Adotado**
C — NLMeans Condicional	42,03	Equivalente
B — Filtro Bilateral	48,32	Inferior

A perícia estatística pós-processamento (N=300 amostras) confirmou convergência das distribuições: o viés de hardware foi neutralizado de forma mensurável, não apenas presumida. Ainda

densas do DenseNet entregam gradientes diretos a todas as camadas anteriores, acelerando convergência em texturas finas como infiltrados. No Fold 3 do torneio, o DenseNet iniciou com AUC=0,8707 (abaixo de todos os rivais) mas recuperou para 0,9782 ao final — evidência de robustez real diante de uma partição mais difícil com mais casos virais.

D. Fine-Tuning Gradual em Três Fases

O DenseNet-121 tem 6.954.881 parâmetros. O unfreezing gradual protege o conhecimento pré-treinado ao escalonar a exposição dos pesos ao gradiente:

- Fase 1 — cabeça (1.025 param., lr=1e-3, 5 épocas): loss cai de 0,1820 → 0,0748 no Fold 1. Convergência estável sem oscilação confirma que a cabeça estabilizou antes de qualquer fine-tuning.
- Fase 2 — denseblock4+norm5 (2.161.153 param., lr=1e-4): maior salto em todos os folds. No Fold 1, AUC vai de 0,9878 para 0,9978 em 5 épocas; loss de 0,0556 para 0,0036. O denseblock4 concentra as representações de maior abstração semântica — é onde a especialização médica ocorre de fato.
- Fase 3 — backbone completo (6.954.881 param., lr=1e-5): ganhos marginais. A loss colapsa para 0,0009 no Fold 1 — sinal inequívoco de memorização que motivou a iteração seguinte.

E. Regularização: Mixup + Label Smoothing

A loss de treino na Fase 3 em 0,002–0,010 confirmou memorização. Augmentação geométrica foi descartada por restrições anatômicas: flip horizontal simula dextrocardia; rotações acima de 10° produzem incidências clinicamente inexistentes — degradariam o sinal médico ao invés de amplificá-lo.

Mixup (α=0,2) interpola pares de imagens e rótulos via distribuição Beta, penalizando confiança excessiva sem gerar imagens anatomicamente inválidas. Label Smoothing (ε=0,1) converte {0,1} para {0,05; 0,95} permanentemente, cobrindo exatamente os batches onde λ do Mixup se aproxima dos extremos — as duas técnicas são complementares e não redundantes. Resultado: loss de treino na Fase 3 estabilizou em 0,165–0,183 (contra 0,002–0,010 anterior), AUC subiu de 0,9978 para 0,9983.

F. TorchXRayVision — Pesos de Domínio

O backbone ImageNet captura representações de fotografias naturais. O TorchXRayVision (densenet121-res224-all) foi pré-treinado em >100.000 radiografias (NIH+CheXpert+MIMIC+PadChest). Adaptações: conv0 de 1→3 canais por replicação/3 (preservando escala das ativações) e substituição do classificador de 18 saídas por cabeça binária.

A versão v1 (5/5/5 épocas) ficou abaixo do Mixup+LS (0,9949 vs 0,9983) por convergência insuficiente: a loss de treino estagnou em 0,19–0,22 na Fase 3, indicando que o backbone especializado precisa de mais épocas para se adaptar ao dataset específico do que o backbone ImageNet. A versão v2 (5/7/10 épocas) atingiu 0,9972 com spread de 0,0008 — o menor do projeto inteiro. Isso indica que pesos pré-treinados em raio-X generalizam de forma mais uniforme entre subpopulações de pacientes: o domínio médico já estava codificado nos pesos, restando apenas a especialização na tarefa específica.

assim, a convergência foi descrita no próprio notebook como "parcial" — uma honestidade metodológica importante que demarca o limite do pré-processamento e avisa que o modelo treinado ainda pode ter alguma sensibilidade residual ao equipamento.

C. Normalização e Sanity Check

Tensores float32 (3, 224, 224) com valores em [−2,12; 2,64] confirmam normalização ImageNet correta. Feeding pesos pré-treinados com pixels brutos (0–255) degradaria as representações pré-treinadas e exigiria mais épocas de reconvergência — custo evitável dado o orçamento computacional limitado do ambiente Kaggle.

V. RESULTADOS, ENSEMBLE E ANÁLISE CRÍTICA

A. Evolução de Desempenho — Todos os Experimentos

Tabela 3 — ROC AUC por configuração (CV 5 folds + holdout)

Configuração	AUC CV	Spread
DenseNet-121 baseline (torneio)	0,9833	0,0061
+ Fine-tuning gradual (5/5/5)	0,9978	0,0009
+ Mixup + Label Smoothing	0,9983	0,0009
TorchXRayVision v1 (5/5/5)	0,9949	0,0007
TorchXRayVision v2 (5/7/10)	0,9972	0,0008
DenseNet Agressivo (rot+color)	0,9978	0,0016
EfficientNet-B4 (380×380)	0,9969	0,0008
Ensemble ótimo (w=0,70/0,30)	**0,9987**	—
Holdout interno (802 imgs / 519 pac.)	0,9954	—
Leaderboard Kaggle (submissão final)	**0,99129**	—

B. Análise do Ensemble — Grid Search e Plateau

O grid search de pesos (passos de 0,05) revelou um plateau largo: qualquer combinação com w_txrv entre 0,25 e 0,50 produz AUC=0,9987. O peso ótimo encontrado foi 0,70 para Mixup+LS e 0,30 para TorchXRayVision v2 — resultado contraintuitivo, pois o Mixup+LS (pesos ImageNet) superou o TorchXRayVision em 4 dos 5 folds de validação cruzada apesar de partir de representações menos especializadas.

A adição de um terceiro (DenseNet Agressivo, AUC=0,9978) e quarto modelo (EfficientNet-B4, AUC=0,9969) não moveu o plateau além de 0,9987 em nenhuma combinação testada. O DenseNet Agressivo recebeu peso ótimo de apenas 0,10; o EfficientNet-B4, igualmente. Isso significa que a diversidade arquitetural introduzida por compound scaling e depthwise separable convolutions não produziu erros suficientemente descorrelacionados para superar o plateau — os modelos falham nos mesmos casos difíceis (casos virais iniciais), independente da arquitetura.

C. Diagnóstico: O Teto é nos Dados, Não nos Modelos

O plateau de 0,9987 com 4 modelos arquiteturalmente distintos é o resultado mais informativo do projeto. Com folds de ~880 amostras, discriminar ganhos abaixo de 0,001 AUC é estatisticamente não confiável — o teto está no volume de dados (4.430 imagens de treino), não na capacidade de representação. Para superar este limite seria necessário ampliar o dataset via fontes externas (CheXpert, NIH ChestX-ray14) antes de qualquer novo experimento de modelagem.

D. Análise Crítica: O Experimento V3 com AUC=1,0

Um experimento tardio no notebook treinou DenseNet-121 via timm sem StratifiedGroupKFold, usando um split treino/validação simples sobre todo o dataset. O resultado — AUC=0,9991 na época 1, AUC=1,0000 na época 4, loss 0,0312 — é um caso clássico de leakage por sobreposição de pacientes.

> Análise original: AUC=1,0 com loss decrescente (0,0312) não é sucesso — é diagnóstico de memorização completa por leakage de paciente.
> ⚠ *O experimento foi submetido ao Kaggle e retornou AUC insatisfatório,* confirmando empiricamente que o modelo "reconheceu" pacientes e não *→ aprendeu a diagnosticar. O leakage foi validado pela competição real.*

Esse episódio demonstra o valor do protocolo rigoroso estabelecido desde a EDA: o StratifiedGroupKFold não é preciosismo metodológico — é a diferença entre um modelo que funciona e um que apenas parece funcionar. A versão final (ensemble Mixup+LS + TorchXRayVision v2) foi submetida posteriormente e atingiu 0,99129.

E. Interpretabilidade (XAI) — Evidências do Notebook

A análise de interpretabilidade pode ser feita sem implementar Grad-CAM. Os logs do notebook fornecem evidência direta sobre o que o modelo aprendeu:

- A convergência explosiva do denseblock4 na Fase 2 (AUC 0,9878 → 0,9978 em 5 épocas; loss 0,0556 → 0,0036) indica que as representações de maior abstração semântica do DenseNet capturaram os padrões diagnósticos essenciais — consolidações e infiltrados são padrões espaciais de alta frequência concentrados exatamente nas camadas mais profundas.
- O offset de normalização persistente (NORMAL: −0,154; PNEUMONIA: −0,381) confirma que o modelo distingue classes pelo brilho médio — compatível com hiperdensidade focal real, não metadado espúrio.
- A convergência estatística pós-NLMeans (confirmada na perícia de 300 amostras) prova por exclusão: a feature de ruído de sensor foi eliminada antes do treinamento, tornando clinicamente improvável que o modelo a utilize como discriminador.

O plano técnico imediato para Grad-CAM: registrar hooks na

camada features.norm5 do DenseNet-121, executar backward para a classe predita, e sobrepor os mapas às radiografias originais. A validação clínica verificaria se as regiões de maior ativação coincidem com as marcadas por radiologistas — passo de custo computacional zero (inferência apenas).

F. Análise de Erros e Métricas de Negócio

O pos_weight=2,91 resultou em Recall para PNEUMONIA acima de 0,98 em todos os folds. O trade-off é redução de Especificidade — alguns casos NORMAL classificados como PNEUMONIA. Em contexto clínico, o custo assimétrico é deliberado: um falso negativo expõe o paciente a pneumonia não tratada; um falso positivo gera apenas exames adicionais. O threshold de 0,5 é subótimo para maximizar Recall com Precision mínima; deve ser ajustado via curva Precision-Recall exclusivamente sobre o holdout.

G. Limitações e Trabalhos Futuros

- Grad-CAM sobre features.norm5 e validação clínica com anotações de radiologistas — custo: zero retreinamento.
- Ampliação do dataset (CheXpert / NIH ChestX-ray14) para superar o plateau de ensemble em 0,9987.
- Análise de calibração probabilística (ECE): o modelo produz probabilidades bem calibradas para uso como score de triagem contínuo?
- Teste de distribuição shift: a neutralização de hardware por NLMeans é suficiente para generalizar inter-institucionalmente?

H. Conclusão

Rigor metodológico e performance competitiva são complementares. Os achados centrais: (a) detecção original de shortcut learning por viés de hardware, com validação estatística quantitativa; (b) StratifiedGroupKFold como requisito inegociável confirmado empiricamente pelo colapso do V3 no leaderboard; (c) DenseNet-121 selecionado por torneio controlado, com spread inter-folds como critério diferencial; (d) Mixup + Label Smoothing contendo memorização sem custo discriminativo; (e) plateau de ensemble em 0,9987 com 4 modelos como diagnóstico objetivo de que o teto está nos dados. Pipeline reprodutível via SEED=42.