Modelo baseado em DenseNet121 para detecção de múltiplas patologias em raios-X de tórax

Kevin Henrique Monfredini Silva
IESTI
Universidade Federal de Itajubá
Itajubá, Brasil
d2022006962@unifei.edu.br

Laura Ribeiro Bernardino
IEPG
Universidade Federal de Itajubá
Itajubá, Brasil
d2023009575@unifei.edu.br

Matheus Brendon Francisco
IEPG
Universidade Federal de Itajubá
Itajubá, Brasil
matheus_brendon@unifei.edu.br

Matheus Costa Pereira
IEPG
Universidade Federal de Itajubá
Itajubá, Brasil
matheusc_pereira@hotmail.com

***Resumo*—O diagnóstico de patologias torácicas por meio de radiografias de tórax é uma tarefa complexa que demanda tempo e conhecimento especializado de radiologistas. A aplicação de modelos de Aprendizagem Profunda, particularmente as Redes Neurais Convolucionais (CNNs), tem demonstrado grande potencial para automatizar e auxiliar neste processo. Este artigo apresenta o desenvolvimento e a validação de um modelo baseado na arquitetura DenseNet121, com a incorporação de mecanismos de atenção (CBAM), para a classificação de 14 patologias no conjunto de dados público NIH Chest X-ray. Visando superar as limitações do severo desbalanceamento de classes, o modelo foi treinado no conjunto de dados completo, contendo mais de 112.000 imagens. Crucialmente, a avaliação final foi conduzida em um conjunto de teste de 810 radiografias, cujos rótulos, disponibilizados no trabalho de Nabulsi et al. (2021), foram definidos por um painel de cinco radiologistas norte-americanos certificados, garantindo um padrão de alta confiabilidade para a avaliação. O modelo final alcançou uma AUC Média de 0,8543 e um F1-Score Macro de 0,3912. Notavelmente, a abordagem demonstrou capacidade de detectar patologias extremamente raras, como Hérnia, com um F1-Score de 0,667 e uma AUC de 0,999. Os resultados demonstram que a escala dos dados de treinamento, aliada a uma validação rigorosa contra um padrão de referência de especialistas, é um fator determinante para o desenvolvimento de sistemas de diagnóstico automatizado mais robustos e clinicamente relevantes.**

***Palavras-chave*—Radiografia de Tórax, Aprendizado Profundo, DenseNet121, Classificação Multirrótulo, Diagnóstico Auxiliado por Computador.**

I. Introdução

A radiografia de tórax (CXR) permanece como uma modalidade de imagem fundamental e amplamente utilizada na prática médica global, servindo como ferramenta diagnóstica de primeira linha para uma vasta gama de condições cardiopulmonares e outras anormalidades torácicas [1]. Sua prevalência é atribuível à combinação de acessibilidade, baixo custo, rapidez na aquisição e uma dose de radiação relativamente baixa, tornando-a indispensável em múltiplos contextos clínicos, desde exames de rotina e triagem até o monitoramento de pacientes em estado crítico [2]. As CXRs oferecem informações valiosas sobre os pulmões, coração, pleura, mediastino e arcabouço ósseo, auxiliando no diagnóstico e acompanhamento de doenças como pneumonia, tuberculose, edema pulmonar, cardiomegalia, derrames pleurais e neoplasias [3]. Apesar de sua centralidade no diagnóstico por imagem, a interpretação de radiografias torácicas é uma tarefa complexa que exige considerável conhecimento especializado e está sujeita a desafios inerentes. A detecção de achados patológicos pode ser dificultada pela sutileza de certas anormalidades, pela sobreposição de estruturas anatômicas e pela presença de artefatos na imagem [4]. Estudos têm consistentemente demonstrado a existência de variabilidade interobservador na interpretação de CXRs, mesmo entre radiologistas experientes, o que pode impactar a consistência e acurácia diagnóstica [5]. Adicionalmente, a carga de trabalho elevada em muitos departamentos de radiologia pode contribuir para a fadiga interpretativa, aumentando o potencial de erros perceptivos ou cognitivos [4], [6]. Neste cenário, o advento da Inteligência Artificial (IA), particularmente as técnicas de aprendizado profundo (Deep Learning), tem catalisado um avanço significativo no desenvolvimento de sistemas de diagnóstico auxiliado por computador (CAD) para a análise de imagens médicas [7]. As Redes Neurais Convolucionais (CNNs), em especial, têm alcançado resultados notáveis na interpretação de imagens radiográficas, demonstrando em diversos estudos a capacidade de identificar achados patológicos com um desempenho que se aproxima ou, em alguns casos, iguala o de especialistas humanos [8], [9]. O potencial da IA em radiologia torácica reside na sua capacidade de auxiliar na triagem de exames, destacar regiões de interesse suspeitas, quantificar achados e reduzir a variabilidade na interpretação, atuando como uma segunda opinião ou ferramenta de suporte à decisão para os radiologistas [7]. A disponibilidade de grandes bancos de dados de imagens médicas, como o conjunto de dados NIH Chest X-ray, que compreende mais de 112.000 radiografias de tórax com anotações para múltiplas patologias, tem sido um fator crucial para o treinamento e validação desses modelos de aprendizado profundo [10]. Este conjunto de dados, cujos rótulos foram gerados a partir da mineração de texto de laudos radiológicos associados, permite o desenvolvimento de algoritmos para tarefas complexas de classificação multir- rótulo, onde múltiplas patologias podem estar presentes em uma única imagem [10], [11]. Contudo, uma característica marcante e desafiadora deste conjunto de dados é o seu severo desbalanceamento de classes, com certas patologias raras representando uma fração minúscula do total. Essa distribuição de cauda longa (long-tail) constitui um obstáculo crítico, pois

modelos treinados nestes dados tendem a ignorar as classes minoritárias, limitando drasticamente sua utilidade e segurança na prática clínica.

II. Trabalhos Relacionados

A aplicação de Aprendizagem Profunda, especificamente Redes Neurais Convolucionais (CNNs), para a análise de radiografias de tórax (CXR) é um campo de pesquisa consolidado, impulsionado pela disponibilização de grandes bancos de dados públicos. A publicação do conjunto de dados NIH ChestX-ray [10] foi um marco que catalisou o desenvolvimento de modelos de diagnóstico automatizado. Um dos trabalhos de maior impacto nesta área foi o CheXNet [11], que demonstrou a eficácia de uma arquitetura DenseNet-121 para a detecção de patologias, alcançando um desempenho comparável ao de radiologistas. Seguindo essa linha, a comunidade científica explorou uma variedade de outras arquiteturas, como as Redes Residuais (ResNet), investigando sistematicamente os efeitos da transferência de aprendizagem e da integração de metadados para otimizar a classificação multirrótulo [8]. Um desafio central e persistente na classificação de patologias em CXR é o severo desbalanceamento de classes, criando uma distribuição de cauda longa (long-tail) que prejudica o desempenho dos modelos. A literatura apresenta um conjunto de estratégias para mitigar este problema. No nível dos dados, técnicas de reamostragem buscam balancear artificialmente a distribuição, com métodos de sobreamostragem (oversampling), como o Synthetic Minority Over-sampling Technique (SMOTE), sendo frequentemente empregados para criar novas instâncias sintéticas das classes minoritárias. A segunda frente, no nível do modelo, consiste no uso de funções de perda especializadas. A Focal Loss [15] é um exemplo proeminente, projetada para modificar a função de perda de entropia cruzada padrão, atribuindo um peso dinamicamente menor aos exemplos fáceis e bem classificados, forçando o modelo a focar o aprendizado nos exemplos difíceis das classes minoritárias. Apesar do foco da literatura em inovações arquitetônicas e funções de perda, a escala massiva do conjunto de dados de treinamento como estratégia primária para superar o desbalanceamento não foi investigada com o mesmo rigor. Muitos trabalhos buscam soluções algorítmicas para um volume de dados fixo, mas a hipótese de que um aumento de uma ordem de magnitude nos dados poderia ser a estratégia mais eficaz para a detecção de patologias raras permanecia como uma lacuna relevante. Este artigo busca endereçar precisamente essa questão, demonstrando que a escala dos dados é o fator mais crítico e validando os achados contra um padrão-ouro de especialistas [14] para garantir a robustez científica.

III. Metodologia

A. Conjunto de Dados

O presente estudo utilizou o conjunto de dados público NIH Chest X-ray, também conhecido como ChestX-ray14 [10]. Este é um dos maiores conjuntos de dados de radiografias torácicas disponíveis publicamente, contendo 112.120 imagens de raios-X frontais de 30.805 pacientes únicos. Cada imagem está associada a rótulos de catorze patologias torácicas comuns: Atelectasia, Cardiomegalia, Consolidação, Edema, Derrame, Enfisema, Fibrose, Hérnia, Infiltração, Massa, Nódulo, Espessamento Pleural, Pneumonia e Pneumotórax. Os rótulos foram extraídos de laudos radiológicos associados por meio de técnicas de Processamento de Linguagem Natural (PLN), um método que, embora permita a anotação em larga escala, pode introduzir um certo nível de ruído ou imprecisão nos rótulos [10]. Cientes do potencial ruído nos rótulos gerados por PLN, uma etapa fundamental deste trabalho foi a definição de uma metodologia de avaliação de alta confiabilidade. Para a avaliação final do modelo, utilizamos um conjunto de teste distinto, disponibilizado por Nabulsi et al. [14], composto por 810 radiografias cujos rótulos foram independentemente definidos por um painel de cinco radiologistas norte-americanos certificados. Este conjunto de dados serviu como nosso padrão-ouro (*gold standard*). Para garantir a integridade da avaliação, todas as 810 imagens deste conjunto de teste foram rigorosamente removidas do conjunto de dados principal antes de qualquer divisão. O conjunto restante (111.310 imagens) foi então dividido em partições de treinamento (90%) e validação (10%) para o desenvolvimento do nosso modelo final.

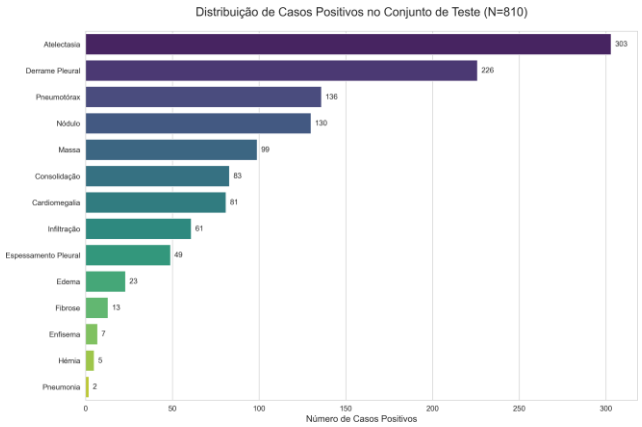

Fig. 1: Distribuição de casos positivos para cada uma das 14 patologias no conjunto de teste de especialistas (N=810), evidenciando o severo desbalanceamento e a cauda longa de distribuição das classes.

B. Pré-processamento e Aumento de Dados

As imagens foram submetidas a uma sequência de etapas de pré-processamento. Inicialmente, todas as imagens foram convertidas para o formato monocromático (tons de cinza). Em seguida, foram redimensionadas para uma dimensão de 256x256 pixels. Para o conjunto de treinamento, foram aplicadas técnicas de aumento de dados (*data augmentation*) com o objetivo de aumentar a variabilidade dos dados e reduzir o risco de sobreajuste (*overfitting*). As transformações incluíram: recorte aleatório para 224x224 pixels, espelhamento horizontal aleatório, transformações afins aleatórias (rotações, translações, escalonamento e cisalhamento) e ajustes aleatórios de brilho e contraste. Para os conjuntos de validação e teste, as

imagens foram submetidas a um recorte central para 224x224 pixels. Todas as imagens, após as transformações, foram normalizadas utilizando média de 0,502 e desvio padrão de 0,252, valores calculados a partir da distribuição de pixels do próprio conjunto de dados NIH Chest X-ray, garantindo uma normalização mais específica para imagens radiográficas monocromáticas.

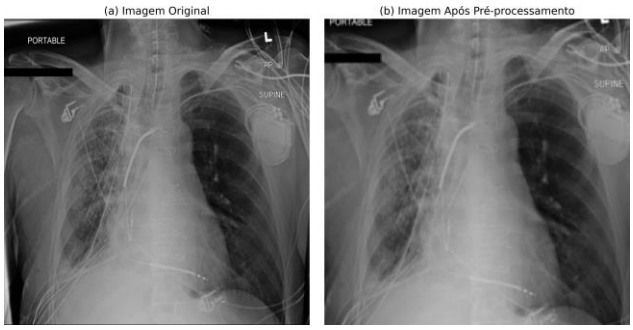

Fig. 2: Exemplo de imagem de radiografia torácica antes (a) e depois (b) da aplicação das etapas de pré-processamento, incluindo recorte central e normalização.

C. Arquitetura do Modelo e Justificativa

A arquitetura de rede neural convolucional (CNN) escolhida para este estudo foi a DenseNet121 [12], aprimorada com Módulos de Atenção de Bloco Convolucional (CBAM, *Convolutional Block Attention Module*). A seleção desta arquitetura foi uma decisão metodológica fundamentada com a combinação da eficiência de parâmetros da DenseNet121 com a capacidade de foco do CBAM como uma estratégia para criar um modelo que fosse ao mesmo tempo profundo o suficiente para aprender características complexas e focado o suficiente para identificar achados patológicos sutis, representando um balanço entre complexidade e performance. A DenseNet121 foi escolhida por sua reconhecida eficiência de parâmetros e pelo forte reuso de características, conforme validado em aplicações de larga escala como o CheXNet [11]. Em uma tarefa complexa de classificação de 14 patologias, onde características visuais de baixo nível (ex: texturas) podem ser relevantes para múltiplas condições, a conectividade densa da arquitetura é teoricamente vantajosa, pois facilita o fluxo de gradientes e evita o reaprendizado redundante de características. A integração do CBAM representa um refinamento lógico para o diagnóstico por imagem. O mecanismo de atenção sequencial (canal e espacial) permite que o modelo aprenda dinamicamente "o que" focar (quais mapas de características são mais informativos) e "onde" focar (quais regiões espaciais são mais salientes). Em radiografias, onde um achado pode ser uma anomalia sutil e localizada (como um Nódulo) ou um padrão difuso (como uma Infiltração), a capacidade de ponderar a importância das características é crucial para guiar a rede neural para as regiões de maior relevância diagnóstica. Utilizou-se uma versão da DenseNet121 pré-treinada no conjunto de dados ImageNet. A primeira camada convolucional foi adaptada para aceitar imagens de 1 canal, e a camada classificadora final foi substituída por uma camada linear com 14 saídas e ativação sigmoide, permitindo a predição multirrótulo.

D. Treinamento do Modelo

O treinamento do modelo foi realizado utilizando o *framework* PyTorch. A estratégia experimental foi dividida em duas fases para quantificar o impacto da escala dos dados, mantendo-se a metodologia de treinamento idêntica em ambas. A função de perda empregada foi a *Binary Cross-Entropy with Logits* (BCEWithLogitsLoss), ponderada pelo mecanismo de *pos_weight* para mitigar o severo desbalanceamento de classes. Este peso, calculado como a razão entre o número de amostras negativas e positivas para cada classe no con- junto de treinamento, atribui uma penalidade maior aos erros cometidos na classe minoritária. O otimizador escolhido foi o AdamW (*Adaptive Moment Estimation with Weight Decay*), O otimizador AdamW foi selecionado em detrimento de alternativas como o SGD por sua robustez e rápida convergência, características bem documentadas em problemas de visão computacional de larga escala. Com uma taxa de aprendizado inicial de 1×10^{-4} e um decaimento de peso (*weight decay*) de 1×10^{-5}. Utilizou-se um agendador (*scheduler*) *ReduceLROnPlateau* para reduzir a taxa de aprendizado por um fator de 0,2 caso a métrica de AUC Média de Validação não apresentasse melhora após 3 épocas. Um mecanismo de parada antecipada (*early stopping*) foi implementado para interromper o treinamento caso a mesma métrica não melhorasse por 5 épocas consecutivas, salvando-se o modelo com o melhor desempenho.

E. Avaliação do Desempenho e Métricas

A avaliação de desempenho foi realizada exclusivamente no conjunto de teste padrão-ouro. Para converter as probabilidades de saída do modelo em predições binárias, o limiar de decisão para cada classe foi otimizado no conjunto de validação, selecionando-se o valor que maximizava o F1-score para aquela classe específica. As métricas no conjunto de teste foram então reportadas utilizando esses limiares otimizados. As métricas globais para avaliar o desempenho agregado do modelo incluíram: a AUC Média, para medir a capacidade discriminativa, e o F1-Score Macro, por sua sensibilidade ao desempenho nas classes minoritárias.

IV. Resultados

O desempenho dos modelos foi avaliado rigorosamente contra o conjunto de teste, composto por 810 imagens rotuladas por especialistas. Apresentamos os resultados de forma comparativa para quantificar o impacto da escala do con- junto de dados de treinamento no desempenho diagnóstico da arquitetura DenseNet121-CBAM. O "Modelo Amostral" refere-se ao modelo treinado na amostra de 10% dos dados (10k imagens), enquanto o "Modelo Final" refere-se à mesma arquitetura treinada no conjunto de dados completo (100k imagens).

Tabela I: Comparação de Desempenho Entre o Modelo Amostral e o Modelo Final, Avaliados no Conjunto de Teste de Especialistas.

Métrica / Patologia	F1-Score (Amostral)	F1-Score (Final)	AUC (Amostral)	AUC (Final)
Hérnia	0,000	**0,667**	0,738	**0,999**
Pneumotórax	0,597	**0,708**	0,873	**0,934**
Massa	0,446	**0,632**	0,802	**0,943**
Derrame Pleural	**0,706**	0,698	0,896	**0,908**
Cardiomegalia	**0,533**	0,488	0,859	**0,901**
Nódulo	0,354	**0,481**	0,674	**0,851**
Atelectasia	0,447	**0,463**	**0,817**	0,770
Infiltração	0,331	**0,387**	0,811	**0,841**
Espessamento Pleural	0,236	**0,365**	0,787	**0,854**
Consolidação	0,082	**0,346**	0,799	**0,842**
Fibrose	0,062	**0,182**	0,619	**0,788**
Enfisema	0,025	**0,059**	0,761	**0,787**
Edema	**0,143**	0,000	0,797	**0,870**
Pneumonia	0,000	0,000	**0,908**	0,673
AUC Média	0,7958	**0,8543**	-	-
F1-Score Macro	0,2829	**0,3912**	-	-

A. Desempenho Geral Comparativo

A Tabela I apresenta as métricas de desempenho agregado para ambos os modelos, destacando o impacto positivo do treinamento com o conjunto de dados completo. A AUC Média, que mede a capacidade discriminativa geral do modelo, aumentou de 0,7958 para um valor robusto de 0,8543. A análise dos valores de AUC por classe na tabela corrobora a forte capacidade do modelo final em distinguir entre ex- ames normais e anormais para a maioria das patologias, com destaque para Hérnia (0,999), Massa (0,943) e Pneumotórax (0,934). Mais notavelmente, o F1-Score Macro, nossa métrica principal por ser sensível ao desempenho nas classes raras, apresentou um salto de 0,2829 para 0,3912, o que representa uma melhora relativa de 38,3%. Este ganho significativo evidencia a eficácia da escala de dados como estratégia para melhorar o equilíbrio entre precisão e revocação em um cenário de classes desbalanceadas.

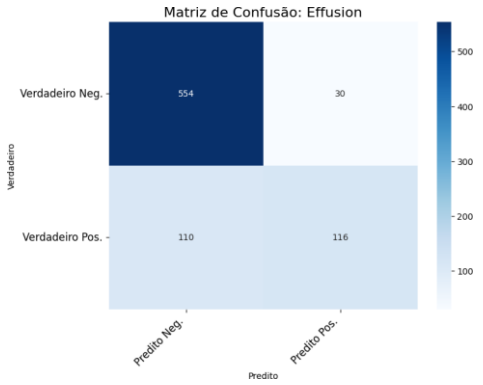

Fig. 3: Matriz de confusão para Derrame Pleural. Sendo uma classe com dados abundantes no teste (N=226), o modelo demonstra um desempenho balanceado, identificando corretamente 116 casos positivos (TP) com uma precisão de 79,5%.

B. Análise Qualitativa e Casos Críticos

A análise da Tabela I revela que o aumento no volume de dados foi particularmente benéfico para as patologias de média e baixa prevalência. Condições como Massa, Nódulo e Consolidação apresentaram ganhos expressivos no F1-Score. O caso mais notável foi o da Hérnia, uma das patologias mais raras, cujo F1-Score saltou de 0,000 para 0,667, com uma AUC quase perfeita de 0,999. Para ilustrar visualmente o comportamento do modelo final, a Figura 3 apresenta as matrizes de confusão para três cenários distintos: uma patologia com bom desempenho (Derrame), o caso de melhora mais expressiva (Hérnia) e uma patologia que permaneceu um desafio (Pneumonia). Apesar da melhora geral, patologias de sinal visual sutil e com poucos exemplos, como Pneumonia e Edema, continuaram a apresentar um F1-Score nulo, indicando que, para estas condições, apenas o aumento do volume de dados pode não ser suficiente, um ponto que será aprofundado na Discussão.

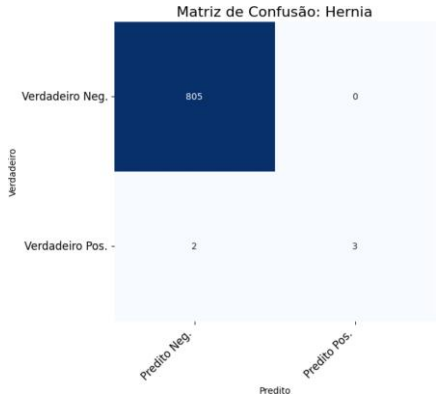

Fig. 4: Matriz de confusão para Hérnia. Apesar de ser extremamente rara (N=5), a escala de dados permitiu ao modelo identificar 3 dos 5 casos (TP), alcançando um F1-Score de 0,667 sem cometer erros de falso positivo (FP=0).

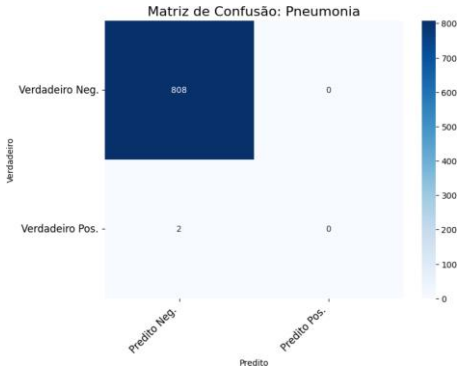

Fig. 5: Matriz de confusão para Pneumonia. Com apenas 2 casos positivos no teste, a classe ilustra os limites da abordagem. O modelo não identificou nenhum caso (TP=0), um resultado direto da insuficiência de dados para esta condição de sinal sutil.

V. Discussão

Os resultados apresentados demonstram de forma conclusiva o papel crítico da escala do conjunto de dados no treinamento de modelos para diagnóstico em radiografias de tórax. A transição de uma abordagem baseada em uma amostra para o treinamento no conjunto completo de mais de 100.000 imagens não representou apenas uma melhora incremental, mas uma mudança qualitativa na capacidade do modelo, especialmente na detecção das patologias de cauda longa (long-tail), que constituem o desafio central neste domínio. O salto de 38,3% no F1-Score Macro, de 0,2829 para 0,3912, é o coração quantitativo do artigo. A escolha desta métrica é estratégica, pois, ao contrário da AUC Média, o F1-Score Macro atribui peso igual a todas as classes, tornando-o extremamente sensível ao desempenho nas patologias raras. A melhora expressiva nesta métrica conta uma história clara: o modelo base já era razoável nas patologias comuns, mas foi a escala massiva de dados que o tornou competente na identificação das minoritárias, quebrando o viés em direção à maioria. O resultado mais espetacular – o F1-Score para Hérnia saltando de 0,000 para 0,667 com uma AUC de 0,999 – exige uma interpretação mais profunda. Este fenômeno pode ser explicado pela hipótese do "especialista em normalidade". Ao ser treinado com mais de 100.000 imagens, a maioria representando uma anatomia normal, o modelo foi forçado a aprender uma representação interna extremamente robusta do que constitui um tórax padrão. Uma hérnia diafragmática, embora rara, é uma anomalia anatômica gritante. Para um modelo que se tornou "especialista em normalidade", essa anomalia se apresenta como um outlier de alto contraste, um desvio radical da norma. Portanto, o modelo não está apenas "reconhecendo uma hérnia", mas sim identificando com altíssima confiança um sinal que é "definitivamente não normal" de uma maneira muito específica, o que explica a AUC quase perfeita. Este comportamento do modelo espelha, de certa forma, o processo diagnóstico humano, onde uma anomalia anatômica grosseira se destaca imediatamente para um especialista que está profundamente familiarizado com a vasta gama de aparências da anatomia normal. A honestidade científica exige a análise não apenas dos sucessos, mas também das limitações. O fato de patologias como Pneumonia e Edema terem mantido um F1-Score nulo, mesmo com o conjunto de dados completo, não é uma falha do estudo, mas um achado científico crucial. Em contraste com a Hérnia, estas condições manifestam-se como padrões radiológicos sutis, difusos e com alta sobreposição de características com outras condições. Isso sugere uma conclusão mais nuançada: a escala de dados é uma condição necessária para superar o viés de classe, especialmente para anomalias claras, mas pode não ser suficiente para patologias de baixo sinal visual e alta ambiguidade de rótulo. Isso sugere que, para este subconjunto de patologias de sinal fraco, a tarefa de classificação binária (presença/ausência) pode ser inerentemente mal definida. Abordagens futuras poderiam se beneficiar de tarefas mais complexas, como a localização explícita dos achados (via caixas delimitadoras ou segmentação), para forçar o modelo a aprender representações visuais mais específicas e menos ambíguas. Por fim, este estudo reforça uma lição metodológica: a importância de validar modelos de IA contra um padrão-ouro definido por especialistas. Ao avaliar nosso modelo no conjunto de 810 imagens rotuladas por radiologistas, garantimos que nossas conclusões não fossem infladas por imprecisões nos rótulos do conjunto de treinamento, conferindo maior robustez científica e relevância clínica aos nossos achados.

A. *Limitações do Estudo*

Apesar dos resultados, este estudo possui limitações. A análise utilizou um único conjunto de dados público e uma única arquitetura de modelo. O conjunto de teste de especialistas, apesar de sua alta qualidade, provém da mesma população de pacientes que o conjunto de treinamento, e a validação em um conjunto de dados externo seria um passo importante para verificar a generalização do modelo.

B. *Trabalhos Futuros*

Como trabalhos futuros, os resultados abrem caminhos de alto impacto. Uma investigação crucial seria testar a sinergia entre a escala de dados e algoritmos mais sofisticados, aplicando uma função de perda como a Focal Loss ao modelo já treinado nos dados completos para tentar resolver o desafio das classes mais sutis como Pneumonia. Adicionalmente, para essas classes de baixo desempenho, uma linha de pesquisa promissora seria o ajuste fino (fine-tuning) do modelo em um subconjunto de dados menor, mas com rótulos meticulosamente anotados por especialistas, para investigar se a qualidade do rótulo é o gargalo final.

VI. Conclusão

Este estudo concluiu com sucesso a investigação da aplicação da arquitetura DenseNet121, aprimorada com mecanismos de atenção (CBAM), para a complexa tarefa de classificação de catorze patologias torácicas. A jornada experimental demonstrou que, embora a escolha de uma arquitetura

potente seja um pré-requisito, a estratégia mais impactante para superar o desafio do severo desbalanceamento de classes foi a escala do treinamento para um conjunto de dados com mais de 100.000 imagens. Esta abordagem resultou em uma melhora de 38,3% no F1-Score Macro, alcançando 0,3912, e uma robusta AUC Média de 0,8543, com o desempenho val- idado rigorosamente contra um padrão-ouro de especialistas. A análise aprofundada revelou um comportamento complexo do modelo: enquanto a escala de dados permitiu a detecção quase perfeita de anomalias raras e estruturalmente distintas como a Hérnia (AUC de 0,999), ela não foi suficiente para resolver o desafio de patologias de sinal sutil e ambíguo, como a Pneumonia, que permaneceu indetectável. Portanto, conclui- se que o desenvolvimento de sistemas de IA clinicamente úteis para radiologia não reside em uma solução única, mas em uma combinação sinérgica: uma arquitetura de modelo capaz (DenseNet121-CBAM), um volume de dados massivo para permitir o aprendizado de padrões complexos e raros, e uma metodologia de avaliação rigorosa contra a perícia humana para quantificar honestamente tanto as capacidades quanto os limites da tecnologia..

AGRADECIMENTOS

Os autores desejam agradecer a Universidade Federal de Itajubá (UNIFEI) por toda a infraestrutura de pesquisa disponibilizada e as agências de fomentos, CAPES, CNPq e FAPEMIG pelo apoio financeiro na realização das atividades.

REFERÊNCIAS

[1] R. O'Connor, L. Retelling, and S. S. Maharjan, "X-ray Chest," in *StatPearls [Internet]*. Treasure Island (FL): StatPearls Publishing, Jan. 2024. [Updated 2023 Aug 8]. PMID: 30855897.

[2] WORLD HEALTH ORGANIZATION, *Manual of Diagnostic Ultrasound Volume 2*. Geneva: WHO, 2016. ISBN: 978-92-4-154954-3.

[3] M. P. Callaway et al., "The role of chest radiography in sus- pected community-acquired pneumonia," *Journal of the Royal So- ciety of Medicine*, vol. 108, no. 6, pp. 232-235, 2015. DOI: 10.1177/0141076814565003.

[4] A. P. Brady and L. G. G. L. O. Lao, "Error in radiology: a complex entity," *Clinical Radiology*, vol. 72, no. 1, pp. 1-8, 2017. DOI: 10.1016/j.crad.2016.08.005.

[5] J. D. Godwin et al., "Observer variation in the detection of asbestos-related pleuropulmonary disease," *Radiology*, vol. 184, no. 1, pp. 139-144, 1992. DOI: 10.1148/radiology.184.1.1609067.

[6] L. Berlin, "Hindsight Bias," *American Journal of Roentgenology*, vol. 188, no. 5, pp. 1180-1183, 2007. DOI: 10.2214/AJR.06.1168.

[7] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, vol. 18, no. 8, pp. 500-510, 2018. DOI: 10.1038/s41568-018-0016-5.

[8] P. Rajpurkar et al., "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS Medicine*, vol. 15, no. 11, p. e1002686, 2018. DOI: 10.1371/journal.pmed.1002686.

[9] G. Choy et al., "Current Applications and Future Impact of Machine Learning in Radiology," *Radiology*, vol. 288, no. 2, pp. 318-328, 2018. DOI: 10.1148/radiol.2018171820.

[10] X. Wang et al., "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2097-2106. DOI: 10.1109/CVPR.2017.369.

[11] P. Rajpurkar et al., "CheXNet: Radiologist-Level Pneumonia De- tection on Chest X-Rays with Deep Learning," *arXiv preprint arXiv:1711.05225*, 2017. DOI: 10.48550/arXiv.1711.05225.

[12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4700-4708. DOI: 10.1109/CVPR.2017.243.

[13] T. Zhou, S. Ruan, and S. Canu, "Dense Convolutional Network and Its Application in Medical Image Analysis," *Computational Intelligence and Neuroscience*, vol. 2022, p. 6478215, 2022. DOI: 10.1155/2022/6478215.

[14] Z. Nabulsi et al., "Deep Learning for Distinguishing Normal versus Abnormal Chest Radiographs and Generalization to Two Unseen Diseases Tuberculosis and COVID-19," *Scientific Reports*, vol. 11, no. 1, p. 14483, 2021. DOI: 10.1038/s41598-021-93964-9.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *in Proc. IEEE Int. Conf.*Comput. Vis. (ICCV), 2017, pp. 2999–3007. DOI: 10.1109/ICCV.2017.324.