UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

CURSO DE BACHARELADO EM ENGENHARIA ELÉTRICA

MIGUEL FELIPE DE ALMEIDA

CLASSIFICAÇÃO EM TRÊS CLASSES (NORMAL, BACTERIANA E VIRAL) EM RADIOGRAFIAS DE TÓRAX USANDO MOBILENETV2 E GRAD-CAM

SÃO CARLOS

2026

MIGUEL FELIPE DE ALMEIDA

CLASSIFICAÇÃO EM TRÊS CLASSES (NORMAL, BACTERIANA E VIRAL) EM RADIOGRAFIAS DE TÓRAX USANDO MOBILENETV2 E GRAD-CAM

Trabalho de Conclusão de Curso apresentado ao Departamento de Engenharia Elétrica da Universidade Federal de São Carlos, como parte dos requisitos para obtenção do título de Bacharel em Engenharia Elétrica.

Orientador: Prof. Dr. Celso Aparecido de França

SÃO CARLOS

2026

Dedico este trabalho à memória de meu tio Luis Felipe e de meu primo Lucas Tadeu.

AGRADECIMENTOS

Primeiramente, agradeço a Deus por todas as vitórias, conquistas e proteção que sempre me proporcionou ao longo da minha vida.

Agradeço também à minha família, especialmente aos meus pais, Sr. Adirson e Sra. Isabel, por todo amor, apoio, dedicação e por sempre me proporcionarem as condições necessárias para seguir em frente. Aos meus irmãos, Vitor e Rafael, agradeço por estarem sempre ao meu lado. Estendo também minha gratidão ao meu tio João e à Dona Cida, que sempre cuidaram de mim com carinho e atenção, e aos meus avós Ruth, Vitória, João e Mário, por todo amor, cuidado e presença ao longo da minha vida.

À minha namorada, Elaine Martins, agradeço por estar comigo em todos os momentos, inclusive nos mais difíceis, sempre me apoiando, incentivando e acreditando na minha capacidade de alcançar meus objetivos.

Agradeço também a todos os meus amigos e colegas pelos momentos compartilhados ao longo dessa trajetória, desde os períodos de estudo antes das provas e os momentos de sufoco, até as alegrias vividas nos laboratórios e na convivência diária. Desejo a todos vocês o melhor e serei sempre muito grato por todas as memórias construídas.

À Projetae, ao meu chefe, supervisores e colegas de trabalho, agradeço pela confiança depositada em mim e pela grande oportunidade de fazer parte dessa empresa, que contribuiu significativamente para o meu crescimento profissional e pessoal.

Por fim, agradeço a todos os professores da UFSCar que contribuíram para a minha formação acadêmica. Em especial, agradeço aos professores Marcelo Suetake e Tatiana Pazelli, e ao meu orientador, Prof. Dr. Celso Aparecido de França, pelos ensinamentos, orientações e exemplos de dedicação, profissionalismo e humanidade.

"As grandes coisas também nascem de
pequenos começos."

Públio Siro

RESUMO

A pneumonia é uma doença respiratória de grande relevância clínica e epidemiológica, cuja investigação inicial frequentemente envolve a análise de radiografias de tórax. Entretanto, a interpretação dessas imagens pode ser desafiadora devido à presença de padrões radiográficos sutis, sobreposição de estruturas anatômicas, variações de qualidade da imagem e semelhanças entre diferentes manifestações pulmonares. Nesse contexto, técnicas de aprendizado profundo têm sido amplamente investigadas como ferramentas de apoio à análise de imagens médicas. Este trabalho teve como objetivo desenvolver e avaliar um sistema de classificação automática de radiografias de tórax em três classes: Normal, Pneumonia Bacteriana e Pneumonia Viral, utilizando aprendizado por transferência com a MobileNetV2, uma arquitetura de rede neural convolucional leve, e análise visual das predições por meio do Grad-CAM, técnica de interpretabilidade baseada em mapas de ativação. Para isso, foi utilizada uma base pública de radiografias de tórax, reorganizada em três categorias a partir da estrutura e da nomenclatura dos arquivos. As imagens foram divididas em conjuntos de treinamento, validação e teste, redimensionadas para o formato de entrada da MobileNetV2 e submetidas a um fluxo de pré-processamento compatível com a arquitetura adotada. O treinamento foi conduzido em duas etapas: inicialmente com a base convolucional congelada e, posteriormente, com ajuste fino parcial das camadas finais da rede. O desempenho do modelo foi avaliado por meio de acurácia, precisão, revocação, F1-score, matriz de confusão e análise complementar com limiar calibrado. No conjunto de teste, o modelo MobileNetV2 obteve acurácia de 80,13%, com melhor desempenho para a classe Pneumonia Bacteriana e maior dificuldade na distinção entre as classes Normal e Pneumonia Viral. A análise complementar com limiar calibrado elevou a acurácia para 80,45%, representando um ganho discreto em relação ao modelo base. Além disso, os mapas Grad-CAM permitiram observar as regiões da imagem que mais influenciaram as decisões da rede, contribuindo para uma análise qualitativa das predições. Os resultados indicam que a abordagem proposta é tecnicamente viável para apoio à classificação multiclasse de radiografias de tórax, combinando desempenho satisfatório, baixo custo computacional e interpretabilidade visual. Contudo, o modelo não deve ser interpretado como ferramenta diagnóstica autônoma, uma vez que apresenta limitações relacionadas à base de dados, à ausência de validação clínica externa e à dificuldade inerente de separação entre padrões radiográficos semelhantes.

Palavras-chave: pneumonia; radiografia de tórax; aprendizado profundo; MobileNetV2; Grad-CAM.

ABSTRACT

Pneumonia is a respiratory disease of significant clinical and epidemiological relevance, whose initial investigation often involves the analysis of chest radiographs. However, the interpretation of these images can be challenging due to subtle radiographic patterns, overlapping anatomical structures, variations in image quality, and similarities among different pulmonary manifestations. In this context, deep learning techniques have been widely investigated as support tools for medical image analysis. This work aimed to develop and evaluate an automatic system for classifying chest X-rays into three classes: Normal, Bacterial Pneumonia, and Viral Pneumonia, using transfer learning with MobileNetV2, a lightweight convolutional neural network architecture, and visual analysis of predictions through Grad-CAM, an interpretability technique based on activation maps. For this purpose, a public chest X-ray dataset was used and reorganized into three categories based on the original file structure and filename patterns. The images were divided into training, validation, and test sets, resized to the MobileNetV2 input format, and submitted to a preprocessing workflow compatible with the adopted architecture. The training process was conducted in two stages: initially with the convolutional base frozen and, subsequently, with partial fine-tuning of the final layers of the network. The model performance was evaluated using accuracy, precision, recall, F1-score, confusion matrix, and a complementary analysis with a calibrated decision threshold. On the test set, the MobileNetV2 model achieved an accuracy of 80.13%, with the best performance observed for the Bacterial Pneumonia class and greater difficulty in distinguishing between the Normal and Viral Pneumonia classes. The complementary analysis with the calibrated threshold increased the accuracy to 80.45%, representing a slight improvement over the baseline model. In addition, the Grad-CAM maps made it possible to observe the image regions that most influenced the network decisions, contributing to a qualitative analysis of the predictions. The results indicate that the proposed approach is technically feasible as a support method for multiclass chest X-ray classification, combining satisfactory performance, low computational cost, and visual interpretability. However, the model should not be interpreted as an autonomous diagnostic tool, since it presents limitations related to the dataset, the absence of external clinical validation, and the inherent difficulty of separating visually similar radiographic patterns.

Keywords: pneumonia; chest X-ray; deep learning; MobileNetV2; Grad-CAM.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas.
CNN	*Convolutional Neural Network* – Rede Neural Convolucional.
CSV	*Comma-Separated Values* – Valores separados por vírgula.
F1-score	Métrica de desempenho que combina precisão e revocação em um único valor.
GPU	*Graphics Processing Unit* – Unidade de Processamento Gráfico.
Grad-CAM	*Gradient-weighted Class Activation Mapping* – Mapeamento de ativação de classe ponderado por gradientes.
IoU	*Intersection over Union* – Interseção sobre união.
RAM	*Random-Access Memory* – Memória de acesso aleatório.
ReLU	*Rectified Linear Unit* – Unidade Linear Retificada.
RGB	*Red, Green and Blue* – Vermelho, verde e azul.
TCC	Trabalho de Conclusão de Curso.
UFSCar	Universidade Federal de São Carlos.

LISTA DE SÍMBOLOS

A_{ij}^{k}	Ativação do canal k na posição espacial (i, j).
c	Classe de interesse considerada no cálculo do Grad-CAM.
FN	Falso negativo.
FP	Falso positivo.
(i, j)	Coordenadas espaciais no mapa de características.
k	Índice do mapa de características.
$L_{\text{Grad-CAM}}^{c}$	Mapa de ativação Grad-CAM gerado para a classe c.
t	Limiar de decisão calibrado utilizado na análise complementar.
TN	Verdadeiro negativo.
TP	Verdadeiro positivo.
y^{c}	Pontuação associada à classe de interesse c.
Z	Número total de posições espaciais do mapa de características.
α_{k}^{c}	Coeficiente de importância do mapa de características k para a classe c.

SUMÁRIO

1 INTRODUÇÃO

Este capítulo apresenta uma visão geral do tema desenvolvido neste trabalho, destacando a relevância da pneumonia como problema de saúde pública e o potencial das técnicas de aprendizado profundo aplicadas à análise de radiografias de tórax. Além disso, são apresentadas a contextualização do problema, a justificativa da pesquisa, os objetivos do estudo e a organização da monografia.

1.1 Contextualização

As doenças respiratórias representam um importante desafio para os sistemas de saúde, especialmente em razão de sua elevada incidência e de seu impacto sobre diferentes faixas etárias. Entre essas doenças, a pneumonia destaca-se por sua relevância clínica e epidemiológica, podendo causar quadros de gravidade variável e demandar diagnóstico e tratamento adequados (World Health Organization, 2025).

Na prática clínica, a investigação inicial da pneumonia costuma ser apoiada pela análise de radiografias de tórax, uma vez que esse exame é amplamente disponível, rápido e de menor custo quando comparado a métodos de imagem mais complexos. No entanto, a interpretação dessas imagens pode ser desafiadora, pois depende da experiência do avaliador, da qualidade do exame e da presença de padrões radiográficos sutis ou sobrepostos. Essa dificuldade torna-se ainda mais evidente quando se busca distinguir imagens normais de casos associados à pneumonia bacteriana ou viral (National Heart, Lung, and Blood Institute, 2022a; MSD Manual Professional Edition, 2024b).

Nesse contexto, o avanço das técnicas de aprendizado profundo, especialmente das Redes Neurais Convolucionais (*Convolutional Neural Networks* – CNNs), tem possibilitado o desenvolvimento de sistemas computacionais capazes de aprender padrões visuais diretamente a partir das imagens. Esses modelos vêm sendo amplamente explorados em aplicações de análise de imagens médicas, incluindo tarefas de detecção e classificação de achados em radiografias de tórax (LITJENS et al., 2017; CALLI et al., 2021).

Assim, a classificação automática de radiografias de tórax surge como uma abordagem promissora para auxiliar a análise de exames, fornecendo uma avaliação computacional complementar. No caso deste trabalho, o foco está na classificação multiclasse entre radiografias *Normais*, casos de *Pneumonia Bacteriana* e casos de *Pneumonia Viral*, utilizando a arquitetura MobileNetV2 com aprendizado por transferência e interpretação visual por meio do Grad-CAM.

1.2 Justificativa

Este trabalho justifica-se pela necessidade de investigar métodos computacionais capazes de auxiliar, de forma padronizada e objetiva, a análise de radiografias de tórax. Embora a radiografia seja amplamente utilizada na avaliação inicial de doenças pulmonares, sua interpretação pode apresentar variabilidade, especialmente em casos com achados discretos ou padrões visualmente semelhantes. No contexto da pneumonia, a distinção entre imagens normais, pneumonia bacteriana e pneumonia viral representa uma tarefa desafiadora, mas relevante para o desenvolvimento de sistemas de apoio à decisão (GARG et al., 2019; FRANQUET, 2001).

A classificação multiclasse de radiografias de tórax apresenta interesse por ir além da simples identificação da presença ou ausência de pneumonia. Ao buscar diferenciar entre três categorias, o modelo pode fornecer uma saída mais informativa e contribuir para uma análise computacional mais detalhada do exame. Essa abordagem é coerente com o avanço recente de aplicações baseadas em redes neurais convolucionais em imagens médicas, nas quais o aprendizado profundo tem demonstrado potencial para extrair automaticamente características relevantes e identificar padrões complexos (LITJENS et al., 2017; CALLI et al., 2021).

Do ponto de vista metodológico, o uso de aprendizado por transferência também reforça a pertinência deste estudo. Essa estratégia permite utilizar uma rede previamente treinada em grandes bases de imagens e adaptá-la a uma tarefa específica, reduzindo o custo computacional e favorecendo o treinamento em cenários nos quais a quantidade de dados rotulados é limitada. Em imagens médicas, essa característica é especialmente importante, pois a obtenção de bases extensas e bem anotadas

depende de processos especializados (MORID; BORJALI; FIOL, 2021).

Nesse cenário, a arquitetura MobileNetV2 mostra-se adequada à proposta deste trabalho por combinar bom desempenho com menor custo computacional. Por se tratar de uma rede neural convolucional leve, baseada em blocos do tipo *inverted residual* e *linear bottleneck*, sua utilização permite explorar uma solução eficiente para classificação de imagens, mantendo uma estrutura compatível com aplicações que demandam menor uso de memória e processamento (SANDLER et al., 2018).

Além do desempenho preditivo, este trabalho também considera a importância da interpretabilidade em modelos de aprendizado profundo aplicados a imagens médicas. Para isso, utiliza-se o Grad-CAM como ferramenta complementar de análise qualitativa, permitindo visualizar regiões da radiografia que mais influenciaram a decisão do modelo. Essa etapa é relevante porque contribui para uma avaliação mais crítica das predições, especialmente em um contexto no qual a transparência das decisões é fundamental (SELVARAJU et al., 2017; SAPORTA et al., 2022).

Dessa forma, o presente estudo busca contribuir para a investigação de uma abordagem computacional baseada em MobileNetV2, aprendizado por transferência e Grad-CAM para a classificação multiclasse de radiografias de tórax. A proposta reúne relevância social, por tratar de um problema associado à saúde pública, e relevância técnica, por aplicar conceitos de aprendizado profundo e interpretabilidade em um problema de visão computacional médica.

1.3 Objetivos

1.3.1 Objetivo Geral

Desenvolver e avaliar um sistema de classificação automática de radiografias de tórax em três classes: *Normal*, *Pneumonia Bacteriana* e *Pneumonia Viral*, utilizando aprendizado por transferência com a arquitetura MobileNetV2 e interpretação visual das predições por meio do Grad-CAM.

1.3.2 Objetivos Específicos

- Organizar e preparar a base de dados de radiografias de tórax, estruturando-a em conjuntos de treinamento, validação e teste;
- Realizar o pré-processamento das imagens, incluindo redimensionamento, normalização e adequação ao formato de entrada da MobileNetV2;
- Desenvolver um modelo de classificação multiclasse a partir da MobileNetV2 pré-treinada, adicionando camadas de classificação adequadas ao problema proposto;
- Incorporar técnicas de aumento de dados (*data augmentation*) ao processo de treinamento, com o objetivo de ampliar a variabilidade das amostras e reduzir a tendência ao sobreajuste;
- Definir e executar uma estratégia de treinamento baseada em aprendizado por transferência, contemplando o congelamento inicial da rede base e o ajuste fino parcial (*fine-tuning*) das camadas finais;
- Avaliar o desempenho do modelo por meio de métricas de classificação, como acurácia, precisão, revocação e F1-score, incluindo análise por classe e matriz de confusão;
- Estabelecer como referência experimental de desempenho uma acurácia mínima próxima de 80% no conjunto de teste, utilizando essa referência como critério auxiliar para avaliar a viabilidade técnica da abordagem proposta;
- Gerar e analisar mapas de ativação Grad-CAM para a interpretação qualitativa das predições realizadas no conjunto de teste;

Dessa forma, os objetivos definidos contemplam desde a preparação da base de dados até a avaliação quantitativa e qualitativa do modelo, permitindo analisar não apenas o desempenho da MobileNetV2, mas também o comportamento das predições e a interpretabilidade visual das decisões por meio do Grad-CAM (SELVARAJU et al., 2017).

1.4 Estrutura do Trabalho

O presente trabalho está organizado em cinco capítulos. No **Capítulo 1**, são apresentados a contextualização do problema, a justificativa da pesquisa, os objetivos do estudo e a organização geral da monografia.

O **Capítulo 2** apresenta a fundamentação teórica necessária ao desenvolvimento do trabalho. Inicialmente, são discutidos aspectos relacionados às doenças respiratórias, à pneumonia e ao uso de radiografias de tórax como ferramenta diagnóstica. Em seguida, são abordados os trabalhos correlatos e os principais conceitos de aprendizado profundo utilizados na pesquisa, incluindo Redes Neurais Convolucionais, aprendizado por transferência, arquitetura MobileNetV2 e interpretabilidade por Grad-CAM.

No **Capítulo 3**, são descritos os procedimentos metodológicos e as etapas de desenvolvimento do sistema proposto. Esse capítulo contempla a descrição da base de dados, o pré-processamento das imagens, a configuração do modelo, a estratégia de treinamento, o protocolo de avaliação e a aplicação do Grad-CAM para análise qualitativa das predições.

O **Capítulo 4** apresenta e discute os resultados obtidos. São analisadas as métricas de desempenho no conjunto de teste, as curvas de treinamento e validação, a matriz de confusão, a análise complementar com limiar calibrado, os mapas de ativação gerados pelo Grad-CAM e a comparação com trabalhos relacionados.

Por fim, o **Capítulo 5** apresenta as considerações finais do trabalho, retomando as principais conclusões, as limitações identificadas e as possibilidades de aprimoramento em pesquisas futuras.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são apresentados os principais conceitos e fundamentos que sustentam o desenvolvimento deste trabalho. Inicialmente, discute-se a pneumonia e sua análise por meio de radiografias de tórax, destacando os aspectos mais relevantes para a proposta de classificação automática. Em seguida, são abordados trabalhos correlatos e os fundamentos teóricos relacionados ao uso de aprendizado profundo, com ênfase em Redes Neurais Convolucionais, *transfer learning*, arquitetura MobileNetV2 e técnicas de interpretabilidade aplicadas a imagens médicas.

2.1 O Problema Estudado

O problema abordado neste trabalho está relacionado à classificação automática de radiografias de tórax, com foco na identificação de imagens normais e de casos associados à pneumonia bacteriana ou viral. Essa tarefa é relevante porque a pneumonia pode apresentar manifestações radiográficas variadas, e sua interpretação depende de fatores como qualidade da imagem, experiência do avaliador e características clínicas do paciente.

Embora a radiografia de tórax seja um exame amplamente utilizado na investigação de doenças respiratórias, a distinção entre diferentes padrões pulmonares nem sempre é simples. Alterações sutis, sobreposição de estruturas anatômicas e semelhanças entre algumas condições podem dificultar a análise visual, especialmente quando se busca diferenciar casos normais de padrões associados à pneumonia.

Nesse contexto, métodos baseados em aprendizado profundo surgem como uma alternativa de apoio à análise dessas imagens, permitindo que modelos computacionais aprendam padrões visuais diretamente a partir dos dados. Assim, antes de apresentar as técnicas utilizadas, esta seção discute brevemente as doenças respiratórias e a pneumonia, destacando sua relação com a análise de radiografias de tórax e com o problema de classificação estudado.

2.1.1 Doenças Respiratórias e Pneumonia

As doenças respiratórias compreendem um amplo conjunto de condições que afetam as vias aéreas e o parênquima pulmonar, podendo comprometer a ventilação, as trocas gasosas e a oxigenação do organismo. Entre essas condições, a pneumonia destaca-se por sua relevância clínica e epidemiológica, sendo reconhecida como um importante problema de saúde pública em diferentes faixas etárias e contextos assistenciais (World Health Organization, 2025).

De forma geral, a pneumonia pode ser definida como um processo infeccioso que acomete os pulmões e provoca inflamação dos alvéolos, estruturas responsáveis pelas trocas gasosas. Em decorrência da infecção, esses espaços podem se preencher com líquido ou pus, dificultando a passagem adequada de oxigênio e favorecendo o surgimento de manifestações como tosse, febre, dor torácica, fadiga e falta de ar (World Health Organization, 2025; National Heart, Lung, and Blood Institute, 2022b). A gravidade do quadro pode variar de acordo com fatores como idade, presença de comorbidades, agente etiológico envolvido e rapidez no diagnóstico e no início do tratamento (MSD Manual Professional Edition, 2024b).

Do ponto de vista etiológico, a pneumonia pode ser causada por diferentes microrganismos, incluindo bactérias, vírus e fungos (MSD Manual Professional Edition, 2024b; National Heart, Lung, and Blood Institute, 2022b). No contexto deste trabalho, assumem maior relevância as pneumonias de origem bacteriana e viral, pois ambas podem produzir alterações visíveis em radiografias de tórax, mas nem sempre são facilmente diferenciáveis apenas pela inspeção visual. Em termos clínicos e radiológicos, pode haver sobreposição entre padrões de apresentação, o que torna a interpretação mais desafiadora e reforça a necessidade de análise especializada (MSD Manual Professional Edition, 2024a; The Radiology Assistant, 2014). Embora alguns achados radiográficos possam sugerir determinadas etiologias, a radiografia de tórax, de forma isolada, geralmente não é suficiente para determinar com segurança o agente causador da infecção (MSD Manual Professional Edition, 2024a). A Tabela 1 apresenta uma síntese das principais categorias de pneumonia consideradas quanto ao agente etiológico.

Tabela 1: Classificação geral da pneumonia quanto ao agente etiológico.

Categoria	**Descrição geral**
Pneumonia bacteriana	Associada à infecção por bactérias, podendo apresentar consolidações pulmonares e evolução clínica variável.
Pneumonia viral	Relacionada à infecção por vírus respiratórios, com padrões clínicos e radiológicos que podem se sobrepor aos de outras etiologias.
Pneumonia fúngica	Menos frequente, ocorrendo com maior relevância em pacientes imunocomprometidos ou em contextos específicos.

Fonte: Elaborada pelo autor com base em (MSD Manual Professional Edition, 2024b; National Heart, Lung, and Blood Institute, 2022b).

2.1.2 Pneumonia Bacteriana e Pneumonia Viral

A pneumonia é uma infecção que acomete o parênquima pulmonar e compromete a função respiratória por meio de alterações inflamatórias nos alvéolos e no interstício pulmonar. Entre seus agentes etiológicos mais frequentes, destacam-se bactérias e vírus, responsáveis por quadros que podem apresentar manifestações clínicas semelhantes, mas que diferem, em certa medida, quanto aos padrões de imagem observados e à conduta terapêutica adotada (World Health Organization, 2025; Centers for Disease Control and Prevention, 2026; SATTAR; SHARMA, 2024; FREEMAN et al., 2023).

No contexto deste trabalho, a principal relevância dessa distinção está na análise de radiografias de tórax. De modo geral, a pneumonia bacteriana tende a estar associada a opacidades mais focais, consolidações segmentares ou lobares e, em muitos casos, à presença de broncograma aéreo (FRANQUET, 2001; GARG et al., 2019; The Radiology Assistant, 2014). Já a pneumonia viral costuma estar relacionada a padrões mais difusos, bilaterais ou intersticiais, podendo também apresentar acometimento multifocal e, em determinados casos, áreas em vidro fosco (KOO et al., 2020; FRANQUET, 2001).

Apesar dessas tendências, a diferenciação entre pneumonia bacteriana e viral não é direta. Existe sobreposição entre os padrões radiográficos, e apresentações atípicas podem ocorrer em ambas as classes. Assim, a imagem isoladamente nem sempre é

suficiente para determinar com segurança a etiologia da infecção, exigindo interpretação conjunta com informações clínicas e laboratoriais (World Health Organization, 2025; Centers for Disease Control and Prevention, 2026; SATTAR; SHARMA, 2024; FREEMAN et al., 2023; MSD Manual Professional Edition, 2024a).

Sob a perspectiva de visão computacional, esse cenário é particularmente relevante. A existência de padrões visuais recorrentes, combinada com a presença de sobreposição entre classes, torna a tarefa de classificação desafiadora e, ao mesmo tempo, adequada ao uso de técnicas de aprendizado profundo. Em outras palavras, trata-se de um problema em que há características radiográficas potencialmente discriminativas, mas cuja identificação pode ser dificultada pela variabilidade entre pacientes, pela qualidade das imagens e pela semelhança entre diferentes manifestações pulmonares.

Dessa forma, compreender as diferenças gerais e as zonas de interseção entre pneumonia bacteriana e viral é importante não apenas do ponto de vista clínico, mas também para fundamentar o desenvolvimento de sistemas automáticos capazes de classificar radiografias de tórax em categorias como normal, pneumonia bacteriana e pneumonia viral.

A Figura 1 apresenta exemplos de radiografias de tórax pertencentes às classes normal, pneumonia bacteriana e pneumonia viral, ilustrando padrões visuais frequentemente associados a cada categoria.

Figura 1: Exemplos de radiografias de tórax das classes normal, pneumonia bacteriana e pneumonia viral.

Fonte: Adaptada de Mooney (2018).

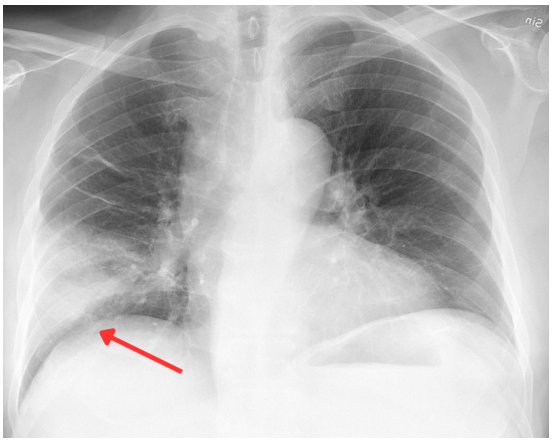
2.1.3 Radiografia de Tórax como Ferramenta Diagnóstica

A radiografia de tórax é um dos exames de imagem mais utilizados na avaliação inicial de pacientes com suspeita de pneumonia, principalmente por sua ampla disponibilidade, rapidez de aquisição e menor custo em comparação com métodos mais complexos, como a tomografia computadorizada (National Heart, Lung, and Blood Institute, 2022a; MSD Manual Professional Edition, 2024b). Na prática clínica, esse exame é empregado para identificar alterações pulmonares compatíveis com infecção, estimar a extensão do acometimento e auxiliar na detecção de possíveis complicações associadas, como derrame pleural e pneumotórax (GARG et al., 2019; The Radiology Assistant, 2014).

Além das limitações diagnósticas, o próprio método apresenta restrições técnicas relevantes. Entre elas, destacam-se a sobreposição de estruturas anatômicas, artefatos de aquisição e menor capacidade de detalhamento em determinadas regiões do tórax (The Radiology Assistant, 2014). Essas características tornam a interpretação dependente da experiência do observador e contribuem para variabilidade na análise, especialmente em casos limítrofes ou com achados sutis.

A Figura 2 apresenta um exemplo de radiografia de tórax com consolidação lobar no pulmão direito e broncograma aéreo, achados frequentemente associados à pneumonia bacteriana.

Figura 2: Radiografia de tórax com consolidação no lobo inferior direito e broncograma aéreo, achados compatíveis com pneumonia bacteriana.



Fonte: Adaptada de Radiopaedia.org (2024).

Assim, a radiografia de tórax permanece como uma ferramenta importante na investigação inicial da pneumonia, tanto pelo seu valor clínico quanto por sua relevância como base para o desenvolvimento de métodos computacionais de apoio ao diagnóstico. Ao mesmo tempo, suas limitações reforçam a necessidade de abordagens automáticas que auxiliem na interpretação das imagens e na distinção entre padrões visualmente semelhantes.

2.2 Trabalhos Correlatos

Esta seção apresenta trabalhos e abordagens relacionados ao uso de aprendizado profundo na análise de radiografias de tórax, com ênfase na classificação de pneumonia e na aplicação de redes neurais convolucionais em imagens médicas. A revisão dos estudos correlatos permite contextualizar a proposta deste trabalho em relação ao avanço recente da área, destacando o papel das bases públicas, das arquiteturas convolucionais, do aprendizado por transferência e dos métodos de interpretabilidade.

2.2.1 Aplicações de Aprendizado Profundo em Radiografias de Tórax

O avanço do aprendizado profundo ampliou significativamente o uso de radiografias de tórax em sistemas de apoio à análise de imagens médicas. Entre as abordagens mais empregadas, destacam-se as redes neurais convolucionais (*Convolutional Neural Networks* – CNNs), capazes de aprender representações diretamente a partir dos pixels da imagem, sem depender exclusivamente de descritores definidos manualmente (LITJENS et al., 2017; CALLI et al., 2021). Essa característica favoreceu sua aplicação em tarefas como detecção de anormalidades torácicas, classificação de patologias pulmonares e triagem automatizada de exames (WANG et al., 2017; IRVIN et al., 2019; ALAPAT; MENON; ASHOK, 2022).

No contexto das radiografias de tórax, a consolidação dessa área esteve associada à disponibilização de bases públicas em larga escala. O conjunto ChestX-ray8 tornou-se uma referência ao reunir grande volume de radiografias anotadas com

múltiplas condições torácicas (WANG et al., 2017). Posteriormente, bases como CheXpert contribuíram para o amadurecimento do campo ao incorporar rótulos de incerteza e protocolos de avaliação mais estruturados (IRVIN et al., 2019). Além disso, bases específicas voltadas à pneumonia, como *Chest X-Ray Images (Pneumonia)*, popularizaram estudos de classificação entre imagens normais e exames com achados compatíveis com pneumonia bacteriana e viral (MOONEY, 2018).

No caso da pneumonia, o uso de CNNs mostra-se promissor porque essa condição pode se manifestar por padrões radiográficos nem sempre triviais, especialmente em situações com opacidades discretas, acometimento multifocal ou sobreposição de estruturas anatômicas. Modelos treinados para essa tarefa procuram identificar padrões visuais associados a consolidações, infiltrados e outras alterações compatíveis com processo infeccioso, explorando relações espaciais difíceis de formalizar por regras manuais (RAJPURKAR et al., 2017; ALAPAT; MENON; ASHOK, 2022).

Um dos fatores que impulsionou essas aplicações foi o uso de aprendizado por transferência. Em vez de treinar uma rede profunda inteiramente do zero, utilizam-se pesos previamente ajustados em grandes bases de imagens, adaptando-os posteriormente ao domínio médico. Essa estratégia é especialmente útil em contextos nos quais a quantidade de exames rotulados é limitada (MORID; BORJALI; FIOL, 2021). A partir desse princípio, arquiteturas convolucionais passaram a ser amplamente exploradas em estudos com radiografias de tórax. Enquanto modelos mais complexos costumam apresentar elevada capacidade de extração de características, arquiteturas leves, como a MobileNetV2, tornam-se atrativas quando se busca menor custo computacional, menor uso de memória e menor tempo de inferência (SANDLER et al., 2018).

Outro aspecto importante é a interpretabilidade. Em aplicações médicas, além da predição, é desejável compreender quais regiões da imagem influenciaram a decisão do modelo. Nesse contexto, métodos como Grad-CAM (*Gradient-weighted Class Activation Mapping*) ganharam destaque por gerarem mapas de ativação que evidenciam áreas visualmente relevantes para a classe prevista (SELVARAJU et al., 2017). Em radiografias de tórax, esses mapas podem auxiliar na verificação qualitativa de

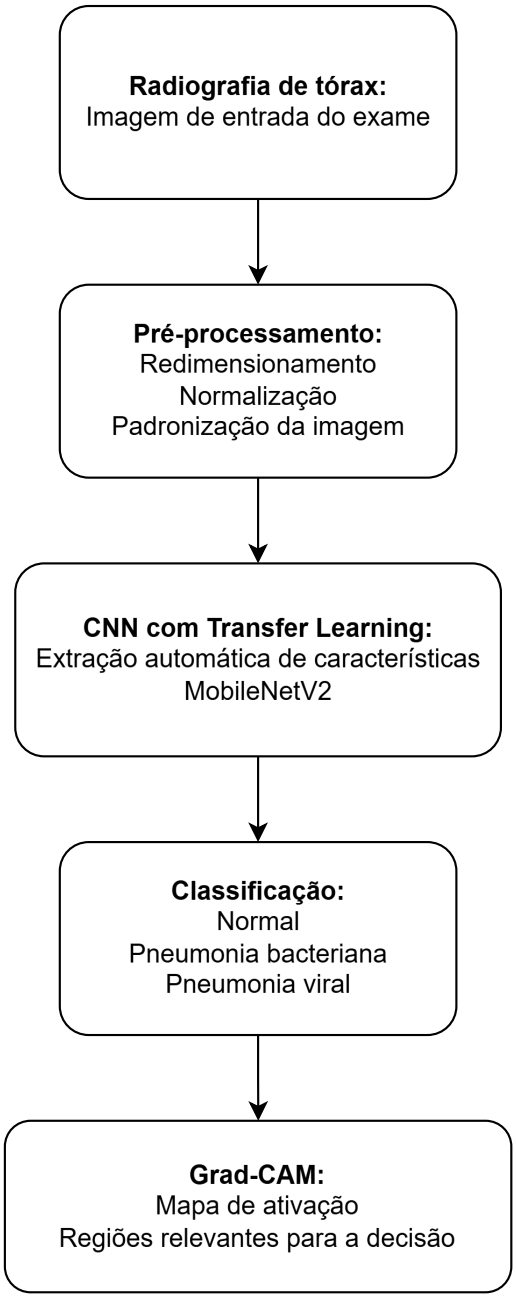
que a rede está concentrando sua atenção em regiões pulmonares compatíveis com alterações relevantes, e não em padrões espúrios da base de treinamento (SAPORTA et al., 2022).

Apesar dos resultados promissores, ainda existem limitações importantes. Entre elas, destacam-se a variabilidade da qualidade das imagens, diferenças entre equipamentos e protocolos de aquisição, ruídos na rotulagem, desbalanceamento entre classes e dificuldade de generalização entre instituições distintas (WANG et al., 2017; IRVIN et al., 2019; CALLI et al., 2021). Por esse motivo, esses sistemas devem ser compreendidos como ferramentas de apoio à decisão, e não como substitutos da análise médica especializada.

Assim, as aplicações de aprendizado profundo em radiografias de tórax configuram um campo de grande relevância para a engenharia e para a informática em saúde. A combinação entre bases públicas, arquiteturas convolucionais, estratégias de aprendizado por transferência e métodos de interpretabilidade permitiu o desenvolvimento de modelos capazes de auxiliar a triagem e a classificação de achados torácicos, incluindo pneumonia. Nesse cenário, o estudo dessas abordagens fornece a base conceitual para compreender a escolha da MobileNetV2 e do Grad-CAM no desenvolvimento deste trabalho (SANDLER et al., 2018; SELVARAJU et al., 2017).

A Figura 3 apresenta, de forma esquemática, o fluxo geral de aplicação de aprendizado profundo em radiografias de tórax, contemplando o pré-processamento da imagem, a extração automática de características por CNN, a classificação final e a geração do mapa de ativação para apoio interpretativo.

Figura 3: Fluxo geral de aplicação de aprendizado profundo em radiografias de tórax.



Fonte: Elaborada pelo autor (2026).

2.2.2 Estudos sobre Classificação de Pneumonia

A classificação automática de pneumonia em radiografias de tórax tem sido amplamente investigada na literatura, principalmente com o uso de redes neurais convolucionais, aprendizado por transferência e combinações de modelos profundos. Esse interesse decorre do potencial dessas abordagens para apoiar a triagem de exames, auxiliar a análise de imagens médicas e reduzir a dependência exclusiva da interpretação visual humana em cenários de grande volume de dados (RAJPURKAR et al., 2017; KUNDU et al., 2021; ALAPAT; MENON; ASHOK, 2022).

De modo geral, os estudos da área convergem em alguns aspectos metodológicos. O primeiro deles é a necessidade de pré-processamento adequado das imagens, incluindo redimensionamento, normalização e, em muitos casos, aumento de dados. O segundo é o uso de aprendizado por transferência como estratégia para melhorar a eficiência do treinamento. O terceiro é a adoção de métricas além da acurácia, como precisão, revocação e F1-score, permitindo uma análise mais completa do comportamento do modelo, especialmente em bases desbalanceadas (MORID; BORJALI; FIOL, 2021; KUNDU et al., 2021; ALAPAT; MENON; ASHOK, 2022).

Com o objetivo de organizar as abordagens discutidas na literatura, a Tabela 2 sintetiza estudos e referências relevantes sobre classificação de pneumonia e análise de radiografias de tórax, destacando suas contribuições e sua relação com este trabalho.

Tabela 2: Síntese de estudos relacionados à classificação de pneumonia e análise de radiografias de tórax.

Referência	**Abordagem principal**	**Contribuição do estudo**	**Relação com este trabalho**
Rajpurkar et al. (2017)	CNN profunda aplicada a radiografias de tórax	Apresenta o uso de aprendizado profundo para detecção de pneumonia em radiografias, evidenciando o potencial de CNNs em tarefas de apoio à análise médica.	Reforça a viabilidade do uso de redes convolucionais para identificação de padrões associados à pneumonia.
Kundu et al. (2021)	*Ensemble* de modelos profundos	Investiga a combinação de múltiplas arquiteturas de aprendizado profundo para melhorar a robustez da detecção de pneumonia.	Mostra que combinações de modelos podem elevar o desempenho, embora com maior custo computacional.
Alapat et al. (2022)	Revisão sobre redes neurais aplicadas à pneumonia	Discute diferentes abordagens baseadas em redes neurais para detecção de pneumonia em radiografias de tórax.	Fornece base para compreender estratégias recorrentes da literatura, como pré-processamento, CNNs e avaliação por métricas de classificação.
Morid et al. (2021)	Revisão sobre aprendizado por transferência em imagens médicas	Analisa o uso de modelos pré-treinados em bases gerais de imagens e sua adaptação para tarefas médicas.	Sustenta a escolha do aprendizado por transferência como estratégia adequada ao contexto de bases médicas limitadas.
Saporta et al. (2022)	Avaliação de métodos de saliência em radiografias de tórax	Investiga métodos de interpretabilidade visual em modelos aplicados a radiografias, destacando a importância e as limitações das explicações visuais.	Dialoga diretamente com o uso do Grad-CAM para análise qualitativa das predições do modelo.

Fonte: Elaborada pelo autor com base em Rajpurkar et al. (2017), Kundu et al. (2021), Alapat, Menon e Ashok (2022), Morid, Borjali e Fiol (2021) e Saporta et al. (2022).

A partir da Tabela 2, observa-se que Rajpurkar et al. (2017) exploraram o uso de uma CNN profunda para detecção de pneumonia em radiografias de tórax, demonstrando o potencial das redes convolucionais nesse tipo de tarefa. Kundu et al. (2021), por sua vez, investigaram uma abordagem baseada em *ensemble*, combinando diferentes modelos profundos para aumentar a robustez da classificação. Já Alapat et al. (2022) apresentaram uma revisão das principais estratégias baseadas em redes neurais aplicadas à detecção de pneumonia, destacando etapas recorrentes como pré-processamento, escolha da arquitetura e avaliação por métricas de classi-

ficação. Morid et al. (2021) discutiram especificamente o uso de aprendizado por transferência em imagens médicas, fundamentando a adaptação de modelos pré-treinados para bases com menor quantidade de dados rotulados. Por fim, Saporta et al. (2022) analisaram métodos de interpretabilidade visual em radiografias de tórax, aspecto diretamente relacionado ao uso do Grad-CAM neste trabalho.

2.2.3 Limitações e lacunas identificadas

Apesar dos avanços obtidos na classificação de pneumonia em radiografias de tórax, a literatura ainda apresenta limitações recorrentes que dificultam a aplicação prática desses modelos em ambientes clínicos reais. Entre os problemas mais frequentemente relatados, destaca-se o desbalanceamento de classes em bases públicas amplamente utilizadas, como o *Chest X-Ray Images (Pneumonia)*, nas quais a quantidade de imagens de pneumonia costuma superar a de casos normais. Essa distribuição desigual pode induzir o treinamento de modelos enviesados, favorecendo a classe majoritária e comprometendo o desempenho em cenários mais próximos da prática clínica (MOONEY, 2018; ALAPAT; MENON; ASHOK, 2022).

Outra limitação importante refere-se à generalização dos modelos. Embora diversos trabalhos relatem resultados elevados em bases de teste controladas, o desempenho tende a se reduzir quando os modelos são aplicados a imagens obtidas em contextos distintos, com variações de equipamento, protocolos de aquisição, posicionamento do paciente e características populacionais. Além disso, a forte dependência de bases anotadas por especialistas constitui um obstáculo adicional, uma vez que a disponibilidade de imagens com diagnóstico confiável ainda é limitada em muitos contextos hospitalares, especialmente em instituições com menor infraestrutura (LITJENS et al., 2017; CALLI et al., 2021; MORID; BORJALI; FIOL, 2021).

A interpretabilidade também permanece como um desafio relevante. Em muitos casos, redes neurais profundas atingem bom desempenho preditivo, porém com baixa transparência quanto às regiões da imagem que efetivamente contribuíram para a decisão do modelo. Essa característica reduz a confiabilidade do uso clínico dessas abordagens, já que profissionais da saúde tendem a exigir justificativas visuais ou

diagnósticas mais claras para apoiar a adoção de sistemas automáticos. Soma-se a isso a dificuldade de classificação em casos mais sutis ou ambíguos, como diferentes padrões de pneumonia bacteriana, viral ou atípica, cujas características radiográficas podem ser sobrepostas ou pouco evidentes, especialmente em bases com pouca diversidade e desbalanceamento entre classes (SAPORTA et al., 2022; SELVARAJU et al., 2017; ALAPAT; MENON; ASHOK, 2022).

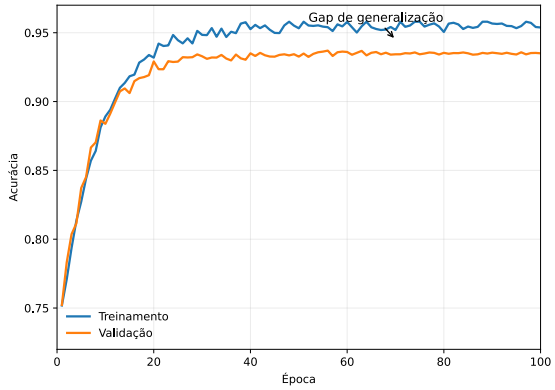
Do ponto de vista computacional, embora estratégias de *transfer learning* reduzam parte do custo de treinamento, arquiteturas mais profundas e complexas ainda podem demandar recursos computacionais elevados para ajuste fino e validação adequada. Tal aspecto limita a reprodutibilidade e a adoção desses métodos em contextos com infraestrutura restrita. Nesse sentido, observa-se na literatura a necessidade de investigações voltadas ao desenvolvimento de modelos mais leves, à integração entre imagens e dados clínicos, ao uso de estratégias que preservem a privacidade dos dados e à realização de validações multicêntricas em contextos reais. Também se destaca a importância de métricas mais alinhadas à prática clínica, como sensibilidade em casos leves, robustez entre bases distintas e tempo de inferência, em vez da análise isolada da acurácia global (MORID; BORJALI; FIOL, 2021; CALLI et al., 2021; SANDLER et al., 2018).

Diante dessas limitações, percebe-se a existência de espaço para propostas que conciliem desempenho, menor custo computacional e maior interpretabilidade. Nesse contexto, a utilização de uma arquitetura leve como a MobileNetV2, associada a uma técnica de interpretação visual como o Grad-CAM, mostra-se coerente com a busca por soluções mais eficientes e potencialmente mais adequadas a cenários de aplicação prática (SANDLER et al., 2018; SELVARAJU et al., 2017).

Entre as lacunas discutidas, a generalização merece destaque, pois modelos de aprendizado profundo podem apresentar desempenho elevado durante o treinamento, mas desempenho inferior quando avaliados em dados não vistos ou provenientes de contextos distintos. A Figura 4 ilustra, de forma conceitual, esse comportamento, no qual o desempenho obtido durante o treinamento pode ser superior ao observado em validação ou teste. Essa diferença reforça a importância de avaliar os modelos em

conjuntos independentes e, quando possível, em bases externas ao conjunto utilizado no desenvolvimento (GOODFELLOW; BENGIO; COURVILLE, 2016; LITJENS et al., 2017; CALLI et al., 2021).

Figura 4: Ilustração conceitual da evolução da acurácia em treinamento e validação ao longo das épocas, evidenciando um possível gap de generalização entre as curvas.



Fonte: Elaborada pelo autor (2026).

2.3 Fundamentos de Aprendizado Profundo

O aprendizado profundo constitui uma subárea do aprendizado de máquina baseada no uso de redes neurais artificiais com múltiplas camadas, capazes de aprender representações progressivamente mais abstratas a partir dos dados. Em problemas de visão computacional, essa abordagem tornou-se especialmente relevante por permitir que o próprio modelo extraia características diretamente das imagens, reduzindo a dependência de descritores definidos manualmente e possibilitando maior adaptação a diferentes tarefas de classificação.

No contexto deste trabalho, os fundamentos de aprendizado profundo são importantes para compreender a escolha de redes neurais convolucionais como base para a análise de radiografias de tórax. Essas redes são particularmente adequadas para imagens, pois exploram relações espaciais entre pixels e aprendem padrões visuais em diferentes níveis de complexidade, desde bordas e texturas até estruturas mais específicas associadas à tarefa de classificação.

Dessa forma, esta seção apresenta os principais conceitos relacionados às redes

neurais convolucionais, ao aprendizado por transferência e à arquitetura MobileNetV2, que compõem a base técnica do modelo desenvolvido neste estudo.

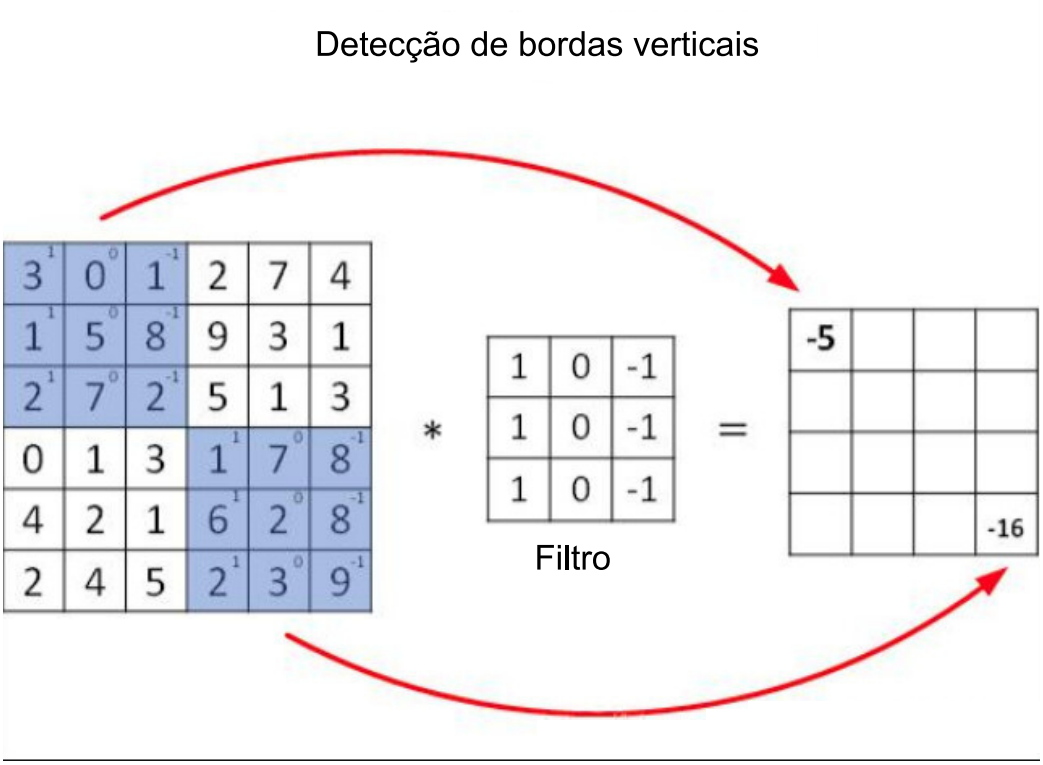
2.3.1 Redes Neurais Convolucionais

As Redes Neurais Convolucionais (*Convolutional Neural Networks* – CNNs) constituem uma das principais arquiteturas de aprendizado profundo aplicadas ao processamento e à classificação de imagens. Diferentemente das redes neurais totalmente conectadas, as CNNs foram desenvolvidas para explorar a estrutura espacial dos dados visuais por meio de conexões locais e compartilhamento de pesos. Essa característica reduz a quantidade de parâmetros a serem ajustados e torna o modelo mais eficiente para tarefas de visão computacional (LECUN et al., 1998; LECUN; BENGIO; HINTON, 2015).

De forma geral, uma CNN é composta por etapas sucessivas de extração e transformação de características, seguidas de uma etapa final de classificação. Nas primeiras camadas, a rede tende a identificar padrões visuais mais simples, como bordas, linhas, contrastes e texturas. À medida que a informação percorre camadas mais profundas, esses padrões passam a ser combinados em representações mais complexas, permitindo que o modelo reconheça estruturas relevantes para a tarefa proposta (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; LECUN; BENGIO; HINTON, 2015).

A operação central dessa arquitetura é a convolução, na qual filtros treináveis percorrem a imagem de entrada e geram mapas de características (*feature maps*). Esses filtros atuam como detectores de padrões locais, permitindo que a rede aprenda automaticamente quais regiões e estruturas da imagem são mais relevantes. Em cada posição, os valores cobertos pelo filtro são multiplicados pelos pesos correspondentes e somados, produzindo um novo valor no mapa de características. A Figura 5 apresenta um exemplo ilustrativo dessa operação aplicada à detecção de bordas verticais (LECUN et al., 1998; LECUN; BENGIO; HINTON, 2015).

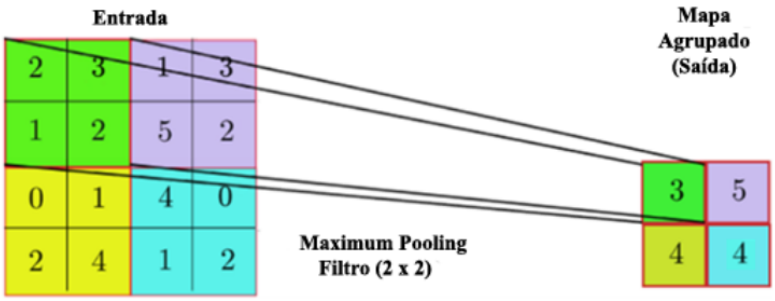
Figura 5: Exemplo ilustrativo da operação de convolução aplicada à detecção de bordas verticais.



Fonte: Adaptada de Bgraysea (2021).

Após a convolução, costuma-se aplicar uma função de ativação não linear, como a ReLU (*Rectified Linear Unit*), responsável por ampliar a capacidade de representação do modelo. Em seguida, operações de *pooling* são utilizadas para reduzir a dimensionalidade espacial dos mapas de características. Entre essas operações, o *max-pooling* é uma das mais comuns, pois seleciona o maior valor dentro de uma região analisada, preservando as ativações mais relevantes e reduzindo o custo computacional. A Figura 6 apresenta um exemplo dessa operação com filtro 2×2 (LECUN et al., 1998; LECUN; BENGIO; HINTON, 2015).

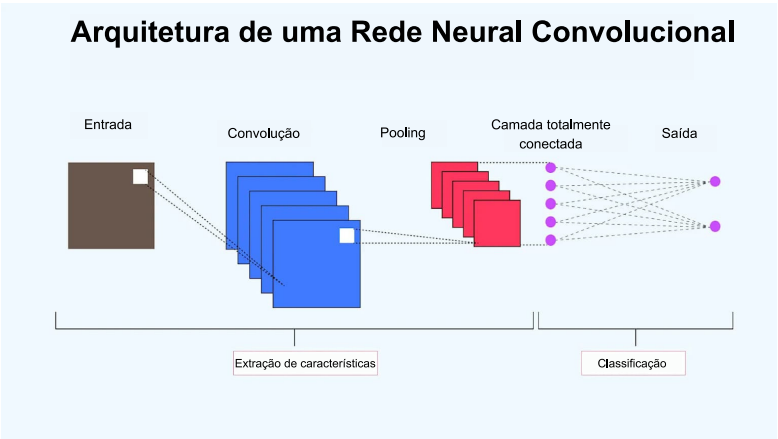
Figura 6: Exemplo ilustrativo da operação de *max-pooling* com filtro 2×2.



Fonte: Lima (2019).

Na etapa final, os mapas de características são reorganizados em formato vetorial e encaminhados para camadas densas, responsáveis pela decisão de classificação. Essa organização em blocos torna as CNNs particularmente adequadas para problemas em que a identificação automática de padrões visuais é essencial, como ocorre na análise de imagens médicas. A Figura 7 apresenta uma visão geral dessa arquitetura, destacando o fluxo entre a imagem de entrada, as etapas de extração de características e a classificação final.

Figura 7: Representação simplificada da estrutura de uma rede neural convolucional.



Fonte: Adaptada de Jain e Srihari (2023).

A partir dessa representação, observa-se que a CNN pode ser compreendida como uma sequência de blocos responsáveis por transformar a imagem de entrada em características cada vez mais abstratas. Para complementar essa descrição, a Tabela 3 resume os principais componentes dessa arquitetura e suas funções.

No contexto deste trabalho, as CNNs são relevantes porque permitem extrair

Tabela 3: Principais componentes de uma rede neural convolucional.

Componente	**Função**
Convolução	Extrai características locais da imagem, como bordas, texturas e padrões visuais.
Ativação ReLU	Introduz não linearidade ao modelo, ampliando sua capacidade de representação.
Pooling	Reduz a dimensionalidade espacial dos mapas de características e o custo computacional.
Flatten	Converte os mapas de características em um vetor unidimensional para as etapas finais da rede.
Camada densa	Combina as características extraídas para realizar a decisão final de classificação.

Fonte: Elaborada pelo autor (2026).

automaticamente características visuais de radiografias de tórax, apoiando a identificação de padrões associados às classes normal, pneumonia bacteriana e pneumonia viral. Entretanto, seu desempenho depende da qualidade da base de dados, do equilíbrio entre as classes e da capacidade de generalização do modelo, aspectos que justificam o uso de estratégias como aumento de dados, validação adequada e aprendizado por transferência (LITJENS et al., 2017; CALLI et al., 2021).

2.3.2 *Transfer Learning*

O *transfer learning* é uma estratégia amplamente utilizada em aprendizado profundo para reaproveitar conhecimentos previamente adquiridos em uma tarefa e aplicá-los em um novo problema relacionado. Em vez de treinar uma rede neural inteiramente do zero, utiliza-se um modelo previamente treinado como ponto de partida, o que reduz o custo computacional, acelera o treinamento e tende a melhorar o desempenho quando há quantidade limitada de dados rotulados (LE et al., 2022; MORID; BORJALI; FIOL, 2021).

Essa abordagem é especialmente relevante em imagens médicas, pois a construção de grandes bases anotadas exige tempo, infraestrutura e validação especializada. Em radiografias de tórax, por exemplo, a disponibilidade de imagens rotuladas em larga escala costuma ser menor do que em bases gerais de imagens naturais. Nesse

cenário, o reaproveitamento de modelos pré-treinados torna-se uma solução prática e tecnicamente vantajosa (LE et al., 2022; MORID; BORJALI; FIOL, 2021).

Em tarefas de classificação de imagens, o *transfer learning* pode ser aplicado de diferentes formas. Uma das estratégias mais comuns é a extração de características (*feature extraction*), na qual as camadas convolucionais da rede permanecem congeladas e apenas a etapa final de classificação é ajustada para a nova tarefa. Outra possibilidade é o *fine-tuning*, em que parte das camadas profundas também é atualizada, permitindo maior adaptação às características específicas da nova base de dados (LE et al., 2022; MORID; BORJALI; FIOL, 2021).

No contexto da classificação de pneumonia em radiografias de tórax, essa estratégia apresenta vantagens importantes. As camadas iniciais da rede tendem a aprender padrões visuais mais gerais, como bordas, contrastes e texturas, enquanto as camadas mais profundas podem ser ajustadas para captar padrões mais específicos do domínio clínico, como opacidades e regiões de consolidação. Dessa forma, o *transfer learning* permite combinar conhecimento previamente aprendido com adaptação ao problema de interesse (KIM et al., 2022).

Estudos recentes indicam que essa abordagem pode alcançar desempenho competitivo mesmo em cenários com disponibilidade limitada de dados. Em radiografias de tórax, estratégias de pré-treinamento e adaptação ao domínio médico têm mostrado potencial para reduzir a necessidade de grandes volumes de imagens rotuladas e favorecer a generalização do modelo em tarefas específicas (KIM et al., 2022).

Apesar de suas vantagens, o uso de *transfer learning* não elimina todos os desafios. Quando a diferença entre o domínio de origem e o domínio-alvo é muito grande, a transferência de conhecimento pode ser menos eficiente, fenômeno frequentemente associado ao *domain shift*. Além disso, a escolha entre congelar camadas e realizar *fine-tuning* mais profundo deve ser feita com cuidado, pois um ajuste excessivo pode levar ao *overfitting*, enquanto uma adaptação insuficiente pode limitar a capacidade do modelo de aprender características relevantes da nova base (LE et al., 2022; MORID; BORJALI; FIOL, 2021).

No contexto deste trabalho, o aprendizado por transferência foi aplicado utilizando a MobileNetV2 pré-treinada como base convolucional. As camadas iniciais e intermediárias da rede foram aproveitadas como extratoras de características visuais, enquanto a cabeça de classificação foi substituída para adequar o modelo às três classes analisadas: Normal, Pneumonia Bacteriana e Pneumonia Viral. Dessa forma, a estratégia adotada permitiu reaproveitar conhecimento visual previamente aprendido e ajustá-lo ao domínio específico das radiografias de tórax, com posterior ajuste fino parcial das camadas finais.

2.3.3 Arquitetura MobileNetV2

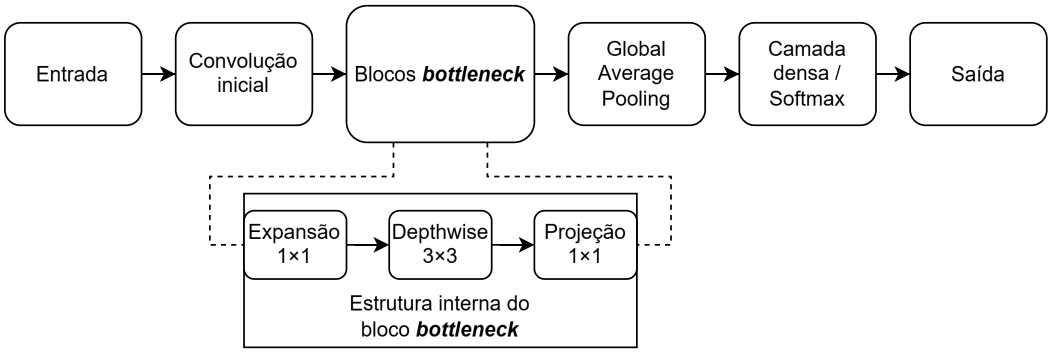
A MobileNetV2 é uma arquitetura de rede neural convolucional proposta por Sandler et al. (2018), desenvolvida para oferecer bom desempenho com baixo custo computacional. Essa arquitetura é especialmente adequada para aplicações com restrições de processamento e memória, pois foi concebida para reduzir o número de parâmetros e operações sem comprometer de forma significativa a capacidade de representação do modelo. Por essas características, tornou-se uma alternativa atrativa em tarefas de visão computacional e, posteriormente, em aplicações envolvendo imagens médicas.

O principal diferencial da MobileNetV2 está no uso de blocos do tipo *inverted residual* com *linear bottleneck*, associados a convoluções separáveis em profundidade. Em vez de empregar convoluções convencionais diretamente sobre todos os canais, a arquitetura organiza o processamento em três etapas principais: expansão de canais por convolução 1×1, convolução espacial separável em profundidade (*depthwise convolution*) e projeção linear novamente por convolução 1×1. Essa estrutura reduz a quantidade de parâmetros e operações computacionais, preservando a capacidade de extração de características relevantes para a classificação (SANDLER et al., 2018).

A Figura 8 apresenta uma representação simplificada da arquitetura da MobileNetV2. Nela, observa-se o fluxo principal da rede, composto por uma convolução inicial, uma sequência de blocos *bottleneck*, a etapa de *global average pooling* e a camada densa responsável pela classificação final. O destaque inferior da figura ilustra

a estrutura interna do bloco *bottleneck*, formado pelas etapas de expansão, convolução *depthwise* e projeção. Esse bloco constitui o principal elemento construtivo da MobileNetV2 e explica sua eficiência em aplicações com restrições de processamento e memória.

Figura 8: Representação simplificada da arquitetura da MobileNetV2, com destaque para a estrutura interna do bloco *bottleneck*.

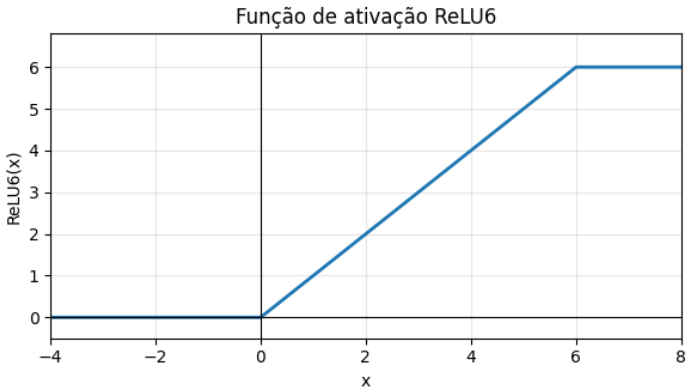


Fonte: Adaptada de Sandler et al. (2018).

Na etapa final da arquitetura, o *global average pooling* resume cada mapa de características gerado pela rede em um único valor médio, reduzindo a dimensão dos dados antes da classificação. Em seguida, a camada densa recebe essas características resumidas e realiza a combinação final das informações extraídas pela MobileNetV2. A função *softmax*, aplicada na saída, transforma os valores finais do modelo em probabilidades associadas às classes consideradas. No caso deste trabalho, essa saída foi adaptada para três classes: Normal, Pneumonia Bacteriana e Pneumonia Viral.

Além dessa estrutura, a MobileNetV2 utiliza a função de ativação ReLU6 em etapas intermediárias da rede. Essa função corresponde a uma variação da ReLU tradicional, porém com a saída limitada ao valor máximo 6, conforme ilustrado na Figura 9. De forma simplificada, sua operação pode ser representada por $\text{ReLU6}(x) = \min(\max(0, x), 6)$. Essa limitação contribui para maior estabilidade numérica em arquiteturas leves e em implementações com menor precisão computacional, como aplicações móveis e embarcadas (SANDLER et al., 2018).

Figura 9: Representação da função de ativação ReLU6.



Fonte: Adaptada de Sandler et al. (2018).

No contexto da classificação de pneumonia em radiografias de tórax, a MobileNetV2 destaca-se por oferecer um equilíbrio entre eficiência computacional e desempenho. Como bases médicas frequentemente possuem tamanho mais limitado do que grandes conjuntos de dados gerais de visão computacional, o uso de uma arquitetura leve e pré-treinada pode favorecer a convergência do modelo, reduzir o custo de treinamento e diminuir a tendência ao sobreajuste. Dessa forma, sua utilização é coerente com estratégias de aprendizado por transferência em imagens médicas, nas quais modelos previamente treinados são adaptados para tarefas específicas com bases de dados mais restritas (MORID; BORJALI; FIOL, 2021; SANDLER et al., 2018; CALLI et al., 2021).

Apesar dessas vantagens, a MobileNetV2 não elimina completamente os desafios da tarefa. Em bases pequenas ou desbalanceadas, a arquitetura ainda pode apresentar perda de desempenho quando ajustada de forma inadequada. Além disso, sua leveza computacional pode implicar menor capacidade representacional em comparação com modelos mais complexos, o que exige escolha cuidadosa da estratégia de treinamento, do nível de *fine-tuning* e dos hiperparâmetros. Dessa forma, a MobileNetV2 pode ser entendida como uma solução de compromisso entre custo computacional, robustez e precisão (SANDLER et al., 2018).

2.4 Interpretabilidade em Modelos de Aprendizado Profundo

Embora modelos de aprendizado profundo apresentem desempenho elevado em diversas tarefas de classificação de imagens, sua aplicação em contextos sensíveis, como o domínio médico, exige não apenas bons resultados quantitativos, mas também mecanismos que permitam compreender de forma mais transparente como as decisões são produzidas. Nesse cenário, a interpretabilidade assume papel relevante ao fornecer subsídios para analisar se o modelo está utilizando informações visualmente coerentes com o problema estudado (ABDAR et al., 2021; SAPORTA et al., 2022).

No caso da classificação de radiografias de tórax, a interpretação das decisões torna-se especialmente importante, uma vez que diferentes condições pulmonares podem apresentar padrões sutis e parcialmente sobrepostos. Assim, investigar quais regiões da imagem influenciam a predição do modelo contribui tanto para a análise de confiabilidade do sistema quanto para a identificação de possíveis vieses e limitações (SAPORTA et al., 2022; ALAPAT; MENON; ASHOK, 2022).

A partir dessa perspectiva, esta subseção discute a importância da interpretabilidade em imagens médicas e apresenta a técnica Grad-CAM, adotada neste trabalho como ferramenta de apoio à análise visual das predições realizadas pela rede neural (SELVARAJU et al., 2017).

2.4.1 Importância da Interpretabilidade em Imagens Médicas

A interpretabilidade tem se tornado um aspecto importante para a adoção de modelos de aprendizado profundo em aplicações médicas, especialmente quando esses modelos são utilizados como ferramentas de apoio à decisão diagnóstica. Em problemas de classificação de imagens, não basta que o sistema produza uma resposta correta; também é necessário compreender quais regiões da imagem influenciaram essa decisão e verificar se o modelo está de fato considerando áreas compatíveis com achados clinicamente relevantes (ABDAR et al., 2021; SAPORTA et al., 2022).

No contexto das radiografias de tórax, essa necessidade torna-se ainda mais evi-

dente, uma vez que diferentes patologias podem apresentar padrões visuais sutis e parcialmente sobrepostos. Assim, espera-se que o modelo concentre sua atenção em regiões pulmonares relevantes, como áreas de opacidade ou consolidação, e não em artefatos, bordas da imagem, marcações externas ou regiões sem relação direta com a condição analisada. Estudos sobre métodos de saliência em radiografias mostram que, mesmo quando os modelos apresentam bom desempenho em métricas de classificação, suas decisões podem ser influenciadas por regiões inadequadas, o que compromete a confiabilidade da interpretação visual (SAPORTA et al., 2022).

Ferramentas de explicabilidade, como o Grad-CAM e outros métodos de saliência, têm sido utilizadas para investigar esse comportamento. Essas técnicas geram mapas de ativação que destacam as regiões mais influentes para a predição, permitindo uma análise visual complementar do processo de decisão do modelo (SELVARAJU et al., 2017). No entanto, a literatura aponta que esses mapas não devem ser interpretados como explicações absolutas ou causais, mas como aproximações visuais úteis para auditoria e análise qualitativa das predições (SAPORTA et al., 2022; ABDAR et al., 2021).

A Tabela 4 resume as principais motivações para o uso de interpretabilidade em modelos de imagens médicas. Esses aspectos são relevantes porque permitem avaliar não apenas o desempenho quantitativo do classificador, mas também a coerência visual das regiões utilizadas pelo modelo durante a tomada de decisão.

Tabela 4: Principais motivações para o uso de interpretabilidade em modelos de imagens médicas.

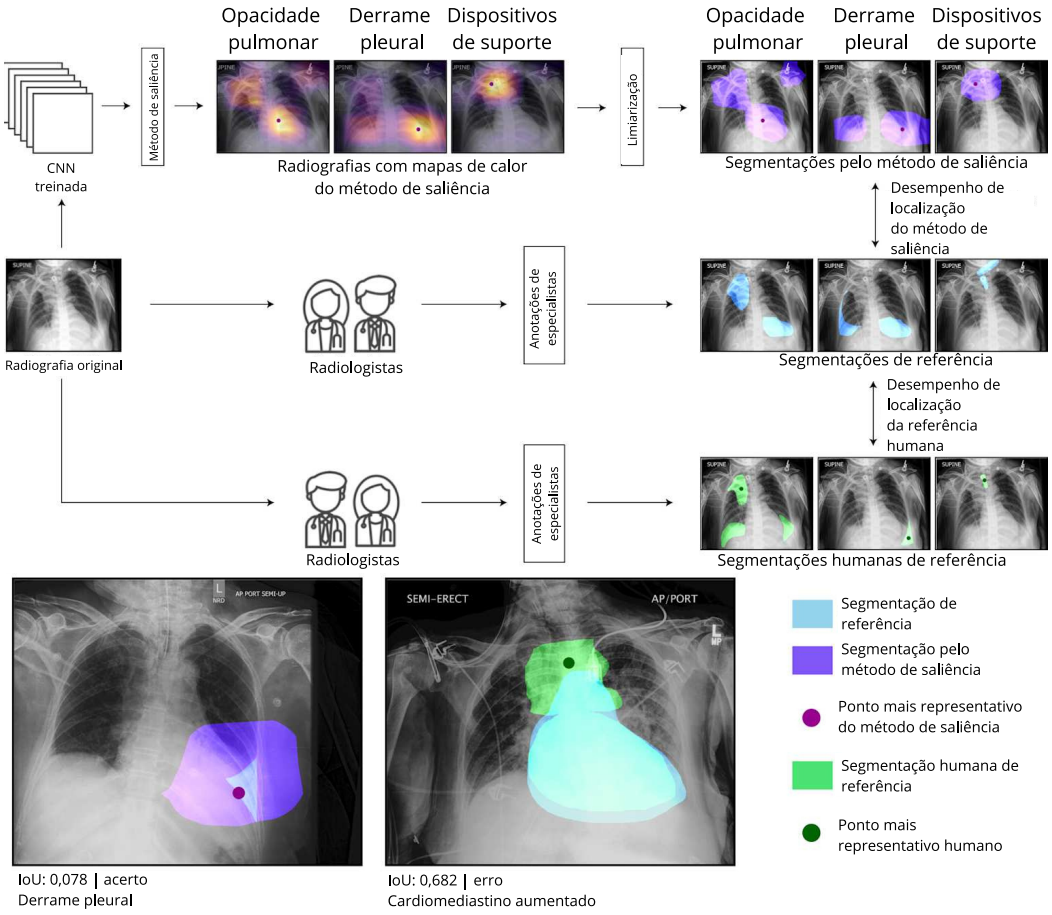
Motivo	**Justificativa**
Coerência visual	Permite verificar se a predição se apoia em regiões compatíveis com a condição analisada.
Transparência do modelo	Contribui para uma análise mais clara das decisões produzidas pelo sistema.
Identificação de vieses	Auxilia na detecção de decisões baseadas em artefatos, marcações externas ou regiões irrelevantes da imagem.
Análise de erros	Facilita a investigação de predições incorretas e a compreensão das limitações do modelo.

Fonte: Elaborada pelo autor com base em Abdar et al. (2021) e Saporta et al. (2022).

Além da identificação qualitativa das regiões ativadas, alguns estudos avaliam quantitativamente a correspondência entre os mapas de saliência e anotações realizadas por especialistas. Nesse tipo de análise, os mapas gerados pelo modelo podem ser processados por técnicas de limiarização (*thresholding*), resultando em segmentações aproximadas das regiões de interesse. Em seguida, essas segmentações são comparadas com marcações de referência, permitindo avaliar o grau de sobreposição entre a região destacada pelo modelo e os achados clínicos esperados (SAPORTA et al., 2022).

A Figura 10 apresenta um exemplo de *pipeline* de avaliação de métodos de interpretabilidade em radiografias de tórax. Nesse contexto, o termo *pipeline* refere-se à sequência organizada de etapas utilizadas para verificar se os mapas gerados pelo modelo estão coerentes com regiões clinicamente relevantes da imagem. Inicialmente, um modelo de rede neural convolucional gera mapas de saliência que indicam as regiões mais influentes para a decisão. Em seguida, esses mapas são processados, por exemplo por limiarização, para obter regiões aproximadas de interesse. Por fim, essas regiões são comparadas com anotações feitas por especialistas, permitindo avaliar a capacidade do método em localizar corretamente os achados esperados.

Figura 10: Pipeline de avaliação de métodos de interpretabilidade em radiografias de tórax.



Fonte: Adaptada de Saporta et al. (2022).

Métricas como a interseção sobre união (*Intersection over Union* – IoU) podem ser empregadas para medir o grau de sobreposição entre as regiões identificadas pelo modelo e aquelas definidas como referência. Resultados experimentais apontam que, em diversos casos, essa sobreposição pode ser limitada, evidenciando discrepâncias entre o comportamento do modelo e o julgamento clínico humano (SAPORTA et al., 2022).

Esses aspectos reforçam que, embora técnicas de interpretabilidade sejam ferramentas importantes para análise e validação de modelos, elas ainda apresentam limitações. Dessa forma, a interpretabilidade deve ser utilizada como complemento à avaliação quantitativa, contribuindo para uma análise mais crítica e completa dos modelos desenvolvidos.

Dentre as técnicas de interpretabilidade mencionadas, destaca-se o Grad-CAM, amplamente utilizado em tarefas de visão computacional médica e adotado neste trabalho para a análise qualitativa das predições do modelo.

2.4.2 Grad-CAM

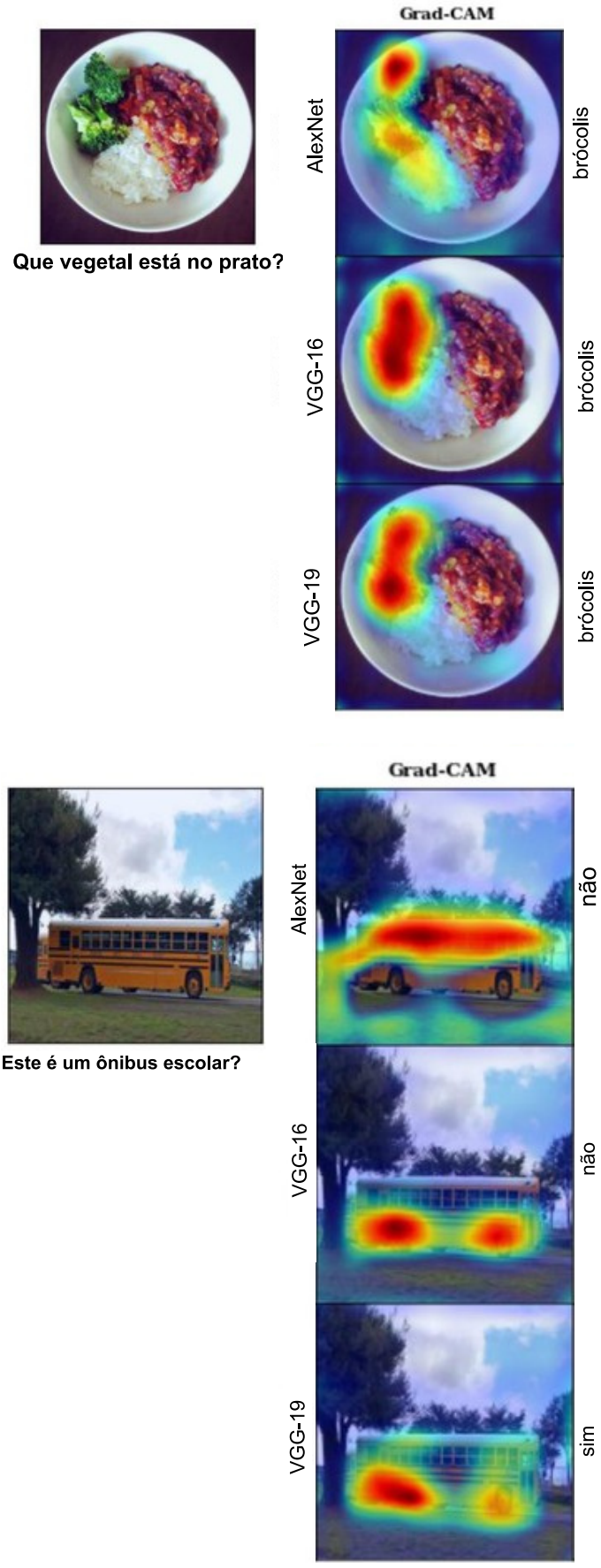
O Grad-CAM (*Gradient-weighted Class Activation Mapping*) é uma técnica de explicabilidade *post-hoc* amplamente utilizada para interpretar decisões de redes neurais convolucionais em tarefas de classificação de imagens. Nesse contexto, o termo *post-hoc* indica que a técnica é aplicada após o treinamento da rede neural, sem modificar sua arquitetura ou interferir no processo de aprendizado. Sua proposta consiste em gerar mapas de ativação que indicam, de forma aproximada, quais regiões da imagem contribuíram mais para a predição de uma classe específica. Em outras palavras, o método busca transformar a saída numérica de uma rede profunda em uma representação visual mais compreensível, o que é particularmente útil em aplicações médicas, nas quais a transparência da decisão é tão importante quanto o resultado predito (SELVARAJU et al., 2017; ABDAR et al., 2021).

Do ponto de vista conceitual, o Grad-CAM utiliza os gradientes associados à classe de interesse e os combina com os mapas de características da última camada convolucional da rede. Essa operação produz um mapa de calor que destaca as regiões mais relevantes para a decisão do modelo, sem exigir alterações na arquitetura nem novo treinamento. Por essa razão, o método pode ser aplicado a diferentes modelos baseados em CNN, incluindo arquiteturas utilizadas em classificação de radiografias de tórax, como ResNet, DenseNet e MobileNetV2 (SELVARAJU et al., 2017).

A Figura 11 ilustra o princípio de funcionamento do Grad-CAM, mostrando como a imagem original pode ser sobreposta por um mapa de calor que evidencia as áreas mais influentes para a predição. Embora a figura apresente exemplos gerais de classificação de imagens, o mesmo princípio pode ser aplicado à análise de radiografias de tórax. Em aplicações de pneumonia, esse tipo de visualização permite verificar se o modelo está concentrando sua atenção em regiões pulmonares compatíveis com

opacidades ou consolidações, em vez de se apoiar em estruturas irrelevantes da imagem. Esse aspecto é especialmente importante em radiologia, pois uma predição correta, porém baseada em pistas inadequadas, não garante confiabilidade clínica (SAPORTA et al., 2022).

Figura 11: Exemplos de visualizações geradas por Grad-CAM e Grad-CAM guiado em diferentes arquiteturas convolucionais.



Fonte: Adaptada de Selvaraju et al. (2017).

Além de contribuir para a validação visual do modelo, o Grad-CAM também é útil para identificação de vieses e análise de erros. Em bases de dados médicas, onde artefatos, marcações e diferenças de aquisição podem influenciar as imagens, a interpretação dos mapas de ativação ajuda a detectar se o sistema está aprendendo correlações espúrias. Estudos recentes apontam que, embora o Grad-CAM seja amplamente aceito por sua intuitividade, sua interpretação deve ser feita com cautela, pois o mapa gerado é uma aproximação e não uma explicação causal completa do comportamento da rede (ABDAR et al., 2021; SAPORTA et al., 2022).

A Tabela 5 resume as principais características do Grad-CAM no contexto de aplicações em imagens médicas.

Tabela 5: Principais características do Grad-CAM.

Característica	**Descrição**
Tipo de técnica	Explicabilidade *post-hoc* para redes neurais convolucionais.
Saída visual	Gera mapas de calor sobrepostos à imagem original.
Vantagem principal	Indica regiões mais relevantes para a decisão da rede.
Uso em imagens médicas	Permite verificar se o modelo observa regiões visualmente coerentes com a condição analisada.
Limitação	Produz aproximações visuais, não explicações causais completas.

Fonte: Elaborada pelo autor (2026).

2.5 Síntese do capítulo

Com base nos conceitos apresentados ao longo desta seção, observa-se que técnicas de aprendizado profundo, aliadas a métodos de interpretabilidade como o Grad-CAM, constituem ferramentas promissoras para a análise de imagens médicas. No contexto deste trabalho, tais abordagens são empregadas não apenas para realizar a classificação multiclasse de radiografias de tórax, mas também para permitir a interpretação visual das decisões do modelo, contribuindo para maior transparência e alinhamento com critérios clínicos. A partir desses fundamentos, o próximo capítulo apresenta a metodologia adotada, abrangendo a descrição da base de dados, o pré-processamento das imagens, a arquitetura do modelo, bem como o processo de treinamento e avaliação.

3 METODOLOGIA

Este capítulo apresenta as etapas adotadas para o desenvolvimento do sistema de classificação automática de radiografias de tórax proposto neste trabalho. Inicialmente, descreve-se a base de dados utilizada, sua origem, organização em três classes e distribuição entre os subconjuntos de treinamento, validação e teste. Em seguida, são detalhados os procedimentos de pré-processamento aplicados às imagens, incluindo redimensionamento, preparação do *pipeline* de dados e técnicas de aumento de dados empregadas durante o treinamento.

Na sequência, apresenta-se a arquitetura do modelo adotado, baseada na MobileNetV2 com aprendizado por transferência, bem como as adaptações realizadas para a classificação multiclasse entre *Normal*, *Pneumonia Bacteriana* e *Pneumonia Viral*. Também são descritas a estratégia de treinamento em duas etapas, o ajuste fino parcial da rede, o protocolo de avaliação, as métricas utilizadas e a aplicação do Grad-CAM como ferramenta de interpretabilidade visual.

Por fim, são apresentados o ambiente computacional utilizado e os cuidados adotados para favorecer a reprodutibilidade dos experimentos. Dessa forma, este capítulo reúne os elementos metodológicos necessários para compreender como o modelo foi construído, treinado, avaliado e interpretado, servindo como base para a análise dos resultados apresentada no capítulo seguinte.

3.1 Base de Dados

A base de dados utilizada neste trabalho foi o conjunto *Chest X-Ray Images (Pneumonia)*, disponibilizado publicamente na plataforma Kaggle (MOONEY, 2018). Nesta seção, são descritas sua origem, sua organização em três classes diagnósticas e a divisão adotada entre os subconjuntos de treinamento, validação e teste.

3.1.1 Origem e aquisição

O conjunto de dados utilizado é composto por radiografias de tórax (*chest X-ray*) obtidas a partir de um *dataset* público disponibilizado na plataforma Kaggle. Esse conjunto é amplamente empregado em estudos de classificação de pneumonia e, em sua forma original, encontra-se organizado em duas classes principais: *Normal* e *Pneumonia*. Ao todo, a base contém 5.864 imagens distribuídas entre os subconjuntos de treinamento, validação e teste (MOONEY, 2018).

Para adequar o conjunto de dados ao objetivo deste trabalho, que consiste na classificação multiclasse de radiografias de tórax, realizou-se uma reestruturação dos diretórios da base. A classe originalmente identificada como *Pneumonia* foi subdividida em duas categorias: *Pneumonia Bacteriana* e *Pneumonia Viral*. Para isso, foi desenvolvido um *script* em Python para a triagem automática das imagens originalmente contidas na pasta genérica de pneumonia, utilizando padrões textuais presentes nos nomes dos arquivos, como `bacteria` e `virus`, como critério para separação das categorias.

É importante destacar que essa separação foi realizada a partir da nomenclatura e da organização original dos arquivos disponibilizados no conjunto de dados. Assim, os rótulos utilizados neste trabalho não correspondem a uma nova validação clínica realizada pelo autor, mas sim ao reaproveitamento da classificação já indicada na base pública. Além disso, essa etapa consistiu apenas na redistribuição dos arquivos em novos diretórios, sem qualquer alteração no conteúdo visual das imagens.

3.1.2 Organização dos subconjuntos e rotulagem

O conjunto de dados foi dividido em subconjuntos independentes de treinamento, validação e teste, garantindo que o ajuste dos parâmetros, a seleção de hiperparâmetros e a estimativa final de generalização do modelo ocorressem de maneira isolada. Para viabilizar a rotulagem automática, as imagens foram separadas nas três classes do estudo: `Bacteria`, `Normal` e `Viral`. Essa organização foi realizada por meio de um procedimento automatizado em Python, que utilizou a identificação contida nos

nomes dos arquivos originais para estruturar os diretórios correspondentes. Dessa forma, foi possível aplicar a função `image_dataset_from_directory` da biblioteca TensorFlow/Keras, que infere os rótulos diretamente a partir dessa estrutura de pastas durante o carregamento das amostras.

No experimento final, as classes foram carregadas na seguinte ordem: `BACTERIA`, `NORMAL` e `VIRAL`. A distribuição final da base, após a reorganização em três classes, é apresentada na Tabela 6.

Tabela 6: Distribuição das imagens após a reorganização da base de dados.

Subconjunto	**Bacteria**	**Normal**	**Viral**	**Total**
Treinamento	2338	1149	1153	4640
Validação	200	200	200	600
Teste	242	234	148	624
Total	**2780**	**1583**	**1501**	**5864**

Fonte: Elaborada pelo autor (2026).

Observa-se que o conjunto de treinamento apresenta desbalanceamento entre as classes, com maior quantidade de imagens da classe `Bacteria` em relação às classes `Normal` e `Viral`. Esse aspecto foi considerado na análise dos resultados, especialmente na interpretação das métricas por classe e da matriz de confusão, uma vez que distribuições desiguais podem influenciar o comportamento do classificador. Ainda assim, a avaliação final foi realizada em um conjunto de teste independente, preservado durante todo o processo de treinamento.

3.1.3 Composição e características das imagens

Por se tratar de uma base composta por exames reais, observam-se variações de contraste, ruído, nitidez, escala, posicionamento do paciente e recorte da região torácica. Essas diferenças são características comuns em bases de imagens médicas e podem influenciar o desempenho de modelos de aprendizado profundo, especialmente quando há variação na qualidade de aquisição e na apresentação dos achados radiográficos (MOONEY, 2018; LITJENS et al., 2017; CALLI et al., 2021).

Em sua natureza, radiografias de tórax apresentam predominância de tons de

cinza. Entretanto, como a MobileNetV2 utilizada neste trabalho foi inicializada com pesos pré-treinados em uma base de imagens coloridas, a entrada do modelo foi configurada com três canais. Dessa forma, as imagens foram carregadas no formato RGB durante o processo de leitura, garantindo compatibilidade com a dimensão de entrada esperada pela arquitetura adotada (SANDLER et al., 2018; MORID; BORJALI; FIOL, 2021).

No presente trabalho, esse procedimento foi realizado diretamente no *pipeline* de carregamento das imagens, por meio da função `image_dataset_from_directory`. Assim, cada imagem foi representada com dimensão espacial padronizada e três canais de entrada, sem alteração manual do conteúdo visual original. Essa etapa teve apenas a finalidade de adequar o formato dos dados à arquitetura MobileNetV2 e ao processo de treinamento definido para o classificador.

3.1.4 Estrutura de diretórios e rotulagem

O conjunto de dados foi organizado em diretórios separados por subconjunto e por classe, seguindo a estrutura `train`, `val` e `test`. Dentro de cada subconjunto, as imagens foram distribuídas em pastas correspondentes às três categorias adotadas neste trabalho: `Bacteria`, `Normal` e `Viral`. Essa organização foi escolhida por ser compatível com rotinas automatizadas de carregamento de imagens e por favorecer a reprodutibilidade do experimento.

Foi utilizado o recurso do TensorFlow/Keras para criação automática de conjuntos de imagens a partir de diretórios, por meio do método `image_dataset_from_directory`. Nessa abordagem, os rótulos das imagens são atribuídos automaticamente com base na pasta à qual cada arquivo pertence. Dessa forma, uma imagem armazenada na pasta `BACTERIA`, por exemplo, é associada à classe `Bacteria`; o mesmo procedimento ocorre para as classes `NORMAL` e `VIRAL`. Esse critério reduz a necessidade de rotulagem manual durante a etapa de leitura e garante correspondência direta entre a estrutura dos arquivos e as classes utilizadas no treinamento.

A organização em diretórios também permitiu o carregamento direto das imagens

como um objeto do tipo `tf.data.Dataset`. O `tf.data` é um módulo do TensorFlow utilizado para construir fluxos eficientes de entrada de dados, organizando o carregamento, o processamento e o envio das amostras ao modelo durante o treinamento. Neste trabalho, essa estrutura permitiu ler as imagens a partir das pastas, atribuir os rótulos automaticamente com base nos diretórios, redimensionar as imagens para o formato de entrada da MobileNetV2, agrupar as amostras em lotes e encaminhá-las ao *pipeline* de treinamento, validação e teste. Com isso, o processo de leitura dos dados tornou-se mais padronizado, reprodutível e adequado à arquitetura experimental adotada.

3.1.5 Distribuição das amostras e desbalanceamento

A distribuição das imagens após a reorganização da base evidencia um desbalanceamento entre as classes, especialmente no conjunto de treinamento. Conforme apresentado na Tabela 6, a classe *Pneumonia Bacteriana* concentra 2.338 imagens de treinamento, enquanto as classes *Normal* e *Pneumonia Viral* apresentam 1.149 e 1.153 imagens, respectivamente. Esse comportamento indica que a classe bacteriana possui aproximadamente o dobro de amostras em relação às demais categorias nesse subconjunto.

O desbalanceamento entre classes é um aspecto relevante em problemas de classificação de imagens médicas, pois pode influenciar o processo de aprendizagem do modelo e favorecer a classe majoritária. Nesses casos, a análise do desempenho não deve se limitar à acurácia global, uma vez que essa métrica pode mascarar dificuldades específicas em classes menos representadas. Por esse motivo, neste trabalho, os resultados foram avaliados também por meio de precisão, revocação, F1-score e matriz de confusão, permitindo observar o comportamento do classificador em cada categoria (SOKOLOVA; LAPALME, 2009; ALAPAT; MENON; ASHOK, 2022).

Embora a base apresente desbalanceamento no conjunto de treinamento, o conjunto de validação foi mantido equilibrado, com 200 imagens por classe, o que favorece o acompanhamento do desempenho do modelo durante o treinamento. Já o conjunto de teste preservou a distribuição original disponível após a reorganização,

sendo utilizado apenas na etapa final de avaliação. Na configuração final adotada, esse desbalanceamento não foi corrigido artificialmente por pesos de classe, sendo seu impacto analisado posteriormente por meio das métricas por classe e da matriz de confusão. Dessa forma, a interpretação dos resultados considerou não apenas o desempenho médio do modelo, mas também as diferenças observadas entre as classes *Bacteria*, *Normal* e *Viral*.

3.1.6 Estratégia de divisão: treinamento, validação e teste

A divisão do conjunto de dados em subconjuntos independentes de treinamento, validação e teste foi adotada para separar adequadamente as etapas de ajuste, monitoramento e avaliação final do modelo. O subconjunto de treinamento foi utilizado para atualizar os parâmetros internos da rede neural; o subconjunto de validação foi empregado para acompanhar o desempenho durante o treinamento, apoiar a seleção da melhor configuração e monitorar possíveis sinais de sobreajuste; e o subconjunto de teste foi reservado exclusivamente para a avaliação final do classificador (GOODFELLOW; BENGIO; COURVILLE, 2016; LITJENS et al., 2017).

A manutenção de um conjunto de teste isolado é importante para estimar de forma mais confiável a capacidade de generalização do modelo em dados não vistos. Dessa forma, as métricas obtidas na etapa final refletem o comportamento do classificador em imagens que não participaram do ajuste dos pesos nem da escolha da melhor configuração durante o treinamento.

No presente trabalho, essa separação foi mantida ao longo de todo o processo experimental. Assim, o conjunto de teste foi utilizado apenas após a definição do modelo final, preservando sua função como referência independente para análise do desempenho da MobileNetV2 na classificação das três classes consideradas.

3.1.7 Considerações sobre a base de dados

Embora o *dataset* utilizado seja amplamente empregado em estudos de classificação de pneumonia, é importante ressaltar que os dados provêm de uma fonte pública

específica e apresentam limitações inerentes a bases médicas de imagens. Entre essas limitações, destacam-se variações na qualidade das radiografias, diferenças de contraste, posicionamento e recorte da região torácica, além do desbalanceamento entre as classes após a reorganização em *Normal*, *Pneumonia Bacteriana* e *Pneumonia Viral* (MOONEY, 2018; LITJENS et al., 2017; CALLI et al., 2021).

Essas características podem influenciar o processo de treinamento e limitar a generalização do modelo para imagens obtidas em outros equipamentos, instituições ou populações. Por esse motivo, a interpretação dos resultados não deve se restringir à acurácia global, sendo necessário considerar também o desempenho por classe, a matriz de confusão e os principais padrões de erro observados no conjunto de teste (SOKOLOVA; LAPALME, 2009; ALAPAT; MENON; ASHOK, 2022).

Dessa forma, a base de dados utilizada mostrou-se adequada para o desenvolvimento e a avaliação experimental do modelo proposto. No entanto, por se tratar de uma base pública específica, os resultados devem ser interpretados com cautela, especialmente em uma tarefa multiclasse com padrões radiográficos visualmente semelhantes.

3.2 Pré-processamento e Preparação dos Dados

Após a organização da base de dados em três classes diagnósticas, foi definido um *pipeline* de preparação das imagens com o objetivo de padronizar as entradas do modelo e tornar o processo de treinamento mais consistente e reprodutível. As etapas adotadas contemplam o carregamento das imagens, o redimensionamento para a dimensão de entrada da MobileNetV2, a aplicação do pré-processamento compatível com essa arquitetura, a organização dos dados por meio de `tf.data` e o uso de aumento de dados aplicado exclusivamente ao conjunto de treinamento.

3.2.1 Redimensionamento e Normalização

O carregamento das imagens foi realizado por meio do método `image_dataset_from_directory`, configurado com dimensão de entrada igual a 224×224 *pixels* e

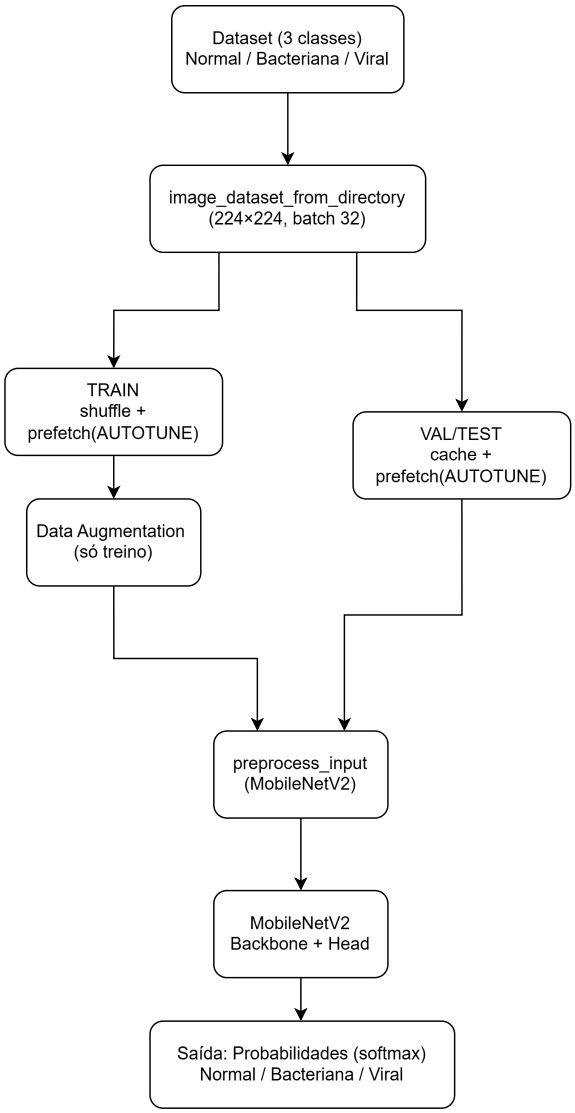
`batch_size` igual a 32. Essa etapa permitiu padronizar o tamanho das radiografias e organizar as amostras em lotes, mantendo a compatibilidade com a entrada esperada pela arquitetura MobileNetV2.

Após o carregamento, aplicou-se, no *pipeline* de pré-processamento, a função `preprocess_input` da MobileNetV2. Essa função é responsável por adequar a escala dos valores de *pixel* ao padrão utilizado pela arquitetura pré-treinada. Essa etapa é importante porque a rede base foi inicializada com pesos previamente ajustados em uma base de imagens de larga escala, de modo que a entrada deve seguir uma padronização compatível com o modelo original (SANDLER et al., 2018; MORID; BORJALI; FIOL, 2021).

Dessa forma, o redimensionamento garantiu uniformidade dimensional entre as imagens, enquanto o pré-processamento assegurou compatibilidade numérica com a MobileNetV2. A separação dessas etapas contribuiu para um fluxo de dados mais organizado, reprodutível e adequado ao treinamento do classificador.

A Figura 12 apresenta o fluxo geral de preparação dos dados e encaminhamento das imagens para a MobileNetV2. O diagrama diferencia o tratamento aplicado ao conjunto de treinamento, que inclui embaralhamento, aumento de dados e otimização do carregamento por meio de `prefetch`, daquele aplicado aos conjuntos de validação e teste, que permanecem sem transformações aleatórias. Nesse contexto, o `prefetch` corresponde a um recurso do TensorFlow que antecipa a preparação dos próximos lotes de imagens enquanto o modelo processa o lote atual, ao passo que o parâmetro `AUTOTUNE` permite que o próprio TensorFlow ajuste automaticamente essa preparação conforme os recursos computacionais disponíveis.

Figura 12: Fluxo do *pipeline* de preparação dos dados e encaminhamento das imagens para a MobileNetV2.



Fonte: Elaborada pelo autor (2026).

A partir do fluxo apresentado na Figura 12, as subseções seguintes detalham a organização do fluxo de dados e o aumento de dados aplicado durante o treinamento.

3.2.2 Pipeline de Dados com `tf.data`

Para tornar o processo de leitura das imagens mais eficiente e reduzir possíveis gargalos entre o armazenamento dos dados e o treinamento do modelo, foi utilizado um *pipeline* baseado na estrutura `tf.data`. Essa organização permitiu carregar, preparar e entregar os lotes de imagens de forma padronizada ao modelo, favorecendo maior estabilidade durante as etapas de treinamento, validação e teste.

No conjunto de treinamento, foi aplicado o embaralhamento das amostras por meio da operação `shuffle`, utilizando *buffer* de 1.000 amostras, semente fixa e re-embaralhamento a cada época. Essa estratégia contribui para reduzir a dependência da ordem original dos arquivos e favorece uma apresentação mais variada dos dados ao modelo durante o aprendizado.

Além disso, empregou-se a técnica de *prefetching* com configuração automática por meio de `AUTOTUNE`. Essa etapa permite que o preparo dos próximos lotes ocorra de forma sobreposta à execução do modelo, reduzindo o tempo de espera entre iterações e melhorando a eficiência computacional do treinamento.

Para os conjuntos de validação e teste, foi adotado o armazenamento em `cache` seguido de *prefetching* com `AUTOTUNE`. Como esses subconjuntos não recebem transformações aleatórias, o uso de `cache` evita leituras repetidas do disco e torna o processo de validação e avaliação final mais estável e eficiente. Dessa forma, o *pipeline* foi estruturado de modo a otimizar o fluxo de dados sem alterar o conteúdo das imagens avaliadas.

3.2.3 Aumento de Dados (*Data Augmentation*)

Para reduzir a tendência ao sobreajuste (*overfitting*) e aumentar a capacidade de generalização do classificador, foi implementada uma etapa de aumento de dados

aplicada dinamicamente ao conjunto de treinamento. Essa estratégia consiste em gerar variações artificiais das imagens originais por meio de transformações controladas, preservando a classe da amostra e ampliando a diversidade visual apresentada ao modelo durante o aprendizado (SHORTEN; KHOSHGOFTAAR, 2019).

No presente trabalho, foram utilizadas transformações geométricas aleatórias implementadas como camadas de pré-processamento integradas ao próprio modelo. A camada `RandomFlip` foi utilizada para realizar espelhamento horizontal das imagens, a `RandomRotation` para aplicar pequenas rotações aleatórias e a `RandomZoom` para introduzir variações controladas de aproximação ou afastamento da imagem. Como essas operações foram incorporadas ao modelo, elas foram aplicadas dinamicamente durante o treinamento, sem alterar permanentemente os arquivos originais da base de dados.

O aumento de dados foi aplicado exclusivamente ao conjunto de treinamento. Os conjuntos de validação e teste foram mantidos sem transformações aleatórias, de modo a preservar a consistência da avaliação experimental. Essa separação é importante porque permite que o modelo seja exposto a maior variabilidade durante o aprendizado, enquanto seu desempenho é avaliado em imagens não modificadas artificialmente.

A adoção dessa estratégia contribui para que o classificador aprenda características mais robustas diante de pequenas variações de escala, posicionamento e orientação das radiografias. No contexto deste trabalho, isso é relevante porque as imagens da base apresentam diferenças naturais de aquisição, enquadramento e qualidade visual, o que torna desejável que o modelo não dependa excessivamente de padrões fixos de posição ou recorte.

3.3 Arquitetura do Modelo

A rede MobileNetV2 foi selecionada como base convolucional do modelo por apresentar boa relação entre desempenho e custo computacional. Essa característica está associada ao uso de convoluções separáveis em profundidade (*depthwise*

separable convolutions) e de blocos residuais invertidos com gargalos lineares (*inverted residuals with linear bottlenecks*), que permitem reduzir o número de parâmetros e operações sem comprometer significativamente a capacidade de extração de características (SANDLER et al., 2018).

As convoluções separáveis em profundidade constituem um tipo de convolução fatorada em etapas, projetada para reduzir o custo computacional em relação à convolução convencional. Nessa abordagem, a operação é dividida em duas fases principais. Primeiro, a convolução *depthwise* aplica filtros espaciais separadamente sobre cada canal ou mapa de características da imagem, permitindo identificar padrões locais, como bordas, texturas e variações de intensidade. Em seguida, uma convolução pontual 1×1, também chamada de *pointwise convolution*, combina as informações entre os diferentes canais. Essa separação entre a extração espacial e a combinação entre canais reduz significativamente a quantidade de operações computacionais, contribuindo para tornar a MobileNetV2 uma arquitetura mais leve e eficiente.

Já os blocos residuais invertidos com gargalos lineares correspondem à unidade básica da MobileNetV2. Em cada bloco, a entrada passa inicialmente por uma convolução 1×1 que expande o número de canais, seguida por uma convolução separável em profundidade, responsável pela extração espacial de características, e, por fim, por uma nova convolução 1×1 que reduz novamente a dimensionalidade. Essa estrutura é denominada "invertida" porque parte de uma representação mais compacta, expande temporariamente para um espaço de maior dimensão e retorna a uma representação reduzida. Além disso, a saída do gargalo é linear, isto é, não utiliza função de ativação não linear nessa etapa, o que contribui para preservar informações em espaços de baixa dimensionalidade e favorece o uso da conexão residual quando as dimensões de entrada e saída são compatíveis (SANDLER et al., 2018).

No contexto deste trabalho, essa eficiência é relevante porque as radiografias precisam ser processadas em uma tarefa multiclasse, envolvendo a distinção entre imagens normais, pneumonia bacteriana e pneumonia viral. A utilização da Mobile-

NetV2 permite aproveitar uma base convolucional capaz de extrair padrões visuais discriminativos das imagens, mantendo menor custo computacional em relação a arquiteturas mais pesadas. Assim, a rede atua principalmente como extratora de características, fornecendo representações visuais que posteriormente são utilizadas pelas camadas finais de classificação.

Além disso, a MobileNetV2 é particularmente adequada ao uso em estratégias de aprendizado por transferência, pois pode ser inicializada com pesos previamente treinados em grandes bases de imagens e posteriormente adaptada ao domínio específico das radiografias de tórax. No presente trabalho, essa arquitetura foi empregada como base convolucional pré-treinada para a tarefa de classificação multiclasse entre imagens *Normal*, *Pneumonia Bacteriana* e *Pneumonia Viral* (MORID; BORJALI; FIOL, 2021; SANDLER et al., 2018).

3.3.1 Aprendizado por Transferência (*Transfer Learning*)

Empregou-se a técnica de aprendizado por transferência, inicializando a MobileNetV2 com pesos pré-treinados no conjunto ImageNet (`weights="imagenet"`). Essa estratégia permite reutilizar representações visuais genéricas aprendidas em larga escala, como bordas, texturas e formas, adaptando-as posteriormente à tarefa específica de classificação de radiografias de tórax.

Essa abordagem é especialmente útil em aplicações com imagens médicas, pois bases rotuladas nesse domínio tendem a ser mais restritas quando comparadas a grandes bases gerais de visão computacional. Dessa forma, o uso de uma rede pré-treinada contribui para maior estabilidade no treinamento, redução do custo computacional e melhor aproveitamento das informações disponíveis no conjunto de dados (MORID; BORJALI; FIOL, 2021).

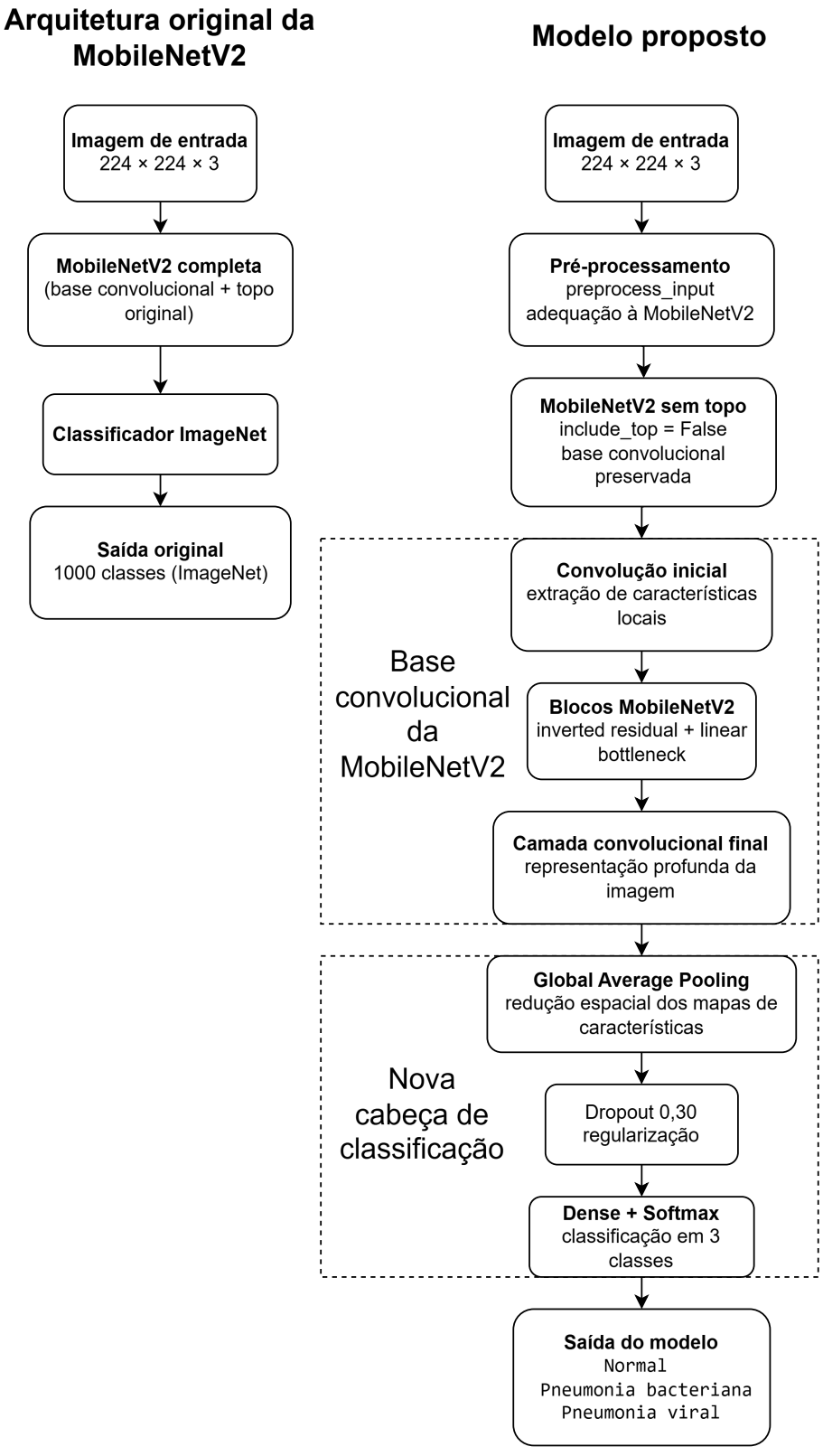
3.3.2 Adaptação da MobileNetV2 para Classificação em Três Classes

A MobileNetV2 foi incorporada ao modelo com remoção da camada de classificação original (`include_top=False`), mantendo-se sua estrutura convolucional

como base para extração de características das imagens. Sobre essa base, foi adicionada uma nova cabeça de classificação (*classification head*) adequada ao problema multiclasse proposto.

A Figura 13 apresenta a comparação entre a arquitetura original da MobileNetV2 e a adaptação realizada neste trabalho. Na configuração original, a rede possui uma cabeça de classificação voltada ao conjunto ImageNet, com saída para 1000 classes. No modelo proposto, essa cabeça original é removida, preservando-se a base convolucional da MobileNetV2 e adicionando-se uma nova cabeça de classificação voltada às três classes analisadas neste estudo.

Figura 13: Comparação entre a arquitetura original da MobileNetV2 e a adaptação proposta para a classificação de radiografias de tórax em três classes.



Fonte: Elaborada pelo autor (2026).

A cabeça de classificação foi composta por uma camada de agrupamento médio global (*Global Average Pooling*), responsável por reduzir os mapas de características gerados pela MobileNetV2 a uma representação vetorial compacta. Em seguida, foi aplicada uma camada de *dropout* com taxa de 0,30, utilizada como estratégia de regularização para reduzir a tendência ao sobreajuste. Por fim, foi adicionada uma camada densa com três neurônios, correspondentes às três classes do problema, e ativação *softmax*, responsável por converter as saídas do modelo em probabilidades associadas às classes *Normal*, *Pneumonia Bacteriana* e *Pneumonia Viral*.

As etapas de aumento de dados e o pré-processamento específico da MobileNetV2 foram integrados ao fluxo do próprio modelo, conforme descrito na Seção 3.2. Essa organização permitiu manter o processo de preparação das imagens associado diretamente à arquitetura utilizada, favorecendo a padronização do treinamento e da avaliação experimental.

3.4 Estratégia de Treinamento e Ajuste Fino

O treinamento do classificador foi conduzido em duas etapas, seguindo uma estratégia comum em aplicações de aprendizado por transferência. Essa abordagem permite, inicialmente, aproveitar as representações visuais previamente aprendidas por uma rede convolucional treinada em larga escala e, em seguida, ajustar parcialmente a arquitetura ao domínio específico das radiografias de tórax (MORID; BORJALI; FIOL, 2021).

Na primeira etapa, descrita na Seção 3.4.4, a base convolucional da MobileNetV2 foi mantida congelada, de modo que apenas a cabeça de classificação adicionada ao modelo fosse treinada. Essa fase teve como objetivo adaptar as camadas finais à tarefa multiclasse proposta, sem alterar os pesos previamente aprendidos pelo *backbone*. Na segunda etapa, apresentada na Seção 3.4.5, realizou-se o ajuste fino parcial (*fine-tuning*) por meio do descongelamento das camadas finais da MobileNetV2, utilizando uma taxa de aprendizado menor para permitir atualizações mais suaves dos pesos da rede (SANDLER et al., 2018; MORID; BORJALI; FIOL, 2021).

3.4.1 Configurações Gerais do Treinamento

Para favorecer a reprodutibilidade dos experimentos, foi fixada uma semente de aleatoriedade igual a 42 (`SEED = 42`), utilizando a função `tf.keras.utils.set_random_seed`. Em processos computacionais que envolvem operações aleatórias, como embaralhamento das amostras, aumento de dados e inicialização de alguns procedimentos internos do treinamento, a definição de uma semente fixa permite controlar a sequência pseudoaleatória gerada. Dessa forma, diferentes execuções do código tendem a seguir condições iniciais semelhantes, reduzindo variações indesejadas nos resultados. Além disso, adotou-se uma configuração padronizada de entrada, com imagens redimensionadas para 224×224 *pixels* e agrupadas em lotes de 32 amostras, conforme descrito na etapa de pré-processamento.

As principais configurações de treinamento, também chamadas de hiperparâmetros, incluíram o tamanho do lote (*batch size*), as taxas de aprendizado, o número máximo de épocas, a taxa de *dropout*, o número de camadas liberadas no ajuste fino e os critérios automáticos de controle do treinamento. Entre esses elementos, a taxa de aprendizado teve papel central, pois controlou a intensidade das atualizações dos pesos durante o treinamento. Esses hiperparâmetros foram definidos com o objetivo de equilibrar estabilidade, capacidade de aprendizado e redução da tendência ao sobreajuste.

O modelo foi compilado utilizando o otimizador Adam e a função de perda *Sparse Categorical Cross-Entropy*, adequada ao uso de rótulos inteiros em problemas de classificação multiclasse. Essa função quantifica a diferença entre as probabilidades previstas pelo modelo e os rótulos reais das imagens. Como métrica principal de acompanhamento, adotou-se a acurácia categórica esparsa, implementada pela métrica `SparseCategoricalAccuracy`, que calcula a proporção de amostras classificadas corretamente considerando a classe de maior probabilidade prevista pelo modelo.

Na primeira fase do treinamento, foi utilizada taxa de aprendizado igual a 1×10^{-4}, com o *backbone* da MobileNetV2 congelado. Na segunda fase, corres-

pondente ao ajuste fino parcial, a taxa de aprendizado foi reduzida para 1×10^{-5}, permitindo atualizar parte das camadas finais da rede de forma mais controlada. Na configuração final do treinamento, não foram aplicados pesos de classe.

O acompanhamento do processo foi realizado por meio do registro do histórico de treinamento em arquivos CSV, permitindo posterior análise das curvas de acurácia e perda. Também foram empregados mecanismos automáticos de controle e seleção do modelo, como salvamento da melhor época, parada antecipada e redução adaptativa da taxa de aprendizado, descritos na subseção seguinte.

3.4.2 Controle de Sobreajuste e Seleção do Modelo

Para controlar a tendência ao sobreajuste e selecionar automaticamente a melhor configuração durante o treinamento, foram utilizados mecanismos de monitoramento baseados no desempenho do conjunto de validação. Em ambas as fases de treinamento, foi empregada a rotina `ModelCheckpoint`, configurada para monitorar a acurácia de validação (`val_accuracy`) e salvar apenas a época com melhor desempenho.

Como o treinamento foi conduzido em duas fases, os melhores modelos foram salvos separadamente para cada etapa. Após a conclusão das duas fases, comparou-se a melhor acurácia de validação obtida em cada uma delas, selecionando-se para a avaliação final o modelo associado ao maior valor de acurácia de validação. Os arquivos correspondentes foram armazenados como `best_model_phase1.keras` e `best_model_phase2.keras`.

Além disso, utilizou-se a técnica de parada antecipada (`EarlyStopping`), configurada com paciência de 8 épocas e restauração dos melhores pesos. Neste contexto, uma época corresponde a uma passagem completa pelo conjunto de treinamento. Assim, a paciência de 8 épocas indica que o treinamento poderia ser interrompido caso não houvesse melhora no desempenho de validação durante esse número consecutivo de passagens completas pelos dados. Essa estratégia reduz o risco de continuidade desnecessária do treinamento e de ajuste excessivo aos dados de treinamento.

Também foi empregado o recurso `ReduceLROnPlateau`, responsável por reduzir a taxa de aprendizado quando a perda de validação (`val_loss`) apresentava estagnação. Essa estratégia favorece uma convergência mais suave, permitindo ajustes menores nos pesos da rede quando o treinamento se aproxima de regiões de menor variação da função de perda. Por fim, o histórico de cada fase foi registrado em arquivos CSV, permitindo a análise posterior das curvas de acurácia e perda apresentadas nos resultados.

3.4.3 Considerações sobre o Desbalanceamento entre Classes

Conforme apresentado na distribuição da base de dados, o conjunto de treinamento apresenta desbalanceamento entre as classes, com maior quantidade de imagens da classe *Pneumonia Bacteriana*. Esse ponto é importante porque uma distribuição desigual pode influenciar o aprendizado do modelo e tornar a acurácia global insuficiente para avaliar o desempenho de forma completa.

Durante os testes preliminares, foram avaliadas alternativas para lidar com esse desbalanceamento, incluindo o uso de pesos de classe. No entanto, na configuração final adotada neste trabalho, optou-se por realizar o treinamento sem pesos externos na função de perda, isto é, com `class_weight=None`. Com isso, o impacto do desbalanceamento foi analisado posteriormente, a partir das métricas por classe e da matriz de confusão.

Essa escolha permitiu avaliar o comportamento do modelo de forma mais transparente, observando não apenas o desempenho médio, mas também possíveis diferenças entre as classes `Bacteria`, `Normal` e `Viral`. As seções seguintes apresentam as duas fases do treinamento: primeiro com o *backbone* da MobileNetV2 congelado e, em seguida, com ajuste fino parcial (*fine-tuning*) das camadas finais da rede.

3.4.4 Treinamento Inicial da Cabeça de Classificação (Fase 1)

Na primeira fase, utilizou-se a MobileNetV2 com pesos pré-treinados no ImageNet e com a camada de classificação original removida (`include_top=False`). O

backbone convolucional foi mantido congelado (`trainable=False`), de modo que apenas a nova cabeça de classificação fosse treinada.

Essa etapa teve como objetivo adaptar o classificador final à tarefa proposta, preservando as representações visuais já aprendidas pela MobileNetV2. Assim, o modelo passou a combinar as características extraídas pela rede base para distinguir entre as classes *Normal*, *Pneumonia Bacteriana* e *Pneumonia Viral*, sem alterar os pesos do *backbone* nessa primeira fase (MORID; BORJALI; FIOL, 2021; SANDLER et al., 2018).

O treinamento inicial foi realizado com o otimizador Adam, taxa de aprendizado igual a 1×10^{-4} e função de perda *Sparse Categorical Cross-Entropy*. Como métrica de acompanhamento, utilizou-se a acurácia categórica esparsa. Essa fase foi configurada para até 12 épocas, isto é, até 12 passagens completas pelo conjunto de treinamento, mantendo os mecanismos de salvamento do melhor modelo, parada antecipada e redução adaptativa da taxa de aprendizado.

3.4.5 Ajuste Fino Parcial do *Backbone* (Fase 2)

Após o treinamento inicial da cabeça de classificação, realizou-se o ajuste fino parcial (*fine-tuning*) para adaptar parte da MobileNetV2 ao domínio das radiografias de tórax. Nessa etapa, foram liberadas para treinamento apenas as últimas 80 camadas da rede, enquanto as camadas anteriores permaneceram congeladas.

Essa estratégia busca equilibrar adaptação e preservação do conhecimento previamente aprendido. As camadas iniciais de uma rede convolucional costumam representar padrões mais gerais, como bordas, texturas e contrastes, enquanto as camadas mais profundas tendem a capturar características mais específicas da tarefa. Por isso, o descongelamento parcial permite ajustar a parte final da arquitetura ao problema estudado, sem modificar excessivamente os pesos obtidos no pré-treinamento (MORID; BORJALI; FIOL, 2021; SANDLER et al., 2018).

Durante o ajuste fino, as camadas de *Batch Normalization* do *backbone* foram mantidas congeladas. Essa decisão foi adotada para preservar a estabilidade do

treinamento, evitando alterações nas estatísticas internas dessas camadas durante a adaptação a uma base menor que a utilizada no pré-treinamento original.

Na segunda fase, o modelo foi recompilado com o otimizador Adam e taxa de aprendizado reduzida para 1×10^{-5}, permitindo atualizações mais suaves dos pesos treináveis. A função de perda permaneceu como *Sparse Categorical Cross-Entropy*, e a métrica de acompanhamento continuou sendo a acurácia categórica esparsa.

O ajuste fino foi configurado para até 20 épocas, isto é, até 20 passagens completas pelo conjunto de treinamento, mantendo mecanismos de controle equivalentes aos utilizados na fase anterior. Esse limite foi maior do que o adotado na primeira fase, configurada para até 12 épocas, porque no ajuste fino parte das camadas finais da MobileNetV2 passou a ser atualizada, exigindo um processo de adaptação mais gradual ao domínio das radiografias de tórax. Ainda assim, em ambas as fases, esses valores corresponderam a limites máximos de treinamento, uma vez que a parada antecipada poderia interromper o processo caso não houvesse melhora no desempenho de validação. Ao final do processo, comparou-se a melhor acurácia de validação obtida nas duas fases, selecionando-se para a avaliação final o modelo com melhor desempenho no conjunto de validação.

3.5 Protocolo de Avaliação e Métricas

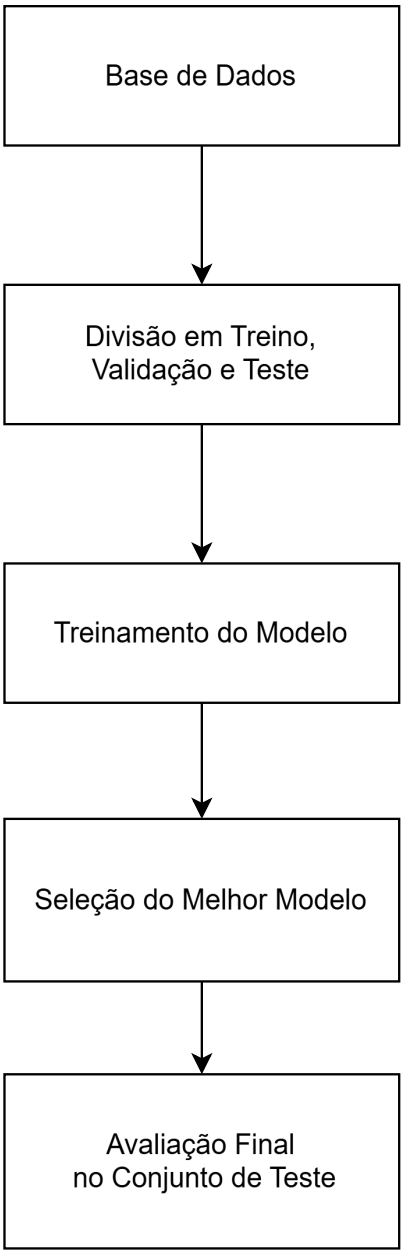
Esta seção apresenta o protocolo utilizado para avaliar o desempenho do modelo. A avaliação foi organizada a partir da separação da base em conjuntos de treinamento, validação e teste, mantendo o conjunto de teste isolado até a etapa final. Também são descritas a avaliação principal do modelo, a análise complementar com limiar calibrado e as métricas utilizadas para interpretar os resultados.

Essa organização permite analisar o modelo não apenas pela acurácia global, mas também pelo comportamento em cada classe. Dessa forma, torna-se possível observar os acertos, os erros e as principais confusões entre as categorias *Pneumonia Bacteriana*, *Normal* e *Pneumonia Viral*.

A Figura 14 apresenta, de forma resumida, o protocolo geral de avaliação adotado

neste trabalho.

Figura 14: Diagrama de blocos do protocolo geral de avaliação adotado no desenvolvimento do classificador.



Fonte: Elaborada pelo autor (2026).

3.5.1 Particionamento da base de dados

A avaliação do modelo foi conduzida utilizando o conjunto de teste, mantido separado das etapas de treinamento, validação, seleção do modelo e calibração complementar do limiar de decisão. Essa separação teve como objetivo estimar a capacidade de generalização do classificador em imagens não utilizadas no ajuste dos pesos da rede nem na definição das configurações do experimento (GOODFELLOW; BENGIO; COURVILLE, 2016; LITJENS et al., 2017).

No presente trabalho, a avaliação final foi realizada em 624 radiografias do conjunto de teste, distribuídas entre as classes *Pneumonia Bacteriana*, *Normal* e *Pneumonia Viral*. Esse conjunto também não foi utilizado na definição do limiar calibrado, preservando sua função como referência independente para análise do desempenho final.

3.5.2 Avaliação principal do modelo

A avaliação principal foi realizada utilizando diretamente a saída da rede neural, sem aplicação de regras externas de pós-processamento. Para cada imagem do conjunto de teste, o modelo produziu um vetor de probabilidades por meio da função *softmax*, e a classe final foi definida como aquela de maior probabilidade.

Essa configuração foi adotada como resultado principal por representar o comportamento direto da MobileNetV2 treinada. Assim, as métricas obtidas nessa etapa refletem a capacidade de classificação do modelo em sua forma base, sem ajustes posteriores na decisão final.

3.5.3 Análise complementar com limiar calibrado

Além da avaliação principal, foi realizada uma análise complementar com calibração de limiar a partir do conjunto de validação. O objetivo foi verificar se uma regra simples de pós-processamento poderia reduzir parte das confusões observadas entre as classes *Normal* e *Pneumonia Viral*.

Para isso, foram obtidas as probabilidades previstas pelo modelo no conjunto de validação e testados diferentes valores de limiar no intervalo de 0,10 a 0,60. A regra foi aplicada apenas aos casos em que a predição original era *Pneumonia Viral*. Nesses casos, se a probabilidade atribuída à classe *Normal* fosse maior ou igual ao limiar definido, a amostra era reclassificada como *Normal*.

O valor selecionado foi $t = 0{,}42$. Após essa escolha, o limiar permaneceu fixo e foi aplicado ao conjunto de teste apenas como análise complementar. Dessa forma, evitou-se utilizar o conjunto de teste para ajustar a regra de decisão.

Essa calibração não altera a arquitetura nem os pesos da rede neural. Trata-se apenas de uma etapa adicional de pós-processamento da saída probabilística do modelo. Por esse motivo, os resultados com limiar calibrado foram interpretados como complementares, enquanto a avaliação principal permaneceu baseada na predição direta da MobileNetV2.

3.5.4 Métricas de desempenho

O desempenho do modelo foi avaliado por meio de acurácia, precisão, revocação e F1-score. Essas métricas permitem analisar o classificador de forma mais completa do que a acurácia isolada, especialmente em um problema multiclasse com distribuição desigual entre as categorias (SOKOLOVA; LAPALME, 2009).

Para o cálculo dessas métricas, consideram-se os conceitos de verdadeiros positivos (*true positives* – TP), falsos positivos (*false positives* – FP), falsos negativos (*false negatives* – FN) e verdadeiros negativos (*true negatives* – TN). No contexto multiclasse, esses valores podem ser interpretados a partir de uma abordagem um-contra-todos, na qual cada classe é analisada em relação às demais.

A acurácia representa a proporção total de predições corretas. A precisão indica, entre as amostras classificadas como pertencentes a uma classe, quantas realmente pertencem a ela. A revocação, também chamada de *recall* ou sensibilidade, mede a capacidade do modelo de identificar corretamente as amostras reais de uma classe. Já o F1-score combina precisão e revocação em uma única métrica, sendo útil para ava-

liar o equilíbrio entre falsos positivos e falsos negativos (SOKOLOVA; LAPALME, 2009).

$$\text{Acurácia} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{Precisão} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Revocação} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$F1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{4}$$

As métricas foram calculadas para cada classe individualmente e também apresentadas por meio das médias macro e ponderada. A média macro atribui o mesmo peso a todas as classes, enquanto a média ponderada considera o suporte de cada categoria no conjunto avaliado. Essa distinção é relevante porque a base apresenta desbalanceamento entre as classes.

Além das métricas numéricas, utilizou-se a matriz de confusão para analisar a distribuição dos acertos e erros entre as classes. Essa representação permite identificar quais categorias foram mais confundidas pelo classificador e contribui para uma leitura mais detalhada do desempenho do modelo (CHICCO; JURMAN, 2020).

3.6 Interpretabilidade com Grad-CAM

Com o objetivo de conferir maior transparência às predições do classificador e complementar a análise quantitativa dos resultados, empregou-se a técnica *Gradient-weighted Class Activation Mapping* (Grad-CAM), proposta por Selvaraju et al. (2017). Essa abordagem permite gerar mapas de ativação que indicam, de forma visual, quais regiões da imagem exerceram maior influência sobre a decisão do modelo

em relação a uma classe de interesse.

No presente trabalho, o Grad-CAM foi aplicado ao modelo final baseado na MobileNetV2, utilizando a camada `out_relu` como referência para a extração dos mapas de ativação. Essa camada corresponde à ativação final da base convolucional da MobileNetV2 e foi escolhida por preservar informações espaciais relevantes antes da etapa de classificação. Dessa forma, os mapas gerados permitem analisar qualitativamente quais regiões da radiografia contribuíram de maneira mais expressiva para a predição do classificador.

O procedimento adotado consiste, inicialmente, em realizar a propagação direta (*forward pass*) da imagem de entrada no modelo, obtendo-se as probabilidades associadas às três classes consideradas: *Normal*, *Pneumonia Bacteriana* e *Pneumonia Viral*. Em seguida, calcula-se o gradiente da pontuação da classe de interesse em relação aos mapas de características da última camada convolucional selecionada. Esses gradientes são agregados espacialmente por média global, produzindo coeficientes que representam a importância relativa de cada mapa de ativação para a classe analisada (SELVARAJU et al., 2017).

Os coeficientes de importância α_k^c são calculados conforme a Equação 5:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{5}$$

A partir desses coeficientes, o mapa Grad-CAM é obtido por meio da combinação ponderada dos mapas de ativação, seguida da aplicação da função ReLU, conforme apresentado na Equação 6:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{6}$$

Nas Equações 5 e 6, Z representa o número total de posições espaciais do mapa de características, A_{ij}^k corresponde à ativação do canal k na posição espacial (i, j), y^c denota a pontuação associada à classe de interesse c, e $L_{\text{Grad-CAM}}^c$ representa o

mapa de ativação produzido para essa classe. A aplicação da função ReLU permite destacar apenas as regiões com contribuição positiva para a decisão do modelo (SELVARAJU et al., 2017).

A Tabela 7 resume a configuração técnica adotada para a geração dos mapas Grad-CAM neste trabalho.

Tabela 7: Configuração técnica adotada para geração dos mapas Grad-CAM.

Parâmetro	**Configuração adotada**
Modelo base	MobileNetV2 pré-treinada no ImageNet
Camada alvo	`out_relu`
Agregação dos gradientes	Média global espacial
Normalização do mapa	Intervalo $[0, 1]$ com ajuste por percentil na visualização
Mapa de cores	`jet`
Classes analisadas	*Pneumonia Bacteriana*, *Normal* e *Pneumonia Viral*

Fonte: Elaborada pelo autor (2026).

Após sua obtenção, o mapa Grad-CAM foi normalizado e redimensionado para as dimensões da radiografia original, permitindo sua sobreposição à imagem de entrada na forma de mapa de calor. Na visualização adotada, utilizou-se o mapa de cores `jet`, no qual tons mais frios indicam menores níveis relativos de ativação, enquanto tons mais quentes indicam regiões com maior contribuição para a predição do modelo.

Os mapas Grad-CAM foram gerados a partir de imagens pertencentes ao conjunto de teste, uma vez que esse subconjunto foi mantido isolado durante as etapas de treinamento e validação. Para a análise qualitativa, foram considerados exemplos envolvendo tanto acertos quanto erros de classificação, possibilitando observar o comportamento do modelo em diferentes situações de decisão.

Essa análise visual permite verificar se o classificador tende a concentrar sua atenção em regiões compatíveis com o campo pulmonar ou se apresenta ativações em áreas menos informativas, como bordas, regiões externas ao tórax ou artefatos da imagem. Esse aspecto é relevante porque, em aplicações médicas, mapas de ativação podem auxiliar na auditoria qualitativa das decisões do modelo, mas devem ser interpretados com cautela, uma vez que não correspondem a uma validação clínica definitiva (SAPORTA et al., 2022).

Assim, o Grad-CAM foi utilizado como ferramenta complementar às métricas quantitativas, contribuindo para uma interpretação visual das decisões do classificador. Dessa forma, a técnica permitiu relacionar o desempenho numérico do modelo com uma análise qualitativa das regiões da radiografia que exerceram maior influência sobre as predições, sem substituir a avaliação objetiva por métricas de desempenho nem a interpretação especializada de um profissional da área da saúde.

3.7 Ambiente Computacional e Reprodutibilidade

Os experimentos computacionais deste trabalho foram conduzidos em ambiente *Python*, utilizando bibliotecas amplamente empregadas em tarefas de aprendizado profundo, processamento numérico, visualização de dados e avaliação de modelos de classificação. As principais ferramentas utilizadas foram o TensorFlow 2.15.1, empregado na construção, treinamento e ajuste fino da rede neural convolucional; o scikit-learn 1.6.1, utilizado no cálculo das métricas de desempenho; o NumPy 1.26.4, empregado em operações numéricas e manipulação de arranjos; e o Matplotlib 3.9.4, adotado para a geração de gráficos, matrizes de confusão e visualizações associadas aos resultados. O ambiente foi executado em Python 3.9.13.

A implementação foi realizada no Visual Studio Code, com execução do código em formato de *notebook*. Essa organização permitiu estruturar o fluxo experimental em etapas sucessivas, contemplando a preparação dos dados, a construção do modelo, o treinamento em duas fases, a avaliação quantitativa e a geração dos mapas Grad-CAM. Os experimentos foram executados em um computador com sistema operacional Windows 10, processador AMD Ryzen 7 5700X 8-Core Processor, 32 GB de memória RAM e GPU NVIDIA GeForce RTX 3070.

As figuras esquemáticas e diagramas utilizados ao longo do trabalho foram elaborados com o auxílio da ferramenta draw.io, buscando padronizar a representação visual dos fluxos metodológicos, da organização dos dados e da arquitetura experimental adotada. As demais visualizações quantitativas, como gráficos de treinamento, matrizes de confusão e exemplos de mapas de ativação, foram geradas a partir dos próprios resultados computacionais obtidos no ambiente Python.

Com o objetivo de favorecer a reprodutibilidade dos resultados, adotou-se semente aleatória fixa igual a 42 nas rotinas pertinentes de inicialização e embaralhamento dos dados. Essa prática contribui para reduzir a variabilidade entre diferentes execuções, especialmente em etapas que envolvem embaralhamento de dados, inicialização de pesos e operações aleatórias de aumento de dados.

Adicionalmente, a configuração da arquitetura empregada, os critérios de treinamento em duas etapas, o protocolo de avaliação e o procedimento de interpretabilidade com Grad-CAM foram documentados de forma explícita. Essa descrição busca permitir a repetição do fluxo experimental em condições computacionais equivalentes, assegurando maior transparência metodológica, consistência experimental e rastreabilidade dos resultados apresentados.

Embora pequenas variações possam ocorrer devido a diferenças de hardware, versões de bibliotecas e implementações internas das rotinas utilizadas, a configuração descrita estabelece um cenário de referência suficientemente especificado para permitir comparações coerentes com estudos futuros.

3.8 Síntese do capítulo

Ao longo deste capítulo, foram apresentadas as principais etapas metodológicas adotadas para o desenvolvimento deste trabalho. Inicialmente, descreveu-se a organização da base de dados, incluindo sua origem, composição, estrutura de diretórios, distribuição das amostras e limitações associadas ao uso de uma base pública de radiografias de tórax.

Em seguida, foram detalhadas as etapas de pré-processamento e preparação dos dados, contemplando o redimensionamento das imagens, o pré-processamento compatível com a MobileNetV2, a organização do *pipeline* com `tf.data` e o aumento de dados aplicado exclusivamente durante o treinamento. Também foi apresentada a arquitetura do modelo, baseada na MobileNetV2 pré-treinada, com adaptação da cabeça de classificação para o problema multiclasse proposto.

O capítulo também descreveu a estratégia de treinamento em duas fases, com-

posta pelo treinamento inicial da cabeça de classificação e pelo ajuste fino parcial do *backbone*, além dos mecanismos de controle de sobreajuste e seleção automática do melhor modelo. Na sequência, foram definidos o protocolo de avaliação, as métricas de desempenho utilizadas e a análise complementar com limiar calibrado.

Por fim, apresentou-se o uso do Grad-CAM como ferramenta de interpretabilidade visual, bem como o ambiente computacional empregado e os cuidados adotados para favorecer a reprodutibilidade experimental. Com isso, foram definidos os procedimentos necessários para a construção, avaliação e interpretação do modelo, servindo como base para a análise dos resultados apresentada no capítulo seguinte.

4 RESULTADOS

Este capítulo apresenta e discute os resultados obtidos a partir do modelo MobileNetV2 desenvolvido para a classificação multiclasse de radiografias de tórax. A análise contempla tanto o desempenho quantitativo do classificador quanto aspectos qualitativos relacionados ao comportamento das predições, aos principais padrões de erro e à interpretabilidade visual das decisões da rede neural.

Inicialmente, são apresentadas as métricas de desempenho obtidas no conjunto de teste, incluindo acurácia, precisão, revocação e F1-score para cada classe. Em seguida, são analisadas as curvas de treinamento e validação, a matriz de confusão, a análise complementar com limiar calibrado e os mapas de ativação gerados por meio do Grad-CAM. Por fim, os resultados obtidos são comparados com estudos relacionados, permitindo contextualizar o desempenho alcançado em relação a outras abordagens aplicadas à classificação de pneumonia em radiografias de tórax.

Dessa forma, a análise dos resultados não se limita à acurácia global do modelo, mas busca compreender de maneira mais detalhada como a rede se comporta diante das classes *Pneumonia Bacteriana*, *Normal* e *Pneumonia Viral*. Essa abordagem permite identificar diferenças de desempenho entre as categorias, avaliar os principais padrões de confusão e discutir as limitações ainda presentes na separação entre classes visualmente semelhantes.

4.1 Desempenho do Modelo MobileNetV2

Nesta seção são apresentados os resultados obtidos pelo modelo MobileNetV2 na tarefa de classificação multiclasse de radiografias de tórax. A avaliação foi realizada em um conjunto de teste independente, composto por 624 imagens distribuídas entre as classes *Pneumonia Bacteriana*, *Normal* e *Pneumonia Viral*. O modelo avaliado corresponde à configuração final obtida após o treinamento em duas etapas, com congelamento inicial do *backbone* e posterior ajuste fino parcial das camadas finais da rede.

A configuração considerada como resultado principal corresponde ao modelo sem aplicação de regras externas de pós-processamento. Dessa forma, avalia-se diretamente o comportamento da MobileNetV2 treinada, sem ajustes posteriores na decisão final. No conjunto de teste, o modelo obteve acurácia de 80,13% e perda de 0,5990. A Tabela 8 apresenta as principais métricas obtidas para cada classe.

Tabela 8: Desempenho do modelo MobileNetV2 no conjunto de teste.

Classe	**Precisão**	**Recall**	**F1-score**	**Suporte**
Pneumonia Bacteriana	0,8531	0,8636	0,8583	242
Normal	0,9167	0,7521	0,8263	234
Pneumonia Viral	0,6150	0,7770	0,6866	148
Acurácia		0,8013		624
Média macro	0,7949	0,7976	0,7904	624
Média ponderada	0,8204	0,8013	0,8056	624

Fonte: Elaborada pelo autor (2026).

A classe *Pneumonia Bacteriana* apresentou o desempenho mais equilibrado, com precisão de 0,8531, recall de 0,8636 e F1-score de 0,8583. Esses valores indicam que o modelo conseguiu identificar boa parte das imagens dessa classe, mantendo equilíbrio entre acertos e falsas classificações.

A classe *Normal* obteve a maior precisão entre as três categorias, atingindo 0,9167. Isso mostra que, quando o modelo classificou uma radiografia como normal, a predição esteve correta na maior parte dos casos. Por outro lado, o recall de 0,7521 indica que uma parcela das imagens normais foi atribuída a outras classes, ponto que é analisado com mais detalhe na matriz de confusão.

A classe *Pneumonia Viral* apresentou a maior dificuldade para o classificador. Embora o recall tenha sido de 0,7770, a precisão foi de 0,6150, indicando que parte das imagens classificadas como virais pertencia, na verdade, a outras categorias. Esse comportamento sugere uma maior sobreposição entre os padrões associados à pneumonia viral e as demais classes, especialmente a classe *Normal*.

De maneira geral, a acurácia de 80,13% e o F1-score ponderado de 0,8056 indicam um desempenho satisfatório para a proposta do trabalho, considerando a referência experimental estabelecida nos objetivos, correspondente a uma acurácia mínima

próxima de 80% no conjunto de teste. Essa interpretação também leva em conta o uso de uma arquitetura leve e a natureza multiclasse do problema. Ainda assim, as diferenças entre as métricas por classe mostram que o desempenho não foi uniforme. Por isso, a análise não deve se limitar à acurácia global, sendo complementada pelas curvas de treinamento, pela matriz de confusão e pela avaliação dos principais padrões de erro nas próximas subseções.

4.2 Curvas de Treinamento e Validação

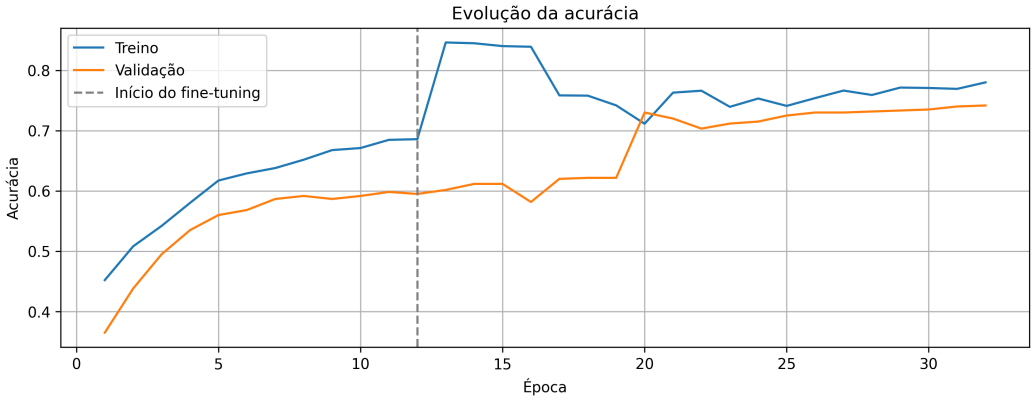
As curvas de treinamento e validação permitem acompanhar o comportamento do modelo ao longo das épocas e verificar a estabilidade do processo de aprendizagem. Neste trabalho, o treinamento da MobileNetV2 foi realizado em duas etapas: inicialmente com o *backbone* congelado e, posteriormente, com ajuste fino parcial das camadas finais da rede. Essa estratégia é comum em abordagens de aprendizado por transferência, especialmente em cenários com bases médicas de tamanho limitado (MORID; BORJALI; FIOL, 2021; SANDLER et al., 2018).

A Figura 15 apresenta a evolução da acurácia durante o treinamento. Na primeira fase, com o *backbone* congelado, o modelo ajustou apenas a cabeça de classificação adicionada à MobileNetV2, alcançando melhor acurácia de validação em torno de 59,83%. Após o início do *fine-tuning*, indicado pela linha tracejada, houve aumento da acurácia de treino e a melhor acurácia de validação chegou a aproximadamente 74,17%, indicando ganho de desempenho com o ajuste parcial das camadas finais.

Ainda que a acurácia de treino tenha permanecido superior à acurácia de validação em parte do processo, especialmente após o início do ajuste fino, a curva de validação apresentou melhora e manteve-se relativamente estável nas épocas finais. Esse comportamento sugere que o modelo aprendeu características relevantes do conjunto de treinamento sem apresentar divergência acentuada entre treino e validação (GOODFELLOW; BENGIO; COURVILLE, 2016).

A Figura 16 apresenta a evolução da função de perda durante o treinamento. Na primeira fase, observa-se redução progressiva da perda de treino e de validação,

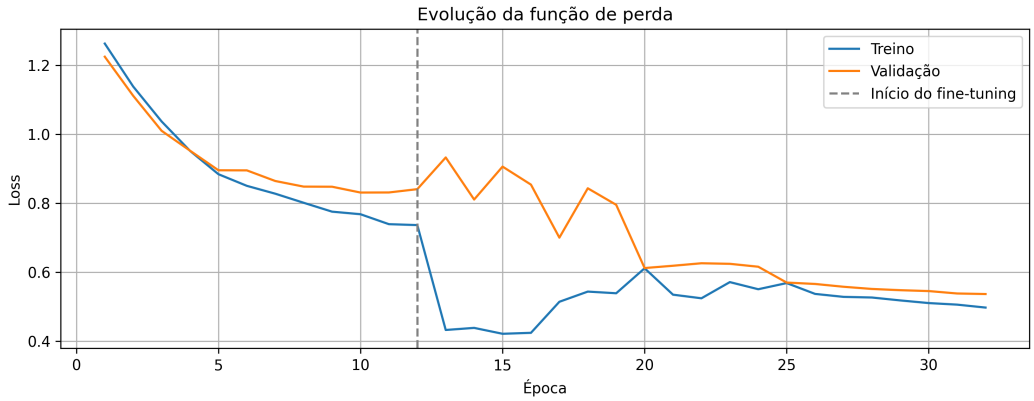
Figura 15: Evolução da acurácia de treino e validação ao longo do processo de treinamento da MobileNetV2.



Fonte: Elaborada pelo autor (2026).

indicando ajuste consistente da cabeça de classificação. No início do *fine-tuning*, aparecem oscilações mais evidentes na perda de validação, comportamento esperado quando parte das camadas pré-treinadas passa a ser atualizada.

Figura 16: Evolução da função de perda de treino e validação durante o treinamento da MobileNetV2.



Fonte: Elaborada pelo autor (2026).

Nas épocas finais, a perda de validação apresentou tendência de estabilização, aproximando-se da perda de treino. A diferença residual entre as curvas é compatível com a natureza da tarefa, que envolve classes visualmente próximas e imagens médicas com variações de aquisição, contraste e posicionamento (LITJENS et al., 2017; CALLI et al., 2021).

De modo geral, as curvas indicam que a estratégia em duas etapas contribuiu

para a evolução do treinamento. A primeira fase permitiu ajustar o classificador final sem alterar as representações convolucionais pré-treinadas, enquanto o *fine-tuning* proporcionou ganho adicional de desempenho, contextualizando os resultados obtidos posteriormente no conjunto de teste.

4.3 Matriz de Confusão e Análise dos Erros

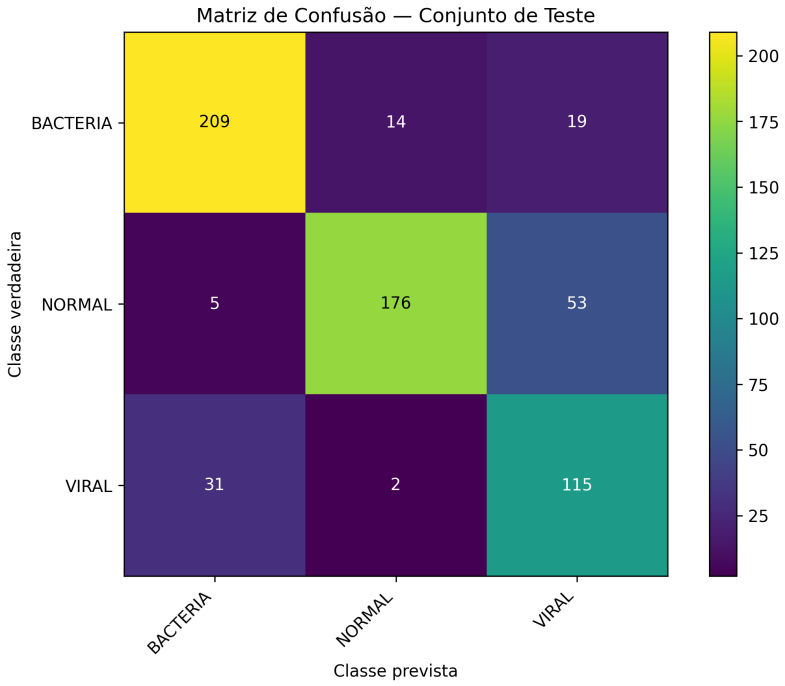
A matriz de confusão permite analisar o desempenho do classificador de forma mais detalhada do que a acurácia global, pois mostra como os acertos e erros se distribuem entre as classes avaliadas. Em problemas multiclasse, essa análise é especialmente importante, uma vez que um mesmo valor de acurácia pode esconder comportamentos diferentes entre as categorias. Neste trabalho, a matriz de confusão auxilia na identificação das principais confusões entre as classes *Pneumonia Bacteriana*, *Normal* e *Pneumonia Viral* (CHICCO; JURMAN, 2020; SOKOLOVA; LAPALME, 2009).

A Figura 17 apresenta a matriz de confusão em valores absolutos para o conjunto de teste. Observa-se que a classe *Pneumonia Bacteriana* foi a mais bem reconhecida pelo modelo, com 209 imagens classificadas corretamente de um total de 242. A classe *Normal* apresentou 176 acertos em 234 imagens, enquanto a classe *Pneumonia Viral* apresentou 115 acertos em 148 imagens.

A análise dos erros mostra que as confusões se concentraram principalmente entre as classes *Normal* e *Pneumonia Viral*. Ao todo, 53 imagens normais foram classificadas como virais. Esse comportamento indica uma dificuldade do modelo em separar radiografias normais de imagens associadas à pneumonia viral, possivelmente devido à presença de padrões visuais sutis, variações de contraste, posicionamento do paciente ou características anatômicas interpretadas como alterações pulmonares discretas. Essa dificuldade é coerente com a literatura sobre radiografias de tórax, que destaca a variabilidade das imagens médicas e a sobreposição de achados entre diferentes condições pulmonares (LITJENS et al., 2017; CALLI et al., 2021).

Outro erro relevante ocorreu entre as classes *Pneumonia Viral* e *Pneumonia Bac-*

Figura 17: Matriz de confusão do conjunto de teste em valores absolutos.



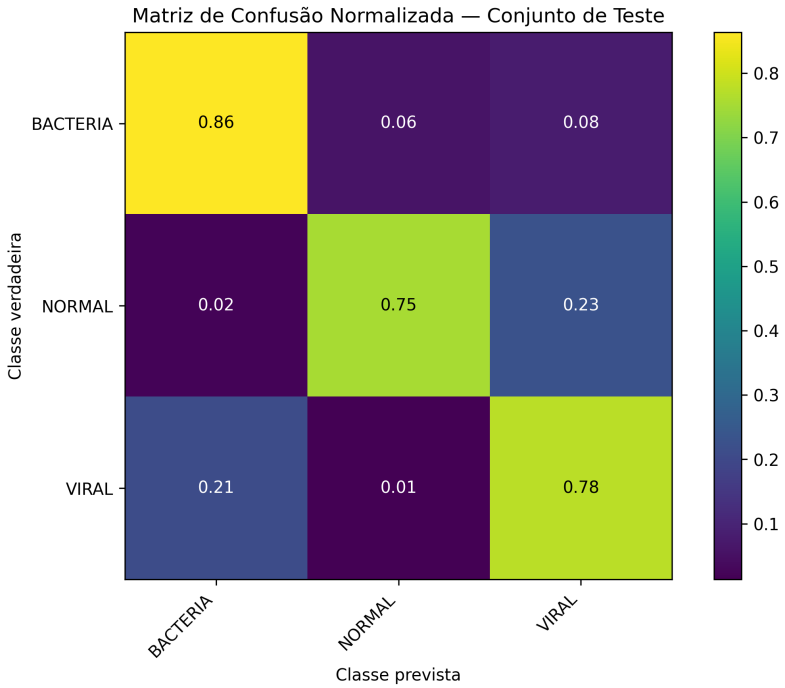
Fonte: Elaborada pelo autor (2026).

teriana, com 31 imagens virais classificadas como bacterianas. Esse resultado sugere que, em parte dos casos, o modelo associou padrões de pneumonia viral a características mais próximas da classe bacteriana. Essa confusão também é compatível com a natureza do problema, uma vez que radiografias de tórax podem apresentar sobreposição visual entre diferentes etiologias de pneumonia, principalmente quando os achados são discretos, difusos ou multifocais (GARG et al., 2019; FRANQUET, 2001; KOO et al., 2020).

A Figura 18 apresenta a matriz de confusão normalizada por classe verdadeira, permitindo observar a proporção de acertos e erros em cada categoria. Nessa representação, a classe *Pneumonia Bacteriana* apresentou aproximadamente 86% de acerto, enquanto as classes *Normal* e *Pneumonia Viral* apresentaram cerca de 75% e 78%, respectivamente.

A matriz normalizada reforça que o melhor desempenho proporcional ocorreu na classe *Pneumonia Bacteriana*, enquanto a maior dificuldade permaneceu relacionada à distinção entre *Normal* e *Pneumonia Viral*. Ainda assim, a predominância dos valores na diagonal principal mostra que o modelo conseguiu aprender padrões rele-

Figura 18: Matriz de confusão normalizada do conjunto de teste.



Fonte: Elaborada pelo autor (2026).

vantes das três classes, embora ainda apresente limitações na separação de categorias visualmente próximas.

Do ponto de vista da aplicação, esses resultados mostram que a MobileNetV2 apresenta potencial como ferramenta de apoio à classificação de radiografias de tórax, mas não deve ser interpretada como um sistema autônomo de diagnóstico. Os erros observados, principalmente entre as classes *Normal* e *Pneumonia Viral*, reforçam a importância de analisar o modelo por diferentes perspectivas. Por esse motivo, as próximas subseções apresentam a análise complementar com limiar calibrado e, posteriormente, a interpretação visual das predições por meio do Grad-CAM (SELVARAJU et al., 2017; SAPORTA et al., 2022).

4.4 Análise Complementar com Limiar Calibrado

Além da avaliação principal do modelo MobileNetV2, foi realizada uma análise complementar com ajuste de limiar a partir do conjunto de validação. O objetivo foi verificar se uma regra simples de pós-processamento poderia reduzir parte

das confusões observadas entre as classes *Normal* e *Pneumonia Viral*, identificadas anteriormente na matriz de confusão.

A avaliação principal considerou diretamente a classe de maior probabilidade prevista pela rede neural, sem regras externas de pós-processamento. Entretanto, como algumas predições apresentaram probabilidades próximas entre as classes *Normal* e *Pneumonia Viral*, foi testada uma regra complementar calibrada exclusivamente no conjunto de validação. Dessa forma, evitou-se escolher o limiar com base no conjunto de teste, preservando a validade da avaliação final.

O limiar selecionado na validação foi de 0,42. Com essa regra, a acurácia no conjunto de teste passou de 80,13% para 80,45%. Também houve leve aumento no F1-score macro, de 0,7904 para 0,7936, e no F1-score ponderado, de 0,8056 para 0,8086. A Tabela 9 apresenta a comparação entre o modelo base e a configuração com limiar calibrado.

Tabela 9: Comparação entre o modelo base e a análise complementar com limiar calibrado.

Configuração	**Acurácia**	**F1-score macro**	**F1-score ponderado**
Modelo base	80,13%	0,7904	0,8056
Modelo com limiar calibrado	80,45%	0,7936	0,8086

Fonte: Elaborada pelo autor (2026).

Apesar da melhora observada, o ganho foi discreto, correspondendo a 0,32 ponto percentual de acurácia. Na matriz de confusão, o principal efeito da regra foi a correção de duas imagens da classe *Normal* que anteriormente haviam sido classificadas como *Pneumonia Viral*. Assim, o limiar calibrado não modificou substancialmente o comportamento geral do classificador, mas indicou que parte dos erros entre essas duas classes estava associada a decisões próximas da fronteira de classificação.

Por esse motivo, o resultado principal deste trabalho permanece sendo o modelo MobileNetV2 sem regra de pós-processamento. A configuração com limiar calibrado é tratada apenas como análise complementar, pois seu ganho foi limitado e não alterou de forma significativa o perfil de desempenho do modelo. Ainda assim, essa análise contribui para compreender melhor a natureza dos erros entre *Normal* e

Pneumonia Viral, reforçando a importância de avaliar não apenas a acurácia global, mas também o comportamento por classe.

4.5 Interpretabilidade com Grad-CAM

Além da avaliação quantitativa, este trabalho utilizou o Grad-CAM (*Gradient-weighted Class Activation Mapping*) como ferramenta de interpretabilidade do modelo. Essa técnica gera mapas de ativação sobre a imagem original, indicando as regiões que mais contribuíram para a decisão da rede neural. Em modelos convolucionais aplicados a imagens médicas, esse tipo de análise auxilia na inspeção qualitativa das regiões consideradas pelo classificador durante a predição (SELVARAJU et al., 2017; SAPORTA et al., 2022).

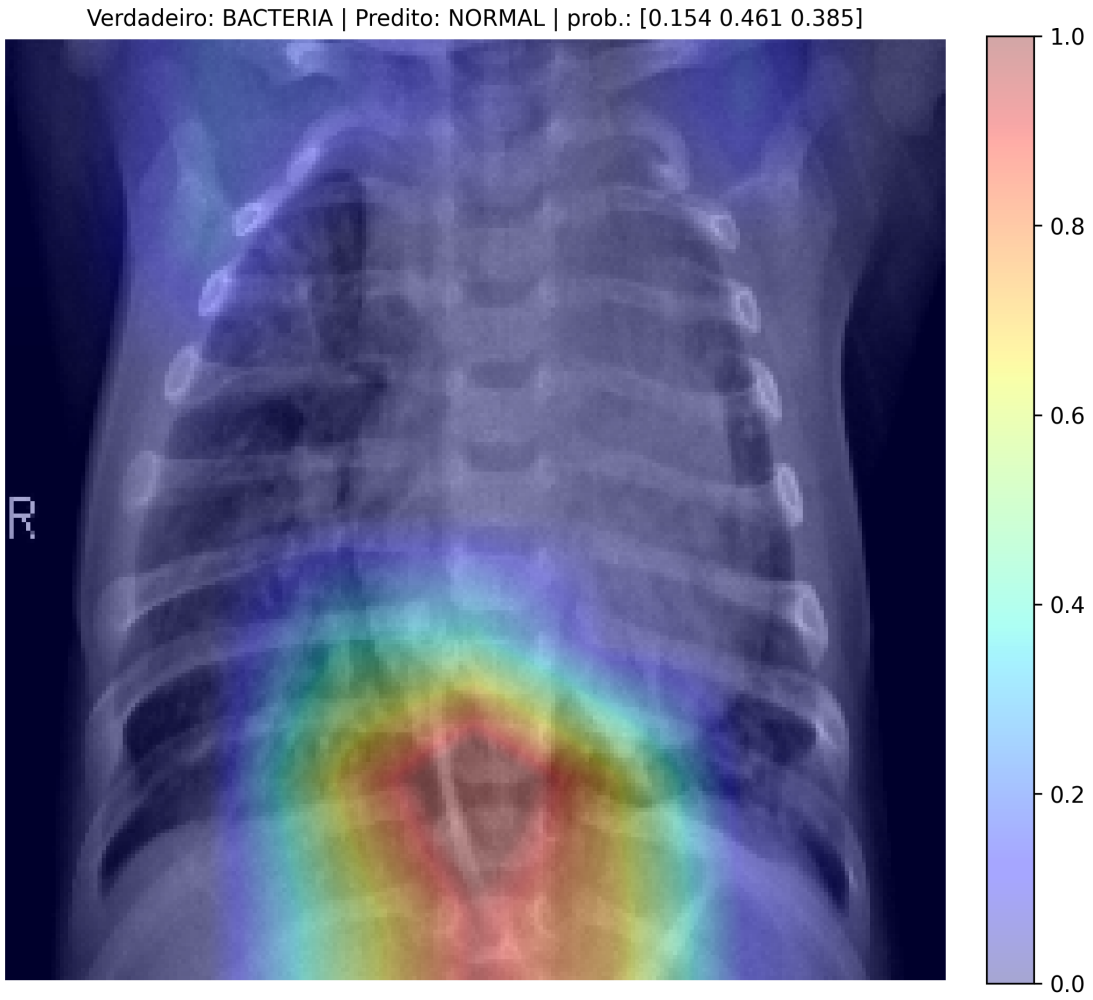
O Grad-CAM é uma técnica *post-hoc*, aplicada após o treinamento do modelo, sem necessidade de alterar sua arquitetura ou realizar novo treinamento. O método utiliza os gradientes da classe de interesse em relação aos mapas de características de uma camada convolucional profunda, gerando um mapa de calor que destaca as áreas mais relevantes para a decisão. Nas visualizações, regiões em tons mais quentes indicam maior contribuição relativa para a predição (SELVARAJU et al., 2017).

No presente trabalho, o Grad-CAM foi aplicado ao modelo final baseado na MobileNetV2. A camada alvo selecionada foi a `out_relu`, correspondente à ativação final da base convolucional da MobileNetV2. Essa escolha é adequada ao objetivo da análise, pois utiliza mapas de características profundos ainda capazes de preservar informação espacial suficiente para projeção sobre a radiografia original.

A Figura 19 apresenta um exemplo individual de Grad-CAM obtido a partir de uma imagem do conjunto de teste. Nesse tipo de análise, a interpretação não deve considerar apenas a classe prevista, mas também a localização das regiões ativadas. Quando o mapa de calor se concentra majoritariamente sobre o campo pulmonar, a decisão do modelo torna-se visualmente mais plausível. Por outro lado, ativações em regiões periféricas, bordas da imagem ou áreas externas ao tórax podem indicar

limitações do classificador ou dependência de padrões pouco relacionados à condição analisada.

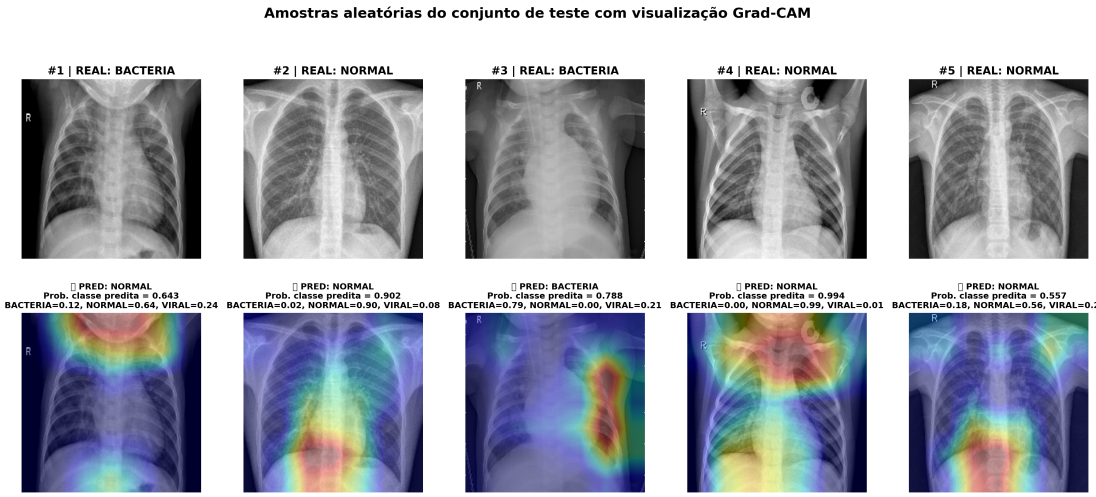
Figura 19: Exemplo individual de Grad-CAM aplicado a uma radiografia de tórax do conjunto de teste.



Fonte: Elaborada pelo autor (2026).

Além da análise individual, também foi gerada uma visualização em grade com amostras aleatórias do conjunto de teste, conforme apresentado na Figura 20. Essa visualização permite comparar diferentes casos simultaneamente, observando a imagem original, a classe verdadeira, a predição do modelo e o respectivo mapa de ativação. Com isso, torna-se possível analisar padrões de acerto e erro de forma complementar à matriz de confusão.

Figura 20: Amostras aleatórias do conjunto de teste com visualização Grad-CAM.



Fonte: Elaborada pelo autor (2026).

Observa-se que a Figura 20 apresenta tanto predições corretas quanto um caso de erro. Na primeira amostra, por exemplo, a imagem pertence à classe *Pneumonia Bacteriana*, mas foi classificada pelo modelo como *Normal*. Esse exemplo é relevante porque mostra uma limitação real do classificador e reforça que a análise não se restringe apenas aos casos favoráveis. Dessa forma, a visualização permite observar o comportamento do modelo de maneira mais transparente, incluindo situações em que a decisão da rede não corresponde ao rótulo verdadeiro.

Nos exemplos analisados, o modelo frequentemente atribuiu maior relevância a regiões localizadas no campo pulmonar, o que indica coerência visual em parte das decisões tomadas pela MobileNetV2. Entretanto, também foram observadas ativações em regiões próximas ao diafragma, bordas da radiografia e áreas de contraste elevado. Esse comportamento mostra que, embora o Grad-CAM contribua para a interpretação do modelo, seus mapas não devem ser entendidos como segmentações precisas da lesão ou como prova definitiva da justificativa clínica da predição.

Essa cautela é importante porque métodos de saliência apresentam limitações na interpretação de redes profundas. Diferentes técnicas de explicabilidade podem destacar regiões distintas para uma mesma imagem, e um mapa visualmente plausível não garante, por si só, que o modelo tenha aprendido uma relação causal com a doença analisada. Assim, o Grad-CAM deve ser interpretado como uma ferramenta

auxiliar de inspeção qualitativa, e não como validação clínica isolada do modelo (SAPORTA et al., 2022).

De modo geral, a aplicação do Grad-CAM complementa os resultados quantitativos apresentados nas subseções anteriores. Enquanto as métricas e a matriz de confusão indicam o desempenho numérico do classificador, os mapas de ativação permitem observar se as regiões destacadas pela rede são compatíveis com a tarefa proposta. Dessa forma, a interpretabilidade contribui para uma análise mais transparente do modelo, especialmente em um problema sensível como a classificação de radiografias de tórax.

4.6 Comparação com Trabalhos Similares

A comparação com trabalhos similares permite contextualizar o desempenho obtido neste estudo em relação à literatura sobre classificação de pneumonia em radiografias de tórax. Entretanto, essa comparação deve ser interpretada com cautela, pois os estudos diferem quanto à base de dados, número de classes, arquitetura utilizada, estratégia de divisão dos dados e critérios de avaliação. Dessa forma, a Tabela 10 não tem o objetivo de estabelecer uma comparação absoluta entre os modelos, mas sim de situar os resultados deste trabalho frente a abordagens já exploradas na literatura (KUNDU et al., 2021; CHOUHAN et al., 2020).

Tabela 10: Comparação entre o desempenho obtido neste trabalho e estudos relacionados.

Trabalho	**Arquitetura**	**Nº de classes**	**Base de dados**	**Acurácia**
Este trabalho	MobileNetV2	3	Chest X-Ray Images (Kaggle)	80,13%
Kundu et al. (2021)	*Ensemble* de modelos profundos	2	Kermany e RSNA	98,81% e 86,85%
Chouhan et al. (2020)	*Transfer learning* com *ensemble*	2	Chest X-Ray Images	96,39%

Fonte: Elaborada pelo autor com base em Kundu et al. (2021) e Chouhan et al. (2020).

Observa-se que os trabalhos de Kundu et al. (2021) e Chouhan et al. (2020) reportam acurácias superiores às obtidas neste estudo. No entanto, ambos foram desenvolvidos em cenários distintos, com destaque para tarefas de classificação binária e, em alguns casos, uso de *ensembles* de modelos profundos. Essas diferenças tornam a comparação direta limitada, especialmente porque o presente trabalho aborda uma tarefa multiclasse, envolvendo as classes *Pneumonia Bacteriana*, *Normal* e *Pneumonia Viral* (KUNDU et al., 2021; CHOUHAN et al., 2020).

A classificação em três classes é mais desafiadora do que a separação binária entre pneumonia e normal, pois exige que o modelo diferencie não apenas a presença de alterações pulmonares, mas também padrões associados a categorias com possível sobreposição visual. Essa dificuldade foi observada nos próprios resultados deste trabalho, principalmente nas confusões entre *Normal* e *Pneumonia Viral*.

Além disso, a proposta adotou a MobileNetV2, uma arquitetura leve, com menor custo computacional e estrutura mais simples em comparação a abordagens baseadas em *ensembles*. Embora modelos combinados possam alcançar desempenho superior, eles tendem a apresentar maior complexidade de treinamento, maior custo de inferência e menor simplicidade de implantação. Assim, a acurácia de 80,13% deve ser analisada considerando o equilíbrio entre desempenho, leveza arquitetural e interpretabilidade visual (SANDLER et al., 2018).

Outro aspecto importante é que a avaliação realizada neste trabalho não se limitou à acurácia global. Foram analisadas métricas por classe, matriz de confusão, curvas de treinamento, limiar calibrado e mapas de ativação Grad-CAM. Essa abordagem permitiu compreender melhor o comportamento do classificador, identificando as classes mais difíceis, os principais padrões de erro e as regiões que influenciaram parte das decisões da rede (SELVARAJU et al., 2017; SAPORTA et al., 2022).

Portanto, embora os valores de acurácia sejam menores que os reportados por alguns estudos da literatura, os resultados obtidos são compatíveis com a proposta deste trabalho. O modelo apresentou desempenho satisfatório para a tarefa multiclasse considerada, uma vez que atingiu a referência experimental previamente esta-

belecida nos objetivos, correspondente a uma acurácia mínima próxima de 80% no conjunto de teste. Além disso, manteve baixa complexidade computacional e possibilitou análise visual por Grad-CAM. Dessa forma, a solução desenvolvida mostra-se tecnicamente viável como ferramenta de apoio e investigação em classificação automática de radiografias de tórax, respeitadas as limitações da base utilizada e a ausência de validação clínica externa.

4.7 Síntese do capítulo

Este capítulo apresentou os principais resultados obtidos pela MobileNetV2 na classificação multiclasse de radiografias de tórax. No conjunto de teste, o modelo alcançou acurácia de 80,13%, com desempenho mais consistente para a classe *Pneumonia Bacteriana* e maior dificuldade na separação entre *Normal* e *Pneumonia Viral*.

As curvas de treinamento e validação indicaram que a estratégia em duas etapas, composta pelo treinamento inicial da cabeça de classificação e pelo ajuste fino parcial da MobileNetV2, contribuiu para a evolução do aprendizado. A matriz de confusão, por sua vez, permitiu identificar os principais padrões de erro, especialmente entre classes visualmente próximas.

A análise complementar com limiar calibrado apresentou ganho discreto, elevando a acurácia de 80,13% para 80,45%, sem alterar de forma significativa o comportamento geral do classificador. Por esse motivo, o resultado principal permaneceu associado à MobileNetV2 sem regra externa de pós-processamento.

Por fim, os mapas Grad-CAM possibilitaram uma análise visual complementar das decisões da rede, permitindo observar regiões de maior ativação nas radiografias analisadas. A comparação com trabalhos similares ajudou a contextualizar o desempenho obtido, considerando diferenças entre bases de dados, número de classes, arquiteturas e critérios de avaliação. A partir desses resultados, o capítulo seguinte apresenta as considerações finais, as limitações do estudo e as possibilidades de trabalhos futuros.

5 CONSIDERAÇÕES FINAIS

Este capítulo apresenta as considerações finais do trabalho, retomando os principais resultados obtidos, as limitações observadas e as possibilidades de continuidade da pesquisa. A discussão final considera a avaliação quantitativa, a análise dos erros e a interpretação visual por Grad-CAM no contexto da classificação multiclasse de radiografias de tórax.

5.1 Conclusões

Este trabalho teve como objetivo desenvolver e avaliar um sistema de classificação automática de radiografias de tórax em três classes: *Normal*, *Pneumonia Bacteriana* e *Pneumonia Viral*. Para isso, foi utilizada a arquitetura MobileNetV2 com aprendizado por transferência, treinamento em duas etapas e interpretação visual por meio do Grad-CAM. De modo geral, os objetivos propostos foram atendidos, uma vez que a base foi organizada, o modelo foi desenvolvido, treinado, avaliado e interpretado por meio de métricas quantitativas e mapas de ativação.

No conjunto de teste, composto por 624 imagens, o modelo alcançou acurácia de 80,13%. O melhor desempenho ocorreu na classe *Pneumonia Bacteriana*, com recall de 86,36% e F1-score de 0,8583. A classe *Normal* apresentou elevada precisão, atingindo 91,67%, enquanto a classe *Pneumonia Viral* apresentou maior dificuldade de discriminação. Esse resultado indica que a separação entre radiografias normais e imagens associadas à pneumonia viral foi um dos principais desafios do classificador.

A matriz de confusão mostrou que os erros se concentraram principalmente entre as classes *Normal* e *Pneumonia Viral*, além de algumas confusões entre *Pneumonia Viral* e *Pneumonia Bacteriana*. Esse comportamento é compatível com a natureza da tarefa, pois diferentes padrões radiográficos podem apresentar sobreposição visual, especialmente em casos com achados discretos ou pouco evidentes.

A estratégia de treinamento em duas etapas mostrou-se adequada para a proposta do trabalho. O treinamento inicial da cabeça de classificação permitiu adaptar

o modelo à tarefa multiclasse, enquanto o ajuste fino parcial da MobileNetV2, associado à redução da taxa de aprendizado, contribuiu para melhorar o desempenho da rede de forma controlada. A análise complementar com limiar calibrado apresentou ganho discreto, elevando a acurácia de 80,13% para 80,45%, sem alterar de forma significativa o comportamento geral do classificador. Por esse motivo, o resultado principal permaneceu sendo o modelo sem regra externa de pós-processamento.

A aplicação do Grad-CAM complementou a avaliação quantitativa ao permitir a inspeção visual das regiões mais relevantes para as decisões da rede. Em parte dos exemplos, os mapas de ativação apresentaram coerência com regiões do campo pulmonar. Entretanto, também foram observadas ativações em áreas periféricas, próximas ao diafragma ou em regiões de maior contraste. Assim, o Grad-CAM mostrou-se útil como ferramenta auxiliar de análise qualitativa, mas não deve ser interpretado como validação clínica isolada.

Conclui-se, portanto, que a MobileNetV2, associada ao aprendizado por transferência e ao Grad-CAM, representa uma alternativa tecnicamente viável para apoio à classificação multiclasse de radiografias de tórax. A principal contribuição deste trabalho está na construção e avaliação de uma abordagem computacional leve, que atingiu a referência experimental previamente estabelecida de acurácia mínima próxima de 80% no conjunto de teste, além de permitir análise por classe e interpretação visual das predições. Ainda assim, o sistema não deve ser tratado como ferramenta diagnóstica autônoma, mas como uma proposta experimental de apoio e investigação, com espaço para aprimoramentos futuros.

5.2 Limitações do Trabalho

Embora os resultados obtidos indiquem a viabilidade técnica da abordagem proposta, algumas limitações devem ser consideradas. A primeira está relacionada ao conjunto de dados utilizado. Trata-se de uma base pública específica, reorganizada em três classes a partir da estrutura e da nomenclatura original dos arquivos. Assim, a separação entre *Pneumonia Bacteriana* e *Pneumonia Viral* não corresponde a uma nova validação clínica realizada neste trabalho, mas ao reaproveitamento dos

rótulos disponibilizados na própria base.

Além disso, a distribuição das amostras não é totalmente equilibrada, especialmente no conjunto de treinamento, no qual a classe *Pneumonia Bacteriana* apresenta maior quantidade de imagens. Esse desbalanceamento pode influenciar o aprendizado da rede e contribuir para diferenças de desempenho entre as classes. Por esse motivo, os resultados foram interpretados não apenas pela acurácia global, mas também pelas métricas por classe e pela matriz de confusão.

Outra limitação importante está relacionada à generalização. O modelo foi treinado, validado e testado em uma única base de radiografias de tórax, o que pode limitar sua aplicação a imagens obtidas em outros hospitais, equipamentos, protocolos de aquisição ou populações. Em cenários reais, variações de contraste, posicionamento, qualidade da imagem e características dos pacientes podem alterar o comportamento do classificador.

Também deve ser considerado que a distinção entre *Normal*, *Pneumonia Bacteriana* e *Pneumonia Viral* é naturalmente desafiadora quando baseada apenas na radiografia. A matriz de confusão mostrou maior dificuldade na separação entre *Normal* e *Pneumonia Viral*, além de algumas confusões entre *Pneumonia Viral* e *Pneumonia Bacteriana*. Esse comportamento é compatível com a sobreposição visual existente entre determinados padrões radiográficos.

Do ponto de vista computacional, a escolha da MobileNetV2 trouxe vantagens relacionadas à simplicidade arquitetural e ao menor custo de processamento. No entanto, por ser uma arquitetura leve, sua capacidade de representação pode ser inferior à de redes mais profundas ou de abordagens em *ensemble*. Dessa forma, há um compromisso entre desempenho, custo computacional e facilidade de implementação, o que justifica a comparação com outras arquiteturas em estudos futuros.

Por fim, este trabalho não realizou validação clínica com especialistas da área médica. As análises foram conduzidas por meio de métricas computacionais, matriz de confusão e mapas Grad-CAM, o que é adequado para a avaliação técnica do modelo, mas não substitui a análise profissional. Além disso, o Grad-CAM deve

ser interpretado apenas como ferramenta auxiliar de inspeção qualitativa, pois seus mapas não representam segmentações precisas nem garantem explicações causais para as decisões da rede.

5.3 Trabalhos Futuros

Como continuidade deste trabalho, uma primeira possibilidade consiste em ampliar a base de dados utilizada. A inclusão de radiografias provenientes de diferentes instituições, equipamentos e protocolos de aquisição permitiria avaliar melhor a capacidade de generalização do modelo em cenários distintos daquele utilizado no treinamento.

Também seria relevante comparar a MobileNetV2 com outras arquiteturas de aprendizado profundo, como DenseNet, ResNet e EfficientNet. Embora a MobileNetV2 tenha sido escolhida por sua leveza e simplicidade computacional, arquiteturas mais profundas ou com diferentes estratégias de conexão entre camadas poderiam ser avaliadas para verificar possíveis ganhos de desempenho, especialmente nas classes com maior dificuldade de separação.

Outra linha de continuidade envolve o estudo de estratégias para reduzir as confusões entre *Normal*, *Pneumonia Bacteriana* e *Pneumonia Viral*. Para isso, poderiam ser investigadas técnicas de balanceamento de dados, funções de perda alternativas, aumento de dados mais direcionado e ajustes mais detalhados dos hiperparâmetros do modelo, como taxa de aprendizado, tamanho do lote, número de camadas liberadas no ajuste fino e número máximo de épocas.

A incorporação de etapas de segmentação ou recorte automático da região pulmonar também representa uma possibilidade de melhoria. Como alguns mapas Grad-CAM indicaram ativações em regiões próximas ao diafragma, bordas da imagem ou áreas de maior contraste, delimitar previamente os pulmões poderia reduzir a influência de regiões menos relevantes para a classificação.

Estudos futuros também poderiam explorar modelos multimodais, combinando radiografias com informações clínicas complementares, como idade, sintomas, histó-

rico do paciente e exames laboratoriais. Essa abordagem se aproxima mais da prática médica, na qual a interpretação da imagem normalmente é feita em conjunto com outros dados do paciente.

Por fim, recomenda-se aprofundar a análise de interpretabilidade. Além do Grad-CAM, outras técnicas de explicabilidade poderiam ser comparadas, e os mapas de ativação poderiam ser avaliados por profissionais da área da saúde. Essa etapa permitiria verificar com maior rigor se as regiões destacadas pelo modelo apresentam coerência com achados radiológicos relevantes, contribuindo para aumentar a confiabilidade da abordagem.

Referências

ABDAR, M. et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. **Information Fusion**, v. 76, p. 243–297, 2021.

ALAPAT, D. J.; MENON, M. V.; ASHOK, S. A review on detection of pneumonia in chest x-ray images using neural networks. **Journal of Biomedical Physics and Engineering**, v. 12, n. 6, p. 551–558, 2022.

Bgraysea. **Are convolutions actually multiplying over nxn grids?** 2021. Fast.ai Course Forums. Acesso em: 11 maio 2026. Disponível em: <https://forums.fast.ai/t/are-convolutions-actually-multiplying-over-nxn-grids/85137>.

CALLI, R. et al. Deep learning for chest x-ray analysis: A survey. **Medical Image Analysis**, v. 72, p. 102116, 2021.

Centers for Disease Control and Prevention. **About Pneumonia**. 2026. Acesso em: 26 abr. 2026. Disponível em: <https://www.cdc.gov/pneumonia/about/index.html>.

CHICCO, D.; JURMAN, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. **BMC Genomics**, v. 21, n. 1, p. 6, 2020.

CHOUHAN, V. et al. A novel transfer learning based approach for pneumonia detection in chest x-ray images. **Applied Sciences**, v. 10, n. 2, p. 559, 2020.

FRANQUET, T. Imaging of pneumonia: Trends and algorithms. **European Respiratory Journal**, v. 18, n. 1, p. 196–208, 2001.

FREEMAN, A. M. et al. **Viral Pneumonia**. 2023. StatPearls [Internet]. Acesso em: 26 abr. 2026. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK513286/>.

GARG, M. et al. Spectrum of imaging findings in pulmonary infections. part 1: Bacterial and viral. **Polish Journal of Radiology**, v. 84, p. e205–e213, 2019.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016.

IRVIN, J. et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. [S.l.: s.n.], 2019. v. 33, n. 1, p. 590–597.

KIM, J. et al. Simplified transfer learning for chest radiography models using less data. **Radiology**, v. 305, n. 1, p. e220019, 2022.

KOO, H. J. et al. Radiographic and ct features of viral pneumonia. **RadioGraphics**, v. 40, n. 3, p. 719–739, 2020.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2012. v. 25.

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.

LE, S. et al. Transfer learning for medical image classification: A literature review. **BMC Medical Imaging**, v. 22, n. 1, p. 69, 2022.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, 2015.

LECUN, Y. et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998.

LIMA, K. K. d. S. **Capítulo 1**. 2019. Repositório Institucional da Universidade Federal Rural do Semi-Árido. Acesso em: 11 maio 2026. Disponível em: <https://repositorio.ufersa.edu.br/server/api/core/bitstreams/fc5ee196-e67a-472f-8d99-d9ee8cf3f1b2/content>.

LITJENS, G. et al. A survey on deep learning in medical image analysis. **Medical Image Analysis**, v. 42, p. 60–88, 2017.

MOONEY, P. **Chest X-Ray Images (Pneumonia)**. 2018. Acesso em: 26 abr. 2026. Disponível em: <https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia>.

MORID, M. A.; BORJALI, A.; FIOL, G. D. A scoping review of transfer learning research on medical image analysis using imagenet. **Computers in Biology and Medicine**, v. 128, p. 104115, 2021.

MSD Manual Professional Edition. **Community-Acquired Pneumonia**. 2024. Acesso em: 26 abr. 2026. Disponível em: <https://www.msdmanuals.com/professional/pulmonary-disorders/pneumonia/community-acquired-pneumonia>.

MSD Manual Professional Edition. **Overview of Pneumonia**. 2024. Acesso em: 26 abr. 2026. Disponível em: <https://www.msdmanuals.com/professional/pulmonary-disorders/pneumonia/overview-of-pneumonia>.

National Heart, Lung, and Blood Institute. **Pneumonia – Diagnosis**. 2022. Acesso em: 26 abr. 2026. Disponível em: <https://www.nhlbi.nih.gov/health/pneumonia/diagnosis>.

National Heart, Lung, and Blood Institute. **What is Pneumonia?** 2022. Acesso em: 26 abr. 2026. Disponível em: <https://www.nhlbi.nih.gov/sites/default/files/publications/what_is_pneumonia.pdf>.

Radiopaedia.org. **Lobar Pneumonia**. 2024. Acesso em: 26 abr. 2026. Disponível em: <https://radiopaedia.org/articles/lobar-pneumonia>.

RAJPURKAR, P. et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. **arXiv preprint arXiv:1711.05225**, 2017. Disponível em: <https://arxiv.org/abs/1711.05225>.

SANDLER, M. et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2018. p. 4510–4520.

SAPORTA, A. et al. Benchmarking saliency methods for chest x-ray interpretation. **Nature Machine Intelligence**, v. 4, p. 867–878, 2022.

SATTAR, S. B. A.; SHARMA, S. **Bacterial Pneumonia**. 2024. StatPearls [Internet]. Acesso em: 26 abr. 2026. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK513321/>.

SELVARAJU, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: **Proceedings of the IEEE International Conference on Computer Vision**. [S.l.: s.n.], 2017. p. 618–626.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. **Journal of Big Data**, v. 6, n. 1, p. 60, 2019.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information Processing & Management**, v. 45, n. 4, p. 427–437, 2009.

The Radiology Assistant. **Chest X-Ray – Lung Disease**. 2014. Acesso em: 26 abr. 2026. Disponível em: <https://radiologyassistant.nl/chest/chest-x-ray/lung-disease>.

WANG, X. et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2017. p. 2097–2106.

World Health Organization. **Pneumonia**. 2025. Acesso em: 26 abr. 2026. Disponível em: <https://www.who.int/health-topics/pneumonia>.

Glossário

Acurácia

Métrica que indica a proporção total de classificações corretas realizadas pelo modelo em relação ao número total de amostras avaliadas.

Aprendizado profundo

Área do aprendizado de máquina baseada em redes neurais com múltiplas camadas, capazes de aprender padrões complexos a partir dos dados.

Aprendizado por transferência

Técnica em que um modelo previamente treinado em uma grande base de dados é reaproveitado e adaptado para uma nova tarefa.

Aumento de dados

Conjunto de técnicas que gera variações artificiais das imagens de treinamento, como rotações, zoom e espelhamento, com o objetivo de melhorar a generalização do modelo.

Backbone

Parte principal de uma rede neural convolucional responsável pela extração de características das imagens.

Batch size

Quantidade de amostras processadas pelo modelo a cada atualização dos pesos durante o treinamento.

Classificação multiclasse

Problema de classificação no qual o modelo deve escolher uma entre três ou mais classes possíveis.

CNN

Sigla para *Convolutional Neural Network*, ou Rede Neural Convolucional. É uma

arquitetura de rede neural muito utilizada em tarefas de análise e classificação de imagens.

Conjunto de teste

Parte da base de dados utilizada apenas ao final do treinamento para avaliar o desempenho do modelo em imagens não vistas anteriormente.

Conjunto de treinamento

Parte da base de dados utilizada para ajustar os pesos internos da rede neural.

Conjunto de validação

Parte da base utilizada para acompanhar o desempenho do modelo durante o treinamento e auxiliar na escolha da melhor configuração.

Data augmentation

Termo em inglês para aumento de dados. Refere-se à criação de variações artificiais das imagens durante o treinamento.

Dataset

Conjunto de dados utilizado no desenvolvimento do modelo. Neste trabalho, corresponde à base de radiografias de tórax.

Desbalanceamento de classes

Situação em que uma classe possui mais amostras do que outras dentro da base de dados, podendo influenciar o aprendizado do modelo.

Dropout

Técnica de regularização que desativa aleatoriamente parte dos neurônios durante o treinamento, reduzindo a tendência ao sobreajuste.

Feature maps

Mapas de características gerados pelas camadas convolucionais de uma rede neural, representando padrões extraídos da imagem.

Fine-tuning

Etapa de ajuste fino em que parte de uma rede pré-treinada é descongelada e treinada

novamente para se adaptar melhor à nova tarefa.

F1-score

Métrica que combina precisão e revocação em um único valor, sendo útil para avaliar o equilíbrio entre falsos positivos e falsos negativos.

Grad-CAM

Técnica de interpretabilidade que gera mapas de calor sobre a imagem, indicando quais regiões mais contribuíram para a decisão da rede neural.

ImageNet

Grande base de imagens utilizada para pré-treinamento de modelos de visão computacional.

Interpretabilidade

Capacidade de compreender, ainda que parcialmente, quais informações influenciaram a decisão de um modelo de inteligência artificial.

Limiar calibrado

Valor definido a partir do conjunto de validação para ajustar uma regra de decisão complementar sobre as probabilidades previstas pelo modelo.

Matriz de confusão

Tabela que mostra os acertos e erros do modelo, indicando quantas amostras de cada classe foram classificadas corretamente ou confundidas com outras classes.

MobileNetV2

Arquitetura de rede neural convolucional leve, projetada para obter bom desempenho com menor custo computacional.

Normalização

Processo de ajuste da escala dos valores dos pixels para um formato adequado ao modelo utilizado.

Overfitting

Situação em que o modelo aprende excessivamente os padrões do conjunto de trei-

namento e perde capacidade de generalizar para novos dados.

Pipeline

Sequência organizada de etapas utilizadas no processamento dos dados, treinamento, avaliação e interpretação do modelo.

Precisão

Métrica que indica, entre as amostras classificadas como pertencentes a uma determinada classe, quantas realmente pertencem a essa classe.

Recall

Métrica que indica, entre todas as amostras reais de uma classe, quantas foram corretamente identificadas pelo modelo. Também pode ser chamado de revocação ou sensibilidade.

Rede neural convolucional

Tipo de rede neural especializada no processamento de imagens, capaz de extrair padrões visuais por meio de operações de convolução.

Revocação

Métrica que mede a capacidade do modelo de identificar corretamente as amostras reais de uma determinada classe.

Softmax

Função utilizada na camada final de modelos multiclasse para transformar as saídas da rede em probabilidades associadas a cada classe.

Sobreajuste

Tradução de *overfitting*. Ocorre quando o modelo se adapta excessivamente aos dados de treinamento e apresenta pior desempenho em novos dados.

TensorFlow/Keras

Bibliotecas utilizadas para construção, treinamento e avaliação de modelos de aprendizado profundo.

Transfer learning

Termo em inglês para aprendizado por transferência.

Validação

Etapa utilizada para acompanhar o desempenho do modelo durante o treinamento e apoiar decisões como seleção da melhor época ou calibração de limiar.