Avaliação de Desempenho de Redes Neurais Convolucionais no Diagnóstico de Amiloidose Cardíaca

Miguel Elias Silva Rodrigues[1], **Thaís Gaudencio do Rêgo**[1]

[1]Centro de Informática – Universidade Federal da Paraíba (UFPB)
João Pessoa – PB – Brazil

miguelrodrigues@eng.ci.ufpb.br, gaudenciothais@gmail.com

Abstract. *Rare diseases, affecting up to 65 people per 100,000, are largely caused by genetic mutations, often chronic and typically incurable, though manageable. Cardiac amyloidosis (CA), one of these diseases, is marked by the buildup of proteins in heart tissues, leading to stiffness and impaired heart function. Due to its severity and high mortality rate, early diagnosis is crucial. This study aims to evaluate the use of convolutional neural networks (CNNs) to identify this rare disease in a data-scarce context. Five CNN architectures were employed — ResNet-50, VGG16, Xception, MobileNet, and InceptionV3 — trained with a dataset of 138 transthoracic echocardiograms (TTEs). To interpret the network performance, Grad-CAM was applied to generate heatmaps highlighting the most relevant TTE areas for classification. Among the results, MobileNet achieved 75.14% accuracy with a ROC curve of 0.87, while ResNet-50 reached 81.43% accuracy and a ROC curve of 0.93, demonstrating the best performance. In conclusion, despite data limitations, ResNet-50 showed the highest efficacy and generalization capability among the networks evaluated.*

Resumo. *As doenças raras, que afetam até 65 pessoas a cada 100 mil, são majoritariamente causadas por alterações genéticas, sendo crônicas e, em sua maioria, incuráveis, embora tratáveis. A amiloidose cardíaca (AC), uma dessas doenças, é caracterizada pelo acúmulo de proteínas em tecidos do coração, levando à rigidez e comprometimento da função cardíaca. Dada sua gravidade e alta taxa de mortalidade, o diagnóstico precoce é essencial. Este estudo visa avaliar o uso de redes neurais convolucionais (CNNs) para identificar essa doença rara em um contexto de escassez de dados. Foram utilizadas cinco arquiteturas de CNNs — ResNet-50, VGG16, Xception, MobileNet e InceptionV3 — treinadas com um conjunto de 138 ecocardiogramas transtorácicos (ETTs). Para explicar o comportamento das redes, foi aplicado o método Grad-CAM, que gera mapas de calor destacando as áreas mais relevantes dos ETTs para a classificação. Entre os resultados, a MobileNet obteve 75,14% de acurácia e uma curva ROC de 0,87, enquanto a ResNet-50 alcançou 81,43% de acurácia e uma curva ROC de 0,93, demonstrando o melhor desempenho. Conclui-se que, mesmo com as limitações de dados, a ResNet-50 apresentou a maior eficácia e capacidade de generalização entre as redes avaliadas.*

Catalogação na publicação
Seção de Catalogação e Classificação

R696a Rodrigues, Miguel Elias Silva.
Avaliação de desempenho de redes neurais convolucionais para diagnóstico de amiloidose em ecocardiogramas / Miguel Elias Silva Rodrigues. - João Pessoa, 2024.
27 f. : il.

Orientação: Thaís Gaudencio.
TCC (Graduação) - UFPB/CI.

1. Amiloidose cardíaca. 2. Diagnóstico. 3. Redes neurais convolucionais. 4. Classificação. 5. Inteligência artificial. I. Gaudencio, Thaís. II. Título.

UFPB/CI CDU 004.8

Elaborado por Michelle de Kássia Fonseca Barbosa - CRB-738

1. Introdução

A amiloidose cardíaca (AC) é uma doença rara e sistêmica, que resulta da deposição extracelular de fibrilas proteicas, conhecidas como substâncias amiloides. No coração, devido a um erro metabólico, ela causa danos significativos ao músculo. Quando essas fibrilas se depositam no coração, o diagnóstico exige um alto grau de suspeição clínica e deve ser feito precocemente, pois a falta de tratamento pode levar a um agravamento progressivo e potencialmente letal da doença [de Cardiologia, 2021]. Um diagnóstico precoce oferece ao paciente afetado a mais ampla gama de opções de tratamento, que têm um impacto favorável na sobrevivência e/ou previnem a perda potencialmente irreversível da função física e da qualidade de vida. Vários avanços recentes significativos na abordagem diagnóstica, juntamente com a aprovação de terapias eficazes e o engajamento generalizado por sociedades, órgãos reguladores e organizações de defesa, elevaram a AC a uma posição de proeminência diagnóstica com o uso, por exemplo, das técnicas de imagem, que permitem diagnósticos precisos e não invasivos [Kittleson et al., 2023].

A identificação da AC é geralmente desafiadora, exigindo que o médico analise o espessamento da parede do coração e a insuficiência cardíaca, especialmente quando esses sintomas não podem ser atribuídos a outras doenças cardiovasculares. Para confirmar a suspeita de AC, quando não utilizadas técnicas de imagem, é comum realizar uma biópsia das células cardíacas, após descartar outras condições. O ecocardiograma transtorácico (ETT) desempenha um papel crucial ao detalhar características da doença, como a extensão e a gravidade do depósito de proteínas no coração, permitindo assim um tratamento mais específico para cada paciente. Além disso, exames como a cintilografia óssea e a ressonância magnética também podem ser incluídos na avaliação para uma compreensão mais abrangente da condição, porém, a identificação da doença permanece um desafio, mesmo no ambiente profissional [Pfizer, 2023].

Diante desse contexto, a aplicação de técnicas automatizadas na detecção de AC em ETTs possui um grande potencial para aprimorar a eficiência e a precisão do diagnóstico. O uso de algoritmos de aprendizado de máquina para classificar ETTs com o diagnóstico da doença permite uma análise mais rápida e objetiva. No entanto, é importante ressaltar que muitos dos modelos existentes ainda enfrentam limitações em termos de acurácia, precisão, revocação (*recall*), pontuação F1 e função de perda, devido à falta de dados representativos e variados no contexto de uma doença rara. Além disso, a qualidade da captura do ETT pode ser crucial em um diagnóstico automatizado por imagem, pois imagens de alta qualidade permitem que os algoritmos identifiquem com maior precisão as características relevantes para a detecção da doença, reduzindo o risco de erros causados por ruído ou baixa resolução. Para superar essas insuficiências, o uso de *Grad-CAM* pode proporcionar uma interpretação visual das decisões do modelo [Selvaraju et al., 2019], permitindo que os profissionais de saúde compreendam melhor as características das imagens que influenciam a classificação, aumentando assim a confiança no diagnóstico automatizado e auxiliando a tomada de decisão profissional.

Este estudo tem como objetivo comparar a eficácia de cinco diferentes modelos de aprendizado de máquina na classificação de ETTs entre as classes "amiloidose" e "não-amiloidose": *VGG16*, *ResNet-50*, *MobileNet*, *Xception* e *InceptionV3*. Para essa análise, as métricas de desempenho, incluindo acurácia, precisão, revocação, pontuação F1 e função de perda, serão avaliadas para cada modelo. Além disso, a interpretação das

decisões dos modelos será facilitada pelo uso do *Grad-CAM*, que fornece visualizações que destacam as regiões relevantes das imagens que influenciam a classificação. Ademais, as métricas de cada modelo serão analisadas em conjunto com a sua matriz de confusão e a área sob a curva característica de operação do receptor (AUC-ROC), para uma maior compreensão do seu comportamento e eficácia. Esta comparação não apenas permitirá identificar o modelo potencialmente mais eficaz, mas também oferecerá *insights* sobre as características das imagens que são mais significativas para o diagnóstico da amiloidose, contribuindo para a melhoria das práticas de diagnóstico automatizado.

Por fim, este trabalho contribui significativamente para o avanço do diagnóstico automatizado de AC, apresentando uma abordagem robusta para a avaliação e comparação de diferentes modelos em *deep learning* (DL). Ao focar nas métricas de desempenho e na interpretação visual fornecida pelo *Grad-CAM*, a pesquisa não busca apenas identificar o modelo que melhor se adapta às necessidades clínicas, mas também elucidar as características das imagens ecocardiográficas que são determinantes na classificação. Através dessa análise, espera-se proporcionar uma base sólida para futuras investigações na área e aprimorar a acurácia do diagnóstico.

2. Trabalhos relacionados

Nesta seção, abordaremos pesquisas que utilizam e revisam as ferramentas aqui utilizadas (CNN e DL) e o seu estado atual com o objetivo do diagnóstico da AC.

Recentemente, Ahmadi-Hadad e colaboradores (2024) revisaram sistematicamente o uso de inteligência artificial (IA) no diagnóstico de AC, destacando o potencial da DL para melhorar o reconhecimento dessa doença, que é frequentemente subdiagnosticada. A revisão analisou 10 estudos que utilizaram modelos de IA para processar dados de exames laboratoriais, registros médicos, eletrocardiogramas (ECG), ecocardiografias transtorácicas, ressonância magnética cardíaca (CMR) e cintilografia com compostos marcados (WBS) no diagnóstico de AC. Os resultados mostraram que os modelos de IA apresentaram desempenho diagnóstico comparável, e em alguns casos superior, ao de cardiologistas especializados, indicando que essas técnicas podem oferecer ferramentas valiosas para a identificação precoce da AC. Esses achados corroboram a crescente aplicabilidade de métodos baseados em IA, como as redes neurais convolucionais (CNNs), em tarefas de diagnóstico, incluindo a AC, foco central deste trabalho [Ahmadi-Hadad et al., 2024].

Um outro estudo, elaborado por Vrudhula e colaboradores (2024), avaliou o impacto da seleção de casos e controles no desempenho de modelos de IA baseados em ECG para a triagem de AC. Utilizando uma coorte de aproximadamente 1,3 milhão de ECGs, os pesquisadores testaram diferentes definições de casos, incluindo pacientes diagnosticados com amiloidose, pacientes sem AC e pacientes atendidos em clínicas especializadas. Os resultados indicaram que a generalizabilidade dos modelos variou significativamente ao serem aplicados em uma população geral, com uma AUC-ROC variando de 0,467 a 0,898, dependendo da definição de casos utilizada. Embora modelos treinados com dados mais curados tenham mostrado melhor desempenho em coortes de teste similares, os resultados sugerem que é possível treinar modelos eficazes, mesmo em instituições sem clínicas especializadas em AC. Esses achados são relevantes para este trabalho, pois destacam a importância da seleção de dados no desempenho de CNNs aplicadas ao diagnóstico de

AC [Vrudhula et al., 2024].

No ano de 2020, Martini e colaboradores apresentaram um estudo no qual utilizaram DL para análise automática de imagens de ressonância magnética cardiovascular em pacientes com suspeita de amiloidose. A abordagem DL atingiu uma AUC-ROC de 0.982, indicando alta precisão diagnóstica, comparável a algoritmos de *machine learning* (ML) que simulam a interpretação de um operador experiente. Este estudo demonstra a eficácia de técnicas de IA no diagnóstico de amiloidose, a partir de dados de imagem, destacando a capacidade de modelos de DL em capturar padrões relevantes de forma autônoma [Martini et al., 2020].

Um estudo recente de Delbarre e colaboradores em 2023 desenvolveu e validou um modelo de DL baseado em CNNs, para detectar captação cardíaca anormal em cintilografias ósseas, associada ao risco de AC transtirretina. Utilizando uma base de dados hospitalar extensa com mais de 3.000 imagens, o modelo atingiu uma sensibilidade (*recall*) de 98,9%. A abordagem proposta, assim como o presente trabalho, reforça o potencial das CNNs para diagnósticos precisos em cenários médicos, fornecendo uma ferramenta eficaz no diagnóstico precoce de AC [Delbarre et al., 2023].

A aplicação de CNNs no diagnóstico de AC também foi explorada por Agibetov e colaboradores em 2021, que desenvolveram um algoritmo baseado em IA para detectar padrões associados à doença em imagens de CMR. O estudo, que envolveu dados de 502 pacientes, incluindo 82 diagnosticados com AC, revelou que as CNNs poderiam identificar com precisão os padrões de imagem associados à condição, alcançando um *recall* de 94%. Os resultados, obtidos através de uma avaliação com validação cruzada de 10 grupos, destacam o potencial da IA para estabelecer um caminho diagnóstico totalmente automatizado, o que é particularmente valioso em centros com baixa frequência de casos de doenças de armazenamento miocárdico. Assim, a pesquisa não só reforça a importância das CNNs na área de diagnóstico médico, como também se alinha com os objetivos do presente trabalho, que visa avaliar o desempenho de CNNs no diagnóstico de AC, contribuindo para a redução de diagnósticos incorretos e melhorando a precisão clínica [Agibetov et al., 2021].

O estudo de Xiang Yu e colaboradores (2021) foca no desenvolvimento de uma rede diagnóstica semi-automática baseada em DL para detectar hipertrofia ventricular esquerda (HVE) em ETTs. Com 1610 ETTs coletados retrospectivamente, abrangendo pacientes com diferentes condições (doença cardíaca hipertensiva, cardiomiopatia hipertrófica e AC), os modelos de DL ResNet-18 para classificação visual, *ResNet-50* para detecção de HVE e *U-net++* para segmentação, foram usados durante o estudo. A rede atingiu uma AUC-ROC de 0,98 na detecção de HVE, com sensibilidade de 94% e especificidade de 91,6%. Além disso, o modelo final conseguiu distinguir as quatro condições com uma AUC média de 0,91, demonstrando a capacidade de identificar HVE e suas etiologias latentes com alta precisão [Yu et al., 2022].

Desta forma, este trabalho se assemelha aos demais estudos apresentados na tabela no objetivo comum de avaliar técnicas de IA, como o DL, no diagnóstico da AC. Assim como os outros, utiliza arquiteturas de CNNs, como *ResNet-50*, *InceptionV3*, *VGG16*, *Xception* e *MobileNet*, mas se diferencia pela aplicação de métodos de observabilidade, como *Grad-CAM*, que não são utilizados nas outras pesquisas. As métricas de avaliação,

incluindo acurácia, precisão, *recall* e AUC-ROC, são semelhantes às empregadas nos estudos revisados. Além disso, este trabalho se baseia em um conjunto de 138 ETTs, alinhando-se à tendência de utilizar dados clínicos reais para treinar e validar modelos.

A análise dos trabalhos supracitados destaca a crescente aplicação de CNNs e técnicas de aprendizagem profunda no diagnóstico de AC, com resultados promissores em termos de sensibilidade e especificidade, demonstrando avanços significativos na detecção precoce e precisa dessa condição e reforçando o potencial dessas ferramentas no auxílio ao diagnóstico médico, especialmente em instituições com infraestrutura limitada. Ao analisar as pesquisas em questão, identificam-se modelos, bases de dados e métricas utilizados tanto nos estudos revisados, quanto neste trabalho, como detalhado na Tabela 1. Esses achados ressaltam a relevância das metodologias utilizadas no presente trabalho, que objetiva avaliar o desempenho de diferentes arquiteturas de CNN no diagnóstico de AC.

Autor	**Objetivo**	**Algoritmo**	**Dataset**	**Métricas**
Martini et al., 2020	Avaliar o desempenho de técnicas de *machine learning* e DL baseadas em ressonância magnética no diagnóstico de AC.	*Global CNN* (Autoria própria) e *Gradient Boosting Machine* (GBM). *Grad-CAM* para observabilidade	CMR de 206 pacientes do *Fondazione Toscana Gabriele Monasterio* (Pisa, Itália)	Acurácia, *recall*, especificidade, VPP, VPN, AUC-ROC
Agibetov, et al., 2021	Aprimorar o diagnóstico de AC em imagens de ressonância magnética utilizando DL em CNNs.	VGG16	CMR de 502 pacientes, gerando um total de 356.675 imagens	AUC-ROC, *recall* e especificidade
Delbarre, et al., 2023	Desenvolver e validar um modelo baseado em deep learning que detecta automaticamente captação cardíaca significativa em WBS, a fim de identificar pacientes em risco de AC.	CNN de autoria própria	Coorte do centro de referência francês para a AC no Hospital Universitário Henri Mondor (HM-UH) e coorte independente do Hospital Universitário de Lille (L-UH)	Acurácia, precisão, *recall*, *F1-score*, especificidade, AUC-ROC

Autor	Objetivo	Algoritmo	Dataset	Métricas
Vrudhula, et al., 2024	Avaliar o desempenho de modelos de IA baseados em formas de onda de ECG para a triagem de AC.	EfficientNet (Modificado)	ECGs de pacientes que receberam atendimento no Cedars-Sinai Medical Center entre 2005 e 2022	AUC-ROC
Yu et al., 2022	Desenvolver uma rede de diagnóstico semi-automática baseada em algoritmos de DL para detectar a HVE em ETTs.	ResNet-18, *ResNet-50* e U-net++	Banco de dados de ecocardiografia do Primeiro Hospital Afiliado da Universidade de Zhejiang, abrangendo relatórios entre janeiro de 2018 e dezembro de 2020	*Recall*, especificidade e AUC-ROC
Ahmadi-Hadad, et al., 2024	Revisar sistematicamente a aplicação de modelos de IA no diagnóstico da AC.	VGG16, SVM com kernel RBF. *Global CNN*, GBM	Datasets variados, para cada artigo revisado	PPV, AUC-ROC e *F1-score*
Este trabalho	Avaliar o desempenho de técnicas de DL baseadas em ETTs no diagnóstico de AC.	ResNet-50, InceptionV3, VGG16, Xception e MobileNet *Grad-CAM* para observabilidade	138 ETTs fornecidos pelo Prof. Dr. Marcelo Dantas Tavares de Melo da Universidade Federal da Paraíba	Acurácia, precisão, *recall*, *F1-score*, AUC-ROC

Tabela 1. Resumo dos trabalhos relacionados

3. Metodologia

Nesta seção, é apresentada a metodologia empregada neste estudo, que abrange a descrição do conjunto de dados utilizado, as técnicas de pré-processamento aplicadas, o ambiente de desenvolvimento adotado, as métricas de avaliação definidas e as arquiteturas de CNNs implementadas para a classificação das imagens de ETTs.

3.1. Conjunto de dados

O conjunto de ETTs utilizado neste estudo foi fornecido pelo Professor Doutor Marcelo Dantas Tavares de Melo, do Departamento de Medicina Interna da Universidade Federal da Paraíba (UFPB), contendo 138 ETTs em vídeo, únicos para cada paciente, que posteriormente tiveram seus *frames* extraídos.

Dentre os ETTs fornecidos, foi realizada uma segregação aleatória utilizando um *script Python* em sua versão 3.10, na qual 71% dos ETTs foram dedicados ao treinamento dos modelos, totalizando 98 ETTs utilizados, enquanto 26 (19%) foram destinados à validação do modelo e 14 (10%) para o teste do modelo. Além disso, cada subconjunto de dados foi igualmente dividido, com 50% dos casos representando amiloidose e 50% representando casos em que a doença não foi diagnosticada.

Após essa divisão inicial, cada ETT foi ainda subdividido em 25 *frames* por vídeo, utilizando a biblioteca *OpenCV* na linguagem *Python*, totalizando 3.450 imagens de ETTs, das quais 2.450 foram dedicadas ao treinamento dos modelos, 650 à validação e 350 ao processo de testes.

3.2. Pré-processamento dos dados

Após a fragmentação dos vídeos dos ETTs em *frames*, o pré-processamento das imagens utilizadas neste estudo foi realizado através de uma função que utilizou a biblioteca *Keras* em sua versão 3.4.1, integrada ao *Tensorflow* em sua versão 2.17.0, para gerar lotes de imagens a partir dos conjuntos de treino, validação e teste, com o objetivo de otimizar o processo de treinamento de um modelo de classificação. Para isto, foi utilizado um *batch size* de 32 amostras a serem processadas em cada iteração, além de um redimensionamento das imagens para o tamanho padrão de 224 *pixels* de altura por 224 *pixels* de largura.

Inicialmente, as imagens foram normalizadas em todos os conjuntos (treinamento, validação e teste) com a técnica de reescala da biblioteca *Keras*, que consiste em dividir os valores dos pixels por 255, ajustando-os para uma faixa entre 0 e 1. Essa normalização é essencial para garantir que os valores de entrada sejam uniformes, o que facilita o aprendizado dos modelos de ML.

Para o conjunto de treinamento, aplicou-se uma técnica de aumento de dados (*data augmentation*), que visa aumentar a variabilidade das imagens e evitar o sobreajustamento do modelo utilizando a função *ImageDataGenerator* da biblioteca *Keras*. Os parâmetros selecionados para esta função incluiram uma série de transformações aleatórias: rotação das imagens (*rotation_range*) em até 20 graus, deslocamentos horizontais e verticais (*width_shift_range* e *height_shift_range*) de até 20% da largura ou altura, transformações de cisalhamento (*shear_range*), *zoom* de até 20%, além de espelhamento horizontal (*horizontal_flip*) aleatório. Essas modificações aleatórias nas imagens ajudam o modelo

a generalizar melhor, tornando-o mais robusto contra variações de posição, tamanho e orientação das imagens de entrada. Após as transformações, os espaços em branco gerados por elas foram preenchidos com os valores de pixels mais próximos, utilizando o método de preenchimento mais próximo.

No entanto, para os conjuntos de validação e teste, não foi aplicado aumento de dados, uma vez que esses conjuntos são utilizados para avaliar o desempenho real do modelo e precisam ser mantidos consistentes. O pré-processamento desses conjuntos limitou-se à normalização dos pixels.

Por fim, as imagens de cada conjunto foram organizadas em lotes, sendo que o diretório de treino foi configurado para usar os lotes no modo "categorical", que é apropriado para problemas de classificação multiclasse. As imagens de validação e teste foram geradas de maneira ordenada (sem embaralhamento) para garantir consistência durante as avaliações do modelo.

3.3. Arquiteturas de CNNs

As CNNs revolucionaram a área de visão computacional, permitindo avanços significativos em tarefas de classificação de imagens. Essas redes foram amplamente utilizadas em problemas complexos devido à sua capacidade de extrair características hierárquicas das imagens, tornando-as ideais para detecção e diagnóstico em imagens médicas [LeCun et al., 2015]. Dentre as diversas arquiteturas desenvolvidas, destacam-se modelos que balanceiam profundidade, eficiência computacional e desempenho, como a *ResNet-50, VGG16, InceptionV3, Xception* e *MobileNet*. Cada um desses modelos apresenta características próprias que os tornam adequados para diferentes cenários, justificando sua escolha em tarefas de classificação complexas, como no diagnóstico de amiloidose em ETTs.

A *ResNet-50* é uma arquitetura baseada em conexões residuais, introduzida para mitigar o problema de gradiente desaparecendo em redes muito profundas. Seu diferencial está na capacidade de aprender representações robustas sem a degradação do desempenho conforme a rede se torna mais profunda. Esse modelo mostrou excelente desempenho na classificação de imagens no ImageNet, tornando-se um padrão para tarefas complexas na área de visão computacional [He et al., 2015]. A escolha da *ResNet-50* para o diagnóstico de amiloidose se justifica por sua habilidade de extrair características profundas das imagens de ETTs, o que é essencial para distinguir padrões sutis relacionados à doença.

A arquitetura *VGG16* se destaca pela simplicidade de sua arquitetura, que consiste em blocos de convoluções pequenas (3x3) empilhadas, seguidas de camadas totalmente conectadas. Apesar de ser mais profundo que muitas redes anteriores, a *VGG16* manteve uma estrutura homogênea que facilitou sua implementação e análise. Seu impacto foi notável, sendo amplamente utilizada como base para outros modelos de classificação de imagens [Simonyan and Zisserman, 2015]. A *VGG16* foi escolhida devido à sua eficácia comprovada em tarefas de classificação de imagens médicas, sendo adequada para a detecção de anomalias em ETTs que indicam amiloidose.

A arquitetura *InceptionV3* introduziu o conceito de blocos de *Inception*, nos quais convoluções de diferentes tamanhos são aplicadas em paralelo para capturar múltiplas escalas de informação na mesma camada. Essa inovação permitiu à rede ser mais eficiente

em termos de parâmetros, mantendo alto desempenho em desafios de visão computacional, como a competição *ImageNet* [Szegedy et al., 2015]. Sua habilidade em capturar detalhes em diferentes escalas é ideal para analisar a complexidade estrutural dos ETTs, tornando-a uma escolha promissora para identificar os sinais de amiloidose.

O *Xception*, uma evolução do *Inception*, utiliza convoluções separáveis em profundidade para melhorar a eficiência computacional e o desempenho. A ideia central é que convoluções em profundidade podem separar o aprendizado espacial do aprendizado de características, o que resultou em uma arquitetura que excedeu as capacidades da *Inception* em várias tarefas [Chollet, 2017]. A escolha do *Xception* se deve ao seu equilíbrio entre precisão e eficiência, sendo altamente aplicável para identificar variações sutis nos ETTs que indicam o desenvolvimento de amiloidose.

Por fim, o *MobileNet* foi desenvolvido com o objetivo de trazer eficiência para dispositivos móveis e sistemas com restrições de *hardware*. Ele também se baseia em convoluções separáveis em profundidade, oferecendo um equilíbrio entre precisão e custo computacional. Sua arquitetura leve permite que seja usado em aplicações de visão em tempo real, mesmo em dispositivos com baixa capacidade de processamento [Howard et al., 2017]. A escolha do *MobileNet* se justifica pela necessidade de diagnósticos rápidos e precisos, o que pode facilitar a análise de ETTs em ambientes com recursos computacionais limitados, como hospitais com infraestrutura básica.

Esses modelos oferecem um amplo espectro de opções para a tarefa de classificação de imagens, desde arquiteturas mais profundas e precisas, até aquelas otimizadas para eficiência computacional, sendo adequados para o cenário e escopo de teste deste trabalho.

3.4. Criação dos modelos

Todos os modelos foram desenvolvidos utilizando uma abordagem padronizada, com base em arquiteturas pré-treinadas em conjuntos de dados amplos, como o *ImageNet*. Essa estratégia assegura que cada modelo inicie o treinamento com um conhecimento prévio sólido, resultando em melhor desempenho nas tarefas de classificação. A mesma configuração de parâmetros foi aplicada a todas as arquiteturas, garantindo consistência nos resultados.

Durante o processo de treinamento, cada modelo foi enriquecido com camadas adicionais para maximizar sua capacidade de classificação. A estrutura incluiu uma camada de *pooling* global que facilita a agregação das características extraídas, seguida por uma camada densa com 256 neurônios, que utiliza a função de ativação *RELU* para introduzir não linearidade. A camada de saída foi configurada para classificar as imagens em duas categorias, utilizando a função *softmax*.

A otimização foi realizada com o otimizador *Adam*, com uma taxa de aprendizado definida em 0,0001, e a função de perda utilizada foi a entropia cruzada categórica. O treinamento foi conduzido por 50 épocas, permitindo que os modelos fossem ajustados para melhor desempenho em um conjunto de dados de validação.

As tecnologias empregadas neste processo incluíram o uso do *TensorFlow* e *Keras*, que proporcionaram uma interface intuitiva e poderosa para a construção e treinamento destas redes neurais. Além disso, bibliotecas como *NumPy* e *Matplotlib* foram

utilizadas para manipulação de dados e visualização de resultados, respectivamente. O mecanismo de monitoramento foi implementado com *callbacks* que permitiram salvar automaticamente o modelo com melhor desempenho, de acordo com a precisão de validação, garantindo que a versão mais eficiente fosse preservada.

Após o treinamento, cada modelo foi avaliado em um conjunto de teste separado. A acurácia obtida variou conforme a arquitetura utilizada, e relatórios detalhados foram gerados, incluindo a matriz de confusão e o relatório de classificação.

Os resultados foram complementados por gráficos que mostraram a evolução da acurácia e a redução da perda ao longo das épocas de treinamento, permitindo uma visualização clara do desempenho de cada modelo. Essa abordagem metódica assegurou que cada modelo fosse otimizado para as tarefas de classificação de imagens, proporcionando uma base sólida para comparações de desempenho entre diferentes arquiteturas.

3.5. Métricas de avaliação

Para avaliar o desempenho das CNNs no diagnóstico de AC, foram empregadas diversas métricas e métodos que possibilitaram uma análise detalhada da eficácia dos modelos. Primeiramente, acompanhou-se a evolução da acurácia tanto no conjunto de treino, quanto no de validação. A acurácia reflete a proporção de predições corretas em relação ao total de predições realizadas e foi monitorada ao longo das épocas para identificar possíveis tendências de *overfitting* ou *underfitting*. Juntamente com a acurácia, o gráfico de perda (*loss*) foi gerado, permitindo a análise de como o erro do modelo variou no treino e na validação. A função de perda mede o quão distante as predições do modelo estão dos valores reais, sendo essencial para entender como o ajuste do modelo evolui ao longo do tempo.

Além dos gráficos de acurácia e *loss*, foi gerada a matriz de confusão para o conjunto de testes. A matriz de confusão oferece uma visão clara da performance do modelo ao organizar as predições em quatro categorias: verdadeiros positivos (TP), falsos positivos (FP), verdadeiros negativos (TN) e falsos negativos (FN). A partir dessa matriz, foi possível calcular métricas importantes como a precisão (*precision*), o *recall* (sensibilidade), a pontuação F1 e a AUC-ROC.

A precisão (ou precisão positiva) mede a proporção de verdadeiros positivos em relação ao total de predições positivas realizadas pelo modelo, sendo definida pela seguinte fórmula:

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

O *recall* (ou sensibilidade) calcula a proporção de verdadeiros positivos em relação ao total de casos que realmente pertencem à classe positiva, ou seja, mede a capacidade do modelo de identificar corretamente os casos de amiloidose. A fórmula do *recall* é dada por:

$$\textit{Recall} = \frac{TP}{TP + FN}$$

Uma métrica derivada da precisão e do *recall* é a pontuação F1, que é a média

harmônica entre essas duas métricas. A pontuação F1 é útil quando há a necessidade de equilibrar a preocupação entre falsos positivos e falsos negativos. Sua fórmula é:

$$F1 = 2 \times \frac{\text{Precisão} \times \textit{Recall}}{\text{Precisão} + \textit{Recall}}$$

Adicionalmente, foi calculada a acurácia média no conjunto de testes, que é uma métrica geral que mede a proporção de todas as predições corretas (tanto para a classe "amiloidose" quanto para a classe "não-amiloidose"). A fórmula para a acurácia é:

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

Outra métrica importante utilizada foi a AUC-ROC. A AUC-ROC mede a habilidade do modelo de distinguir entre as classes positivas e negativas. Quanto maior a área sob a curva ROC, melhor a capacidade do modelo de discriminar entre as duas classes.

Essas métricas foram geradas juntamente com a análise da matriz de confusão, para assim fornecer uma avaliação abrangente do desempenho dos modelos, permitindo uma análise crítica e detalhada da capacidade das CNNs em identificar a AC a partir dos ETTs.

3.6. Geração do mapa de ativação utilizando *Grad-CAM*

Com o intuito de interpretar de maneira visual as decisões tomadas pelos modelos ao classificar ETTs entre 'amiloidose' e 'não-amiloidose', utilizou-se a técnica *Grad-CAM*. O *Grad-CAM* permite a visualização das áreas da imagem que mais influenciaram a predição do modelo, destacando as regiões de maior relevância para a classificação.

A utilização da técnica *Grad-CAM* no contexto deste trabalho visa aumentar a transparência das CNNs empregadas na classificação dos ETTs. O *Grad-CAM* possibilita a geração de explicações visuais ao destacar as regiões mais importantes da imagem que contribuíram para a predição do modelo. A escolha pelo *Grad-CAM* se justifica por sua aplicabilidade em diversos modelos baseados em CNNs, sem a necessidade de modificar a arquitetura ou realizar re-treinamento. A técnica também se destaca por oferecer uma visualização fiel ao modelo original, permitindo que se entenda o comportamento da rede neural em cada predição, além de ser robusta em cenários de imagens adversariais [Selvaraju et al., 2020]. Além disso, o *Grad-CAM* auxilia na interpretação de falhas do modelo e na identificação de possíveis vieses nos dados, o que é crucial em aplicações de diagnóstico médico, como o presente estudo.

O processo iniciou com o preprocessamento das imagens de entrada, redimensionadas para um tamanho fixo de 224x224 pixels, compatível com os modelos utilizados. Após o redimensionamento, as imagens foram transformadas em *arrays* numéricos, prontos para serem processados pelas CNNs.

Em seguida, para gerar o mapa de ativação utilizando o *Grad-CAM*, foi necessário identificar a última camada convolucional de cada modelo. Esta camada é essencial, pois suas ativações estão diretamente relacionadas aos padrões identificados nas imagens. A partir das ativações desta camada, o gradiente da predição da classe de maior probabilidade foi calculado em relação a cada um dos filtros convolucionais. Esse gradiente

permitiu quantificar a importância de cada filtro para a decisão final do modelo. Dessa forma, foi possível gerar um mapa de calor que visualiza quais partes da imagem tiveram maior influência na predição.

O mapa de calor gerado foi então sobreposto à imagem original, utilizando uma coloração apropriada que destacasse as áreas mais importantes para o modelo. Para tal, utilizou-se uma coloração do tipo "*jet*", que facilita a visualização ao aplicar tons mais quentes (vermelhos e amarelos) nas regiões de maior relevância e tons mais frios (azul) nas regiões de menor importância. Essa sobreposição permitiu uma análise visual clara das áreas de atenção do modelo ao fazer a classificação. Por fim, essa abordagem foi aplicada a todos os modelos analisados neste trabalho, gerando mapas de ativação para um exemplo de imagem de ETT testada, a fim de analisar o comportamento analítico de cada modelo.

4. Resultados e discussões

Nesta seção, são apresentados os resultados e as discussões dos experimentos em cada arquitetura de rede neural, inicialmente de maneira generalizada, e em seguida, cada arquitetura sendo abordada particularmente em sua subseção.

4.1. Métricas de avaliação

A Tabela 2 apresenta as métricas de desempenho de diferentes arquiteturas de CNNs aplicadas à classificação de ETTs para o diagnóstico de amiloidose. A acurácia, a precisão, o *recall* e a pontuação F1 foram avaliadas para cada arquitetura.

Arquitetura	**Acurácia (%)**	**Precisão média**	***Recall* médio**	**Pontuação F1 média**
ResNet-50	81,43	0,82	0,81	0,81
InceptionV3	78,86	0,79	0,79	0,79
VGG16	77,43	0,82	0,77	0,77
Xception	77,14	0,78	0,77	0,77
MobileNet	75,14	0,75	0,75	0,75

Tabela 2. Métricas de avaliação de classificação

A *ResNet-50* obteve o melhor desempenho geral, com uma acurácia de 81,43%, precisão média de 0,82, *recall* médio de 0,81 e pontuação F1 média de 0,81. Essas métricas indicam que o modelo que utilizou essa arquitetura equilibrou bem a precisão e a capacidade de recuperação de exemplos positivos, sendo a mais eficaz na detecção de amiloidose, com um bom compromisso entre evitar falsos positivos e falsos negativos.

O *InceptionV3* também apresentou resultados competitivos, com uma acurácia de 78,86%, e métricas de precisão, *recall* e F1 em 0,79. Embora inferior à ResNet-50, essa arquitetura ainda oferece uma boa capacidade de classificação. ' O modelo da arquitetura *VGG16* teve um desempenho notável em termos de precisão (0,82), mas com um *recall* (0,77) ligeiramente mais baixo, indicando que essa arquitetura tem maior tendência a identificar corretamente exemplos positivos, porém pode deixar de identificar alguns casos de amiloidose (falsos negativos).

O modelo da arquitetura *Xception* obteve uma acurácia similar à VGG16 (77,14%), com resultados inferiores nas demais métricas (0,78 de precisão, 0,77 de *recall* e F1). Isso sugere um desempenho geral equilibrado, mas sem se destacar em nenhuma métrica específica.

Por fim, a *MobileNet* apresentou o menor desempenho, com 75,14% de acurácia e uma performance uniforme nas demais métricas (0,75), o que indica uma menor capacidade de classificação em relação às outras arquiteturas.

Esses resultados demonstram que a *ResNet-50*, dentre as arquiteturas base testadas, é a arquitetura mais adequada para a tarefa de diagnóstico de amiloidose, enquanto as demais redes, embora eficazes, podem apresentar limitações em termos de acurácia ou equilíbrio entre precisão e *recall* para o conjunto de dados limitado.

4.2. Acurácia, Função de Perda e Matriz de Confusão

Nesta subseção, serão avaliados os gráficos de evolução da acurácia ao longo das épocas em conjunto com a evolução da função de perda (*loss*) para cada arquitetura.

4.2.1. *ResNet-50*

A evolução da acurácia do *ResNet-50* na Figura 1 mostra um comportamento no qual a acurácia no conjunto de treinamento rapidamente atinge valores próximos de 1,0, evidenciando que o modelo está aprendendo a classificar as imagens de ETTs com precisão, forte indício de um *overfitting* devido à quantidade limitada de dados para a complexidade arquitetural do modelo. Ainda assim, embora a acurácia no conjunto de validação apresente variações, há uma tendência crescente ao longo das épocas, o que sugere que o modelo está progressivamente melhorando sua capacidade de generalização.

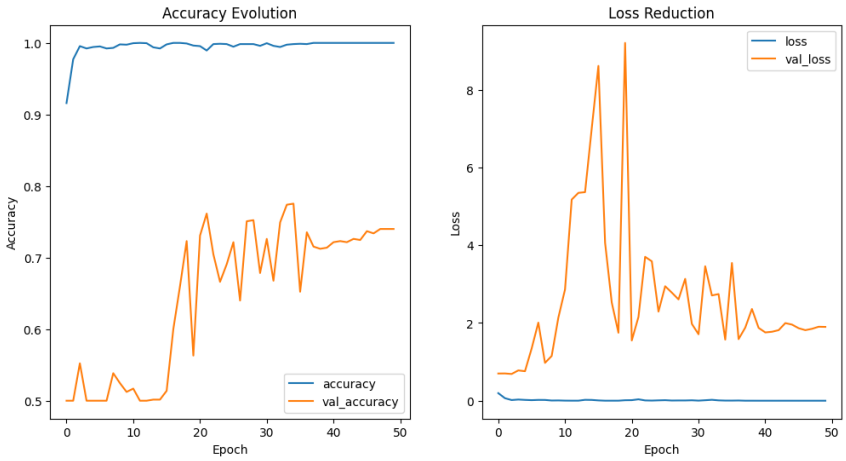

Figura 1. Evolução de acurácia à esquerda e *loss* à direita para o modelo *ResNet-50*

Por outro lado, a análise do *loss*, ou *loss reduction*, revela um cenário de estabilidade no treinamento e de oscilação na validação. O *loss* no conjunto de treinamento atinge valores baixos rapidamente e se mantém estável, o que é um bom indicador de que o modelo está conseguindo ajustar suas previsões de forma eficaz. No entanto, a perda

no conjunto de validação apresenta flutuações significativas, especialmente nas primeiras épocas, sugerindo que o modelo está enfrentando dificuldades para se ajustar completamente aos dados de validação. Apesar disso, há uma estabilização gradual do *loss* nas últimas épocas, indicando que o modelo está aprendendo a lidar com as diferenças entre os conjuntos de treinamento e validação.

Analisando ainda a matriz de confusão evidenciada na Figura 2, é perceptível, em conjunto com as matrizes de confusão dos outros modelos, que a arquitetura *ResNet-50* foi a que menos obteve falsos negativos de diagnóstico de amiloidose, o que apresenta um menor risco de uso da rede. Isso ocorre pelo fato de uma pessoa com amiloidose não ser diagnosticada apresentar um perigo maior do que uma pessoa sem amiloidose, sendo diagnosticada erroneamente, pois a pessoa que possui amiloidose necessita de tratamento precoce, e um falso diagnóstico negativo pode atrasar esse processo.

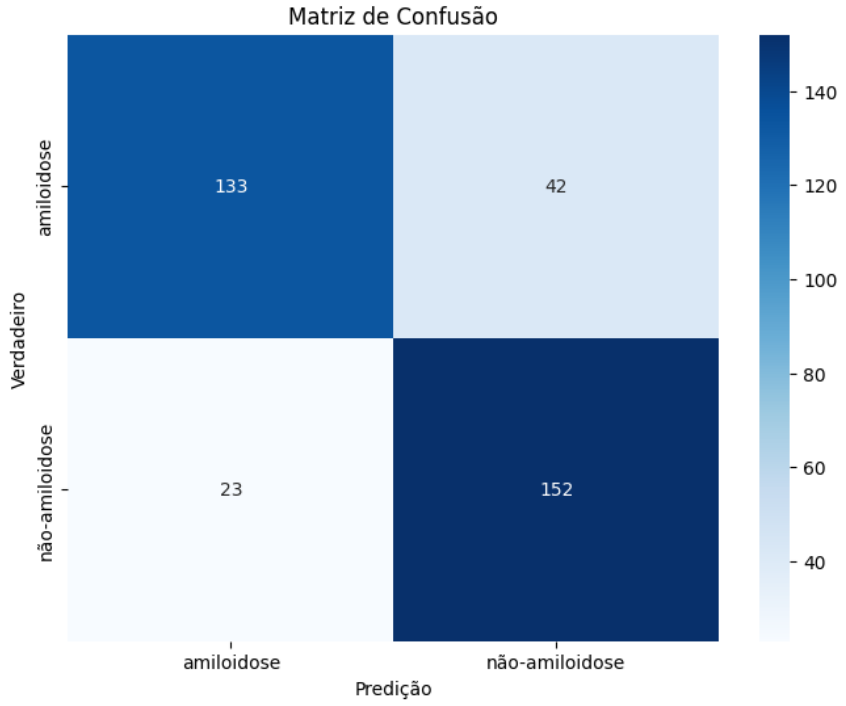

Figura 2. Matriz de confusão do modelo *ResNet-50*

Por fim, essa combinação de resultados indica um progresso contínuo do modelo na tarefa de classificação de ETTs em ”amiloidose” e ”não-amiloidose”. Dessa forma, embora as oscilações no *loss* e na acurácia de validação possam parecer preocupantes, à primeira vista, o fato de que ambas as métricas apresentam uma melhoria consistente sugere que o modelo está se ajustando de maneira eficaz às nuances dos dados, além de possíveis melhoras na capacidade do modelo de diagnosticar corretamente a doença.

4.2.2. *InceptionV3*

No modelo *InceptionV3*, a evolução da acurácia mostrada na Figura 3 apresenta uma clara distinção entre o comportamento no conjunto de treinamento e no conjunto de validação. A acurácia do treinamento atinge rapidamente valores próximos de 1,0, mostrando que

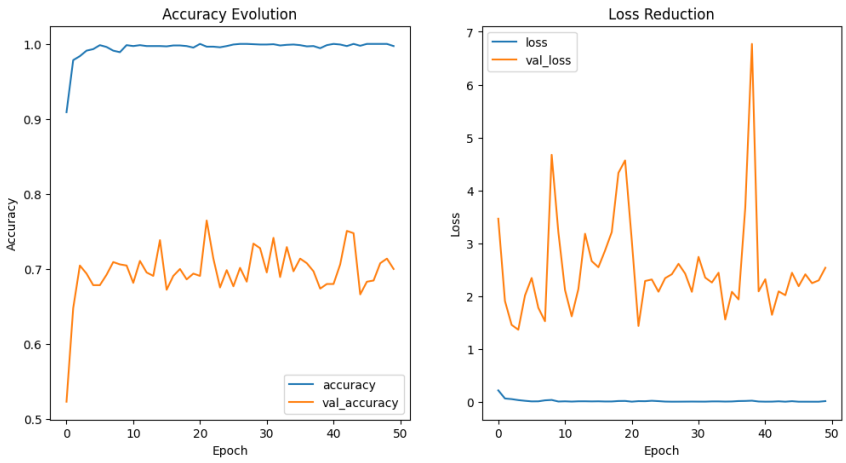

Figura 3. Evolução de acurácia à esquerda e função de perda à direita para o modelo *InceptionV3*

o modelo está aprendendo os padrões dos dados de maneira eficaz e consistente. No entanto, a acurácia da validação, embora não tão estável quanto no treinamento, cresce de forma significativa nas primeiras épocas e se estabiliza em torno de 0,7. Isso indica que o modelo está conseguindo classificar bem as amostras de validação, apesar de uma ligeira variabilidade e uma evidente limitação diante da variabilidade amostral.

Já no gráfico de redução da função de perda, o comportamento entre o *loss* de treinamento e o de validação é divergente. O *loss* no treinamento se mantém baixo e estável ao longo de todo o processo, sugerindo que o modelo está se ajustando bem aos dados, generalizando de forma eficaz. Por outro lado, o *loss* de validação é mais irregular, com picos e quedas consideráveis. Isso pode ser atribuído à natureza dos dados de validação, que podem conter amostras mais complexas que necessitem de um conjunto de treinamento mais amplo, causando variações no desempenho do modelo. Apesar das oscilações, a função de perda na validação segue uma tendência geral de redução com o passar das épocas.

Ao analisar a matriz de confusão apresentada na Figura 4, fica evidente que o modelo *InceptionV3* apresenta uma capacidade semelhante ao *ResNet-50* de classificação de ETTs, pois enquanto o *ResNet-50* obteve 42 falsos negativos, de 175 casos de amiloidose no conjunto de testes, o *Inception* obteve 49 falsos negativos.

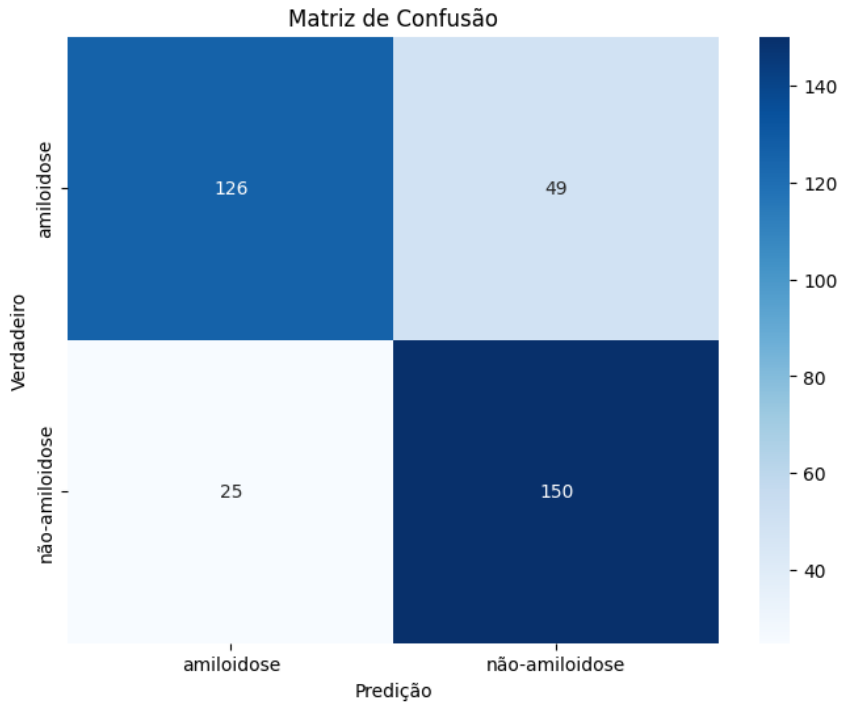

Figura 4. Matriz de confusão do modelo *InceptionV3*

De maneira geral, esta arquitetura mostra um progresso significativo no aprendizado ao longo das épocas, embora tenha um desempenho inferior comparado ao modelo *ResNet-50*. As oscilações no conjunto de validação, tanto na acurácia, quanto na função de perda, indicam que há variação nas dificuldades apresentadas pelas amostras de validação, mas o modelo, mesmo assim, é capaz de manter uma boa performance geral. A alta acurácia e o baixo e estável *loss* no treinamento, junto com a tendência de melhora nas métricas de validação, sugerem que, com ajustes adicionais, como regularização e mais dados de treino, o modelo pode alcançar um desempenho ainda melhor e mais estável na classificação de ETTs.

4.2.3. *VGG16*

No modelo *VGG16*, a acurácia no conjunto de treinamento atinge rapidamente valores próximos a 1,0, como é mostrada na Figura 5 o que pode parecer um bom sinal, mas na verdade pode indicar que o modelo está superajustando os dados de treinamento. Isso porque a acurácia no conjunto de validação apresenta oscilações significativas, estabilizando-se em torno de 0,75 com grande variabilidade ao longo das épocas. Isso sugere que, apesar de o modelo estar aprendendo bem os padrões do conjunto de treinamento, ele não consegue generalizar de maneira consistente para novos dados, resultando em um desempenho subótimo na validação.

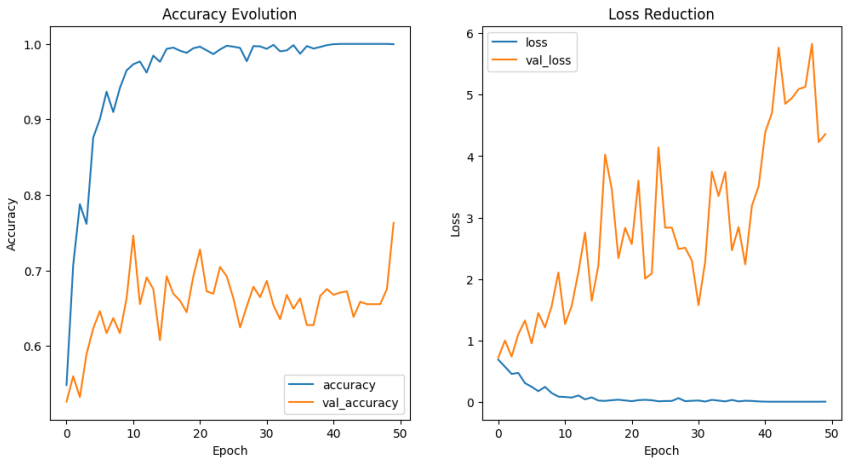

Figura 5. Evolução de acurácia à esquerda e função de perda à direita para o modelo *VGG16*

O comportamento do *loss* reforça essa análise. No conjunto de treinamento, o *loss* cai drasticamente nas primeiras épocas e permanece em valores baixos, o que confirma que o modelo está minimizando o erro para os dados de treinamento. No entanto, o *loss* de validação, é extremamente variável, apresentando picos ao longo de todo o treinamento, o que pode indicar que o modelo está aprendendo ruídos específicos dos dados de validação em vez de padrões generalizáveis. Essa inconsistência no *loss* sugere um problema de generalização e a possibilidade de *overfitting*, levando em consideração também que o *loss* apresenta uma tendência crescente, o que indica que o modelo está cada vez mais distante da resposta correta para a classificação do ETT, como será visto na Figura 6.

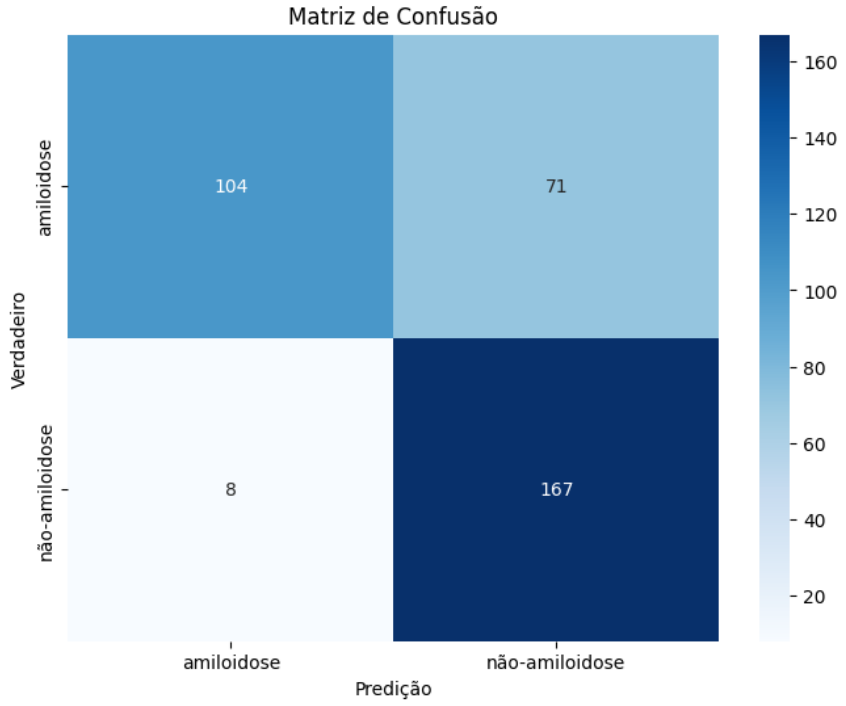

Figura 6. Matriz de confusão do modelo *VGG16*

Ao analisar a matriz de confusão para esse modelo, levando em consideração o conjunto de testes, o *VGG16* foi a arquitetura com a maior taxa de falsos negativos do espaço amostral de arquiteturas, com 71 falsos negativos, como evidencia a Figura 6, o que indica que esse modelo foi o que mais encontrou dificuldades para diagnosticar corretamente a AC.

Em resumo, os gráficos indicam que o *VGG16* está claramente superajustado aos dados de treinamento limitado, com acurácia quase perfeita e *loss* muito baixo nesse conjunto. No entanto, o desempenho no conjunto de validação é instável, o que aponta para uma dificuldade significativa do modelo em generalizar.

4.2.4. Xception

Os gráficos da Figura 7 mostram a evolução da acurácia e do *loss* durante o treinamento e validação do modelo *Xception* de classificação. No primeiro gráfico, a acurácia do conjunto de treino rapidamente atinge um valor próximo de 1,0 após algumas épocas e se mantém estável. Em contraste, a acurácia de validação é muito mais variável, com valores entre 0,6 e 0,8, e nunca se estabiliza, o que indica que o modelo está tendo dificuldades em generalizar para dados novos.

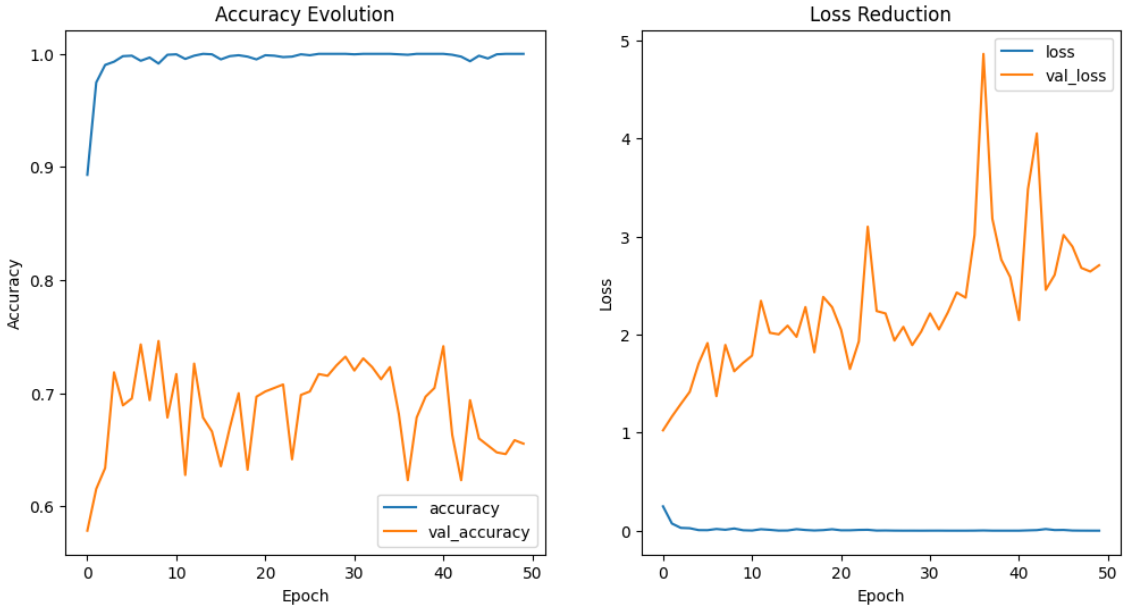

Figura 7. Evolução de acurácia à esquerda e função de *loss* à direita para o modelo *Xception*

O segundo gráfico, que mostra o *loss*, reforça essa análise. Enquanto o *loss* no conjunto de treino diminui rapidamente e se mantém muito baixa, a perda no conjunto de validação é bastante alta e instável, com uma tendência crescente, picos grandes e flutuações frequentes. Isso sugere que o modelo não está conseguindo minimizar efetivamente o erro nos dados de validação, levando a um comportamento errático.

Como observado na Figura 8, o modelo possui mais casos em que erroneamente classificou como não-amiloidose do que o *InceptionV3*, *MobileNet* e *ResNet-50*, entre-

tanto, ficou à frente do *VGG16*, que possui o maior número de casos de falsos negativos das arquiteturas testadas.

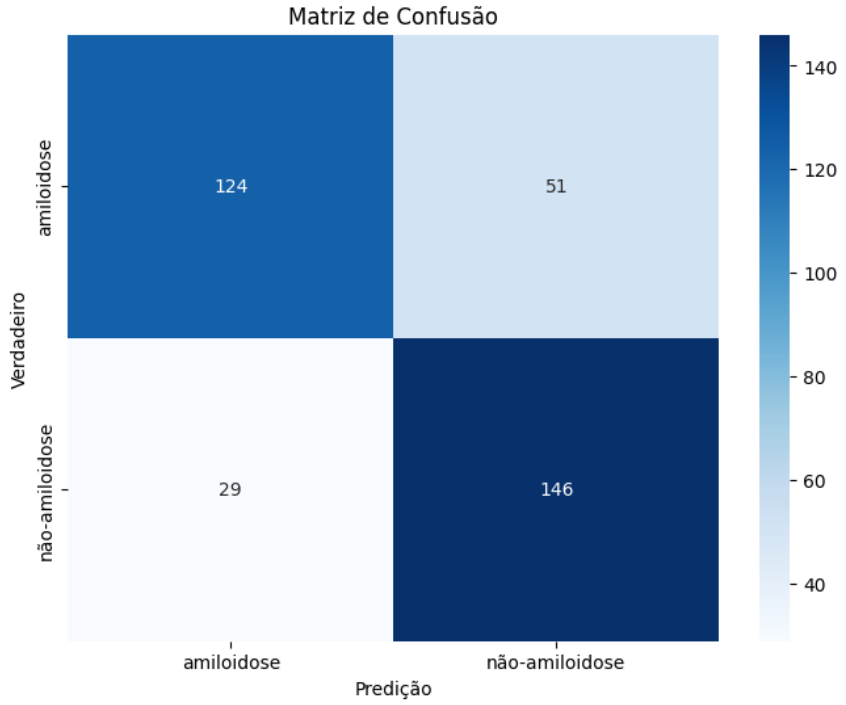

Figura 8. Matriz de confusão do modelo *Xception*

Por fim, a interpretação dos gráficos da Figura 7 sugere que o modelo pode estar com *overfitting*. A elevada acurácia no conjunto de treino, junto com a alta perda e variabilidade na acurácia de validação, indicam que o modelo está aprendendo padrões muito específicos do conjunto de treino, mas não está generalizando bem para novos dados. A instabilidade da perda de validação sugere que ajustes no modelo ou nas técnicas de regularização, ou um conjunto de dados para treinamento mais amplo, são necessários para melhorar a performance em dados de validação.

4.2.5. *MobileNet*

No gráfico de acurácia do modelo *MobileNet* exibido na Figura 9, observamos que a acurácia do conjunto de treino atinge rapidamente valores muito próximos de 1,0, assim como nos casos anteriores, indicando que o modelo está aprendendo bem nos dados de treino. No entanto, a acurácia no conjunto de validação varia entre 0,55 e 0,75 ao longo das épocas, sem sinais claros de estabilização, sugerindo que o modelo enfrenta dificuldades para manter uma performance consistente em novos dados.

O *loss* reforça essa interpretação. A perda no conjunto de treino é extremamente baixa e se mantém estável após as primeiras épocas, o que é comum em situações de superajuste (*overfitting*). A perda no conjunto de validação, por outro lado, flutua entre 2 e 2,5, com picos pronunciados, sugerindo que o modelo não está conseguindo minimizar o erro de forma eficiente em relação aos dados de validação. Essa combinação de resultados

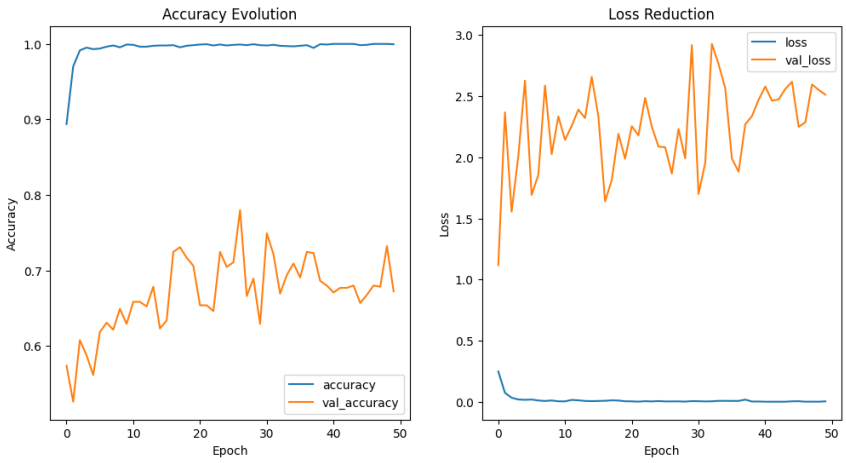

Figura 9. Evolução de acurácia à esquerda e *loss* à direita para o modelo *MobileNet*

sugere que o MobileNet também sofre de *overfitting*, ou seja, a performance no treino é excelente, mas a generalização para dados não vistos é limitada.

Entretanto, ao analisar a matriz de confusão na Figura 10, a predição do conjunto de testes mostra um desempenho promissor. O modelo acertou 133 casos de "amiloidose" e 130 de "não-amiloidose", o que indica que ele consegue identificar corretamente a maioria dos casos em ambas as categorias. Esses resultados mostram que o modelo tem boa capacidade de reconhecer corretamente a presença e a ausência da condição.

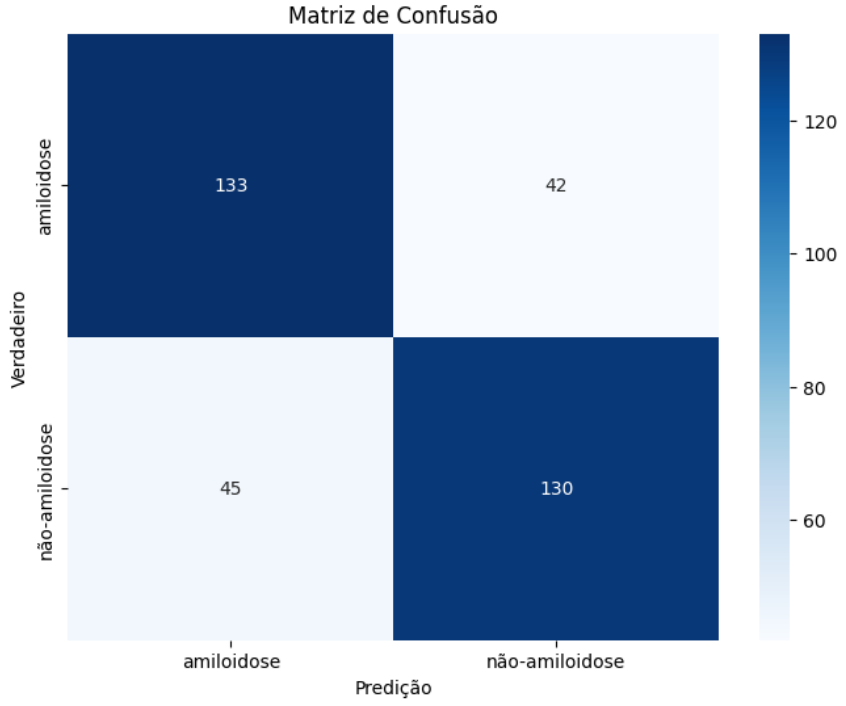

Figura 10. Matriz de confusão do modelo *MobileNet*

Embora existam 42 falsos negativos e 45 falsos positivos, esses valores são re-

lativamente baixos em comparação com o número de acertos, aproximando o *MobileNet* do *ResNet-50* que foi o mais performático, das arquiteturas testadas. Isso sugere que o modelo está próximo de uma boa generalização, além de evidenciar que a taxa de erros não é alarmante e poderia ser ainda mais refinada com ajustes no treinamento, como regularização adicional ou otimização do limiar de decisão.

4.3. AUC-ROC

Nesta seção, são apresentadas as AUC-ROC das cinco arquiteturas de redes neurais profundas, aplicadas à tarefa de classificação de AC em ETTs: *ResNet-50*, *InceptionV3*, *VGG16*, *Xception* e *MobileNet*. As AUC-ROC permitem avaliar o desempenho de cada modelo em termos de separação entre as classes. A seguir, discutimos os resultados obtidos por cada uma das arquiteturas, como mostrados na Figura 11.

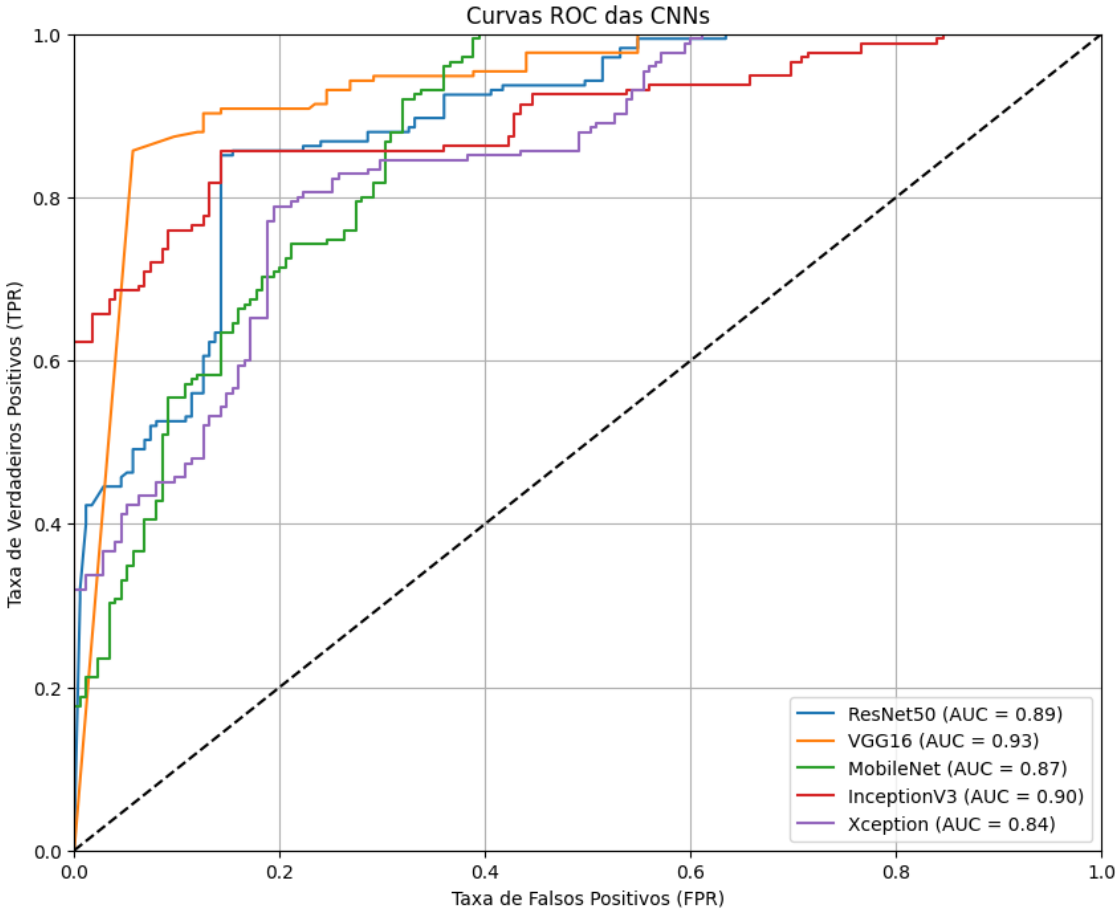

Figura 11. AUC-ROC para cada arquitetura

Em resumo, as curvas indicam que os modelos com maior AUC-ROC, como *VGG16* obtendo 0,93 e *InceptionV3* obtendo 0,90, são mais eficazes em separar as classes de forma consistente, enquanto *ResNet-50* se destaca por seu bom equilíbrio entre sensibilidade e especificidade com uma AUC-ROC de 0,89. Por fim, *Xception* e *MobileNet*, com AUC-ROC menores de 0,87, têm uma performance razoável, mas com uma maior probabilidade de falsos positivos e negativos.

4.4. Mapa de ativação utilizando Grad-CAM

De acordo com a Figura 12, os mapas de calor gerados por diferentes arquiteturas, destacam as regiões do coração que cada modelo considerou mais relevantes para a decisão de classificação, com foco principal em áreas como o ventrículo esquerdo e o septo interventricular, que são comumente afetadas pela deposição de amiloide [Kittleson et al., 2020]. A comparação das ativações entre as arquiteturas permite avaliar o comportamento de cada modelo na detecção de padrões associados à amiloidose.

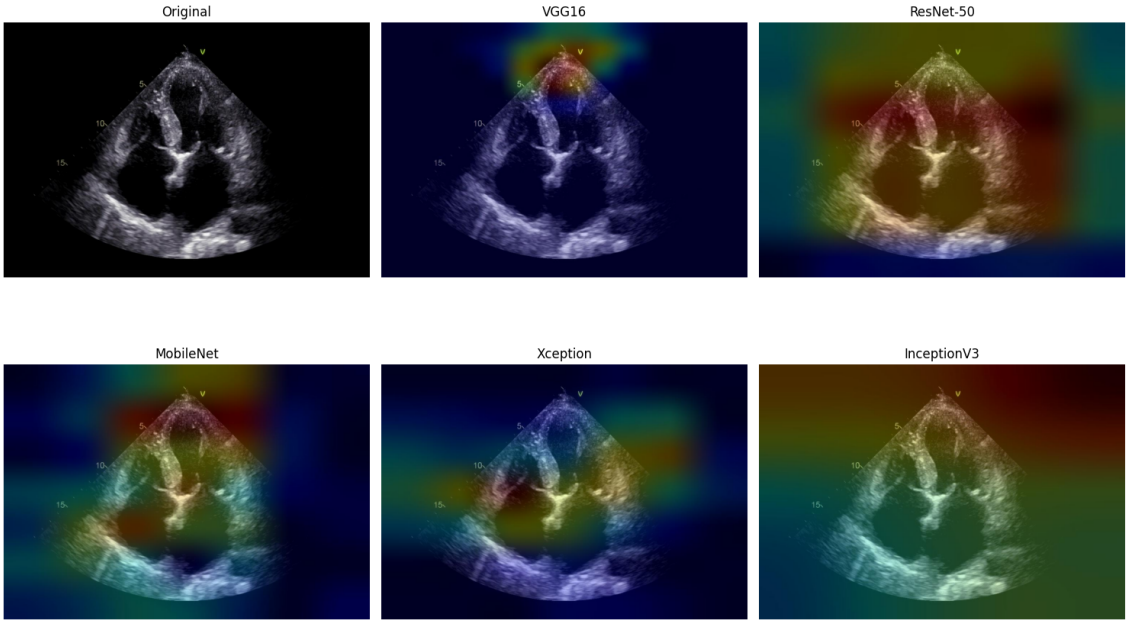

Figura 12. Mapa de ativação das CNNs

No mapa de calor da *ResNet-50*, as ativações estão concentradas nas regiões do ventrículo esquerdo, especialmente no ápice, e nas proximidades do septo interventricular. Essas áreas são fundamentais na detecção de AC, uma condição que frequentemente se manifesta por espessamento ventricular e alterações na função do ventrículo esquerdo [Kittleson et al., 2020]. O foco preciso nas estruturas mais impactadas pela amiloidose sugere que o modelo está capturando de maneira eficaz os sinais relevantes para o diagnóstico. Isso é confirmado pela sua acurácia superior em relação às outras arquiteturas, indicando um bom desempenho no diagnóstico.

O mapa de calor do *InceptionV3* mostra ativações espalhadas em várias regiões do coração, incluindo o ventrículo esquerdo e o septo interventricular. No entanto, o modelo também ativa regiões periféricas que não são tão relevantes para a detecção da doença, como bordas ao redor da imagem. Esse padrão de ativação mais amplo sugere que o *InceptionV3* está captando tanto características importantes, quanto ruído, o que pode resultar em uma performance ligeiramente inferior à da *ResNet-50*.

Na *VGG16*, as ativações estão mais concentradas na parte superior da imagem, principalmente ao redor do átrio esquerdo e da valva mitral, enquanto parece negligenciar partes essenciais como o ventrículo esquerdo e o septo interventricular. A falta de foco nas regiões mais impactadas pela doença pode explicar a menor acurácia e *recall* do modelo,

já que ele pode estar deixando de capturar características relevantes para o diagnóstico preciso da amiloidose.

A *Xception* apresenta ativações em várias áreas do ventrículo esquerdo e no septo interventricular, mas as ativações estão mais dispersas, incluindo regiões menos críticas. A falta de foco nas áreas principais que apresentam espessamento, ou alterações funcionais no coração, pode indicar que o modelo está processando informações irrelevantes junto com as relevantes. Isso se reflete em seu desempenho médio, sugerindo que a dispersão das ativações prejudica sua capacidade de detectar a amiloidose com alta precisão.

Por fim, o mapa de calor do *MobileNet* indica uma ativação difusa, com foco em várias regiões do ventrículo esquerdo, ventrículo direito, e nas bordas da imagem. Essa dispersão sugere que o modelo está captando informações de áreas que não são diretamente relacionadas à AC, o que afeta negativamente sua capacidade de identificar corretamente a doença. A falta de foco nas áreas essenciais, como o ventrículo esquerdo e o septo interventricular, onde a amiloidose geralmente se manifesta com mais clareza, pode explicar sua performance inferior, com a menor acurácia e pontuações entre os modelos.

5. Conclusão

Este trabalho buscou avaliar o desempenho de diferentes arquiteturas de CNNs no contexto de dados limitados, concentrando-se no comportamento das redes nessas condições, contribuindo para o avanço do diagnóstico precoce da AC. O objetivo geral foi verificar se, em um contexto escasso de dados, como é o caso de uma doença rara, os modelos de redes neurais performam satisfatoriamente em uma tarefa de DL, utilizando uma tarefa de classificação entre duas classes: "amiloidose" e "não-amiloidose", compostas por imagens de ETTs que continham ou não a doença.

Embora a progressão da função de perda e da acurácia nas fases de treinamento e validação sejam indicativos de *overfitting*, é essencial analisar o desempenho do modelo com outras métricas para obter uma avaliação mais completa. Assim, foram empregadas a matriz de confusão, a AUC-ROC e o *Grad-CAM* para observar o comportamento do modelo em detalhes. Essas ferramentas permitiram avaliar que, apesar dos indícios de *overfitting*, o modelo alcançou resultados consideráveis, reforçando a importância de uma análise ampla que vá além das métricas tradicionais para compreender o desempenho real das redes na tarefa de classificação de ETTs.

O objetivo foi alcançado, uma vez que as métricas coletadas, em conjunto com o mapa de calor com a ativação das redes, foram analisadas e foi verificado, rede a rede, quais arquiteturas conseguiram uma boa performance com base na acurácia, precisão, *recall* e pontuação F1. Neste cenário, a rede que melhor performou foi a *ResNet-50*, com base na acurácia. Em seguida, vieram a *InceptionV3*, *VGG16*, *Xception* e *MobileNet*. Entretanto, cada rede apresentou suas particularidades comportamentais e, além de regiões de ativação diferentes para um mesmo ETT, tendo *ResNet-50* mais coerente com a literatura relacionada, observando regiões esperadas no diagnóstico desta doença.

Ademais, as análises realizadas não apenas evidenciam a eficácia das CNNs na classificação de imagens médicas, mas também ressaltam a importância de considerar a escolha da arquitetura da rede em função do problema específico em questão. Essa pesquisa poderá servir como base para futuras investigações, estimulando o desenvolvimento de modelos mais robustos e a integração de técnicas de aprendizado de máquina em práticas clínicas, com o objetivo de melhorar o diagnóstico e o tratamento de doenças raras como a AC.

5.1. Trabalhos futuros

Embora os resultados tenham apresentado soluções viáveis para a utilização de DL no diagnóstico de AC, existem margens para o aprimoramento da acurácia dos modelos, redução do *loss* de validação e entendimento ainda mais profundo sobre o comportamento das arquiteturas utilizando o *Grad-CAM*. Para isto, as seguintes abordagens podem ser sugeridas:

- Utilização da técnica *grid search* para o ajuste individual de parâmetros de cada arquitetura, objetivando extrair o comportamento mais adequado para cada modelo
- Coleta de ETTs contendo outros casos de AC, a fim de aprimorar a generalização dos modelos
- Explorar mapa de ativação em diferentes ETTs e observar o comportamento em diferentes situações

Referências

Agibetov, A., Kammerlander, A., Duca, F., Nitsche, C., Koschutnik, M., Donà, C., Dachs, T.-M., Rettl, R., Stria, A., Schrutka, L., Binder, C., Kastner, J., Agis, H., Kain, R., Auer-Grumbach, M., Samwald, M., Hengstenberg, C., Dorffner, G., Mascherbauer, J., and Bonderman, D. (2021). Convolutional neural networks for fully automated diagnosis of cardiac amyloidosis by cardiac magnetic resonance imaging. *Journal of Personalized Medicine*, 11(12).

Ahmadi-Hadad, A., De Rosa, E., Di Serafino, L., et al. (2024). Artificial intelligence as a tool for diagnosis of cardiac amyloidosis: A systematic review. *Journal of Medical and Biological Engineering*, 44:499–513.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

de Cardiologia, S. B. (2021). Diagnóstico precoce e tratamento adequado são fundamentais para pacientes com amiloidose cardíaca. Disponível em: `https://www.portal.cardiol.br/br/post/diagn%c3%b3stico-precoce-e-tratamento-adequado-s%c3%a3o-fundamentais-para-pacientes-com-amiloidose-card%c3%adaca`. Acesso em 24 de setembro de 2024.

Delbarre, M.-A., Girardon, F., Roquette, L., Blanc-Durand, P., Hubaut, M.-A., Éric Hachulla, Semah, F., Huglo, D., Garcelon, N., Marchal, E., Esper, I. E., Tribouilloy, C., Lamblin, N., Duhaut, P., Schmidt, J., Itti, E., and Damy, T. (2023). Deep learning on bone scintigraphy to detect abnormal cardiac uptake at risk of cardiac amyloidosis. *JACC: Cardiovascular Imaging*, 16(8):1085–1095.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*. Tech report.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*. Submitted on 17 Apr 2017.

Kittleson, M. M., Maurer, M. S., Ambardekar, A. V., Bullock-Palmer, R. P., Chang, P. P., Eisen, H. J., Nair, A. P., Nativi-Nicolau, J., Ruberg, F. L., behalf of the American Heart Association Heart Failure, O., and of the Council on Clinical Cardiology, T. C. (2020). Cardiac amyloidosis: Evolving diagnosis and management: A scientific statement from the american heart association. *Circulation*, 142(1):e7–e22.

Kittleson, M. M., Ruberg, F. L., Ambardekar, A. V., Brannagan, T. H., Cheng, R. K., Clarke, J. O., Dember, L. M., Frantz, J. G., Hershberger, R. E., Maurer, M. S., Nativi-Nicolau, J., Sanchorawala, V., and Sheikh, F. H. (2023). 2023 acc expert consensus decision pathway on comprehensive multidisciplinary care for the patient with cardiac amyloidosis. *Journal of the American College of Cardiology*, 81(11):1076–1126.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Martini, N., Aimo, A., Barison, A., Della Latta, D., Vergaro, G., Aquaro, G. D., Ripoli, A., Emdin, M., and Chiappino, D. (2020). Deep learning to diagnose cardiac amyloi-

dosis from cardiovascular magnetic resonance. *Journal of Cardiovascular Magnetic Resonance*, 22(1):84.

Pfizer (2023). Amiloidose cardíaca. Disponível em: `https://www.pfizer.com.br/sua-saude/doencas-raras/amiloidose-cardiaca`. Acesso em 24 de setembro de 2024.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2019). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision (IJCV)*, 128(2):336–359.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567*.

Vrudhula, A., Stern, L., Cheng, P. C., Ricchiuto, P., Daluwatte, C., Witteles, R., Patel, J., and Ouyang, D. (2024). Impact of case and control selection on training artificial intelligence screening of cardiac amyloidosis. *JACC: Advances*, 3(9, Part 2):100998. AI in Cardiology: Improving Outcomes for All Focus Issue.

Yu, X., Yao, X., Wu, B., Zhou, H., Xia, S., Su, W., Wu, Y., and Zheng, X. (2022). Using deep learning method to identify left ventricular hypertrophy on echocardiography. *The International Journal of Cardiovascular Imaging*, 38(4):759–769.