Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Departamento de Computação e Sistemas

Fusão de Redes Neurais Convolucionais e Vision Transformers para Classificação de Câncer de Pele em Imagens Médicas

Mikael Vieira Magalhães

TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:
Eduardo da Silva Ribeiro

Julho, 2026
João Monlevade–MG

Mikael Vieira Magalhães

Fusão de Redes Neurais Convolucionais e Vision Transformers para Classificação de Câncer de Pele em Imagens Médicas

Orientador: Eduardo da Silva Ribeiro

Monografia apresentada ao curso de Sistemas de Informação do Instituto de Ciências Exatas e Aplicadas, da Universidade Federal de Ouro Preto, como requisito parcial para aprovação na Disciplina "Trabalho de Conclusão de Curso II".

Universidade Federal de Ouro Preto

João Monlevade

Julho de 2026

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

M188f Magalhães, Mikael Vieira.
Fusão de redes neurais convolucionais e vision transformers para classificação de câncer de pele em imagens médicas. [manuscrito] / Mikael Vieira Magalhães. - 2026.
49 f.

Orientador: Prof. Dr. Eduardo da Silva Ribeiro.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Graduação em Sistemas de Informação .

1. Aprendizado profundo (Aprendizado de máquina). 2. Inteligência artificial. 3. Pele - Câncer - Classificação. 4. Processamento de Imagem Assistido por Computador. 5. Redes Neurais Convolucionais. 6. Visão computacional. I. Ribeiro, Eduardo da Silva. II. Universidade Federal de Ouro Preto. III. Título.

CDU 004.8:616.5-006.6

Bibliotecário(a) Responsável: Flavia Reis - CRB6/2431

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE COMPUTACAO E SISTEMAS

FOLHA DE APROVAÇÃO

Mikael Vieira Magalhães

Fusão de Redes Neurais Convolucionais e Vision Transformers para Classificação de Câncer de Pele em Imagens Médicas

Monografia apresentada ao Curso de Sistemas de Informação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de bacharel em Sistemas de Informação

Aprovada em 30 de Julho de 2026

Membros da banca

Dr. Eduardo da Silva Ribeiro - Orientador (Universidade Federal de Ouro Preto)
Dr. Talles Henrique de Medeiros (Universidade Federal de Ouro Preto)
Dr. Luiz Carlos Bambirra Torres (Universidade Federal de Ouro Preto)

Eduardo da Silva Ribeiro, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 26/08/2026

Documento assinado eletronicamente por **Eduardo da Silva Ribeiro**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 26/08/2026, às 16:33, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **1153244** e o código CRC **2A0DE653**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3808-0819 - www.ufop.br

Dedico este trabalho aos meus pais, pelo amor, apoio e incentivo incondicionais, e à minha família, que sempre acreditou no meu potencial e esteve ao meu lado em todos os momentos desta caminhada.

Agradecimentos

Agradeço primeiramente a minha família, por me apoiarem e proverem um ambiente em que eu possa dedicar meu tempo aos estudos e pelo incentivo, apoio e compreensão durante toda a minha trajetória acadêmica. Sem vocês, esta conquista não seria possível.

Ao meu orientador, pela dedicação, paciência e pelos valiosos ensinamentos compartilhados ao longo do desenvolvimento deste trabalho. Sua orientação foi imprescindível para a realização desta pesquisa.

Aos professores do curso de Sistemas de Informação da Universidade Federal de Ouro Preto, pelos conhecimentos transmitidos durante a graduação e pela contribuição para minha formação acadêmica e profissional.

Aos meus amigos e colegas, pelo companheirismo, pelas conversas, pela troca de experiências e pelo apoio durante esta jornada.

Por fim, agradeço à Universidade Federal de Ouro Preto pela oportunidade de crescimento pessoal e profissional proporcionada ao longo desses anos.

"Science is more than a body of knowledge; it is a way of thinking."

— Carl Sagan (1934 – 1996),
in: The Demon-Haunted World: Science as a Candle in the Dark.

Resumo

O câncer de pele é o mais frequente no Brasil e no mundo, tornando fundamental o desenvolvimento de métodos para auxiliar o diagnóstico da doença. Neste trabalho, propõe-se a arquitetura híbrida SWF-Net, baseada na fusão em nível de características entre uma Rede Neural Convolucional (ResNet-50) e um Vision Transformer, para a classificação de câncer de pele em imagens médicas. O objetivo da pesquisa é avaliar se a estratégia de fusão proposta proporciona desempenho superior em comparação a seus resultados isolados. A pesquisa caracteriza-se como aplicada, utilizando técnicas de aprendizado profundo e os conjuntos de dados BloodMNIST, empregado na etapa piloto, e DermaMNIST, utilizado na avaliação final. Os modelos foram avaliados por meio das métricas de acurácia, AUC macro e matrizes de confusão. Os resultados demonstraram que a SWF-Net alcançou acurácia de 87,03% e AUC macro de 0,9638 no DermaMNIST, superando o desempenho da ResNet-50 isolada (84,94%) e do Vision Transformer treinado com pesos iniciados aleatoriamente (70,37%), indicando que a estratégia de fusão proposta é eficaz para melhorar o desempenho da classificação de lesões dermatológicas.

Palavras-chaves: Câncer de pele. Aprendizado profundo. Redes Neurais Convolucionais. Vision Transformers. Classificação de imagens médicas.

Abstract

Skin cancer is the most common type of cancer in Brazil and worldwide, making the development of methods to support its diagnosis essential. This work proposes the SWF-Net hybrid architecture, based on feature-level fusion between a Convolutional Neural Network (ResNet-50) and a Vision Transformer for skin cancer classification in medical images. The objective of this research is to evaluate whether the proposed fusion strategy provides superior performance compared to the individual models. This applied research employs deep learning techniques and the BloodMNIST and DermaMNIST datasets, used in the pilot and final evaluation stages, respectively. The models were evaluated using accuracy, macro AUC, and confusion matrices. The results showed that SWF-Net achieved an accuracy of 87.03% and a macro AUC of 0.9638 on the DermaMNIST dataset, outperforming the standalone ResNet-50 (84.94%) and the Vision Transformer trained from scratch (70.37%). These findings indicate that the proposed fusion strategy is effective in improving the performance of skin lesion classification.

Key-words: Skin Cancer, Deep Learning, Convolutional Neural Networks, Vision Transformers, Medical Image Classification.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

As doenças dermatológicas representam um importante desafio para a saúde pública, especialmente devido à elevada incidência de lesões com potencial maligno. Segundo estimativas do Instituto Nacional de Câncer (INCA), o câncer de pele não melanoma segue como o tumo maligno de maior incidência no Brasil, com projeção de aproximadamente 781 mil novos casos por ano entre 2026 e 2028, considerando o conjunto de tipos mais frequentes Instituto Nacional de Câncer (INCA) [2026]. Entre essas enfermidades, o câncer de pele destaca-se pela sua alta frequência e pela necessidade de diagnóstico precoce para aumentar as chances de sucesso no tratamento. Nesse contexto, técnicas de inteligência artificial e aprendizado de máquina têm sido amplamente empregadas para auxiliar na identificação e classificação dessas lesões. Essas abordagens possibilitam a análise automatizada de imagens médicas, contribuindo para a detecção precoce do câncer de pele e de seus diferentes tipos, além de fornecer suporte aos profissionais de saúde durante o processo diagnóstico. A disponibilidade de conjuntos de dados dermatológicos públicos e padronizados, como o HAM10000, atuou como um fator determinanete para viabilizar o treinamento e a comparação entre os modelos Tschandl et al. [2018].

Com o avanço das tecnologias de predição e classificação, novas técnicas foram desenvolvidas e aprimoradas para análise de imagens de forma automatizada, possibilitando a criação de ferramentas de apoio para os profissionais da saúde. A combinação de arquiteturas convulacionais com arquiteturas baseadas em atenção tem se consolidado promissoramente na visão computacional aplicada em imagens médicas, impulsionando ganhos de desempenho em diversas tarefas de classificação Haruna et al. [2025].

1.1 O problema de pesquisa

Apesar dos avanços recentes, parte das abordagens de classificação de lesões dermatológicas ainda se baseia no uso isolado de Redes Neurais Convolucionais (CNNs) ou de Vision Transformers (ViTs). As CNNs são reconhecidas por sua capacidade de extrair características locais, como bordas, texturas e padrões de contorno das lesões He et al. [2016], enquanto os ViTs se destacam pela capacidade de capturar relações globais entre diferentes regiões da imagem por meio de mecanismos de atenção Dosovitskiy et al. [2020]. Essa diferença sugere que as duas arquiteturas possuem características complementares, e que sua combinação pode gerar representações mais robustas do que o uso isolado de qualquer uma delas.

Entretanto, ainda há uma lacuna na literatura quanto a estratégias de fusão que explorem de forma equilibrada e adaptativa a contribuição de cada arquitetura durante

o treinamento, especialmente em cenários de dados médicos limitados, como o conjunto DermaMNIST Yang et al. [2023]. Parte dos trabalhos existentes utiliza estratégias estáticas de fusão, como concatenação direta de características ou média de predições (*late fusion*), tratando as arquiteturas como modelos independentes que apenas têm suas representações ou decisões combinadas ao final do processo Haruna et al. [2025]. Em contraste, estratégias que incorporam a fusão como parte da própria arquitetura, permitindo que os ramos sejam ajustados conjuntamente e influenciem mutuamente o aprendizado durante o treinamento, ainda são menos exploradas.

Diante disso, este trabalho investiga a seguinte questão de pesquisa: *é possível melhorar a acurácia da classificação de lesões dermatológicas por meio de uma estratégia de fusão estocástica em nível de características entre uma ResNet-50 e um Vision Transformer, em comparação ao uso isolado de cada arquitetura?*

1.2 Objetivos

O presente trabalho tem como objetivo geral propor e avaliar uma arquitetura híbrida de fusão em nível de características entre uma ResNet-50 e um Vision Transformer, denominada SWF-Net, para a classificação de lesões dermatológicas nos conjuntos de dados BloodMNIST e DermaMNIST Yang et al. [2023].

Este trabalho possui os seguintes objetivos específicos:

- Avaliar o impacto do aprendizado por transferência (pesos pré-treinados na ImageNet) e da técnica de *oversampling* no desempenho dos modelos ResNet-50 e ViT;
- Desenvolver uma camada de fusão estocástica (*Stochastic Weight Fusion*) capaz de ponderar dinamicamente a contribuição de cada arquitetura por dimensão do vetor de características;
- Implementar e treinar os modelos ResNet-50, ViT e a arquitetura híbrida proposta sobre os conjuntos BloodMNIST (piloto) e DermaMNIST (definitivo);
- Realizar experimentos de ablação, isolando a contribuição de cada ramo da arquitetura ($\lambda = 0$ e $\lambda = 1$), para validar a eficácia da estratégia de fusão;
- Comparar o desempenho da arquitetura proposta com os modelos individuais (ResNet-50 e ViT), utilizando métricas de acurácia e AUC macro.

1.3 Organização do trabalho

O restante deste trabalho é organizado como se segue. O Capítulo 2 apresenta a fundamentação teórica sobre Redes Neurais Convolucionais, Vision Transformers e

estratégias de fusão de características, além de uma revisão dos trabalhos relacionados na área de classificação de lesões dermatológicas. O Capítulo 5 descreve o desenvolvimento da arquitetura SWF-Net (Stochastic Weight Fusion), incluindo as configurações de treinamento e os conjuntos de dados utilizados. O Capítulo 6 apresenta os experimentos conduzidos e os resultados obtidos. Por fim, o Capítulo 7 apresenta as considerações finais, as limitações do trabalho e sugestões para pesquisas futuras.

2 Revisão bibliográfica

A classificação de lesões dermatólogicas por meio de aprendizagem profundo segue sendo um tema aplamente explorado na literatura, com trabalhos que podem ser abordados em três grupos, sendo eles, abordagens baseadas exclusivamente em CNNs, abordagens baseadas em VITs e abordagens híbridas que combinam as duas famílias de arquiteturas

2.1 Abordagens baseadas em CNNs

O uso de CNNs para classificação de câncer de pele é consolidado na literatura desde meados da década de 2010, impulsionado sobretudo pela disponibilização de bases dermatoscópicas públicas, como o HAM10000, que reúne mais de dez mil imagens de lesões pigmentadas categorizadas em sete classes diagnósticas Tschandl et al. [2018], e pelas competições do *International Skin Imaging Collaboration* (ISIC), voltadas especificamente à detecção de melanoma Codella et al. [2019]. Uma revisão sistemática conduzida por Brinker et al. Brinker et al. [2018] reuniu diversos estudos que empregaram CNNs para essa tarefa e identificou que, em vários cenários, o desempenho desses modelos já se aproximava ou superava o de dermatologistas em testes controlados, reforçando o potencial das CNNs como ferramentas de apoio ao diagnóstico. Arquiteturas convolucionais consagradas, como ResNet, VGG e EfficientNet, seguem sendo amplamente utilizadas como *backbones* nesses estudos, geralmente combinadas a técnicas de aprendizado por transferência para compensar o volume limitado de dados rotulados disponíveis na área médica.

Essa dependência exclusiva de características convolucionais, no entanto, também aparece em outros domínios de imagem médica. Yadav et al. Yadav et al. [2025] apresentaram um *framework* de *transfer learning* baseado exclusivamente em uma VGG-16 modificada para classificação multiclasse de tumores cerebrais (glioma, meningioma, pituitária e ausência de tumor) a partir de imagens de ressonância magnética, alcançando 98,5% de acurácia. Entretanto, por depender apenas de características convolucionais, essa abordagem tem dificuldade em capturar relações contextuais entre regiões distantes da imagem, uma limitação que motiva a adição do ramo ViT em modelos híbridos como o proposto neste trabalho.

2.2 Abordagens baseadas em ViTs

Mais recentemente, o Vision Transformer passou a ser investigado como alternativa ou um complemento às CNNs na classificação de imagens. Dagnaw et al. Dagnaw et al. [2024] compararam Vits pré treinados, Swin Transformers, uma abordagem que visa

solucionar o problema de alto custo computacional para imagens de grande tamanho e CNNs pré treinadas na classificação de câncer de pele, onde avaliam os ViTs como abordagens competitivas em desempenho em relação às CNNs.

Estudos similares utilizando a base de dados HAM10000, mostram que ViTs treinados para segmentação e classificação conjunta de lesões apresentam desempenho comparável com arquiteturas convolucionais consolidades, mesmo que, de forma geral, os ViTs requiram volumes de dados maiores de treinamento para atingir valores de alto desempenho, uma limitação relevante em bases médicas de escala reduzida como o DermaMNIST Yang et al. [2023].

2.3 Abordagens híbridas CNN–ViT

Diante das caractéristicas complementarem entre CNNs e ViTs, uma linha de pesquisa crescente propõe arquiteturas híbridas combinando ambos modelos. Haruna et al. Haruna et al. [2025] apresentam uma revisão abrangente dessas estratégias, propondo modelos híbridos em arranjos paralelos e estratégias de fusões como fusão antecipada (*early fusion*) ou tardia (*late fusion*), evidenciando que a escolha da estratégia da fusão é determinante para o ganho de desempenho obtido pela combinação das arquiteturas.

No estudo de imagens médicas utilizando aprendizado de máquina, métodos de fusão foram desenvolvidos combinando redes convolucionais e Vision Transformers, como o proposto por Jayaraman et al. Jayaraman et al. [2025] ao criar o CAFNet, um *framework* que combina uma rede neural convolucional com um Vision Transformer para classificação de tumores cerebrais a partir de ressonância magnética. Diferentemente de abordagens de fusão por concatenação simples, os autores introduziram um módulo de *Cross-Attention Fusion* (CAF) para integrar de forma mais eficaz as características locais da CNN com as características globais do ViT, alcançando uma acurácia de teste de 96,41% em um *dataset* multiclasse de MRI, superando modelos convencionais de *machine learning*, *deep learning* e *transfer learning*. Enquanto o CAFNet utiliza atenção cruzada como mecanismo de fusão determinístico e aprendido, o modelo SWF-Net proposto neste trabalho explora uma abordagem estocástica, na qual o peso de distribuição de cada ramo (λ) é variável, permitindo que a rede utilize ambos os modelos durante o treinamento.

De forma semelhante, Hussain et al. Hussain et al. [2025] propuseram o EFFResNet-ViT, um modelo que combina a extração de características convolucionais da ResNet com os mecanismos de atenção global do ViT para classificação explicável de imagens médicas, alcançando 99,31% de acurácia em um *dataset* de tumores cerebrais (BT CE-MRI) e 92,54% em um *dataset* de imagens retinianas. Assim como o CAFNet, o EFFResNet-ViT aprende uma combinação determinística entre os dois ramos, diferindo da abordagem estocástica adotada pela SWF-Net.

Esses trabalhos evidenciam a tendência de substituição de estratégias de fusão estocásticas, como a concatenação direta ou média de previsão, por mecanismos de ponderação dinâmica entre as redes convulacionais e os modelos baseados em atenção.

3 Redes Neurais Convolucionais e Vision Transformers

Este capítulo apresenta a fundamentação teórica das duas famílias de arquiteturas que compõem os ramos da rede híbrida proposta neste trabalho: as Redes Neurais Convolucionais (CNNs), representadas pela ResNet-50, e os Vision Transformers (ViTs). Além da descrição técnica de cada arquitetura, é apresentada uma revisão dos trabalhos que aplicam essas arquiteturas, isoladamente ou de forma combinada, à classificação de lesões dermatológicas e de imagens médicas em geral.

3.1 Redes Neurais Convolucionais

As Redes Neurais Convolucionais (CNNs) constituem a classe de arquiteturas predominante em tarefas de visão computacional, sendo estruturadas a partir de operações de convolução que extraem características espaciais hierárquicas das imagens de entrada. Nas camadas iniciais dessas redes, tipicamente são aprendidos padrões visuais de baixo nível, como bordas e texturas, ao passo que camadas mais profundas combinam essas representações em estruturas semânticas de maior complexidade. Entretanto, o aumento da profundidade dessas arquiteturas não resulta necessariamente em ganhos desempenho: He et al. He et al. [2016] demonstraram que redes convolucionais muito profundas sofrem de um problema de degradação, no qual o erro de treinamento aumenta à medida que camadas adicionais são empilhadas, fenômeno associado à dificuldade de propagação do gradiente ao longo da rede durante o treinamento.

Para mitigar essa limitação, a arquitetura ResNet introduziu o conceito de aprendizado residual, por meio de conexões de atalho (*skip connections*) que permitem que o sinal de uma camada seja somado diretamente à saída de camadas posteriores, sem passar pelas transformações intermediárias. Dessa forma, cada bloco residual passa a aprender uma função resíduo em relação à sua entrada, em vez de uma transformação completa e não referenciada, o que facilita a otimização de redes com dezenas ou centenas de camadas. A variante ResNet-50, utilizada neste trabalho, é composta por 50 camadas com peso organizadas em blocos residuais do tipo *bottleneck*, e representa um ponto de equilíbrio amplamente adotado na literatura entre profundidade da rede e custo computacional Hussain et al. [2025].

Neste estudo, a ResNet-50 foi utilizada com pesos inicializados a partir do pré-treinamento no conjunto ImageNet Deng et al. [2009], prática comumente denominada aprendizado por transferência (*transfer learning*). Esse procedimento parte do princípio

de que representações visuais aprendidas a partir de um conjunto de dados amplo e diversificado como bordas, texturas e padrões de contorno são, em grande medida, transferíveis para outros domínios de imagem, reduzindo a necessidade de grandes volumes de dados rotulados na tarefa-alvo. A partir dessa inicialização, todos os parâmetros da rede foram ajustados por meio de *fine-tuning* completo, utilizando exclusivamente as imagens do conjunto de dados deste estudo, o que caracteriza esse modelo como baseline para avaliação da capacidade de aprendizado de uma arquitetura convolucional beneficiada por conhecimento prévio.

3.2 Vision Transformers

O Vision Transformer (ViT), proposto por Dosovitskiy et al. Dosovitskiy et al. [2020], adapta ao domínio da visão computacional a arquitetura Transformer, originalmente concebida para tarefas de processamento de linguagem natural e fundamentada no mecanismo de atenção (*self-attention*) Vaswani et al. [2017]. Diferentemente das CNNs, cujo processamento é predominantemente local e hierárquico, o ViT trata a imagem de entrada como uma sequência de unidades discretas: a imagem é particionada em blocos regulares (*patches*), os quais são linearizados e projetados em vetores de características, de forma análoga à representação de tokens em modelos de linguagem. Informações de posição espacial são incorporadas a essa sequência por meio de *embeddings* posicionais, uma vez que o mecanismo de atenção, em sua formulação original, é invariante à ordem dos elementos de entrada.

O mecanismo de autoatenção permite que cada *patch* seja relacionado diretamente a todos os demais patches da imagem, independentemente da distância espacial entre eles, possibilitando a modelagem de dependências globais já nas primeiras camadas da rede. Essa característica contrasta com o comportamento das CNNs, nas quais o campo receptivo é ampliado progressivamente ao longo das camadas. Em contrapartida, o ViT não incorpora, em sua estrutura, os vieses indutivos inerentes às convoluções, como localidade e equivariância à translação, o que o torna, em geral, mais dependente de grandes volumes de dados de treinamento para atingir desempenho competitivo em relação às arquiteturas convolucionais Dosovitskiy et al. [2020].

Neste trabalho, optou-se por treinar o ViT integralmente com pesos iniciados aleatoriamente, sem inicialização a partir de pesos pré-treinados, de modo a isolar e avaliar o desempenho de uma arquitetura baseada em atenção nas mesmas condições experimentais impostas aos demais modelos, sujeita à mesma quantidade limitada de dados de treinamento. Adotou-se, ainda, uma configuração simplificada da arquitetura original, denominada neste trabalho ViT-64, caracterizada pela utilização de imagens de entrada com resolução reduzida (64x64 pixels) e batch size inferior ao empregado na configuração

convencional (ViT-224). Essa adaptação teve como objetivo reduzir o custo computacional do treinamento e o número de patches processados por imagem, adequando o modelo aos recursos computacionais disponíveis, ainda que às custas de uma redução na granularidade das informações visuais capturadas.

4 SWF-Net: Arquitetura de Fusão Estocástica Proposta

O modelo correspondente a arquitetura proposta neste trabalho, denominada SWF-Net, é o resultado da combinação de duas arquiteturas descritas anteriormente: uma ResNet-50 inicializada com pesos pré-treinados no conjunto ImageNet e ajustada por meio de fine-tuning completo; e um Vision Transformer treinado com pesos iniciados aleatoriamente utilizando o conjunto de dados do estudo.

A técnica de fusão selecionada realiza a fusão em nível de características (*feature-level fusion*) entre os vetores do modelo da ResNet-50 e do Vision Transformer. Inicialmente, ambos os vetores são instanciados para a mesma dimensionalidade e normalizados por meio de Layer Normalization. Em seguida, é aplicada a camada Stochastic Weight Fusion (SWF), responsável por combinar as representações utilizando pesos estocásticos gerados para cada dimensão do vetor de características. A fusão é definida pela Equação 4.1.

$$F_{\text{fused}}^{(d)} = \lambda^{(d)} F_{\text{cnn}}^{(d)} + \left(1 - \lambda^{(d)}\right) F_{\text{vit}}^{(d)} \tag{4.1}$$

em que F_{cnn} e F_{vit} representam os vetores de características produzidos pelos ramos da ResNet-50 e do Vision Transformer, respectivamente, enquanto $\lambda^{(d)} \in [0, 1]$ controla a contribuição relativa de cada modelo na dimensão d. Durante o treinamento, os valores de $\lambda^{(d)}$ não são fixos, mas gerados estocasticamente a cada *forward pass*: para cada dimensão d do vetor de características, são amostrados seis valores independentes de uma distribuição uniforme $\mathcal{U}(0, 1)$, cuja média compõe o valor inicial de $\lambda^{(d)}$. Em seguida, esse valor é recentralizado, de modo que a média de $\lambda^{(d)}$ ao longo das dimensões do vetor, para cada amostra do lote, seja igual a 0,5, sendo o resultado então limitado ao intervalo $[0, 1]$. Esse procedimento aproxima o comportamento de uma distribuição concentrada em torno de 0,5 (relacionada à soma de variáveis uniformes, segundo o Teorema Central do Limite), fazendo com que a rede seja exposta, a cada iteração, a pequenas variações estocásticas na contribuição relativa de cada ramo em torno de uma combinação equilibrada, em vez de depender de um peso de fusão fixo. Essa injeção de ruído controlado atua como um mecanismo de regularização, incentivando o aprendizado de representações robustas em ambos os ramos. No SWF-Net, esse comportamento estocástico permanece ativo também na etapa de avaliação, de modo que $\lambda^{(d)}$ continua sendo reamostrado mesmo durante a inferência sobre o conjunto de teste, e não apenas durante o treinamento. Já nos modelos de ablação, o comportamento estocástico é desativado por completo, sendo utilizados dois modelos variantes independentes, com λ fixado em 0 e em 1, respectivamente, isolando o

uso exclusivo do Vision Transformer e da ResNet-50 na representação final.

Para tornar mais concreto o procedimento descrito, apresenta-se a seguir um exemplo simplificado do mecanismo de fusão, considerando vetores de características de dimensionalidade reduzida ($d = 4$), meramente ilustrativos, uma vez que, na prática, os vetores produzidos pelos ramos da ResNet-50 e do Vision Transformer possuem dimensionalidade consideravelmente maior.

Para cada dimensão d, são amostrados seis valores independentes de uma distribuição uniforme $\mathcal{U}(0, 1)$, cuja média compõe o valor bruto de $\lambda^{(d)}$. A Tabela 1 apresenta os valores amostrados e a respectiva média para cada uma das quatro dimensões consideradas.

Tabela 1 – Amostragem dos seis valores uniformes e média resultante por dimensão

d	u_1	u_2	u_3	u_4	u_5	u_6	Média
1	0,72	0,55	0,61	0,48	0,39	0,66	0,5683
2	0,30	0,42	0,51	0,35	0,28	0,44	0,3833
3	0,85	0,77	0,69	0,91	0,73	0,80	0,7917
4	0,20	0,15	0,33	0,25	0,18	0,29	0,2333

A média dos valores brutos de $\lambda^{(d)}$ ao longo das quatro dimensões é igual a 0,4942, e não a 0,5, como requerido pelo procedimento de recentralização. Dessa forma, soma-se a cada dimensão a diferença $(0,5 - 0,4942) = 0,0058$, obtendo-se os valores recentralizados apresentados na Tabela 2. Como todos os valores resultantes já se encontram dentro do intervalo $[0, 1]$, a etapa de limitação (*clipping*) não altera o resultado neste exemplo.

Tabela 2 – Valores de $\lambda^{(d)}$ antes e após a recentralização

d	$\lambda^{(d)}$ bruto	$\lambda^{(d)}$ recentralizado
1	0,5683	0,5742
2	0,3833	0,3892
3	0,7917	0,7976
4	0,2333	0,2392

Considerando vetores de características hipotéticos $F_{\text{cnn}} = [1{,}2;\ -0{,}5;\ 0{,}8;\ 2{,}1]$ e $F_{\text{vit}} = [0{,}3;\ 1{,}1;\ -0{,}2;\ 0{,}4]$, a aplicação da Equação 4.1 com os valores de $\lambda^{(d)}$ recentralizados resulta no vetor fundido apresentado na Tabela 3.

Tabela 3 – Cálculo do vetor de características fundido

d	$F_{\text{cnn}}^{(d)}$	$F_{\text{vit}}^{(d)}$	$\lambda^{(d)}$	$F_{\text{fused}}^{(d)}$
1	1,20	0,30	0,5742	0,8168
2	-0,50	1,10	0,3892	0,4773
3	0,80	-0,20	0,7976	0,5976
4	2,10	0,40	0,2392	0,8066

Observa-se que, nas dimensões em que $\lambda^{(d)}$ assume valores mais próximos de 1 (como na dimensão 3), a contribuição do ramo convolucional predomina na representação

final, ao passo que, nas dimensões em que $\lambda^{(d)}$ se aproxima de 0 (como na dimensão 2), prevalece a contribuição do ramo baseado em atenção. Como os valores de $\lambda^{(d)}$ são reamostrados a cada *forward pass*, um novo conjunto de vetores F_{fused} seria obtido caso o procedimento fosse repetido para a mesma entrada, refletindo a natureza estocástica do mecanismo de fusão.

Após a obtenção do vetor de características fundido F_{fused}, este é encaminhado a uma cabeça de classificação, composta por camadas totalmente conectadas, responsável por projetar essa representação combinada no espaço de saída correspondente às classes do problema. A saída dessa camada é submetida a uma função Softmax, que converte os valores produzidos em uma distribuição de probabilidade sobre as classes possíveis, a partir da qual é obtida a predição final do modelo. Durante o treinamento, essa distribuição de probabilidade é utilizada no cálculo da função de perda de entropia cruzada (*Cross-Entropy Loss*), responsável por guiar a atualização dos pesos da rede, incluindo os parâmetros dos ramos ResNet-50 e Vision Transformer, por meio de retropropagação do gradiente.

Essa abordagem busca explorar a complementaridade entre representações convolucionais e mecanismos de atenção, reduzindo limitações individuais e aumentando a robustez das classificações.

5 Materiais e Métodos

Este capítulo descreve os materiais e métodos conduzidos para o desenvolvimento e avaliação da arquitetura SWF-Net (Stochastic Weight Fusion), abrangendo a definição das configurações de treinamento, os conjuntos de dados utilizados e as estratégias de fusão avaliadas.

Com o objetivo de avaliar o desempenho da arquitetura proposta, foram conduzidos experimentos utilizando três abordagens distintas de classificação de imagens dermatológicas. Os modelos foram selecionados de forma a permitir uma comparação entre arquiteturas individuais e uma estratégia de fusão multimodal baseada em redes neurais convolucionais e Transformers.

5.1 Análise do Dataset

Os conjuntos de dados utilizados neste trabalho foram obtidos do repositório MedMNIST[1] Yang et al. [2023], uma coleção de datasets de imagens médicas padronizados para classificação, amplamente utilizada como benchmark em tarefas de aprendizado de máquina na área da saúde. Inicialmente, o conjunto BloodMNIST foi utilizado como ambiente piloto para validação do pipeline de treinamento e fusão, sendo posteriormente substituído pelo conjunto DermaMNIST, dataset-alvo desta dissertação, para a condução dos experimentos definitivos.

5.2 Treinamento dos Dados

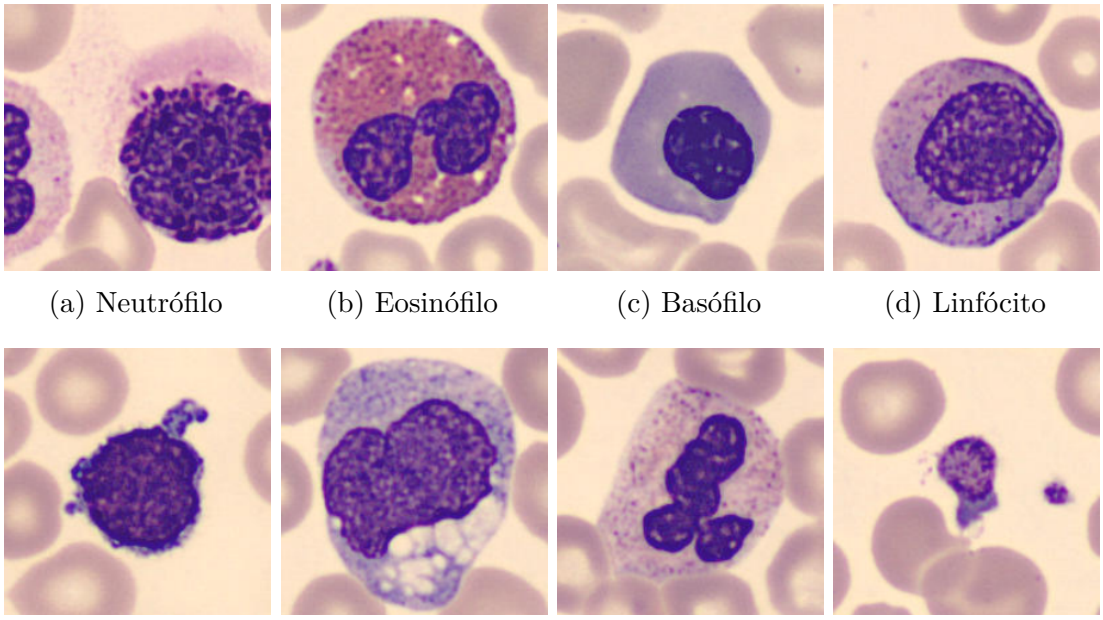
Para realizar o treinamento dos dados, o conjunto DermaMNIST e BloodMNist são fornecidos em uma distribuição dividida em subconjuntos de treino, validação e teste, com proporções de 70%, 10% e 20%, respectivamente, seguindo o particionamento oficial do MedMNIST

5.2.1 BloodMNIST

O conjunto BloodMNIST é composto por imagens de células sanguíneas individuais, capturadas por meio de microscopia Acevedo et al. [2020], e organizado em 8 classes correspondentes a diferentes tipos celulares. Este conjunto foi utilizado exclusivamente na etapa piloto dos experimentos, com o objetivo de validar o pipeline de treinamento antes da migração para o dataset-alvo.

[1] https://medmnist.com/

A Figura 1 apresenta exemplos de imagens para cada uma das classes presentes no conjunto BloodMNIST.



(a) Neutrófilo (b) Eosinófilo (c) Basófilo (d) Linfócito

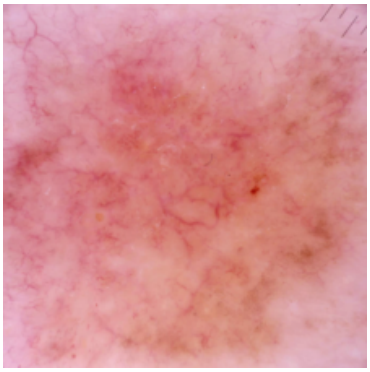
(e) Monócito (f) Granulócito (g) Eritroblasto (h) Plaqueta

Figura 1 – Exemplos das classes presentes no conjunto BloodMNIST

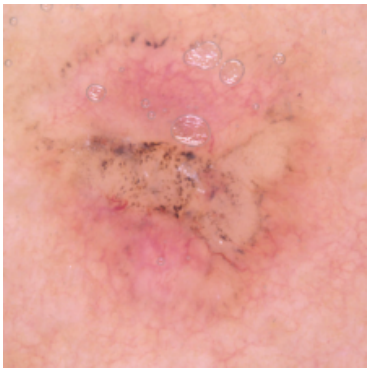
5.2.2 DermaMNIST

O conjunto DermaMNIST é composto por imagens dermatoscópicas organizadas em 7 classes de lesões de pele, derivado do dataset HAM10000 Tschandl et al. [2018]. Este foi o conjunto de dados-alvo utilizado na condução dos experimentos definitivos deste trabalho.

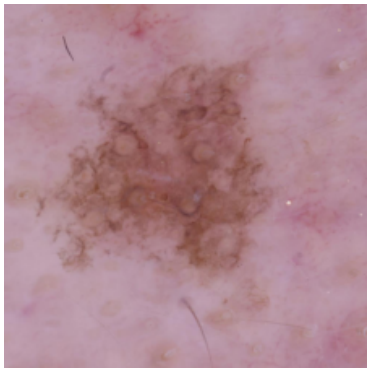
A Figura 2 apresenta exemplos de imagens para cada uma das classes presentes no conjunto DermaMNIST.



(a) Nevo melanocítico

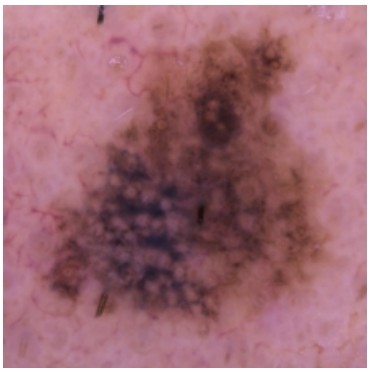


(b) Melanoma

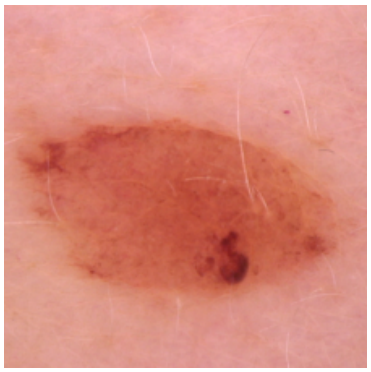


(c) Lesão benigna (ceratose)

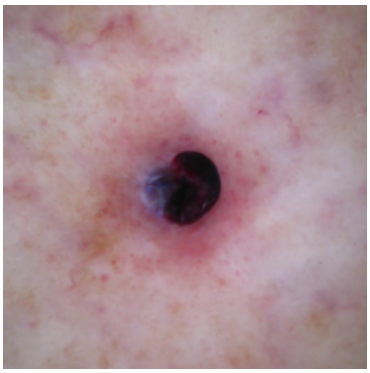
(d) Carcinoma basocelular



(e) Ceratose actínica



(f) Lesão vascular



(g) Dermatofibroma

Figura 2 – Exemplos das classes presentes no conjunto DermaMNIST

5.2.3 Distribuição das classes e oversampling

O conjunto DermaMNIST apresenta forte desbalanceamento entre classes, com a categoria de lesão vascular representando a maioria absoluta das amostras. Esse desbalanceamento pode induzir um viés do modelo em favor da classe majoritária, prejudicando a capacidade de generalização para as demais classes. A Tabela 4 apresenta a quantidade de amostras por classe antes e após a aplicação da técnica de oversampling no conjunto de treinamento. A ténica de oversampling utilizada engloba o uso de flips horizontais e verticais, rotações em 90° e 270°, alterações no brilho e no contraste e uso de crop mais resize das imagens.

Tabela 4 – Distribuição de amostras por classe no conjunto de treinamento do DermaMNIST antes e após oversampling adaptativo

Classe	**Antes do oversampling**	**Após oversampling**
Nevo melanocítico	0228	0774
Melanoma	0359	0774
Lesão benigna (ceratose)	0769	0774
Carcinoma basocelular	0080	0774
Ceratose actínica	0779	0779
Lesão vascular	4693	4693
Dermatofibroma	0099	0774

5.3 Configurações de treinamento

Os experimentos 1 a 3 foram conduzidos utilizando o ambiente Google Colab, com aceleração por GPU. Devido a limitações de tempo de execução contínua e à possibilidade de desconexão da sessão, alguns treinamentos foram interrompidos antes da convergência completa ou do acionamento do mecanismo de *EarlyStopping*, sendo retomados em sessões subsequentes quando necessário. Essa limitação de infraestrutura explica, por exemplo, a interrupção ou retomada de alguns treinamentos em sessões distintas durante a etapa piloto, conduzida com o conjunto BloodMNIST (Seção 5.2.1), antes da migração para o conjunto DermaMNIST (Seção 5.2.2), dataset-alvo dos experimentos definitivos.

A partir do Experimento 4, os treinamentos passaram a ser executados em um computador local com capacidade de processamento adequada para a execução completa dos modelos, eliminando as restrições de tempo de sessão e desconexão observadas anteriormente e permitindo a condução dos treinamentos sem interrupções.

O treinamento dos três modelos (ResNet-50, Vision Transformer e SWF-Net) seguiu uma configuração comum de otimização. Utilizou-se o otimizador Adam Kingma and Ba [2015], com taxa de aprendizado de 1×10^{-3} que posteriormente foi alterada para 1×10^{-4}, e a função de perda *Categorical Crossentropy* com *label smoothing* de 0,1, técnica que suaviza os rótulos *one-hot* e reduz o overconfidence do modelo, contribuindo para uma melhor generalização.

Como estratégia complementar ao oversampling adaptativo (Seção 5.2.3), foram calculados pesos de classe balanceados (*class weights*), utilizados para ponderar a função de perda de acordo com a frequência inversa de cada classe durante o treinamento. Essa combinação busca mitigar tanto o desbalanceamento em nível de dados (oversampling) quanto em nível de otimização (ponderação da perda).

Para o controle do processo de treinamento, foram empregados os seguintes *callbacks*:

- **EarlyStopping**: monitorando a métrica de perda de validação (*val_loss*), com paciência de 15 épocas sem melhora antes da interrupção do treinamento;
- **ReduceLROnPlateau**: monitorando *val_loss*, reduzindo a taxa de aprendizado pela metade (fator de 0,5) após 8 épocas consecutivas sem melhora;
- **ModelCheckpoint**: salvando os pesos do modelo correspondentes à melhor acurácia de validação (*val_accuracy*) observada ao longo do treinamento.

5.4 Métricas de avaliação

Para a avaliação dos modelos, foram utilizadas como métricas principais a acurácia de teste e a área sob a curva ROC (AUC) macro. A escolha da AUC macro como métrica complementar prioritária se justifica pelo desbalanceamento de classes presente no conjunto DermaMNIST (Seção 5.2.3): diferentemente da acurácia, que pode ser dominada pelo desempenho na classe majoritária, a AUC macro calcula a métrica individualmente para cada classe e realiza a média não ponderada entre elas, oferecendo uma medida mais representativa do desempenho do modelo em classes minoritárias. Por esse motivo, a análise de resultados apresentada neste trabalho não se limita à acurácia agregada, incorporando também uma avaliação do desempenho por classe e testes estatísticos de significância (Seção 6.5), de modo a verificar se o ganho de acurácia da arquitetura proposta se sustenta quando observado sob a ótica da AUC por classe.

6 Resultados

Este trabalho foi conduzido em duas etapas principais, utilizando inicialmente o conjunto de dados BloodMNIST como ambiente piloto para validação do pipeline de treinamento e fusão, e posteriormente o conjunto DermaMNIST, dataset-alvo desta dissertação, para a condução dos experimentos definitivos.

6.1 Experimento 1 — Validação piloto com o conjunto BloodMNIST

Na etapa inicial, adotou-se uma taxa de aprendizado de 1×10^{-3} para os treinamentos do ViT-64 e da ResNet-50 sobre o conjunto BloodMNIST. Devido a restrições de uso da infraestrutura de computação empregada, um dos treinamentos foi limitado a 25 épocas, número inferior ao observado em execução posterior do mesmo experimento, que se estendeu por 43 épocas até a ativação do mecanismo de *EarlyStopping*. Nessas condições restritas, a acurácia de validação do ViT-64 manteve-se estável entre 0,48 e 0,50 ao longo das últimas épocas, sem oscilações significativas, acompanhada de uma AUC de 0,84. Ao final do treinamento, obtiveram-se os resultados apresentados na Tabela 5.

Tabela 5 – Resultados do experimento piloto no conjunto BloodMNIST

Modelo	Acurácia de Teste	AUC Macro
ViT-64	0,5082	0,8685
ResNet-50	0,6658	0,9389
Fusão SWF-Net	0,6773	0,9271

Em virtude do elevado custo computacional associado ao tamanho do conjunto BloodMNIST, optou-se pela migração dos experimentos para o conjunto DermaMNIST, de menor dimensão, permitindo ciclos de treinamento mais rápidos e maior número de iterações experimentais.

6.2 Experimento 2 — Experimentos definitivos com o conjunto DermaMNIST

De forma análoga ao observado na etapa piloto, o treinamento inicial sobre o conjunto DermaMNIST com taxa de aprendizado de 1×10^{-3} apresentou desempenho insatisfatório. A redução da taxa de aprendizado para 1×10^{-4} resultou em melhora

expressiva nos resultados. Para esta etapa, adotou-se o ViT-224, configurado com imagens de resolução 224×224 e batch size de 16, conforme apresentado na Tabela 6.

Tabela 6 – Resultados parciais dos experimentos no conjunto DermaMNIST

Modelo	Acurácia de Teste	AUC Macro
ViT-224	0,7037	0.9171
ResNet-50	0.8494	0.9631

6.3 Experimento 3 — ViT-64 versus ViT-224

Para definição dos modelos, utilizamos uma abordagem de nomeação para as alterações nos modelos do ViT, sendo eles o ViT-64 e o ViT-224, referente a dimensão de projeção (embedding dimension), além da diferença entre parâmetros como segue a Tabela 7.

Tabela 7 – Configurações dos modelos Vision Transformer

Parâmetro	**ViT-64**	**ViT-224**
Resolução de entrada	128 × 128	224 × 224
Batch size	8	16
Patches	16	16
Dimensão de projeção (ViT)	64	224
Número de heads	4	4
Camadas transformer	6	6
MLP units	[128, 64]	[256, 224]
Acurácia	0.5082	0.7037
AUC	0.8685	0.9171

Inicialmente, adotou-se uma taxa de aprendizado de 1×10^{-3} para ambos os treinamentos do ViT-64 e da ResNet-50. Observou-se, contudo, que a acurácia de validação (val_acc) permanecia estagnada em torno de 0,014 durante as primeiras épocas do ViT, indicando que o modelo não estava convergindo, comportamento compatível com atualizações de gradiente excessivamente agressivas, que impedem o ajuste fino dos pesos da camada de atenção. Enquanto para a ResNet, o modelo não convergia, atingindo acurácia de em torno de 66%. A partir dessa constatação, reduziu-se a taxa de aprendizado para 1×10^{-4}, o que resultou em convergência estável e melhora consistente da acurácia ao longo das épocas. Esse valor foi, portanto, adotado na configuração final do modelo.

A migração da resolução de 128×128 (ViT-64) para 224×224 (ViT-224) refletiu diretamente nos resultados obtidos entre o experimento piloto e os experimentos definitivos, evidenciando a importância da resolução de entrada para a capacidade de representação do Vision Transformer. Na prática, o ViT-64 obteve acurácia de 50,82%, enquanto o ViT-224

alcançou 70,37%, evidenciando que o aumento da resolução de entrada resultou em uma melhora expressiva no desempenho do modelo

6.4 Experimento 4 — Avaliação do uso de Oversampling e ImageNet

A etapa inicial dos experimentos teve como objetivo avaliar diferentes configurações de treinamento para identificar a estratégia mais adequada para compor a arquitetura híbrida proposta. Para isso, foram analisados o impacto do aprendizado por transferência utilizando pesos pré-treinados no conjunto ImageNet, bem como a influência da técnica de oversampling sobre o desempenho dos modelos.

A fim de validar o uso do Oversampling e da ImageNet, treinamentos preliminares foram realizados, comparando os resultados obtidos como segue na Tabela 8.

Tabela 8 – Resultados dos experimentos preliminares de configuração

Modelo	**ImageNet**	**Oversampling**	**Acurácia**	**AUC**
ResNet-50	X	X	**84,94%**	0,9631
ResNet-50	-	X	72,27%	0,9343
ViT-224	-	X	70,37%	0,9171
ViT-224	-	-	60,45%	0,9060
ResNet-50	X	-	84,64%	**0,9651**
ResNet-50	-	-	66,73%	0,9304

Os resultados apresentados na Tabela 8 deixam claro o peso de cada fator avaliado. O aprendizado por transferência foi o que mais influenciou o desempenho da ResNet-50: com pesos pré-treinados no ImageNet, a acurácia saltou de 72,27% para 84,94% no cenário com oversampling, e de 66,73% para 84,64% sem oversampling, um ganho superior a 12 pontos percentuais em ambos os casos. Diante disso, optou-se por manter o ImageNet como configuração padrão para o ramo convolucional da arquitetura proposta.

Já o oversampling mostrou um efeito mais irregular, variando bastante conforme o modelo já partia ou não de pesos pré-treinados. Na ResNet-50 com ImageNet, a diferença entre os treinamentos apresentou impacto reduzido, de 84,94% contra 84,64%, apenas 0,30%, como se o conhecimento prévio da ImageNet já bastasse para lidar com o desbalanceamento das classes. Sem o ImageNet, porém, o oversampling passou a fazer diferença, elevando a acurácia em 5,54%, de 66,73% para 72,27%. O efeito mais expressivo foi observado no ViT, pois ao remover o oversampling, a acurácia caiu de 70,37% para 60,45%, uma perda de quase 10%. Isso sugere que modelos baseados em atenção, quando treinados com pesos iniciados aleatoriamente, dependem muito mais de um balanceamento adequado das classes do que redes convolucionais pré-treinadas.

6.5 Experimento 5 — Avaliação da estratégia de fusão

Com o objetivo de avaliar a contribuição individual de cada ramo da arquitetura proposta, foram realizados dois experimentos de ablação.

No primeiro experimento, definiu-se $\lambda = 1$, fazendo com que apenas as características extraídas pela ResNet-50 contribuíssem para a classificação. Embora o ramo ViT permanecesse presente na arquitetura, sua contribuição era anulada durante a fusão.

No segundo experimento, definiu-se $\lambda = 0$, de forma que somente as características produzidas pelo Vision Transformer fossem utilizadas na etapa de classificação.

Esses experimentos permitem avaliar o impacto da estratégia de fusão proposta, bem como verificar a contribuição individual de cada arquitetura para o desempenho final do modelo.

Tabela 9 – Resultados finais dos experimentos realizados no conjunto DermaMNIST

Modelo	Acurácia	AUC Macro
ResNet-50 ImageNet	84,94%	0,9631
ViT-224	70,37%	0,9171
SWF-Net (fusão estocástica)	**87,03%**	0,9638
SWF-Net ($\lambda = 1$)	86,48%	**0,9706**
SWF-Net ($\lambda = 0$)	71,42%	0,9088

Observa-se que a arquitetura proposta (SWF-Net), na Tabela 9, apresentou a maior acurácia entre todos os modelos avaliados, alcançando 87,03%. Esse resultado indica que a combinação das características extraídas pela ResNet-50 e pelo Vision Transformer proporciona representações mais discriminativas do que aquelas obtidas pelos modelos individuais.

Os experimentos de ablação reforçam essa conclusão. Quando apenas a ResNet-50 contribui para a fusão ($\lambda = 1$), a acurácia reduz para 86,48%. Por outro lado, quando somente o Vision Transformer participa da classificação ($\lambda = 0$), a acurácia cai para 71,42%, valor próximo ao obtido pelo modelo utilizando apenas o ViT.

Além do desempenho preditivo, avaliou-se também o custo computacional associado ao treinamento de cada uma das três arquiteturas principais (ResNet-50, ViT-224 e SWF-Net com fusão estocástica), conforme apresentado na Tabela 10. Os experimentos de ablação ($\lambda = 1$ e $\lambda = 0$) foram executados a partir da mesma rotina de treinamento da SWF-Net, variando-se apenas o valor fixo de λ na etapa de fusão, e por esse motivo não geraram tempos de treinamento independentes a serem reportados.

Tabela 10 – Tempo de treinamento e número de épocas dos modelos avaliados no conjunto DermaMNIST

Modelo	Épocas	Tempo de Treinamento
ResNet-50 ImageNet	51	6h47min
ViT-224	60	1h23min
SWF-Net (fusão estocástica)	37	5h39min

A ResNet-50 apresentou o maior tempo de treinamento entre os três modelos, totalizando 6h47min ao longo de 51 épocas, seguida pela SWF-Net, com 5h39min em 37 épocas, e pelo Vision Transformer, com 1h23min distribuídos em 60 épocas. Esse resultado chama atenção por não seguir a intuição inicial de ser razoável esperar que a SWF-Net, por incorporar simultaneamente os dois extratores de características (ResNet-50 e Vision Transformer) em uma única arquitetura de fusão, demandasse um tempo de treinamento igual ou superior ao da ResNet-50 isolada.

A análise do tempo médio por época revela a origem real dessa aparente contradição. Individualmente, cada época da SWF-Net é, de fato, mais custosa do que uma época da ResNet-50 isolada, sendo aproximadamente 9min11s contra 7min58s, respectivamente, o que é coerente com o maior número de parâmetros e operações envolvidas no processamento simultâneo dos dois ramos de extração de características. Entretanto, o critério de parada antecipada (*early stopping*) foi acionado de forma consideravelmente mais precoce para a SWF-Net, que convergiu em 37 épocas frente às 51 necessárias pela ResNet-50 treinada isoladamente. Esse menor número de épocas mais do que compensou o maior custo por época, resultando em um tempo total de treinamento inferior para a arquitetura de fusão. Esse comportamento sugere que a combinação estocástica das representações extraídas pelos dois modelos, mediada pelo peso λ variável, proporciona um sinal de aprendizado mais informativo à rede, permitindo que o modelo atinja seu melhor desempenho em validação com um número menor de atualizações de época, mesmo às custas de um processamento mais pesado a cada uma delas.

Vale observar também o caso do Vision Transformer, que apresentou o maior número de épocas entre os três modelos (60) e, ainda assim, o menor tempo total de treinamento (1h23min), em decorrência do custo computacional por época substancialmente menor (1min23s), um reflexo direto da arquitetura mais leve do modelo em relação às demais.

Do ponto de vista prático, esses resultados têm implicações relevantes para a viabilidade da proposta. Um dos questionamentos naturais ao se propor uma arquitetura de fusão que combina duas redes neurais profundas é se o ganho de desempenho obtido justifica o custo computacional adicional em relação aos modelos individuais. Os resultados aqui apresentados indicam que, ao menos no cenário avaliado, esse custo adicional não se concretizou de forma proporcional: a SWF-Net alcançou a maior acurácia entre todos

os modelos comparados (87,03%) com um tempo total de treinamento inferior ao do modelo unimodal mais custoso (ResNet-50), mesmo apresentando um custo por época mais elevado. Esse achado reforça a viabilidade prática da arquitetura proposta, uma vez que o ganho de desempenho não veio acompanhado de um aumento proporcional no custo computacional total de treinamento, característica desejável em cenários de aplicação real, especialmente em ambientes com recursos computacionais limitados, como é comum em contextos clínicos e hospitalares onde a infraestrutura de hardware voltada à inteligência artificial nem sempre está disponível em larga escala.

Vale destacar ainda que essa análise se restringe ao tempo de treinamento, não contemplando o tempo de inferência, que é a métrica de maior relevância prática para um eventual uso do modelo em ambiente de produção ou como ferramenta de apoio ao diagnóstico. Enquanto o tempo de treinamento impacta primariamente o custo do desenvolvimento e da experimentação, o tempo de inferência determina a viabilidade de uso do modelo em tempo real, por exemplo, em um sistema de triagem de lesões cutâneas integrado a um fluxo de atendimento clínico. Essa distinção é discutida com maior profundidade na Seção 7.2, onde são apontadas oportunidades de investigação futura relacionadas à otimização e implantação prática da arquitetura proposta.

A fim de investigar com maior granularidade o comportamento de cada modelo, foi realizado um *breakdown* do desempenho por classe, apresentado na Tabela 11. Essa análise permite identificar não apenas o desempenho médio de cada arquitetura, mas também em quais categorias específicas de lesões cada modelo se destaca ou apresenta maior dificuldade.

Tabela 11 – Comparação de Precisão, Recall e F1-Score por classe entre ResNet-50, ViT-224 e SWF-Net

Classe	SWF-Net		ResNet-50		ViT-224	
	Recall	F1	Recall	F1	Recall	F1
Nevo melanocítico	0,5909	0,6724	**0,6970**	**0,6866**	0,5152	0,3716
Melanoma	**0,9320**	**0,8276**	0,8447	0,8131	0,5534	0,4615
Ceratose benigna	0,6636	0,7392	**0,7682**	**0,7478**	0,5455	0,5195
Carcinoma basocelular	0,6957	0,7442	**0,7826**	**0,8000**	0,6522	0,3704
Ceratose actínica	0,6278	0,6364	**0,6816**	**0,6801**	0,6906	0,4813
Lesões vasculares	**0,9456**	**0,9317**	0,9224	0,9304	0,7122	0,8208
Dermatofibroma	**1,0000**	**0,9355**	0,9310	0,9153	0,8966	0,7429

Observa-se que, apesar da SWF-Net apresentar a maior acurácia geral entre os três modelos (Tabela 9), esse ganho não se traduz em superioridade uniforme entre as classes. Em termos de F1-Score, a SWF-Net obtém o melhor resultado em apenas três das sete classes avaliadas (Melanoma, Lesões vasculares e Dermatofibroma), enquanto a ResNet-50 isolada apresenta o maior F1-Score nas quatro classes restantes (Nevo melanocítico,

Ceratose benigna, Carcinoma basocelular e Ceratose actínica). O Vision Transformer, por sua vez, não obtém o melhor resultado em nenhuma classe, o que é consistente com seu desempenho geral inferior.

Um padrão relevante emerge ao se observar a relação entre o tamanho das classes e o modelo vencedor sendo que a SWF-Net tende a superar a ResNet-50 justamente nas classes de maior representatividade no conjunto de teste, como Lesões vasculares (1341 amostras) e, em menor grau, Melanoma (103 amostras) ao passo que a ResNet-50 mantém vantagem nas classes mais raras, como Carcinoma basocelular (23 amostras) e Nevo melanocítico (66 amostras). Esse comportamento sugere que o mecanismo de fusão estocástica, ao otimizar para uma métrica agregada como a acurácia geral, pode estar favorecendo implicitamente as classes majoritárias em detrimento das classes minoritárias, mesmo com a etapa de *oversampling* aplicada durante o treinamento.

Esse achado é particularmente relevante do ponto de vista clínico, uma vez que o Carcinoma basocelular e o Nevo melanocítico atípico estão entre as categorias em que erros de classificação, sobretudo falsos negativos podem ter maior impacto na condução diagnóstica. Assim, embora a SWF-Net apresente o melhor desempenho médio global, os resultados por classe indicam que a escolha entre as arquiteturas avaliadas não deveria se basear exclusivamente na acurácia agregada, sendo necessário ponderar também o desempenho em classes clinicamente prioritárias. Essa constatação reforça a discussão já apresentada sobre a redução do recall do Carcinoma basocelular na SWF-Net em relação à ResNet-50 isolada, evidenciando que tal comportamento não é um caso isolado, mas parte de um padrão mais amplo de *trade-off* entre desempenho agregado e desempenho por classe.

Com o objetivo de verificar a estabilidade do resultado obtido pela SWF-Net e avaliar a medida com que o comportamento estocástico da camda de fusão influencia na variabilidade das métricas finais, tal treinamento foi posto em análise perante 5 repetições, variando-se a semenete (*seed*) de inicialização. Os resultados de cada execução, bem como a média e o intervalo de confiança de 95% (calculado por meio da distribuição t de Student, com quatro graus de liberdade) Bussab and Morettin [2017], estão apresentados na Tabela 12.

Tabela 12 – Resultados das cinco execuções da SWF-Net com sementes distintas

Execução	Seed	Acurácia	AUC Macro	Tempo de Treinamento
1	42	0,8768	0,9674	8h52min
2	43	0,8683	0,9611	7h02min
3	44	0,8813	0,9647	8h24min
4	45	0,8723	0,9574	9h10min
5	46	0,8574	0,9689	6h07min
Média	—	0,8712	0,9639	7h55min
IC 95%	—	[0,8599; 0,8826]	[0,9581; 0,9697]	—

Podemos observar que os valores de acurácia e AUC Macro reportados na Tabela 9 correspondem, de forma aproximada, à média das cinco execuções (87,12% de acurácia e 0,9639 de AUC Macro), o que reforça a representatividade do resultado principal apresentado ao longo deste trabalho. O desvio entre as execuções foi relativamente baixo, com um intervalo de confiança de apropximadamente 2,3% para a acurácia, indicando que, apesar da variação do resultado nas execuções devido a natureza estocástica do mecanismo de fusão, existe uma razoável diferença entre as inicializações.

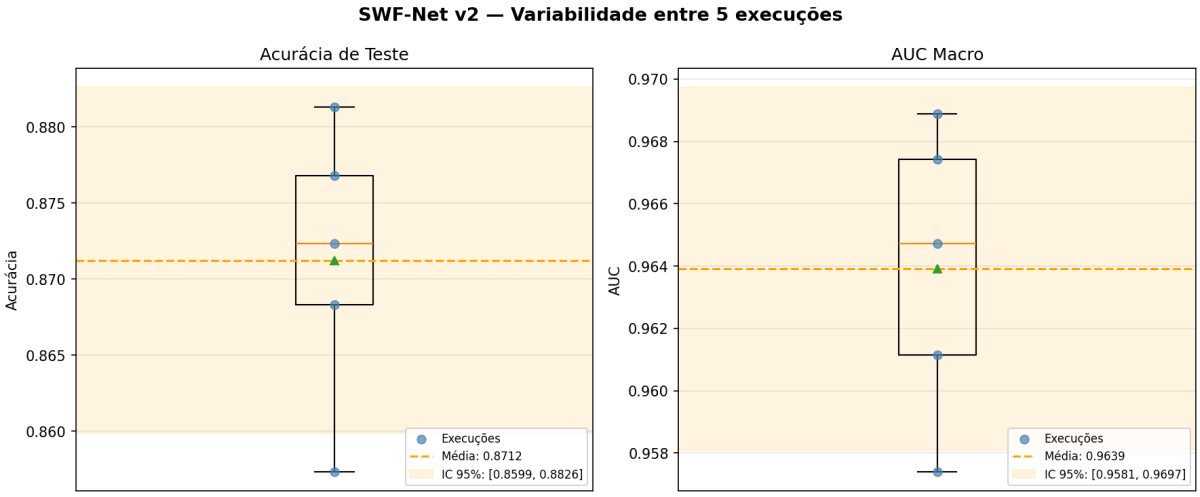


Figura 3 – Variabilidade da acurácia de teste e da AUC Macro da SWF-Net entre as cinco execuções com sementes distintas

A Figura 3 ilustra graficamente a dispersão dos resultados apresentados na Tabela 12. Observa-se que, tanto para a acurácia quanto para a AUC Macro, os valores obtidos nas cinco execuções se distribuem de forma relativamente concentrada em torno da média, com a execução de seed 46 representando o valor mínimo em ambas as métricas e a execução de seed 44 (acurácia) e seed 46 (AUC) representando os valores máximos. A amplitude entre a execução de melhor e de pior desempenho foi de aproximadamente 2,4% para a acurácia e 1,1% para a AUC Macro, o que reforça a interpretação de que, apesar

da natureza estocástica da camada de fusão, o desempenho da SWF-Net não apresenta variações extremas entre diferentes inicializações.

Adicionalmente, a fim de investigar a diferença de desempenho entre a SWF-Net e a Resnet-50, observada na análise de classe, aplicou-se o teste de Wilcoxon pareado por classe (n=7) Demšar [2006], comparando a AUC média das cinco execuções da SWF-Net com a AUC obtida pela ResNet-50, conforme apresentado na tabela 13. Vale-se ressaltar que a comparação apresentada não perfeitamente simétrica, uma vez que a AUC da SWF-Net corresponde a média de cinco execuções, enquanto para a ResNet-50 seja de apenas uma execução, motivo que deve ser considerado na interpretação dos resultados.

Tabela 13 – Comparação de AUC, Recall e F1-Score por classe entre SWF-Net e ResNet-50

Classe	AUC SWF (média)	AUC ResNet	Recall SWF	Recall ResNet	F1 SWF	F1 ResNet
Nevo melanocítico	0,9491	0,9540	0,5909	0,6970	0,6724	0,6866
Melanoma	0,9864	0,9913	0,9320	0,8447	0,8276	0,8131
Ceratose benigna	0,9609	0,9642	0,6636	0,7682	0,7392	0,7478
Carcinoma basocelular	0,9757	0,9903	0,6957	0,7826	0,7442	0,8000
Ceratose actínica	0,9221	0,9331	0,6278	0,6816	0,6364	0,6801
Lesões vasculares	0,9546	0,9567	0,9456	0,9224	0,9317	0,9304
Dermatofibroma	0,9986	0,9998	1,0000	0,9310	0,9355	0,9153

O teste de Wilcoxon aplicado à AUC revelou uma diferença estatisticamente significativa entre os modelos ($p = 0{,}0156$), com a SWF-Net apresentando AUC inferior à ResNet-50 em todas as sete classes avaliadas, ainda que as diferenças individuais sejam pequenas, variando entre 0,12% e 1,46%. Esse resultado é consistente e reforça, sob outra métrica, o padrão já discutido de que o ganho de desempenho da SWF-Net não se distribui uniformemente entre as classes, sendo particularmente concentrado nas classes majoritárias. Já para o Recall, o teste de Wilcoxon não indicou diferença estatisticamente significativa entre os modelos ($p = 0{,}4688$), com a SWF-Net superando a ResNet-50 em três das sete classes (Melanoma, Lesões vasculares e Dermatofibroma) e sendo superada nas quatro restantes, sem um padrão direcional consistente. O mesmo padrão se repete no F1-Score, para o qual o teste também não indicou diferença estatisticamente significativa ($p = 0{,}5781$), com a SWF-Net superando a ResNet-50 exatamente nas mesmas três classes. Ressalva-se, contudo, que todos os testes foram conduzidos com uma amostra pequena (n=7 classes), o que limita o poder estatístico da análise, especialmente no caso do Recall e do F1-Score.

Para investigar como cada arquitetura organiza o espaço de representações latentes, foi realizada uma projeção t-SNE dos embeddings extraídos por cada modelo, permitindo comparar visualmente a separabilidade entre classes obtida pela ResNet-50, pelo ViT médio standalone e pela SWF-Net, conforme ilustrado na Figura 4.

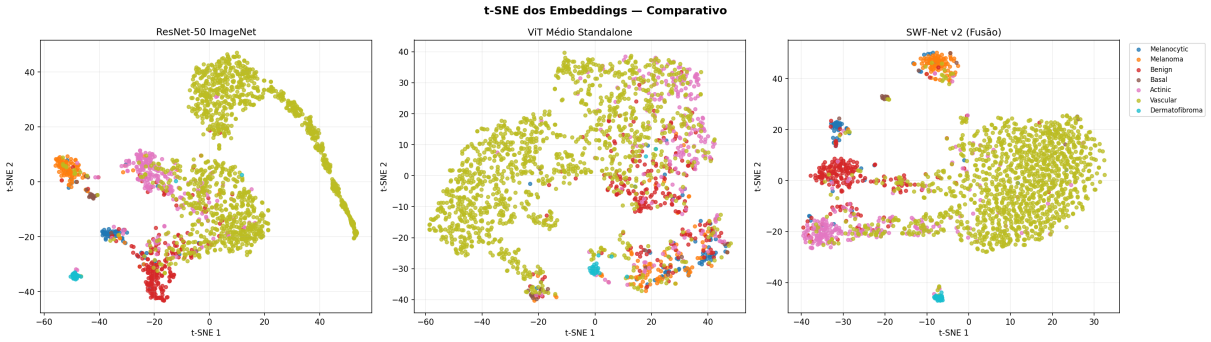


Figura 4 – Projeção t-SNE dos embeddings extraídos pela ResNet-50 ImageNet, pelo ViT médio standalone e pela SWF-Net v2 (fusão), com as sete classes de lesões coloridas conforme legenda

A análise qualitativa dos embeddings via t-SNE (Figura 4) complementa os resultados quantitativos discutidos anteriormente. O ViT médio standalone apresenta a organização menos discriminativa dentre os três modelos, com sobreposição substancial entre praticamente todas as classes, o que é consistente com seu desempenho inferior em Recall e F1-Score. A ResNet-50, por sua vez, organiza a classe majoritária (Lesões vasculares) ao longo de uma estrutura contínua, com as demais classes formando agrupamentos parcialmente sobrepostos. Já a SWF-Net produz o agrupamento mais compacto e mais bem separado para a classe majoritária, condizente com seus valores elevados de Recall e F1 nessa classe. Chama atenção, contudo, a sobreposição entre Nevo melanocítico e Melanoma no cluster superior da SWF-Net: essa confusão espacial ajuda a explicar o padrão observado nas métricas por classe, em que o Recall de Melanoma atinge seu valor mais alto (0,9320) exatamente à custa do Recall de Nevo melanocítico, seu valor mais baixo (0,5909), sugerindo que o modelo tende a resolver a ambiguidade entre essas duas classes clinicamente relacionadas em favor da classe de maior risco clínico. Já Dermatofibroma, tanto na ResNet-50 quanto na SWF-Net, forma um agrupamento pequeno e isolado, refletindo o Recall próximo ou igual a 1,0 observado em ambos os modelos para essa classe minoritária.

6.5.1 Análise das curvas de treinamento e matriz de confusão

Analisando os modelos realizados e o estudo de ablação, utilizamos a matriz de confusão para ajudar a visualização nas diferenças dos modelos e seus resultados que embora próximos, sofrem de diferentes problemas.

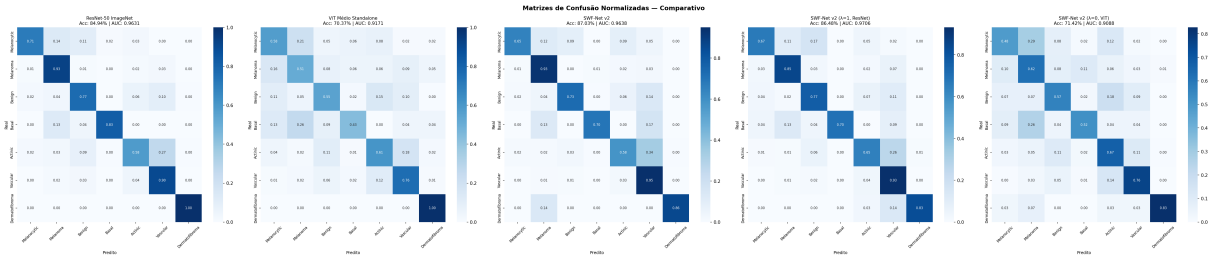


Figura 5 – Matrizes de confusão normalizadas — comparação entre ResNet-50, ViT standalone e SWF-Net (incluindo os experimentos de ablação com $\lambda = 1$ e $\lambda = 0$) no conjunto DermaMNIST

A Figura 5 apresenta as matrizes de confusão normalizadas para os cinco modelos avaliados, permitindo identificar quais classes específicas são mais afetadas pela estratégia de fusão, informação que as métricas agregadas de acurácia e AUC macro não revelam isoladamente.

De forma geral, observa-se que a classe de nevo melanocítico (*Melanocytic*) é a mais desafiadora para todos os modelos, sendo frequentemente confundida com melanoma e lesão benigna do tipo ceratose. A ResNet-50 isolada e a SWF-Net apresentam recall de 0,71 e 0,65, respectivamente, para essa classe, os valores mais baixos entre as sete categorias avaliadas.

A comparação entre a ResNet-50 isolada e a SWF-Net ($\lambda = 0,5$) revela um comportamento não uniforme da fusão: enquanto classes como melanoma (0,93 em ambos os modelos) e lesão vascular (0,90 na ResNet-50 e 0,95 na SWF-Net) mantêm ou melhoram o desempenho com a fusão, a classe de carcinoma basocelular apresenta uma queda expressiva, de 0,83 de recall na ResNet-50 isolada para 0,70 na SWF-Net. Um padrão semelhante ocorre com a ceratose actínica, que na SWF-Net é mais frequentemente confundida com lesão vascular (0,34) do que na ResNet-50 isolada (0,27).

Os experimentos de ablação permitem investigar essa questão com mais profundidade. Para a classe de carcinoma basocelular, tanto a configuração $\lambda = 1$ (apenas ResNet) quanto $\lambda = 0$ (apenas ViT) mantêm recall de 0,70 e 0,52, respectivamente, ou seja, mesmo isolando a contribuição da ResNet-50 (que originalmente atinge 0,83 quando treinada sozinha, fora do contexto da fusão), o desempenho nessa classe específica não se recupera dentro da arquitetura de fusão. Isso sugere que o processo de fusão, embora benéfico para o desempenho agregado, pode introduzir interferência entre ramos em classes específicas, possivelmente pela partilha de parâmetros na camada de classificação final ou por incompatibilidades entre as representações aprendidas por cada ramo para determinadas categorias.

Esse resultado reforça a relevância de se avaliar o desempenho por classe, além das métricas agregadas: uma melhora na acurácia geral (de 84,94% para 87,03%) pode coexistir com perdas de desempenho em classes clinicamente relevantes, como o carcinoma

basocelular, um subtipo de câncer de pele.

Além da comparação dos resultados finais, é relevante analisar o comportamento dos modelos ao longo do treinamento. A Figura 6 apresenta a evolução da acurácia, da perda (*loss*) e da AUC no conjunto de validação, ao longo de até 100 épocas, para a ResNet-50, o ViT standalone e a SWF-Net.

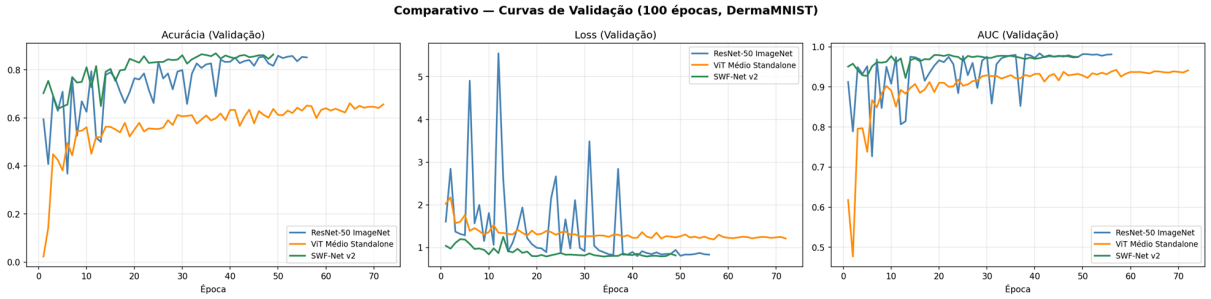


Figura 6 – Curvas de validação (acurácia, loss e AUC) ao longo do treinamento — ResNet-50, ViT e SWF-Net no conjunto DermaMNIST

Observa-se, na Figura 6, que tanto a ResNet-50 quanto a SWF-Net atingem patamares elevados de acurácia já nas primeiras épocas, comportamento esperado em razão do uso de pesos pré-treinados na ImageNet no ramo convolucional de ambos os modelos. O ViT standalone, por sua vez, apresenta convergência consideravelmente mais lenta, com acurácia e AUC inferiores ao longo de praticamente todas as épocas, refletindo a maior dificuldade de uma arquitetura baseada unicamente em atenção para aprender representações discriminativas quando treinada com pesos iniciados aleatoriamente em um conjunto de dados de escala limitada.

Um aspecto que diferencia a SWF-Net da ResNet-50 isolada, no entanto, é a estabilidade da curva de loss. Embora ambas utilizem o mesmo ramo convolucional pré-treinado, a ResNet-50 isolada apresenta picos abruptos de loss ao longo do treinamento (atingindo valores superiores a 5), enquanto a SWF-Net mantém a loss consistentemente baixa e com poucas oscilações. Como essa diferença não pode ser atribuída ao pré-treinamento presente em ambos os modelos, ela sugere que a incorporação do ramo ViT à fusão contribui para uma otimização mais estável, possivelmente por atenuar a sensibilidade do modelo a lotes de dados específicos que provocam atualizações de gradiente mais agressivas na ResNet isolada.

7 Conclusão

Este trabalho teve como objetivo propor e avaliar a arquitetura SWF-Net, uma abordagem híbrida de fusão em nível de características entre uma ResNet-50 e um Vision Transformer, aplicada à classificação de lesões dermatológicas nos conjuntos BloodMNIST e DermaMNIST. A motivação central partiu da lacuna identificada na literatura quanto a estratégias de fusão que ponderem, de forma adaptativa e por dimensão do vetor de características, a contribuição relativa de arquiteturas convolucionais e baseadas em atenção.

Os resultados obtidos indicam que a fusão estocástica proposta contribuiu para um ganho de desempenho em relação aos modelos individuais: a SWF-Net alcançou 87,03% de acurácia e 0,9638 de AUC macro no conjunto DermaMNIST, superando tanto a ResNet-50 isolada (84,94%) quanto o Vision Transformer treinado com pesos iniciados aleatoriamente (70,37%). Os experimentos de ablação, com λ fixado em 0 e em 1, reforçaram essa conclusão ao isolar a contribuição de cada ramo dentro da arquitetura de fusão, evidenciando que o desempenho da arquitetura de fusão depende consideravelmente mais da contribuição do ramo convolucional pré-treinado do que do ramo ViT treinado com pesos iniciados aleatoriamente.

Esse resultado, no entanto, deve ser interpretado à luz das condições específicas de treinamento do Vision Transformer utilizado neste trabalho. Diferentemente da ResNet-50, que se beneficiou dos pesos pré-treinados na ImageNet, o ViT foi treinado com pesos iniciados aleatoriamente, utilizando um volume de dados consideravelmente menor do que o tipicamente empregado no treinamento de arquiteturas baseadas em atenção, conforme discutido na Seção 6.4, onde seu desempenho aumentou em cerca de 10% ao utilizar oversampling no conjunto de dados, referenciando que uma quantidade maior de dados resulta em um desempenho mais aceitável do modelo. Dessa forma, a menor contribuição do ramo ViT observada nos experimentos de ablação pode refletir uma limitação do regime de treinamento adotado para essa arquitetura específica, e não necessariamente uma limitação da estratégia de fusão proposta.

Retomando a questão de pesquisa proposta na Seção 1.1, que buscava investigar se é possível melhorar a acurácia da classificação de lesões dermatológicas por meio de uma estratégia de fusão estocástica em nível de características entre uma ResNet-50 e um Vision Transformer, em comparação ao uso isolado de cada arquitetura, os resultados obtidos permitem responder afirmativamente, ainda que com ressalvas. A análise por classe (Subseção 6.5.1) revelou que o ganho de desempenho agregado não se distribui uniformemente entre as categorias de lesão. Em termos de F1-Score, a SWF-Net superou a

ResNet-50 isolada em apenas três das sete classes avaliadas, sendo ultrapassada nas quatro restantes, nevo melanocítico, ceratose benigna, carcinoma basocelular e ceratose actínica, justamente aquelas com menor número de amostras no conjunto de teste. Nessas classes, o desempenho inferior da SWF-Net se manifestou principalmente por meio de um maior número de falsos negativos em relação à ResNet-50, ou seja, uma queda no recall, o que é particularmente preocupante do ponto de vista clínico, já que casos de lesões malignas ou pré-malignas não identificados representam o tipo de erro de maior custo em um cenário de triagem diagnóstica. Esse achado indica que a estratégia de fusão proposta, embora benéfica de forma geral e superior nas classes com maior representatividade no conjunto de dados, como melanoma e lesões vasculares, pode introduzir interferência entre os ramos ResNet e ViT justamente nas classes minoritárias, o que representa uma limitação a ser observada com cautela antes de uma eventual aplicação prática da arquitetura.

Esse padrão foi avaliado estatisticamente por meio do teste de Wilcoxon pareado por classe ($n = 7$), aplicado tanto ao Recall quanto ao F1-Score obtidos pela SWF-Net e pela ResNet-50, Tabela 11. Em ambos os casos, o teste não indicou diferença estatisticamente significativa entre os dois modelos ($p = 0{,}4688$ para o Recall e $p = 0{,}5781$ para o F1-Score), o que equivale ao padrão misto observado na análise por classe, em que a SWF-Net supera a ResNet-50 em três das sete classes e é superada nas quatro restantes. Esse resultado indica que, embora a SWF-Net apresente uma vantagem consistente na acurácia agregada, essa vantagem não se sustenta como estatisticamente robusta quando o desempenho é avaliado classe a classe, o que deve ser interpretado, contudo, à luz do tamanho reduzido da amostra ($n = 7$) considerado nesse teste, que limita seu poder estatístico.

Diante disso, cabe qualificar em qual direção a SWF-Net se qualifica em superioridade em relação à ResNet-50. Do ponto de vista da acurácia agregada e da AUC macro, a SWF-Net apresenta desempenho consistentemente superior, e essa vantagem se mantém estável entre diferentes sementes de inicialização, conforme discutido na análise de variabilidade na Seção **??**). Entretanto, nessa superioridade dos modelos, existe um comportamento distinto entre os dois tipos de erro de classificação. Os falsos positivos, ou seja, os casos em que uma lesão é incorretamente posta a uma classe, tendem a ter impacto clínico menor, resultando possivelmente em uma investigação adicional desnecessária. Já os falsos negativos, em que uma lesão não é reconhecida como pertencente à sua classe verdadeira, representam o erro de maior perigo em um cenário diagnóstico, por poderem atrasar o início do tratamento. É nesse erro que a SWf-Net supera a ResNet-50, apresentando uma redução de falsos negativos nas classes melanoma, dermatofibroma e na classe de lesões vasculares, classe essa de maior quantidade presente no conjunto de dados, com uma margem considerável sobre as demais. De acordo com essa análise, o uso da SWF-Net em cenários de triagem clínicas, se mostra indicado para essas três classes, das quais o modelo demonstrou maior confiabilidade na redução de faslsos negativos em relação a ResNet-50, reforçando seu uso como uma ferramente de apoio na priorização de

casos suspeitos antes de uma avaliação especializada.

7.1 Limitações do trabalho

Algumas limitações devem ser consideradas na interpretação dos resultados apresentados. Em primeiro lugar, os experimentos definitivos foram conduzidos sobre o conjunto DermaMNIST, cujo volume reduzido de amostras em determinadas classes, como carcinoma basocelular e dermatofibroma, pode restringir a capacidade de generalização dos modelos avaliados, mesmo após a aplicação da técnica de oversampling. Essa limitação de dados pode estar relacionada ao padrão observado na análise das classes, pois em classes minoritárias a SWF-Net apresenta um desempenho inferior ao da ResNet-50, sugerindo que o mecanismo de fusão, ao ser mesclado com um ViT de desempenho inferior, se torna mais sensível à escassez de exemplos de treinamento em relação a modelos como a ResNet-50 isolada.

Em segundo lugar, conforme discutido anteriormente, o Vision Transformer utilizado neste trabalho foi treinado com pesos iniciados aleatoriamente, sem pesos pré-treinados, sobre um volume de dados consideravelmente menor do que o tipicamente empregado no treinamento de arquiteturas baseadas em atenção. Essa condição pode explicar, ao menos parcialmente, a menor contribuição desse ramo observada nos experimentos de ablação, não sendo possível, a partir dos experimentos conduzidos, isolar completamente se essa limitação decorre da arquitetura em si ou das condições específicas de treinamento às quais ela foi submetida.

Em terceiro lugar, restrições de infraestrutura computacional, sobretudo durante a etapa piloto conduzida no Google Colab, limitaram o número de épocas de alguns treinamentos e impediram, em certos casos, a convergência completa antes da ativação do mecanismo de *EarlyStopping*.

Por fim, diferentemente de arquiteturas de fusão convencionais, a camada SWF mantém seu comportamento estocástico tanto durante o treinamento quanto durante a fase de inferência. Assim, o coeficiente de fusão λ não permanece fixo durante os testes, sendo novamente amostrado a cada execução. Como consequência, diferentes inferências sobre o mesmo conjunto de teste podem produzir pequenas variações nas métricas de desempenho. Embora essa característica faça parte da proposta da arquitetura, sua influência sobre a estabilidade dos resultados não foi investigada neste trabalho por meio de múltiplas execuções e testes estatísticos de significância.

7.2 Trabalhos futuros

Como direções para trabalhos futuros, sugere-se: (i) a avaliação da arquitetura proposta utilizando um Vision Transformer pré-treinado, de modo a verificar se um ramo ViT mais competitivo altera o equilíbrio de contribuições observado na camada SWF; (ii) a investigação de mecanismos de fusão que combinem a estratégia estocástica proposta com pesos determinísticos aprendidos por classe, de modo a mitigar a perda de desempenho observada nas classes minoritárias, como carcinoma basocelular e nevo melanocítico; (iii) investigar estratégias para controlar ou fixar o coeficiente de fusão durante a inferência, comparando seu impacto na estabilidade e no desempenho da arquitetura em relação ao comportamento estocástico atualmente adotado; e (iv) a avaliação da arquitetura proposta em conjuntos de dados dermatológicos de maior escala e resolução para além do formato padronizado do MedMNIST e (v) a avaliação do tempo de inferência da arquitetura proposta, de modo a complementar a análise de custo computacional restrita ao treinamento apresentada neste trabalho e verificar sua viabilidade em cenários de uso em tempo real; e (vi) realização de repetições igualitárias para a Resnet-50 e a SWF-Net para a realização de uma comparação adequada em relação ao teste de Wilcoxon.

Referências

Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, 30:105474, 2020. doi: 10.1016/j.dib.2020.105474. Citado na página 25.

Titus J. Brinker, Achim Hekler, Jochen S. Utikal, Niels Grabe, Dirk Schadendorf, Joachim Klode, Carola Berking, Theresa Steeb, Alexander H. Enk, and Christof Von Kalle. Skin cancer classification using convolutional neural networks: systematic review. *Journal of Medical Internet Research*, 20(10):e11936, 2018. Citado na página 16.

Wilton O. Bussab and Pédro A. Morettin. *Estatística Básica*. Saraiva, São Paulo, 9 edition, 2017. Citado na página 36.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael A. Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. Citado na página 16.

Girma H. Dagnaw, Mustapha El Mouhtadi, and Mohammed Mustapha. Skin cancer classification using vision transformers and explainable artificial intelligence. *Journal of Medical Artificial Intelligence*, 7:14, 2024. doi: 10.21037/jmai-24-6. Citado na página 16.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006. Citado na página 38.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. Citado na página 19.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. Citado 2 vezes nas páginas 13 e 20.

Yunusa Haruna, Shiyin Qin, Abdulrahman Hamman Adama Chukkol, Abdulganiyu Abdu Yusuf, Isah Bello, and Adamu Lawan. Exploring the synergies of hybrid convolutional

neural network and vision transformer architectures for computer vision: A survey. *Engineering Applications of Artificial Intelligence*, 144:110057, 2025. doi: 10.1016/j.engappai.2025.110057. Citado 3 vezes nas páginas 13, 14 e 17.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90. Citado 2 vezes nas páginas 13 e 19.

T. Hussain, H. Shouno, M. A. Mohammed, H. A. Marhoon, and T. Alam. Effresnet-vit: A fusion-based convolutional and vision transformer model for explainable medical image classification. *IEEE Access*, 13:54040–54068, 2025. doi: 10.1109/ACCESS.2025.3554184. Citado 2 vezes nas páginas 17 e 19.

Instituto Nacional de Câncer (INCA). Inca estima 781 mil novos casos de câncer por ano no brasil entre 2026 e 2028. https://www.gov.br/inca/pt-br/assuntos/noticias/2026/inca-estima-781-mil-novos-casos-de-cancer-por-ano-no-brasil-entre-2026-e-2028, 2026. Acesso em: 28 ago. 2026. Citado na página 13.

Ganesh Jayaraman, S. Meganathan, S. Sheik Mohideen Shah, M. Anuradha, Ranjeeth Kumar Sundararajan, and R. Rajakumar. A hybrid cnn–vit framework with cross-attention fusion and data augmentation for robust brain tumor classification. *Scientific Reports*, 15:45769, 2025. doi: 10.1038/s41598-025-28636-9. Citado na página 17.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015. Citado na página 28.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5:180161, 2018. doi: 10.1038/sdata.2018.161. Citado 3 vezes nas páginas 13, 16 e 26.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. Citado na página 20.

A. C. Yadav, K. Shah, A. Purohit, and M. H. Kolekar. Computer-aided diagnosis for multi-class classification of brain tumors using CNN features via transfer-learning. *Multimedia Tools and Applications*, 84(31):38959–38982, 2025. doi: 10.1007/s11042-025-20751-z. Citado na página 16.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2 – a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023. doi: 10.1038/s41597-022-01721-8. Citado 3 vezes nas páginas 14, 17 e 25.