A Hybrid Approach Combining CNN and Ensemble Algorithms for Dermoscopic Image Classification

Pedro A. A. Soares[1], **Leandro A. Ensina**[2], **Juliano H. Foleis**[2]

[1]Centro Universitário FAG
Cascavel – PR – Brazil

[2]Departamento Acadêmico de Computação
Universidade Tecnológica Federal do Paraná (UTFPR)
Campo Mourão – PR – Brazil

paasoares@minha.fag.edu.br,{leandroa,julianofoleiss}@utfpr.edu.br

Abstract. *This work presents a hybrid approach for the classification of pigmented skin lesions, including skin cancer types, in dermoscopic images. The technique combines convolutional neural networks (CNNs) as feature extractors and ensemble algorithms as classifiers, along with the introduction of two distinct preprocessing stages for data augmentation. The first stage occurs before training the CNNs, while the second is applied prior to training the classifiers. Experiments conducted with the HAM10000 dataset demonstrate the method's effectiveness, achieving overall F1-Scores above 80%. Additionally, the study suggests directions for future work aimed at improving the method.*

Resumo. *Este trabalho apresenta uma abordagem híbrida para a classificação de lesões cutâneas pigmentadas, incluindo tipos de câncer de pele, em imagens dermatoscópicas. A técnica combina redes neurais convolucionais (CNNs) como extratoras de características e algoritmos ensemble como classificadores, além da introdução de duas etapas distintas de pré-processamento para aumento de dados. A primeira etapa ocorre antes do treinamento das CNNs, enquanto a segunda é aplicada antes do treinamento dos classificadores. Os experimentos realizados com a base HAM10000 evidenciam a eficácia do método, alcançando F1-Scores gerais superiores a 80%. Além disso, o estudo aponta direções para trabalhos futuros, visando o aprimoramento do método.*

1. Introdução

A dermatoscopia é uma técnica diagnóstica não invasiva amplamente utilizada em clínicas dermatológicas, que melhora a identificação de lesões cutâneas pigmentadas benignas e malignas em comparação com o exame a olho nu [Tschandl et al. 2018]. Além disso, imagens dermatoscópicas servem como uma fonte valiosa para o treinamento de algoritmos de aprendizado de máquina para a classificação dessas lesões, auxiliando especialistas na tomada de decisão [Brancaccio et al. 2024].

Neste contexto, técnicas de aprendizado profundo, particularmente Redes Neurais Convolucionais (*Convolutional Neural Networks* – CNNs), têm-se destacado na classificação de imagens dermatoscópicas. Sua capacidade de realizar a extração automática de atributos dos dados elimina a necessidade de engenharia manual de características, um processo que requer conhecimento especializado do domínio para

a identificação de características relevantes. Com isso, as CNNs podem aprender representações complexas diretamente dos dados brutos, o que é particularmente valioso para o reconhecimento de padrões em imagens médicas.

Outro motivo para o sucesso do uso das CNNs na classificação de imagens deve-se, em parte, à capacidade de transferir conhecimento de redes neurais treinadas em grandes bases de dados para aprimorar classificadores em bases menores [Menegola et al. 2017]. Neste contexto, a estratégia denominada *fine-tuning* ajusta redes pré-treinadas em conjuntos genéricos de imagens, resultando em melhores desempenhos para aplicações específicas e reduzindo limitações associadas a conjuntos de dados pequenos, que podem levar ao sobreajuste (*overfitting*). Ao reutilizar modelos treinados em grandes bases de dados, como a ImageNet, e adaptá-los para bases específicas, a transferência de aprendizado pode resultar em melhores desempenhos.

Diversos trabalhos são encontrados no estado da arte para a classificação de imagens dermatoscópicas, empregando tanto algoritmos de aprendizado profundo quanto algoritmos de aprendizado de máquina clássicos. Contudo, ainda existem lacunas a serem exploradas. Alguns trabalhos avaliam suas abordagens em bases de dados que tratam o problema de modo binário (lesão maligna e não maligna) [Benyahia et al. 2022, Chang et al. 2022, Spolaôr et al. 2024], apesar da classificação de imagens dermatoscópicas ser multiclasse. Na prática clínica, o desafio vai além de simplesmente distinguir lesões malignas de benignas. É crucial estabelecer diagnósticos específicos, pois diferentes lesões malignas, como o melanoma e o carcinoma basocelular, exigem abordagens de tratamento e prazos distintos [Tschandl et al. 2018]. Outros trabalhos avaliam seus métodos para um problema multiclasse, mas seus desempenhos demonstram margem para melhorias, especialmente devido ao desbalanceamento intrínseco do problema.

Em resposta, o método proposto apresenta uma abordagem híbrida que combina redes neurais convolucionais (CNNs) e algoritmos de *ensemble* para a classificação de imagens. As CNNs são utilizadas como extratoras de atributos, enquanto os algoritmos *ensemble* atuam como classificadores. As redes avaliadas foram pré-treinadas na ImageNet, com ajustes realizados por meio de *fine-tuning* utilizando o conjunto de treinamento. Adicionalmente, o método incorpora duas etapas de pré-processamento para o aumento de dados (*data augmentation*): a primeira, aplicada antes do treinamento das CNNs, e a segunda, antes do treinamento dos classificadores. As arquiteturas de CNNs utilizadas incluem VGG19, Inception, Xception e ResNet, enquanto os algoritmos *ensemble* avaliados incluem Random Forest, Extra Trees, XGBoost e Adaboost. Os resultados alcançados demonstram boa efetividade do método, com desempenhos competitivos com o estado da arte e perspectivas para melhorias do método para a sequência deste trabalho.

As principais contribuições do trabalho são: (1) a introdução de um pipeline de pré-processamento de aumento de dados em duas etapas; (2) a análise comparativa de diferentes combinações entre arquiteturas de CNNs e algoritmos *ensemble*; e (3) o fornecimento de direcionamentos para pesquisas futuras com base nos resultados obtidos.

2. Trabalhos Relacionados

Diversas pesquisas têm explorado o potencial das CNNs na classificação multiclasse de imagens dermatoscópicas, especialmente utilizando a base de dados HAM10000. [Yanchatuña et al. 2021] propuseram um método que envolveu segmentação prévia das

lesões (remoção de fundos e pelos), aumento de dados e normalização das imagens antes da extração de características por múltiplas CNNs pré-treinadas. Posteriormente, utilizaram um classificador SVM para realizar a classificação final das lesões na base HAM10000, atingindo 0,903 de acurácia com o melhor modelo AlexNet+SVM.

[Shetty et al. 2022] utilizaram a base HAM10000 — selecionando arbitrariamente apenas 100 imagens por classe, totalizando 700 imagens — e aplicaram aumento de dados por espelhamento horizontal antes da separação entre os conjuntos de treino e teste. Essa prática pode introduzir *data leakage*, uma vez que versões artificiais da mesma imagem podem estar presentes em ambos os conjuntos. As características foram extraídas por *Global Feature Descriptors* baseadas em três aspectos: cor (histograma de cor), forma (momentos de Hu) e textura (textura de Haralick), sendo classificadas por algoritmos tradicionais como Support Vector Machine (SVM) e Random Forest. O pipeline foi avaliado com validação cruzada e, embora o modelo Random Forest tenha alcançado um F1-Score de 0,94, tal desempenho é potencialmente superestimado devido ao *data leakage*.

[Afza et al. 2022] apresentaram um pipeline incluindo normalização das imagens, aumento de dados com rotações e translações, e aplicação de filtros de suavização para remoção de ruídos. As CNNs pré-treinadas extraíram características que foram selecionadas por técnicas de seleção de atributos baseadas em relevância. O classificador final utilizado foi uma Extreme Learning Machine (ELM), alcançando uma acurácia de 0,934 na base HAM10000.

[Chang et al. 2022] apresentaram um pipeline usando CNNs (InceptionResNetV2 e MELA-CNN), técnica de *oversampling* (K-Means SMOTE) e classificação com algoritmos clássicos e *ensemble*. Utilizam ambas as bases ISIC2018 e ISIC2019 para treinamento. Avaliado por validação cruzada estratificada, o método alcançou em seu melhor modelo (XGBoost) 0,86 de F1-Score. No entanto, o intuito do trabalho é avaliar a classificação binária (benigno ou melanoma).

[Ajiboye 2024] utilizaram CNNs pré-treinadas para extração de atributos e classificadores *ensemble* baseados em XGBoost na base HAM10000. O pré-processamento incluiu remoção de pelos, redimensionamento, suavização e balanceamento das classes. Os autores reportaram F1-Score de 0,857.

[Sá et al. 2024] propuseram um modelo que utiliza a rede VGG19 como extratora de atributos, sem *fine-tuning*, seguida por classificadores tradicionais como Perceptron, SVM e Regressão Logística. Aplicando validação cruzada 5×2 na base HAM10000, os autores observaram que os classificadores clássicos obtiveram desempenho superior ao uso da VGG19 como classificador direto, com F1-Score máximo de 0,72. Embora os resultados não superem os das abordagens mais avançadas da literatura, eles reforçam a viabilidade de arquiteturas mais simples, com menor custo computacional, especialmente em contextos com recursos limitados.

Apesar dos avanços apresentados, os trabalhos da literatura ainda enfrentam desafios recorrentes, em especial devido ao desbalanceamento inerente do problema de classificação de imagens dermatoscópicas [Tschandl et al. 2018]. Além disso, o estado da arte demonstra oportunidades no desenvolvimento de estratégias que possibilitem melhorar o desempenho de novas abordagens, como na realização de etapas de pré-processamento e a exploração das potencialidades das CNNs como extratoras de carac-

terísticas e algoritmos clássicos como classificadores.

3. Materiais e Métodos

Nesta seção estão descritos os detalhes do método proposto. Na Seção 3.1 está apresentada a base de dados HAM10000. O método proposto é apresentado na Seção 3.2, enquanto o delineamento experimental é abordado na Seção 3.3.

3.1. Base de Dados HAM10000

A base de dados HAM10000 é um conjunto de 10.015 imagens dermatoscópicas de lesões pigmentadas de diferentes populações [Tschandl et al. 2018], contendo uma coleção representativa de todas as categorias diagnósticas importantes no campo das lesões pigmentadas. Essa base foi criada para superar as limitações dos conjuntos de dados disponíveis, que geralmente focavam exclusivamente em lesões melanocíticas (i.e., diferenciação entre melanoma e nevo) e desconsideravam lesões pigmentadas não melanocíticas, apesar de sua alta prevalência na prática clínica.

As classes presentes na base e suas respectivas quantidades de exemplos são:

1. **akiec** (ceratoses actínicas e carcinoma intraepitelial): 327 imagens;
2. **bcc** (carcinoma basocelular): 514 imagens;
3. **bkl** (ceratose benigna): 1.099 imagens;
4. **df** (dermatofibroma): 115 imagens;
5. **mel** (melanoma): 1.113 imagens;
6. **nv** (nevos melanocíticos): 6.705 imagens;
7. **vasc** (lesões vasculares): 142 imagens.

3.2. Método Proposto

O método proposto está organizado em três fases principais: (1) pré-processamento, (2) treinamento das CNNs e (3) treinamento dos classificadores. A Figura 1 demonstra o pipeline do método.

Na primeira fase, o conjunto original de treinamento (CT) passa por dois processos distintos de *data augmentation*. No primeiro processo, as imagens são processadas dinamicamente, alterando as imagens a cada época, com técnicas probabilísticas que incluem giros horizontais (50%), giros verticais (20%), rotações de 90° (30%), ajustes de contraste e claridade limitados a 0,15 (50%), desfoque gaussiano (30%), modificação de matiz e saturação (20%) e CLAHE (30%), formando o primeiro conjunto (CT1). No segundo processo, oito variações por imagem são geradas estaticamente durante o treinamento com rotações (±15° e ±30°), giros horizontais e verticais, ajustes discretos de brilho e contraste e zooms controlados (1,05× e 0,95×), formando o segundo conjunto (CT2). Para isso, utilizamos a biblioteca Albumentations [Buslaev et al. 2020].

A segunda fase consistiu no *fine-tuning* das CNNs pré-treinadas na ImageNet, para o qual o CT1 foi utilizado como conjunto de treinamento. Em seguida, o CT2 foi empregado para que essas CNNs pudessem extrair suas características. Na Seção 3.3 estão descritos maiores detalhes deste processo.

Já na terceira e última fase, o vetor de atributos resultante da fase anterior foi usado para o treinamento dos algoritmos *ensemble*, usados em nosso método como classificadores.

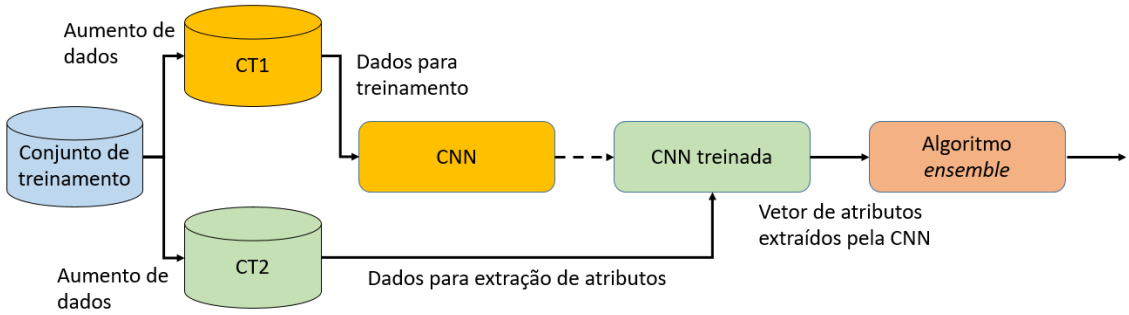

Figura 1. Diagrama do método proposto.

3.3. Protocolo Experimental

Esta seção detalha os procedimentos adotados para a condução de nossos experimentos destinados a avaliar o desempenho e a eficácia do método. Em nossas avaliações, empregamos as seguintes arquiteturas de CNNs para a extração de atributos: VGG19; Inception; ResNet; Xception.

Durante o treinamento das CNNs, aplicamos funções de *callback*, como parada antecipada e redução de taxa de aprendizado em platô (reduzindo automaticamente a taxa de aprendizado quando o desempenho não melhora por múltiplas épocas), garantindo eficiência e estabilidade no treinamento. Além disso, também aplicamos *fine-tuning*, que consiste em descongelar apenas parte das camadas convolucionais superiores de cada rede, o que corresponde aproximadamente a 25% do total de camadas de cada rede. Os hiperparâmetros usados para cada rede estão apresentados na Tabela 1.

Tabela 1. Valores dos hiperparâmetros utilizados na configuração das CNNs.

Hiperparâmetro	**Valor**
Taxa de aprendizado	0,0001
Tamanho do lote	16
Número de épocas	100
Otimizador	Adam
Função de ativação	ReLU
Função de perda	Cross-Entropy
Taxa de *dropout*	0,5

Para os algoritmos *ensemble* utilizados, foi executada uma busca em grade prévia para otimizar o número ideal de estimadores, resultando nas seguintes configurações: Random Forest com 200 estimadores; Extra Trees com 200 estimadores; XGBoost com 200 estimadores; AdaBoost com 150 estimadores.

Cada combinação de CNN e algoritmo *ensemble* foi analisada individualmente para avaliar suas eficiências para classificar o conjunto de dados dermatoscópico HAM10000. Além disso, de modo a estimar a eficácia da solução, as mesmas CNNs utilizadas para extrair atributos também foram utilizadas diretamente como classificadores (*end-to-end*). Dessa maneira, podemos mensurar os desempenhos individuais das CNNs frente aos seus usos como extratoras de características.

Inicialmente, a base HAM10000 foi randomicamente dividida em dois conjuntos: treinamento, contendo 85% dos dados; e teste, com 15% dos dados. Essa divisão foi realizada respeitando a proporção de exemplos para cada classe (i.e., estratificada). Em seguida, utilizamos a técnica de amostragem de validação cruzada estratificada com 5 grupos e 2 repetições (5×2 *stratified cross-validation*) sobre o conjunto de treinamento para que pudéssemos determinar as melhores configurações dos modelos avaliados.

Após a validação cruzada, cada modelo foi treinado 10 vezes com o conjunto completo de treinamento. Esses 10 modelos finais foram, então, individualmente avaliados sobre o conjunto de teste. Portanto, cada combinação de arquitetura e configuração produziu 10 resultados distintos no conjunto de teste. Os experimentos foram analisados pelas métricas *recall*, precisão e F1-Score.

4. Resultados e Discussão

Esta seção tem por objetivo expor, analisar e discutir os resultados obtidos para o método proposto. Devido ao conhecido desbalanceamento das classes no conjunto de dados HAM10000, o F1-Score macro foi adotado como principal métrica de análise de desempenho, por refletir de forma mais adequada o compromisso entre precisão e *recall* em cenários com classes desbalanceadas.

4.1. Desempenho das CNNs em Configuração *end-to-end*

Na primeira etapa dos experimentos, as CNNs foram avaliadas de maneira independente em uma configuração *end-to-end*, isto é, atuando simultaneamente como extratoras de características e classificadoras. A Figura 2 apresenta os resultados obtidos para cada métrica por arquitetura testada para o conjunto de teste.

Analisando a Figura 2, destacam-se particularmente as redes Xception e ResNet, com valores de F1-Score de 0,838 $\pm$ 0,014 e 0,835 $\pm$ 0,01, respectivamente, indicando boa capacidade de generalização e efetividade na captura dos padrões relevantes para o diagnóstico. Em contraste, a Inception obteve F1-Score inferior (0,71 $\pm$ 0,016), sugerindo limitações na sua capacidade de modelar os padrões nas imagens dermatoscópicas. Testes estatísticos realizados pelo Teste de Wilcoxon com intervalo de confiança de 95%, onde comparamos cada par de modelos individualmente, corroboraram o melhor desempenho das redes ResNet e Xception perante as demais para a métrica F1-Score. No entanto, não há diferença estatisticamente significativa entre ResNet e Xception ($p - valor = 0,557$).

A análise detalhada desses resultados revela alguns pontos de interesse. A melhor performance da Xception pode ser atribuída à sua estrutura baseada em separação de convoluções (*depthwise separable convolutions*), que proporciona maior eficiência na captura de relações espaciais complexas com menos parâmetros [Chollet 2017]. A ResNet também apresentou desempenho consistente, combinando um F1-Score elevado com uma das maiores precisões entre as redes avaliadas, o que indica uma menor taxa de falsos positivos. A VGG19, embora tenha obtido um F1-Score ligeiramente inferior (0,796 $\pm$ 0,013), ainda demonstrou competitividade, beneficiando-se do *fine-tuning* em camadas superiores e do uso de *data augmentation* durante o treinamento. Já a Inception apresentou desempenho visivelmente inferior, o que pode estar relacionado a limitações estruturais na captura de padrões mais sutis presentes nas imagens dermatoscópicas.

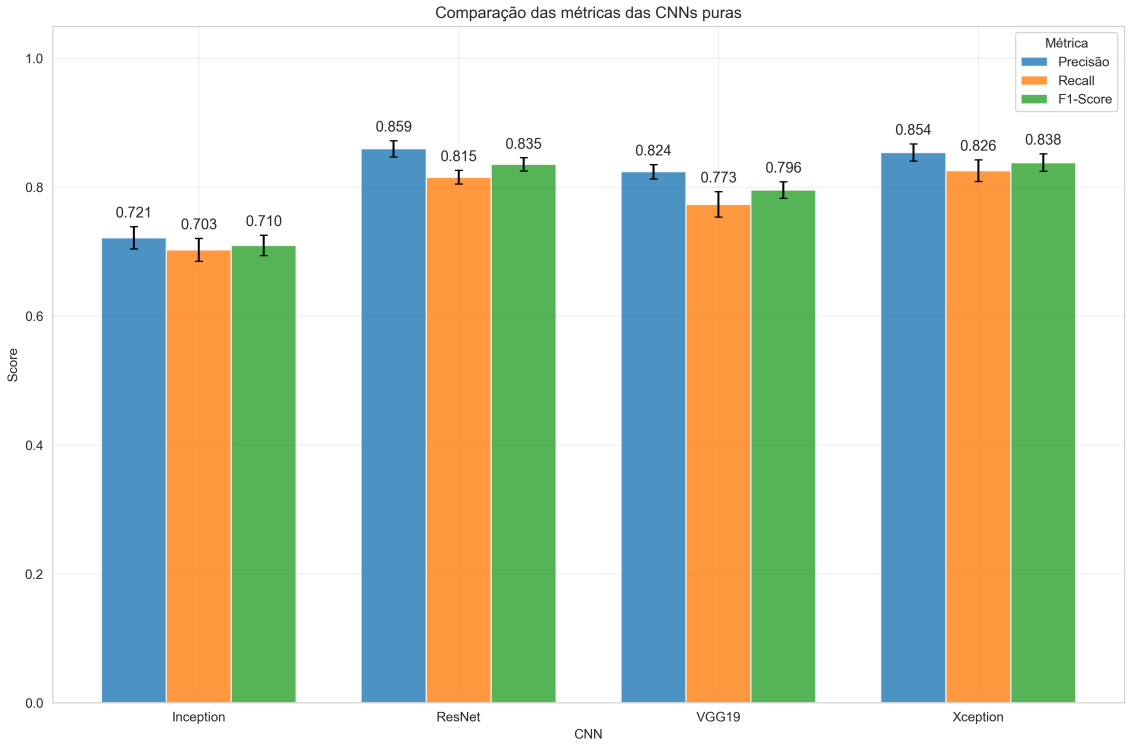

Figura 2. Comparação do desempenho entre as CNNs *end-to-end*.

4.2. Modelos Híbridos: CNN como Extratora + Classificadores *Ensemble*

Esta subseção apresenta os resultados obtidos pelos modelos híbridos, onde as CNNs atuam apenas como extratoras de atributos. A Figura 3 ilustra os resultados comparativos entre as abordagens híbridas e as CNNs puras (*end-to-end*). Os resultados revelam um desempenho marcadamente inferior do AdaBoost frente aos demais modelos para todas as CNNs usadas como extratoras de características, principalmente para sua combinação com a VGG19, em que o F1-Score despenca para 0,32. Em contrapartida, RandomForest, XGBoost e ExtraTrees mostraram-se significativamente mais robustos, com desempenhos muito próximos aos das redes puras e, em eventuais casos, até melhores (como ocorreu para a VGG19, onde estes algoritmos *ensemble* foram estatisticamente superiores, conforme constatado pelo teste de Friedman com intervalo de confiança de 95% seguido pelo pós-teste de Dunn), além de serem mais constantes ao apresentar desvios-padrão menores.

Comparando com o trabalho de [Sá et al. 2024], que utilizou uma abordagem semelhante (CNNs como extratoras e classificadores clássicos), os ganhos são notáveis. Por exemplo, a combinação VGG19 + ExtraTrees no trabalho citado resultou em um F1-Score médio de 0,61, enquanto a mesma combinação híbrida em nosso trabalho alcançou 0,82 — um ganho absoluto de 21 pontos percentuais. Já em questão de melhores modelos híbridos, o trabalho de [Sá et al. 2024] conseguiu, com a combinação VGG19 + SVM, um F1-Score de 0,72 — 11 pontos percentuais absolutos menores que a mesma métrica de nosso melhor modelo. Importante destacar que ambos os trabalhos — este e o de [Sá et al. 2024] — adotaram o mesmo protocolo de validação cruzada estratificada 5×2, conforme recomendado pela literatura para bases desbalanceadas [Roy et al. 2018]. Assim, os ganhos observados podem ser atribuídos às melhorias específicas no pipeline de

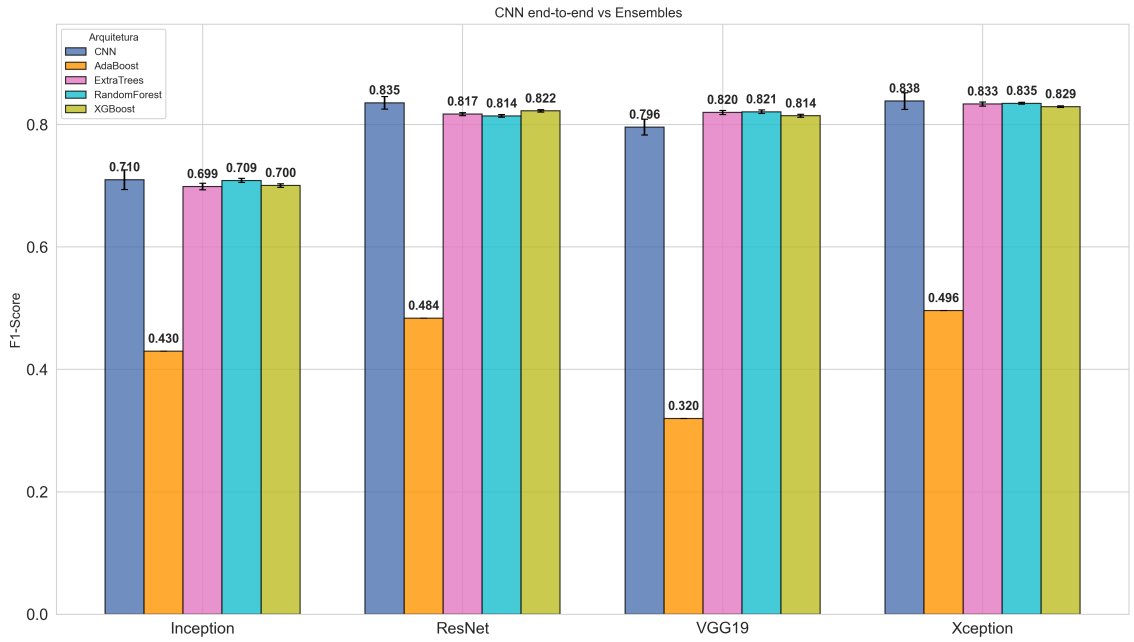

Figura 3. Comparação de F1-score entre todos os classificadores por rede.

pré-processamento e treinamento adotados neste estudo.

O bom desempenho da abordagem proposta decorre de três fatores principais: (i) a liberação parcial das camadas superiores para *fine-tuning*, permitindo especialização nas particularidades dos padrões dermatoscópicos; (ii) o aumento do espaço amostral efetivo por meio do *data augmentation*, reduzindo *overfitting*; e (iii) o uso do segundo processo de pré-processamento, que contribuiu com um incremento médio de 2 pontos percentuais no F1-Score de cada modelo *ensemble*. Comparando o uso e o não uso deste segundo processo pelo teste de Wilcoxon com intervalo de confiança de 95%, nota-se uma melhora estatística geral no desempenho com o uso desta segunda etapa de aumento de dados.

Embora nossos resultados não superem os valores de F1-Score reportados por outros estudos, como os de [Shetty et al. 2022] (F1-Score de 0,94), é fundamental ressaltar que os resultados alcançados por estes autores derivam de um pipeline metodologicamente comprometido. No trabalho de [Shetty et al. 2022], técnicas de aumento de dados foram aplicadas antes da divisão entre treino e teste, o que compromete a independência entre os conjuntos e introduz *data leakage*, superestimando as métricas de desempenho. Além disso, os autores realizaram uma seleção prévia de apenas algumas imagens de toda a base de dados para a realização do aumento de dados, ignorando as demais imagens (até mesmo para o teste). Logo, os critérios adotados para esta separação são questionáveis, sem garantias de que houve uma divisão justa. Em contraste, os resultados apresentados para o nosso método foram obtidos sob condições metodológicas estritas e reprodutíveis, respeitando a separação adequada dos conjuntos.

Outros trabalhos relacionados como os de [Afza et al. 2022] e [Yanchatuña et al. 2021] reportam resultados de acurácia elevados — até 0,934 com ELM e 0,903 com AlexNet + SVM, respectivamente. No entanto, esses estudos utilizam a acurácia como métrica de desempenho, a qual não é muito apropriada para

bases de dados desbalanceadas como a HAM10000, também usada pelos autores.

O estudo de [Chang et al. 2022], por fim, aborda uma tarefa de classificação binária (melanoma *vs* lesão benigna). Apesar de relatar valores elevados de F1-Score (0,86 — próximo ao de nossos resultados), é importante destacar que tarefas binárias — especialmente em contextos controlados — não refletem a complexidade de cenários clínicos reais, onde múltiplos tipos de lesões coexistem com forte desbalanceamento. Assim, tais resultados, embora tecnicamente válidos, não são diretamente comparáveis aos aqui apresentados, que enfrentam o problema em sua forma mais desafiadora: com sete classes e distribuição assimétrica.

Um aspecto crítico que reforça essa interpretação pode ser observado na Figura 4, que mostra a variação do F1-Score por classe para diferentes combinações de modelo. Nota-se que a classe "nv" (nevos melanocíticos), a classe com o maior número de exemplos na base de dados, obteve F1-Score consistentemente alto, independentemente da combinação utilizada. Em contrapartida, classes com poucos exemplos, como "df" (dermatofibroma), "vasc" (lesões vasculares) e "akiec" (ceratoses actínicas), frequentemente obtiveram F1-Scores nulos ou extremamente baixos para determinados cenários.

Uma análise mais minuciosa da Figura 4 revela que o aumento de atributos teve impacto positivo em diversas classes, especialmente nas minoritárias. Por exemplo, na combinação Xception + XGBoost, o F1-Score da classe "vasc" subiu de 0,874 para 0,952 com o uso da segunda etapa de pré-processamento (AF) frente ao uso apenas da primeira etapa (N). Já na combinação VGG19 + RandomForest, o F1-Score da classe "df" aumentou de 0,814 para 0,864 com a segunda etapa de pré-processamento. Embora ainda existam casos com F1-Score próximo de zero (como a classe "df" em diversas combinações com AdaBoost), os resultados mostram que a técnica foi eficaz para mitigar, ao menos parcialmente, os efeitos do desbalanceamento em classes menos representadas. Essa tendência reforça o papel do aumento de atributos como uma ferramenta complementar importante, capaz de ampliar a sensibilidade dos modelos a padrões menos frequentes. No entanto, o impacto estrutural do desbalanceamento da base HAM10000 permanece evidente e demanda abordagens adicionais.

Esse padrão evidencia o impacto direto do desbalanceamento da base HAM10000 sobre o desempenho dos modelos. A capacidade dos classificadores em identificar padrões em classes minoritárias é importante para a viabilidade clínica de qualquer modelo. O uso de técnicas como *oversampling* sintético (e.g., SMOTE ou K-means SMOTE) [Chang et al. 2022] deve ser considerado em futuros trabalhos para mitigar esse problema estrutural. Nesse contexto, propõe-se como direções futuras a aplicação de um pipeline mais completo de pré-processamento dos vetores de características, incorporando normalização e seleção de atributos (como Relief, mRMR e CFS). Complementarmente, faz-se relevante explorar métodos específicos para lidar com o desbalanceamento das classes, seja no nível da imagem ou das características extraídas. Técnicas de *oversampling* para geração de dados sintéticos são alternativas viáveis.

Além disso, outras melhorias no pré-processamento das imagens podem beneficiar significativamente o desempenho dos modelos. Estratégias como a segmentação precisa das lesões (isolando a região de interesse), a remoção automática de artefatos como pelos e a adaptação local de contraste já mostraram resultados promissores na literatura

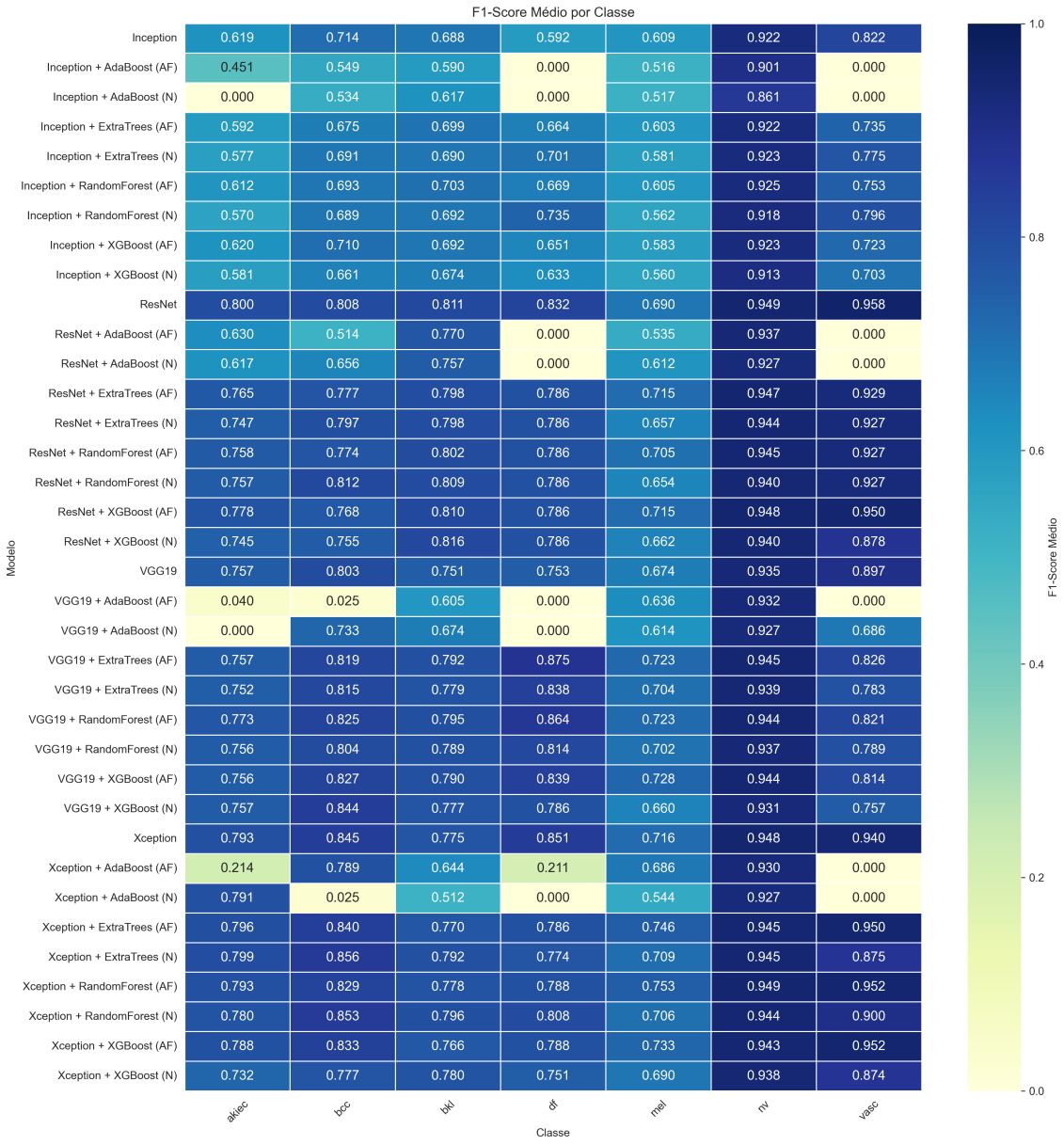

Figura 4. Mapa de calor do F1-Score por classe, indicando o impacto do desbalanceamento das classes.

[Zanddizari et al. 2021, Benjdira et al. 2024]. Tais abordagens não apenas melhoram a qualidade das imagens fornecidas às CNNs, como também aumentam a homogeneidade visual das amostras, reduzindo a complexidade na classificação.

Apesar do desempenho competitivo dos modelos híbridos, os resultados não evidenciam vantagens claras sobre as CNNs usadas isoladamente. Uma possível explicação para essa limitação é a ausência de um maior pré-processamento dedicado aos vetores de características antes da classificação. Técnicas como a seleção de atributos para a redução de dimensionalidade são recorrentes na literatura e frequentemente resultam em melhorias consideráveis [Khan et al. 2021, Alenezi et al. 2023]. Incluiremos tais procedimentos na continuidade deste projeto de modo a alcançarmos melhores desempenhos.

5. Conclusão

O método proposto apresenta uma abordagem híbrida que combina a extração de atributos por redes neurais convolucionais (CNNs) e a classificação por algoritmos de *ensemble*, e implementa um pipeline de pré-processamento em duas etapas. Os resultados alcançados demonstraram boa efetividade da nossa abordagem, com destaque para as CNNs Xception, VGG19 e ResNet e os classificadores Random Forest, XGBoost e Extra Trees, com desempenhos acima de 0,8 para a métrica F1-Score. Contudo, os resultados para a rede Inception e o classificador AdaBoost foram inferiores para todos os cenários avaliados.

Apesar dos desempenhos do uso dos classificadores *ensemble* serem semelhantes ao uso direto das CNNs, em que a rede realizava todo o processo (*end-to-end)*, acreditamos que no prosseguimento deste projeto conseguiremos alcançar resultados superiores. Trabalhos futuros incluem (1) a incorporação de novas técnicas de pré-processamento como a seleção de atributos e a normalização previamente ao treinamento dos classificadores *ensemble*, e (2) a avaliação de outras CNNs, como a EfficientNet.

Ainda, (3) pretendemos avaliar o uso da estratégia da seleção dinâmica de classificadores [Cruz et al. 2018]. Essa técnica permite selecionar, em tempo real, o modelo ou conjunto de modelos mais competentes (acurados) dentro de um *ensemble* para classificar cada novo caso de teste individualmente. Para isso, a competência de cada classificador é estimada no conjunto de treinamento ou de validação, garantindo que apenas os mais adequados sejam utilizados na predição dos padrões de teste. A escolha pela seleção dinâmica justifica-se pelo potencial de melhorar o desempenho em bases de dados desbalanceadas [Roy et al. 2018], como é o caso da base utilizada neste trabalho (HAM10000). Entre as abordagens previstas de uso estão o META-DES e o KNORA [Cruz et al. 2018]. Em suma, nosso objetivo é continuar explorando técnicas de pré-processamento para melhorar a eficiência do método, junto ao uso de seleção dinâmica.

Referências

Afza, F., Sharif, M., Khan, M., Tariq, U., Yong, H.-S., and Cha, J. (2022). Multiclass skin lesion classification using hybrid deep features selection and extreme learning machine. *Sensors*, 22(3):799.

Ajiboye, A. O. (2024). Hybrid skin lesion detection integrating cnn and xgboost for accurate diagnosis. *International Journal of Computer (IJC)*, 53(1):14–71.

Alenezi, F., Armghan, A., and Polat, K. (2023). A multi-stage melanoma recognition framework with deep residual neural network and hyperparameter optimization-based decision support in dermoscopy images. *Expert Systems with Applications*, 215:119352.

Benjdira, B., Ali, M., Koubaa, A., Ammar, A., and Boulila, W. (2024). Dm-ahr: A self-supervised conditional diffusion model for ai-generated hairless imaging for enhanced skin diagnosis applications. *Cancers (Basel)*, 16(17):2947. Published: 2024 Aug 23.

Benyahia, S., Meftah, B., and Lézoray, O. (2022). Multi-features extraction based on deep learning for skin lesion classification. *Tissue and Cell*, 74:101701.

Brancaccio, G., Balato, A., Malvehy, J., Puig, S., Argenziano, G., and Kittler, H. (2024). Artificial intelligence in skin cancer diagnosis: A reality check. *Journal of Investigative Dermatology*, 144(3):492–499.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., and Kalinin, A. A. (2020). Albumentations: Fast and flexible image augmentations. *Information*, 11(2).

Chang, C.-C., Li, Y.-Z., Wu, H.-C., and Tseng, M.-H. (2022). Melanoma detection using xgb classifier combined with feature extraction and k-means smote techniques. *Diagnostics*, 12(7):1747. Published: 19 July 2022.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807.

Cruz, R. M., Sabourin, R., and Cavalcanti, G. D. C. (2018). Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, 41:195–216.

Khan, M. A., Sharif, M., Akram, T., Damaševičius, R., and Maskeliūnas, R. (2021). Skin lesion segmentation and multiclass classification using deep learning features and improved moth flame optimization. *Diagnostics*, 11(5):811. Published: 29 April 2021.

Menegola, A., Fornaciali, M., Pires, R., Bittencourt, F. V., Avila, S., and Valle, E. (2017). Knowledge transfer for melanoma screening with deep learning. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 297–300.

Roy, A., Cruz, R. M., Sabourin, R., and Cavalcanti, G. D. (2018). A study on combining dynamic selection and data preprocessing for imbalance learning. *Neurocomputing*, 286:179–192.

Shetty, B., Fernandes, R., Rodrigues, A. P., Chengoden, R., Bhattacharya, S., and Lakshmanna, K. (2022). Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. *Scientific Reports*, 12(1):18134.

Spolaôr, N., Lee, H. D., Takaki, W. S. R., Coy, C. S. R., and Wu, F. C. (2024). Avaliação de variações da rede profunda efficientnet em bases dermoscópicas. *Journal of Health Informatics*, 16(especial).

Sá, J., Ensina, L., and Jeronymo, D. (2024). Aplicação de redes de aprendizado profundo e algoritmos de aprendizado de máquina para classificar imagens de câncer de pele. In *Anais do XXIV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 651–656, Porto Alegre, RS, Brasil. SBC.

Tschandl, P., Rosendahl, C., and Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):180161.

Yanchatuña, O. P., Pereira, J. P., Pila, K. O., Vásquez, P. A., Veintimilla, K. S., Villalba-Meneses, G. F., Alvarado-Cando, O., and Almeida-Galárraga, D. (2021). Skin lesion detection and classification using convolutional neural network for deep feature extraction and support vector machine. *International Journal on Advanced Science, Engineering and Information Technology*, 11(3):1260–1267.

Zanddizari, H., Nguyen, N., Zeinali, B., and Chang, J. M. (2021). A new preprocessing approach to improve the performance of cnn-based skin lesion classification. *Medical & Biological Engineering & Computing*, 59(5):1123–1131. Epub 2021 Apr 26.