Efficient Ensemble of CNN and Transformer Models for Cassava Leaf Disease Classification

Kaique O. A. dos Santos[1], **Raimundo Correa de Oliveira**[1]

{koads.eng21, rcoliveira}@uea.edu.br

[1]Escola Superior de Tecnologia – Universidade do Estado do Amazonas (UEA)
Manaus – AM – Brasil

Abstract. *This work proposes an ensemble approach for the multiclass classification of cassava leaf diseases, integrating two architectures with complementary characteristics in representational capacity and computational efficiency: CropNet, based on MobileNetV3, and Swin Transformer. The images used were collected in the field, reflecting real-world variations in lighting, background, and positioning. The models were evaluated through stratified $K = 5$ fold cross-validation, using the F1-score macro as the main metric, given the imbalanced class distribution. Data augmentation techniques were dynamically applied during training to improve generalization capacity. The ensemble outperformed the individual models, achieving a mean F1-score macro of 0.8329 and an accuracy of 0.9087. These results demonstrate predictive robustness and indicate potential for accurate visual diagnoses in agricultural contexts with computational constraints.*

Resumo. *Este trabalho propõe uma abordagem de ensemble para a classificação multiclasse de doenças em folhas de mandioca, integrando duas arquiteturas com características complementares em capacidade representacional e eficiência computacional: CropNet, baseada na MobileNetV3, e Swin Transformer. As imagens utilizadas foram coletadas em campo, refletindo variações reais de iluminação, fundo e posicionamento. Os modelos foram avaliados por meio de validação cruzada estratificada com $K = 5$ folds, utilizando o F1-score macro como principal métrica, dada a distribuição desbalanceada entre classes. Técnicas de aumento de dados foram aplicadas dinamicamente durante o treinamento para melhorar a capacidade de generalização. O ensemble superou os modelos individuais, alcançando F1-score macro médio de 0,8329 e acurácia de 0,9087. Esses resultados demonstram robustez preditiva e indicam potencial para diagnósticos visuais precisos em contextos agrícolas com restrições computacionais.*

1. Introdução

A mandioca (*Manihot esculenta* Crantz), planta originária da América do Sul, desempenha um papel estratégico na segurança alimentar global. Estima-se que cerca de 1 bilhão de pessoas dependam dela como uma de suas principais fontes de energia, sendo o segundo alimento energético mais consumido no mundo. Atualmente, a cultura da mandioca está presente em aproximadamente 100 países, com o Brasil ocupando a quinta posição no ranking global de produção, responsável por

5,7% do volume total [Embrapa Mandioca e Fruticultura 2023]. Em 2023, a mandioca movimentou cerca de R$ 19,18 bilhões no Brasil, segundo o Instituto Brasileiro de Geografia e Estatística (IBGE). No entanto, projeções indicam uma redução de 20,9% na área plantada até 2033/34, apesar de ganhos esperados em produtividade [Ministério da Agricultura e Pecuária (MAPA) and Embrapa 2023]. Esse cenário evidencia a necessidade de adoção de soluções tecnológicas que promovam eficiência e sustentabilidade na produção.

Nesse contexto, a agricultura digital tem se destacado como ferramenta para enfrentar os desafios do setor agrícola. Ela envolve a aplicação de tecnologias como sensores, softwares, inteligência artificial e, em particular, visão computacional, com o objetivo de aumentar a eficiência, qualidade, sustentabilidade e rentabilidade da produção agrícola. No Brasil, essas técnicas vêm sendo introduzidas com sucesso na detecção e contagem de frutos, diagnósticos de doenças, detecção precoce de pragas e manejo de cultivos. Por exemplo, [Silva et al. 2024] propuseram um sistema baseado em armadilhas inteligentes e visão computacional para detectar a praga *Spodoptera frugiperda* no cultivo de milho, demonstrando o impacto positivo dessas tecnologias no monitoramento automatizado das lavouras.

Apesar dos avanços recentes, ainda são limitados os estudos que exploram, de forma sistemática, arquiteturas modernas focadas na tarefa de classificação multiclasse de doenças em folhas de mandioca, particularmente com o uso de Vision Transformers e redes convolucionais leves desenvolvidas para aplicações agrícolas. Este trabalho contribui com a investigação do uso de técnicas de visão computacional para a categorização automática de folhas de mandioca em cinco classes: *healthy*, que representa folhas saudáveis, e quatro doenças, que são *Cassava Bacterial Blight (CBB)*, *Cassava Brown Streak Disease (CBSD)*, *Cassava Green Mite (CGM)* e *Cassava Mosaic Disease (CMD)*.

Além dos desafios típicos da classificação visual em ambientes agrícolas, a tarefa é agravada pelo desbalanceamento acentuado do conjunto de dados, com predominância da classe *Cassava Mosaic Disease (CMD)*, que representa mais de 60% das amostras. A ausência de padronização no fundo das imagens introduz ruído visual que prejudica a capacidade de generalização dos modelos. Esse cenário torna relevante a avaliação de arquiteturas que lidem com esse tipo de interferência de forma intrínseca, como os modelos baseados em atenção. A detecção precoce dessas doenças é essencial para mitigar perdas, reduzir o uso excessivo de defensivos e promover uma agricultura mais sustentável e tecnicamente assistida.

Para resolver esse problema, adotou-se uma abordagem baseada em ensemble, combinando os pontos fortes de dois modelos complementares: o *CropNet*, uma rede convolucional leve construída sobre o backbone *MobileNetV3* [Howard et al. 2019], e o *Swin Transformer* [Liu et al. 2021], uma arquitetura baseada em atenção espacial hierárquica. Essa combinação visa melhorar a acurácia e a robustez da classificação multiclasse de doenças em folhas de mandioca.

Essa solução é especialmente relevante para contextos de agricultura familiar e pequenos produtores, que geralmente enfrentam limitações no acesso a tecnologias de alto custo computacional. A abordagem proposta alcançou uma acurácia de 90,87% utilizando apenas um *ensemble* entre MobileNet e Swin Transformer, modelos reconhecidamente

leves e eficientes. Embora existam soluções mais complexas na literatura com desempenho superior, os resultados obtidos demonstram que é possível atingir níveis elevados de acurácia com arquiteturas otimizadas para cenários com restrições computacionais, como aplicações embarcadas na agricultura de precisão.

Para apresentar o que se propõe, este trabalho está organizado como segue. A Seção 2 apresenta os trabalhos relacionados; a Seção 3 descreve os materiais e métodos utilizados; a Seção 4 expõe os resultados obtidos e discute suas implicações; por fim, a Seção 5 traz as considerações finais e perspectivas para trabalhos futuros.

2. Trabalhos relacionados

A detecção precoce e precisa das doenças da mandioca é essencial para um manejo eficaz e para a mitigação dos impactos negativos sobre a produção. Métodos convencionais, como a inspeção visual e os exames laboratoriais, apresentam limitações significativas. A inspeção visual, por exemplo, pode resultar em diagnósticos imprecisos devido à semelhança entre os sintomas de diferentes doenças [Fathima and Bondili 2025]. Já os testes laboratoriais, como a cultura de tecidos e análises bioquímicas, tendem a ser caros e demorados, o que os torna inacessíveis para muitos agricultores de pequeno porte, que representam uma parcela substancial da cadeia produtiva da mandioca em diversas regiões do Brasil [Batista and de Paiva 2019].

Na literatura, diversos estudos têm explorado a aplicação de visão computacional na detecção de doenças foliares em culturas agrícolas variadas. Albuquerque e Guedes [Albuquerque and Guedes 2023], por exemplo, propuseram um modelo para classificação automática de patologias em folhas de café, avaliando diferentes arquiteturas de redes neurais. Entre os modelos testados, a ShuffleNet apresentou desempenho destacado, com *F1-score* médio de 99,88%, valor próximo ao estado da arte, porém com uma quantidade significativamente menor de parâmetros. Esse resultado evidencia uma tendência crescente na busca por soluções computacionais que conciliem desempenho e viabilidade em contextos com restrições de recursos, como aqueles frequentemente encontrados na agricultura de base familiar.

John [John 2022] propõe uma solução computacional acessível para a detecção de doenças em folhas de mandioca, com foco em sua adoção por pequenos produtores. O estudo utiliza o mesmo conjunto de dados adotado neste trabalho, composto por imagens coletadas em campo por agricultores e anotadas por especialistas do National Crops Resources Research Institute (NaCRRI). A abordagem alcançou recall (taxa de verdadeiros positivos) de 92,05%, specificity (taxa de verdadeiros negativos) de 90,98% e Área sob a Curva (AUC, do inglês Area Under the Curve) média de 96,14%. Embora a DenseNet-169 apresente bons resultados, sua aplicação isolada, sem estratégias complementares como atenção ou ensembles, pode limitar a robustez do modelo frente à variabilidade visual presente nas imagens agrícolas.

O trabalho de Costa Junior et al. [Junior et al. 2024] comparou quatro arquiteturas de redes neurais convolucionais (EfficientNet-B3, InceptionV3, ResNet50 e VGG16), aplicadas à classificação multiclasse de doenças em folhas de mandioca, utilizando o mesmo conjunto de dados empregado neste estudo. A EfficientNet-B3 apresentou os melhores resultados, com 87,7% de acurácia, 87,8% de precisão e 87,7% de revocação. Embora o estudo represente um passo importante na avaliação sistemática de CNNs para este

domínio, ele se concentra na análise comparativa entre modelos, sem explorar estratégias complementares como *ensembles* ou mecanismos de atenção. Além disso, aspectos como a variabilidade de fundo, oclusões parciais, iluminação não controlada e a análise dos padrões de erro ainda permanecem pouco discutidos. O presente trabalho busca contribuir nesse sentido, adotando uma abordagem que combina arquiteturas leves e mecanismos de atenção, com foco em robustez e viabilidade computacional para cenários operacionais restritivos.

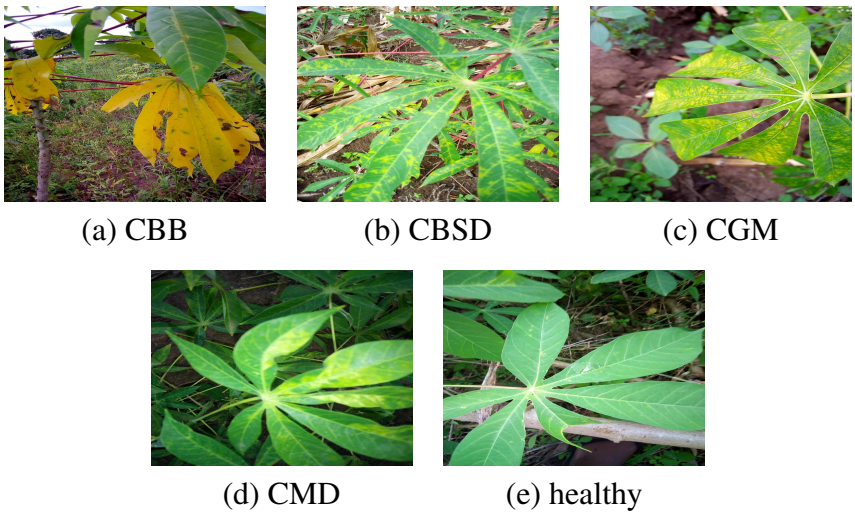
Embora os estudos revisados evidenciem avanços na aplicação de redes neurais para a classificação de doenças em folhas de mandioca, ainda há limitações quanto à robustez em cenários visuais complexos e pouco padronizados, como aqueles encontrados em ambientes agrícolas reais. Propõe-se, portanto, uma solução que integra modelos complementares, uma rede neural convolucional leve e um *vision transformer*, com foco em alta eficiência e resiliência a variações visuais em contextos operacionais restritivos, como os da agricultura de base familiar.

3. Materiais e Métodos

O problema de classificação de doenças em folhas de mandioca foi tratado como uma tarefa de classificação multiclasse supervisionada, com cinco categorias: uma classe saudável e quatro patologias.

3.1. Conjunto de dados

Os dados utilizados neste trabalho são oriundos do conjunto apresentado por Mwebaze *et al*. [Mwebaze et al. 2020], resultado da colaboração entre o Instituto Nacional de Pesquisa de Recursos de Culturas (NaCRRI) e o Laboratório de Inteligência Artificial da Universidade de Makerere, em Kampala. O dataset é composto por 21.367 imagens de folhas da mandioca, majoritariamente coletadas em campo por meio de *crowdsourcing*, com registros feitos por agricultores em seus próprios cultivos e rotulagem conduzida por especialistas. As imagens refletem condições reais de campo, com variações não controladas de iluminação, fundo e posicionamento.



(a) CBB (b) CBSD (c) CGM

(d) CMD (e) healthy

Figura 1. Exemplos visuais das cinco classes do conjunto de dados.

As imagens foram classificadas em cinco categorias: uma correspondente a folhas saudáveis e quatro referentes a doenças foliares específicas, conforme ilustrado na Figura 1 e detalhado na Tabela 1.

Sigla	Descrição (Inglês)	Descrição (Português)	Proporção (%)
CMD	Cassava Mosaic Disease	Doença do Mosaico	61,49%
Healthy	Healthy Leaves	Folhas Saudáveis	12,04%
CGM	Cassava Green Mottle	Vírus do Mosqueado Verde	11,15%
CBSD	Cassava Brown Streak Disease	Doença da Estria Marrom	10,23%
CBB	Cassava Bacterial Blight	Bacteriose da Mandioca	5,08%

Tabela 1. Categorias de doenças foliares da mandioca e folhas saudáveis, com respectivas proporções no conjunto de dados.

Observa-se um desbalanceamento significativo entre as classes, com predominância expressiva da Cassava Mosaic Disease (CMD) e sub-representação da Cassava Bacterial Blight (CBB), enquanto as demais classes apresentam proporções mais próximas entre si.

3.2. Arquiteturas Selecionadas

Este trabalho adota uma abordagem de *ensemble* entre duas arquiteturas com características complementares: o *CropNet*, derivado da *MobileNetV3* [Howard et al. 2019], e o *Swin Transformer* [Liu et al. 2021]. A escolha visa equilibrar leveza computacional e capacidade de representação, atendendo simultaneamente às restrições operacionais de aplicações embarcadas e à complexidade visual decorrente da variabilidade estrutural e de fundo presente em imagens agrícolas.

MobileNetV3 (via CropNet): Arquitetura leve baseada em convoluções separáveis em profundidade e módulos squeeze-and-excitation [Howard et al. 2019]. O modelo CropNet, derivado dessa arquitetura, foi adaptado para tarefas agrícolas com pré-treinamento em conjuntos como iNaturalist [Google AI and TensorFlow Hub Team 2021].

Swin Transformer: Transformer visual hierárquico que utiliza janelas deslocadas (*Shifted WINdows*) para capturar dependências locais e globais de forma eficiente [Liu et al. 2021]. Sua estrutura multiescala favorece a modelagem de padrões espaciais complexos, sendo especialmente útil em contextos de alta variabilidade morfológica e de fundo.

3.3. Pré-processamento

O conjunto de dados apresenta um desbalanceamento acentuado entre as classes, com a *Cassava Mosaic Disease* (CMD) representando 61,49% das amostras e a *Cassava Bacterial Blight* (CBB) apenas 5,08%. Para mitigar esse viés durante o treinamento, foi adotada uma divisão estratificada, mantendo a proporção de classes nos subconjuntos de treino (70%), validação (15%) e teste (15%). O conjunto de teste foi mantido fixo e sem aumento de dados, simulando um cenário real de inferência. Essa escolha metodológica é respaldada por [Sadaiyandi et al. 2023], que demonstram que a estratificação dos dados contribui para maior representatividade das classes minoritárias e melhora a acurácia de modelos de deep learning em cenários de desbalanceamento.

Swin Transformer. As imagens foram redimensionadas para 224×224 pixels e normalizadas de acordo com as médias e desvios padrão do ImageNet [Deng et al. 2009]. Técnicas de aumento de dados (*data augmentation*) foram aplicadas exclusivamente durante o treinamento, por meio de transformações estocásticas realizadas em tempo de execução. As operações incluíram espelhamento horizontal aleatório, rotações de até $\pm 15^{\circ}$ e ajustes de brilho e contraste controlados (*ColorJitter*). Essas transformações foram implementadas de forma dinâmica a cada batch, sem alteração permanente das imagens originais no disco, permitindo a geração contínua de variações amostrais ao longo das épocas. Tal abordagem é amplamente reconhecida por sua eficácia na mitigação de sobreajuste e na melhoria da capacidade de generalização em modelos de visão computacional, especialmente em contextos com dados limitados ou desbalanceados [Shorten and Khoshgoftaar 2019]. Os conjuntos de validação e teste foram submetidos apenas a redimensionamento e normalização, a fim de garantir uma avaliação consistente e livre de artefatos.

CropNet. As imagens foram redimensionadas para 224×224 pixels e normalizadas para o intervalo $[0, 1]$ por meio da divisão dos valores dos pixels por 255. Durante o treinamento, aplicou-se *data augmentation* dinâmico, por meio de transformações estocásticas implementadas com *tf.keras.Sequential*. As operações incluíram espelhamento horizontal aleatório, rotações de até $\pm 15^{\circ}$, além de ajustes de brilho e contraste com `fator=0.2`, o que permite variar aleatoriamente a intensidade dessas propriedades em até 20% para mais ou para menos em relação ao valor original da imagem. As transformações foram realizadas em tempo de execução, sem alterar os arquivos originais no disco, com o objetivo de aumentar a variabilidade amostral e mitigar o sobreajuste ao longo dos *folds*. Para os conjuntos de validação e teste, não foram aplicadas transformações, mantendo apenas o redimensionamento e a normalização. Essa estratégia permite equilibrar leveza computacional com ganho de generalização, especialmente relevante em contextos com dados desbalanceados.

Essa abordagem diferenciada de pré-processamento entre os modelos reflete suas origens arquiteturais e estratégias de treinamento distintas, respeitando as premissas de generalização e eficiência computacional de cada um.

3.4. Estratégias de Ajuste Fino (Fine-tuning)

Ambos os modelos foram treinados em regime supervisionado, com divisão estratificada dos dados em treino (70%), validação (15%) e teste (15%), utilizando *random_state* fixo de 42 para garantir reprodutibilidade. O *batch size* adotado foi de 16 para o *Swin Transformer* e de 32 para o *CropNet*.

CropNet. Foi adotada a estratégia de *feature extraction*, mantendo congelada a base do modelo (MobileNetV3, com aproximadamente 5,2 milhões de parâmetros), previamente treinada com dados botânicos do iNaturalist e ImageNet-21K. Apenas a nova cabeça classificadora foi treinada, composta por uma camada densa com 128 unidades (ativação *ReLU*), seguida de *dropout* (0,3) e uma camada de saída com ativação *softmax*. A otimização foi realizada com o otimizador Adam (taxa de aprendizado de 1×10^{-4}) e a função de perda entropia cruzada esparsa. Utilizou-se o *scheduler ReduceLROnPlateau*

e o critério de parada antecipada com paciência de cinco épocas. A nova cabeça somou aproximadamente 1,5 mil parâmetros treináveis.

Swin Transformer. Foi utilizado o modelo *swin-tiny-patch4-window7-224*, da Microsoft, pré-treinado no ImageNet-1K e adaptado para classificação multiclasse com cinco categorias por meio da substituição da camada final por uma projeção linear. Todo o modelo foi ajustado (*fine-tuning completo*), utilizando o otimizador AdamW com taxa de aprendizado inicial de 5×10^{-5} e decaimento de peso de 0,01, conforme recomendado por Liu et al. [Liu et al. 2021]. A escolha do AdamW se justifica por sua regularização desacoplada, mais eficaz em arquiteturas do tipo Transformer [Loshchilov and Hutter 2019]. Para reduzir a sobreconfiança e melhorar a generalização, foi aplicada a função de perda entropia cruzada com suavização de rótulo (*label smoothing*, $\epsilon = 0{,}1$), conforme proposto por Müller et al. [Müller et al. 2020]. A taxa de aprendizado foi ajustada dinamicamente com o *scheduler ReduceLROnPlateau*. Um critério de parada antecipada com paciência de três épocas foi configurado, mas não foi acionado. O modelo possui aproximadamente 28 milhões de parâmetros.

Infraestrutura. Todos os experimentos foram conduzidos em um ambiente com GPU NVIDIA L40, 16 vCPUs, 250 GB de memória RAM e 50 GB de armazenamento dedicado.

3.5. Estratégia de Ensemble

Com o objetivo de integrar modelos com naturezas arquiteturais distintas, uma CNN leve (CropNet) e um Transformer hierárquico (Swin), foram combinados visando capturar tanto padrões locais quanto relações de longo alcance entre características visuais. Foi adotada uma estratégia de comitê aplicada exclusivamente na etapa de inferência. Cada modelo foi treinado de forma independente, e suas distribuições de probabilidade, obtidas por meio da função *softmax*, foram combinadas por soma simples:

$$p_{\text{final}} = p_{\text{crop}} + p_{\text{swin}}$$

Essa abordagem equivale à atribuição de pesos iguais (50%) para cada modelo, caracterizando uma combinação não paramétrica. A classe final foi determinada por $\hat{y} = \arg\max_c \; p_{\text{final}}^{(c)}$, onde c denota o índice da classe. Embora alternativas como votação majoritária e média ponderada tenham sido consideradas, a soma simples foi escolhida por sua eficácia empírica e estabilidade numérica, além de dispensar ajustes adicionais via validação cruzada. Essa combinação resultou em desempenho superior aos modelos isolados, conforme detalhado na Seção 4.2.

3.6. Métricas e Avaliações

A avaliação dos modelos foi conduzida com base em métricas padrão para classificação multiclasse, com foco em acurácia e *F1-score macro*. Esta última foi priorizada por refletir o equilíbrio do desempenho entre classes, o que é especialmente relevante em cenários com desbalanceamento, como ocorre com frequência em conjuntos de dados agrícolas.

Além do conjunto de teste fixo, foi utilizada validação cruzada estratificada com $K = 5$ *folds* sobre o conjunto de treinamento, visando estimar de forma mais robusta o poder de generalização dos modelos. Em cada rodada, os modelos foram treinados em quatro *folds* e validados no *fold* restante, preservando a proporção original das classes. As métricas de desempenho foram calculadas individualmente em cada *fold* e, ao final, agregadas por meio da média aritmética acompanhada do desvio padrão, o que permite uma avaliação mais estável e confiável, capturando variações entre as partições.

As Equações 1–4 formalizam os cálculos, onde C representa o conjunto de classes, e TP_c, FP_c, FN_c são os verdadeiros positivos, falsos positivos e falsos negativos para a classe c. As métricas de precisão e revocação são utilizadas para qualificar erros por excesso e omissão, respectivamente, ambos críticos para o manejo agronômico, dada a importância de decisões corretas no diagnóstico fitossanitário.

$$\text{Acurácia} = \frac{1}{|C|} \sum_{c \in C} \left(\frac{TP_c + TN_c}{TP_c + FP_c + FN_c + TN_c} \right) \quad (1)$$

$$\text{Precisão} = \frac{1}{|C|} \sum_{c \in C} \left(\frac{TP_c}{TP_c + FP_c} \right) \quad (2)$$

$$\text{Revocação} = \frac{1}{|C|} \sum_{c \in C} \left(\frac{TP_c}{TP_c + FN_c} \right) \quad (3)$$

$$F_1\text{-Score} = 2 \cdot \left(\frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \right) \quad (4)$$

As métricas foram computadas com base nos valores extraídos da matriz de confusão, e a média macro foi utilizada para reduzir o viés causado pela distribuição desigual entre as classes.

4. Resultados e Discussão

Os experimentos foram conduzidos conforme a metodologia descrita na Seção 3, respeitando a divisão dos dados, o pré-processamento, as estratégias de ajuste fino e a abordagem de *ensemble* proposta. Esta seção apresenta os resultados obtidos para cada modelo, individualmente e em combinação, seguidos de uma análise crítica de seu desempenho com base nas métricas discutidas na Subseção 3.6.

4.1. Apresentação dos Resultados

A Tabela 2 apresenta as métricas de desempenho médio e desvio padrão obtidos por meio de validação cruzada estratificada com $K = 5$ *folds*, considerando os três modelos avaliados: Swin Transformer, CropNet e o ensemble. O *F1-score macro* foi utilizado como principal critério comparativo, por refletir o equilíbrio entre precisão e revocação, especialmente relevante em cenários com desbalanceamento de classes.

Observa-se que o *ensemble* superou consistentemente os modelos individuais em todas as métricas avaliadas. Além disso, os baixos desvios padrão indicam estabilidade entre os *folds*, sugerindo que os resultados não são sensíveis à partição dos dados e que o *ensemble* apresenta desempenho robusto e generalizável.

Modelo	Acurácia	Precisão	Revocação	F1 (Macro)
Swin Transformer	0.8671 ± 0.014	0.7773 ± 0.013	0.7614 ± 0.015	0.7737 ± 0.014
CropNet	0.8764 ± 0.012	0.7958 ± 0.011	0.7852 ± 0.010	0.7921 ± 0.012
Ensemble	**0.9087 ± 0.009**	**0.8351 ± 0.010**	**0.8236 ± 0.011**	**0.8329 ± 0.011**

Tabela 2. Métricas médias com desvio padrão por modelo (validação cruzada estratificada com $K = 5$).

A Figura 2 apresenta a matriz de confusão do ensemble, normalizada por classe real, permitindo uma avaliação proporcional do desempenho por categoria. Observa-se que a classe CMD obteve a maior taxa de acerto (97,82%), refletindo tanto sua predominância no conjunto quanto sua morfologia marcante. As demais classes também apresentaram desempenhos satisfatórios: CBSD atingiu 82,98% de acerto e Healthy, 83,42%, ainda que esta última tenha exibido dispersão de erros entre diferentes classes. Por outro lado, a classe CBB alcançou apenas 66,87% de acertos, com confusões concentradas em Healthy (23,31%), e CGM teve 78,21% de acerto, com erros distribuídos entre CMD e Healthy. Esses resultados demonstram a capacidade do *ensemble* em lidar com classes visualmente diversas, mas também revelam desafios persistentes em distinguir padrões mais sutis ou sobrepostos, sugerindo caminhos futuros de aprimoramento, como aumento da diversidade de amostras ou uso de mecanismos de atenção interpretável.

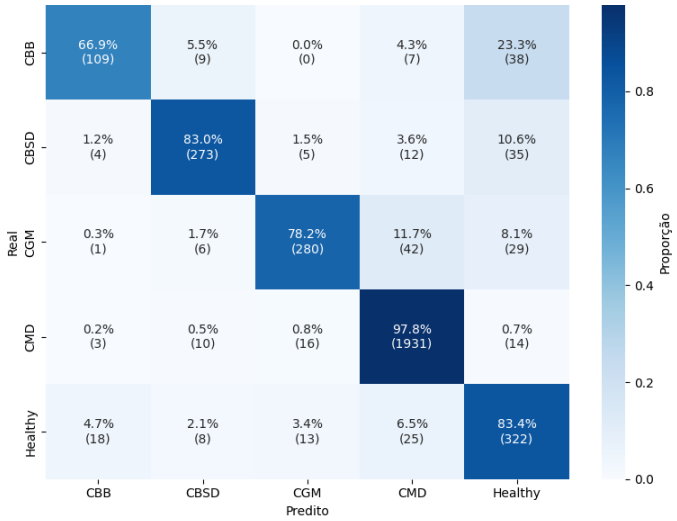


Figura 2. Matriz de confusão do *ensemble* no conjunto de teste.

4.2. Discussão dos Resultados

Os resultados obtidos por meio da validação cruzada demonstram que o *ensemble* superou consistentemente os modelos individuais em todas as métricas avaliadas, apresentando acurácia média de 0,9087 e *F1-score macro* de 0,8329. Esses valores indicam ganhos concretos em equilíbrio entre classes e desempenho geral, particularmente relevantes em um cenário com desbalanceamento severo de dados.

A análise do desvio padrão, que foi inferior a 0,012 em todas as métricas, evidenciando que o desempenho do *ensemble* foi estável entre os *folds*, sugerindo robustez

frente à partição dos dados e capacidade de generalização do modelo proposto. Essa estabilidade é especialmente desejável em contextos agrícolas, onde a variabilidade visual nas imagens pode comprometer a consistência de modelos preditivos.

Os resultados também reforçam a complementaridade das arquiteturas utilizadas: o Swin Transformer, com maior capacidade de representação, contribui para a modelagem de padrões visuais complexos, enquanto o CropNet, por ser leve e eficiente, oferece desempenho sólido em classes minoritárias. A combinação dessas características resultou em ganhos sinérgicos observáveis nas métricas agregadas, sustentando a efetividade da abordagem de *ensemble* proposta tanto em termos de desempenho quanto de estabilidade. Esses achados reforçam seu potencial de aplicação em ambientes agrícolas com restrições computacionais, nos quais robustez e leveza são requisitos críticos.

Uma análise mais refinada das predições por classe revela evidências claras de complementaridade entre os modelos individuais. Por exemplo, o CropNet demonstrou maior sensibilidade em classes com menor representatividade, como CBB, CBSD e CGM, superando o Swin Transformer em termos de revocação nessas categorias. Por outro lado, o Swin mostrou desempenho mais consistente em classes majoritárias, como CMD, evidenciando sua capacidade superior de modelagem em cenários com maior variação morfológica. Essa divergência nos padrões de acerto entre os modelos sugere que seus erros ocorrem de forma não correlacionada, o que é favorável ao uso de técnicas de combinação. A estratégia de *ensemble* adotada, por meio da soma simples das probabilidades, mostrou-se eficaz para explorar essas complementaridades, equilibrando sensibilidade e robustez sem incorrer em custos computacionais adicionais. Tal abordagem reforça sua aplicabilidade em cenários operacionais restritivos, como na agricultura de precisão embarcada.

Além da análise descritiva, foi conduzida uma avaliação estatística para verificar se as diferenças de desempenho observadas entre os modelos eram estatisticamente relevantes. Foram aplicados testes pareados de T-Test sobre os *F1-scores macro* obtidos em cada *fold*, considerando $\alpha = 0{,}05$ como nível de significância. Os resultados indicaram diferenças estatísticas significativas tanto entre o *ensemble* e o Swin Transformer ($p < 0{,}05$) quanto entre o *ensemble* e o CropNet ($p < 0{,}05$). Esses achados foram corroborados pela ANOVA de medidas repetidas aplicada aos três modelos, que também apontou variações significativas entre os grupos ($p < 0{,}05$). Dessa forma, a análise estatística reforça que os ganhos observados com o *ensemble* são consistentes e não atribuíveis ao acaso, evidenciando sua superioridade e estabilidade frente aos modelos individuais.

4.3. Considerações Práticas

A partir de uma perspectiva operacional, a abordagem de *ensemble* propõe um equilíbrio relevante entre desempenho e viabilidade computacional. O Swin Transformer, com aproximadamente 28 milhões de parâmetros, apresenta maior capacidade de representação, mas impõe um custo computacional considerável, especialmente em contextos com infraestrutura limitada. O CropNet, por sua vez, é baseado no MobileNetV3 e conta com cerca de 5,3 milhões de parâmetros, sendo substancialmente mais leve e adequado para dispositivos embarcados ou de borda.

A combinação dessas arquiteturas via *ensemble* foi realizada apenas na etapa de inferência, por soma das probabilidades, evitando o aumento do tempo de treinamento. No entanto, a inferência conjunta demanda a execução de ambos os modelos, o que

pode impactar a latência e o consumo de energia, caso o sistema não adote técnicas de otimização como quantização, poda ou distilação.

Ainda assim, os ganhos médios observados nas métricas, *F1-score macro* de 0,8329 e acurácia de 0,9087, justificam esse custo adicional em aplicações nas quais robustez e acurácia diagnóstica são mais críticas do que resposta em tempo real. Dessa forma, o *ensemble* proposto demonstra viabilidade prática, desde que acompanhado de ajustes técnicos específicos para o ambiente de implantação.

5. Considerações finais

Este trabalho propôs uma estratégia de *ensemble* entre o *CropNet*, uma rede convolucional leve baseada em *MobileNetV3*, e o *Swin Transformer*, um modelo hierárquico baseado em atenção, para a tarefa de classificação multiclasse de doenças em folhas de mandioca. A abordagem combinou arquiteturas com características complementares, buscando maior robustez preditiva frente à variabilidade visual e ao desbalanceamento severo entre classes no conjunto de dados.

Por meio de validação cruzada estratificada com $K = 5$, o *ensemble* demonstrou desempenho médio superior aos modelos isolados, alcançando *F1-score macro* de 0.8329 e acurácia de 0.9087, com baixos desvios padrão entre os *folds*. Esses resultados indicam estabilidade, capacidade de generalização e reforçam a viabilidade da proposta em cenários com restrições computacionais, como na agricultura de base familiar.

Apesar do bom desempenho geral, algumas classes ainda apresentaram padrões de confusão relevantes. Em especial, a CBB obteve taxa de acerto inferior às demais (66,87%), com confusões frequentes com a classe *Healthy*, sugerindo a necessidade de explorar representações mais discriminativas para esses casos. Trabalhos futuros podem incluir o uso de técnicas de explicabilidade, como *Grad-CAM* e mapas de atenção, além de estratégias de compressão e otimização, como poda e quantização, visando à implantação eficiente do *ensemble* em dispositivos embarcados.

6. Agradecimentos

Os autores agradecem à Universidade do Estado do Amazonas (UEA) e à Callidus Academy pelo apoio ao desenvolvimento da pesquisa. Agradecemos, em especial, pelo suporte institucional e técnico que contribuíram significativamente para sua realização.

Referências

Albuquerque, L. and Guedes, E. (2023). Um comparativo de abordagens com redes neurais artificiais para detecção inteligente de patologias na folha do café. In *Anais do XIV Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*, pages 131–140, Porto Alegre, RS, Brasil. SBC.

Batista, L. M. S. and de Paiva, M. S. B. (2019). Sistemas agroflorestais no município de capitão poço: motivações de implementação na comunidade do barro vermelho. Orientador: Prof. Dr. José Sebastião Romano de Oliveira; Coorientadora: Prof.ª Msc. Ana Paula Dias Costa.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. `http://www.image-net.org`. CVPR 2009.

Embrapa Mandioca e Fruticultura (2023). Mandioca. Acesso em: 5 maio 2025.

Fathima, M. and Bondili, H. S. S. (2025). A comprehensive survey on cassava disease detection and classification using deep learning models. In *SCT Proceedings in Interdisciplinary Insights and Innovations*, volume 3, pages 377–393. SCT.

Google AI and TensorFlow Hub Team (2021). Cropnet: Cassava disease classification model. `https://www.tensorflow.org/hub/tutorials/cropnet_cassava`. Model available at TensorFlow Hub. Accessed: 2025-05-25.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Le, Q. V., and Adam, H. (2019). Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324.

John, A. A. (2022). Identification of diseases in cassava leaves using convolutional neural network. In *2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, pages 1–6.

Junior, A. C., da Silva, F., and Rios, R. (2024). Deep learning-based transfer learning for classification of cassava disease. In *Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional*, pages 364–375, Porto Alegre, RS, Brasil. SBC.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ministério da Agricultura e Pecuária (MAPA) and Embrapa (2023). Projeções do agronegócio: Brasil 2022/23 a 2032/33. https://www.gov.br/agricultura/pt-br/assuntos/politica-agricola/todas-publicacoes-de-politica-agricola/projecoes-do-agronegocio/projecoes-do-agronegocio-2022-2023-a-2032-2033.pdf. Acesso em: 5 maio 2025.

Müller, R., Kornblith, S., and Hinton, G. (2020). When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 33, pages 4694–4703.

Mwebaze, E., Mostipak, J., Joyce, Elliott, J., and Dane, S. (2020). Cassava leaf disease classification. `https://www.kaggle.com/competitions/cassava-leaf-disease-classification`. Kaggle Competition Dataset.

Sadaiyandi, J., Arumugam, P., Sangaiah, A. K., and Zhang, C. (2023). Stratified sampling-based deep learning approach to increase prediction accuracy of unbalanced dataset. *Electronics*, 12(21):4423.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48.

Silva, W., Soares, B., Almeida, V., Viana, L., Pastori, P., Magalhães, D., and Rocha, A. (2024). Detecção da praga spodoptera frugiperda no cultivo de milho usando armadilhas inteligentes e visão computacional. In *Anais do XV Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*, pages 61–70, Porto Alegre, RS, Brasil. SBC.