
Avaliação Temporal da Detecção de Deficiências Nutricionais em Folhas de Café por Deep Learning

Hericles Felipe Ferraz



UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Hericles Felipe Ferraz

**Avaliação Temporal da Detecção de Deficiências
Nutricionais em Folhas de Café por Deep
Learning**

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Prof. Dr. Mauricio Cunha Escarpinati

Coorientador: Prof. Dr. Thiago Pirola Ribeiro

Uberlândia

2026

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

F381a
2026

Ferraz, Hericles Felipe, 1995-

Avaliação temporal da detecção de deficiências nutricionais em folhas de café por deep learning [recurso eletrônico] / Hericles Felipe Ferraz. -- 2026.

Orientador: Mauricio Cunha Escarpinati.

Coorientador: Thiago Pirola Ribeiro.

Dissertação (Mestrado) -- Universidade Federal de Uberlândia, Programa de Pós-graduação em Ciência da Computação.

Modo de acesso: Internet.

Disponível em: <http://doi.org/10.14393/ufu.di.2026.11030>

Inclui bibliografia.

Inclui ilustrações.

1. Computação. I. Escarpinati, Mauricio Cunha, 1976-, (Orient.). II. Ribeiro, Thiago Pirola, 1978-, (Coorient.). III. Universidade Federal de Uberlândia. Programa de Pós-graduação em Ciência da Computação. IV. Título.

CDU: 681.3

André Carlos Francisco
Bibliotecário(a)-Documentalista - CRB-6/3408



ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Dissertação, 18/2026, PPGCO				
Data:	17 de Julho de 2026	Hora de início:	09:02	Hora de encerramento:	11:38
Matrícula do Discente:	12412CCP010				
Nome do Discente:	Hericles Felipe Ferraz				
Título do Trabalho:	Avaliação Temporal da Detecção de Deficiências Nutricionais em Folhas de Café por Deep Learning				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Ciência de Dados				
Projeto de Pesquisa de vinculação:	-----				

Reuniu-se por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Thiago Pirola Ribeiro - FACOM/UFU(Coorientador), Bruno Augusto Nassif Travençolo - FACOM/UFU, Marcelo Andrade da Costa Vieira - SEL/USP e Mauricio Cunha Escarpinati - FACOM/UFU, orientador(a) do(a) candidato(a).

Os examinadores participaram desde as seguintes localidades: Marcelo Andrade da Costa Vieira - São Carlos/SP. Os outros membros da banca e o aluno(a) participaram da cidade de Uberlândia.

Iniciando os trabalhos o(a) presidente da mesa, Prof. Dr. Mauricio Cunha Escarpinati, apresentou a Comissão Examinadora e o(a) candidato(a), agradeceu a presença do público, e concedeu ao(à) Discente a palavra para a exposição do seu trabalho.

A seguir o senhor(a) presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o(a) candidato(a). Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato(a):

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.



Documento assinado eletronicamente por **Thiago Pirola Ribeiro, Professor(a) do Magistério Superior**, em 21/07/2026, às 08:43, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Mauricio Cunha Escarpinati, Presidente**, em 21/07/2026, às 09:28, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Bruno Augusto Nassif Travençolo, Professor(a) do Magistério Superior**, em 21/07/2026, às 11:33, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **MARCELO ANDRADE DA COSTA VIEIRA, Usuário Externo**, em 23/07/2026, às 14:50, 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 https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **7467350** e o código CRC **A905301B**.

Aos dias difíceis que me ensinaram a ser forte, e às noites insones que se transformaram em pesquisa e conhecimento.

Agradecimentos

À minha eterna companheira, Letícia, meu porto seguro e minha maior incentivadora. Obrigado por compreender as ausências, acalmar minhas ansiedades e celebrar cada pequena etapa. Este trabalho também é seu.

À minha mãe, Joana, e ao meu pai, Sebastião, por serem meu alicerce e me ensinarem o valor do esforço e acreditarem nos meus sonhos desde o início de tudo.

À minha irmã, Gabrielly, pelo carinho, pelas palavras de incentivo e por sempre estar presente, me apoiando incondicionalmente. E às minhas irmãs, Tatiane e Priscila, pelo carinho e por fazerem parte da minha história.

Ao meu orientador, Prof. Dr. Mauricio Cunha Escarpinati, por ter acreditado neste trabalho desde o princípio e por me guiar com profissionalismo, paciência, generosidade e bom humor. Agradeço por cada ensinamento, direcionamento e correção ao longo desta caminhada.

Ao meu coorientador, Prof. Dr. Thiago Pirola Ribeiro, pela parceria ao longo desta caminhada e pela confiança em compartilhar a base de dados que deu vida a esta pesquisa. Sua parceria foi fundamental para que este trabalho se tornasse possível.

À Faculdade de Computação (FACOM) da Universidade Federal de Uberlândia, pela infraestrutura e pelo ambiente acadêmico que contribuíram para a minha formação e para o desenvolvimento deste estudo.

*“Se sentir que chegou ao seu limite, lembre-se do motivo pelo qual você cerra os punhos,
lembre-se porque resolveu trilhar este caminho e permita que essa memória o carregue
além de seus limites.”*

All Might — Boku no Hero Academia

Resumo

O diagnóstico de deficiências nutricionais na cafeicultura, um pilar do agronegócio brasileiro, é tradicionalmente reativo e subjetivo. Este trabalho propõe uma análise temporal para a detecção precoce de estresses nutricionais, com o objetivo de determinar a partir de que momento os sintomas se tornam visualmente discriminativos para modelos de aprendizado profundo. Para isso, comparou-se sistematicamente o desempenho de três arquiteturas de *deep learning*, sendo elas ResNet50, *Vision Transformer* (ViT-B/16) e um modelo híbrido CNN-ViT, na classificação de seis deficiências nutricionais (N, P, K, Ca, Mg, S) e uma classe de controle, com um conjunto de dados de imagens RGB coletadas ao longo de cinco meses. Os resultados de 98 experimentos demonstram que o desempenho dos classificadores é fortemente dependente do estágio dos sintomas, com melhor desempenho observado na configuração que retém os 40% temporais finais do período experimental. Sob o protocolo padronizado adotado, a arquitetura ResNet50 superou o Híbrido CNN-ViT, que por sua vez superou o ViT-B/16, com acurácia máxima de 77,97% para a arquitetura convolucional com imagens de 512×512 pixels, contra 71,68% do Híbrido e 67,74% do ViT-B/16. A análise exploratória de interpretabilidade via Grad-CAM produziu mapas de ativação visualmente compatíveis com a sintomatologia agrônoma esperada para cada deficiência. Conclui-se que, nas condições controladas do experimento, o viés indutivo de localidade das CNNs mostrou-se mais adequado para a captura de padrões texturais foliares neste domínio, e que a análise temporal é crucial para entender os limites e o potencial da detecção precoce automatizada.

Palavras-chave: Redes Neurais Convolucionais. Vision Transformers. Modelos Híbridos. Interpretabilidade. Deficiências Nutricionais. Diagnóstico Foliar. Agricultura de Precisão.

Abstract

The diagnosis of nutritional deficiencies in coffee cultivation, a cornerstone of Brazilian agribusiness, is traditionally reactive and subjective. This work proposes a temporal analysis for the early detection of nutritional stress, with the objective of determining from which point in time symptoms become visually discriminative for deep learning models. To this end, three deep learning architectures were systematically compared, namely ResNet50, a Vision Transformer (ViT-B/16), and a hybrid CNN-ViT model, in the classification of six nutritional deficiencies (N, P, K, Ca, Mg, S) and a control class, using an RGB image dataset collected over five months. The results of 98 experiments demonstrate that the performance of the classifiers is strongly dependent on the symptom stage, with the best observed performance in the configuration that retains the final 40% of the experimental period. Under the standardized protocol adopted, the ResNet50 architecture outperformed the Hybrid CNN-ViT, which in turn outperformed the ViT-B/16, with the convolutional architecture reaching a maximum accuracy of 77.97% with 512×512 pixel images, against 71.68% for the Hybrid and 67.74% for the ViT-B/16. An exploratory interpretability analysis via Grad-CAM produced activation maps visually consistent with the expected agronomic symptomatology of each deficiency. Under the controlled experimental conditions, the inductive locality bias of CNNs proved more suitable for capturing foliar texture patterns in this domain, and temporal analysis is crucial for understanding the limits and potential of automated early detection.

Keywords: Convolutional Neural Networks. Vision Transformers. Hybrid Models. Interpretability. Nutritional Deficiencies. Foliar Diagnosis. Precision Agriculture..

Lista de ilustrações

Figura 1 – Folhas de café com diferentes deficiências nutricionais	39
Figura 2 – Arquitetura CNN para classificação de imagens	43
Figura 3 – Bloco residual da ResNet	46
Figura 4 – Características aprendidas por uma CNN em diferentes camadas	49
Figura 5 – <i>Scaled Dot-Product</i> e <i>Multi-Head Attention</i>	50
Figura 6 – Visão geral da arquitetura do ViT	53
Figura 7 – Arquitetura do modelo Híbrido CNN-ViT	56
Figura 8 – <i>Pipeline</i> do Grad-CAM	59
Figura 9 – Mapas de atenção do ViT em imagens naturais	61
Figura 10 – Fluxograma do <i>pipeline</i> experimental	70
Figura 11 – Registro fotográfico do experimento controlado	71
Figura 12 – Distribuição do número de amostras por classe	74
Figura 13 – Exemplos de folhas das sete classes	74
Figura 14 – Evolução temporal dos sintomas de deficiência de Ca	80
Figura 15 – Visão geral comparativa das três arquiteturas	82
Figura 16 – Acurácia de ResNet50 e Híbrido em função da fração temporal	96
Figura 17 – Comparação da acurácia entre as três arquiteturas	99
Figura 18 – Impacto da resolução na acurácia do ResNet50	104
Figura 19 – Matriz de confusão do melhor modelo (ResNet50)	108
Figura 20 – Comparação dos mapas Grad-CAM de layer3 e layer4	119
Figura 21 – <i>Grid</i> dos mapas Grad-CAM da layer3 do ResNet50	120
Figura 22 – Acurácia do Híbrido por configuração de <i>patches</i>	121
Figura 23 – Comparação entre Grad-CAM e mapas de atenção por <i>patches</i>	121
Figura 24 – Tempo de treinamento por arquitetura e fração temporal	144
Figura 25 – <i>Precision</i> , <i>Recall</i> e F1-Score por classe (ResNet50)	144
Figura 26 – Matriz de confusão em janela restrita (ResNet50, frac95)	145
Figura 27 – Matriz de confusão do melhor Híbrido CNN-ViT	145
Figura 28 – Comparativo completo entre Grad-CAM e mapas de atenção	148

Lista de tabelas

Tabela 1 – Comparação dos principais trabalhos relacionados	66
Tabela 2 – Distribuição de amostras por classe	73
Tabela 3 – Acurácia do ResNet50 por resolução e fração temporal	94
Tabela 4 – Acurácia do Híbrido CNN-ViT por resolução e fração temporal	94
Tabela 5 – Acurácia do ViT-B/16 por fração temporal	94
Tabela 6 – Custo computacional por arquitetura	102
Tabela 7 – Métricas por classe do melhor modelo (ResNet50)	105
Tabela 8 – Aderência qualitativa entre Grad-CAM e sintomatologia	111
Tabela 9 – Melhor desempenho do Híbrido por grade de <i>patches</i>	112
Tabela 10 – Resultados completos do benchmark (50 experimentos)	141
Tabela 11 – Resultados completos do estudo de ablação (48 experimentos)	142

Lista de siglas

B Boro

Ca Cálcio

CLS Token de Classificação — *Classification Token*

CNN Rede Neural Convolucional — *Convolutional Neural Network*

Fe Ferro

GPU Unidade de Processamento Gráfico — *Graphics Processing Unit*

Grad-CAM *Gradient-weighted Class Activation Mapping*

ICP Espectrometria de Plasma Indutivamente Acoplado — *Inductively Coupled Plasma*

ICPMS Espectrometria de Massas com Plasma Indutivamente Acoplado — *Inductively Coupled Plasma Mass Spectrometry*

K Potássio

LOGO *Leave-One-Group-Out*

Mg Magnésio

N Nitrogênio

P Fósforo

ReLU Unidade Linear Retificada — *Rectified Linear Unit*

ResNet Rede Residual — *Residual Network*

RGB Vermelho, Verde e Azul — *Red, Green, Blue*

S Enxofre

UAV Veículo Aéreo Não Tripulado — *Unmanned Aerial Vehicle*

ViT *Vision Transformer*

XAI Inteligência Artificial Explicável — *Explainable Artificial Intelligence*

Sumário

1	INTRODUÇÃO	25
1.1	Motivação	27
1.2	Objetivos	29
1.2.1	Objetivo Geral	29
1.2.2	Objetivos Específicos	30
1.3	Hipóteses	30
1.4	Contribuições Esperadas	32
1.5	Organização do Trabalho	34
2	FUNDAMENTAÇÃO TEÓRICA	37
2.1	Nutrição Mineral em Plantas de Café	37
2.2	Diagnóstico por Imagens Digitais na Agricultura	40
2.3	Rede Neural Convolucional (CNN)s	42
2.3.1	Fundamentos	42
2.3.2	Arquiteturas Residuais e a Rede Residual (ResNet)	45
2.3.3	Extração Hierárquica de Características	47
2.4	<i>Vision Transformer</i> (ViT)s	48
2.4.1	Mecanismo de Atenção	48
2.4.2	Arquitetura do Vision Transformer	51
2.4.3	Comparação entre CNNs e ViTs	54
2.5	Arquiteturas Híbridas entre Convoluções e Atenção	55
2.6	Avaliação de Desempenho em Classificação	57
2.7	Interpretabilidade e Explicabilidade (Inteligência Artificial Ex- plicável (XAI))	58
2.7.1	<i>Gradient-weighted Class Activation Mapping</i> (Grad-CAM)	58
2.7.2	Mapas de Atenção em Transformers	60
2.7.3	Importância da Interpretabilidade na Agricultura	61

3	TRABALHOS RELACIONADOS	63
4	PROPOSTA METODOLÓGICA	69
4.1	Conjunto de Dados	70
4.2	Pré-processamento	75
4.2.1	Recorte Central Quadrado	75
4.2.2	Aumento de Dados e Normalização	76
4.3	Particionamento Temporal	77
4.4	Arquiteturas de Classificação	81
4.4.1	ResNet50	81
4.4.2	ViT-B/16	83
4.4.3	Modelo Híbrido CNN-ViT	84
4.5	Protocolo de Treinamento	86
4.6	Estratégia de Validação	88
4.7	Métricas de Avaliação	89
4.8	Interpretabilidade	89
5	EXPERIMENTOS E RESULTADOS	91
5.1	Configuração Experimental	92
5.2	Resultados Gerais	94
5.3	Análise do Impacto Temporal (H1)	95
5.4	Análise Comparativa das Arquiteturas (H2)	99
5.4.1	Limitações da comparação arquitetural	100
5.4.2	Fatores técnicos que sustentam a hierarquia observada	101
5.5	Impacto da Resolução	103
5.6	Análise por Classe	105
5.7	Análise de Interpretabilidade (H3)	108
5.8	Estudo de Ablação: Granularidade dos <i>Patches</i>	111
5.9	Discussão	114
6	CONCLUSÃO	123
6.1	Síntese dos Resultados	124
6.2	Contribuições	126
6.2.1	Contribuições Metodológicas	126
6.2.2	Contribuições Empíricas	126
6.2.3	Contribuições Aplicadas	127
6.3	Limitações	128
6.4	Trabalhos Futuros	129
6.5	Considerações Finais	131
	REFERÊNCIAS	133

APÊNDICE A	–	RESULTADOS COMPLETOS DOS EXPERIMENTOS	141
A.1		Benchmark Principal	141
A.2		Estudo de Ablação	141
APÊNDICE B	–	ANÁLISE COMPLEMENTAR DE DESEMPENHO	143
B.1		Custo Computacional por Fração Temporal	143
B.2		Métricas por Classe — Visualização Gráfica	143
B.3		Matrizes de Confusão Complementares	143
APÊNDICE C	–	GALERIA DE MAPAS DE INTERPRETABILIDADE	147
C.1		Grad-CAM versus Atenção por Tamanho de <i>Patch</i>	147

CAPÍTULO 1

Introdução

O café é um dos produtos agrícolas de maior relevância econômica global, com movimentação anual de bilhões de dólares em uma cadeia produtiva que se estende desde pequenos agricultores familiares até conglomerados multinacionais de torrefação e distribuição. O Brasil ocupa posição de destaque nesse cenário como o maior produtor mundial do grão, o que consolida uma tradição cafeeira que remonta ao século XVIII (FERREIRA; ANDRADE; SANTIAGO, 2021). A cafeicultura desempenha um papel estruturante no setor agropecuário brasileiro, com produção distribuída por centenas de municípios (IBGE, 2016) e uma extensa cadeia produtiva que abrange a produção agrícola, o beneficiamento industrial, o transporte e os serviços associados, responsável por centenas de milhares de empregos diretos e indiretos (SESSO; FILHO; PEREIRA, 2021). Essa relevância é particularmente expressiva em Minas Gerais, maior estado produtor, onde o café responde historicamente pela maior parcela da ocupação agrícola entre as lavouras (GANDOLFI; JESUS; GANDOLFI, 2019). A espécie *Coffea arabica*, responsável pela maior parcela da produção nacional, é particularmente valorizada por seus atributos sensoriais superiores e apresenta elevada exigência nutricional ao longo de seu ciclo produtivo (MALAVOLTA, 2006). Nesse sentido, garantir a sanidade nutricional dos cafeeiros constitui não apenas uma questão agronômica, mas também um imperativo econômico e social, dado que perdas de produtividade decorrentes de desequilíbrios nutricionais impactam diretamente a renda de milhares de famílias produtoras e a competitividade do Brasil no mercado internacional.

Apesar da robustez econômica do setor cafeeiro, a produtividade e a qualidade dos grãos são constantemente ameaçadas por uma ampla gama de desafios agronômicos, dentre os quais as deficiências nutricionais se sobressaem pela sua ubiquidade e pelo seu impacto cumulativo sobre o rendimento das lavouras. A carência de macronutrientes essenciais, como Nitrogênio (N), Fósforo (P), Potássio (K), Cálcio (Ca), Magnésio (Mg) e Enxofre (S), compromete funções fisiológicas vitais da planta, tais como a fotossíntese, a síntese de proteínas, a regulação osmótica e a integridade estrutural das membranas celulares (MARSCHNER, 2012). Cada nutriente desempenha papéis bioquímicos espe-

cíficos, e sua deficiência desencadeia sintomas visuais característicos nas folhas. O N, por exemplo, manifesta-se por clorose generalizada nas folhas mais velhas, enquanto a deficiência de K produz necrose marginal, e a carência de Mg gera clorose internerval (MALAVOLTA, 2006). Essas alterações fenotípicas, embora informativas para o diagnóstico, apresentam sobreposições visuais significativas que dificultam a identificação precisa do nutriente deficiente, sobretudo nos estágios iniciais do estresse, quando diferentes deficiências podem produzir padrões de clorose visualmente semelhantes. A complexidade do diagnóstico visual é agravada pelo fato de que múltiplas deficiências podem coexistir em uma mesma planta, e que fatores ambientais como estresse hídrico, excesso de radiação solar e ataque de patógenos podem mimetizar ou mascarar os sintomas nutricionais, o que gera ambiguidades que desafiam até mesmo agrônomos experientes.

O diagnóstico tradicional de deficiências nutricionais em cafeeiros baseia-se em dois pilares complementares, a análise laboratorial de tecido vegetal e a inspeção visual por agrônomos. A análise foliar, embora quantitativamente precisa, envolve procedimentos de coleta, preparação e digestão química das amostras e determinação elementar por técnicas laboratoriais como a Espectrometria de Massas com Plasma Indutivamente Acoplado (ICPMS), o que a caracteriza como um método destrutivo, dependente de infraestrutura especializada e inadequado ao monitoramento contínuo, não destrutivo e em larga escala exigido no campo (HANSEN et al., 2013). A inspeção visual, por sua vez, constitui um método subjetivo, intensivo em mão de obra qualificada e, crucialmente, reativo. Os sintomas visíveis de clorose e necrose se manifestam de forma inequívoca apenas em estágios avançados do estresse nutricional, quando uma parcela significativa das perdas de rendimento já se tornou irreversível (MARSCHNER, 2012). Essa natureza reativa do diagnóstico convencional impõe um paradigma de gestão nutricional que privilegia a correção sobre a prevenção, o que limita a capacidade dos produtores de otimizar o manejo de fertilizantes e minimizar perdas. Em um cenário ideal, o diagnóstico seria realizado de forma precoce, contínua e não destrutiva, o que permitiria a detecção de desequilíbrios nutricionais antes que os sintomas visuais se tornem evidentes ao olho humano e antes que as perdas fisiológicas se tornem irreversíveis. A busca por esse paradigma preventivo constitui um dos principais vetores de inovação na interface entre agronomia e tecnologia da informação.

Nesse contexto, a agricultura de precisão, impulsionada por avanços em tecnologias de sensoriamento remoto e proximal, processamento de imagens e análise de dados em larga escala, oferece um paradigma alternativo para o monitoramento contínuo e automatizado de culturas (SOUSSI et al., 2024). A visão computacional, particularmente o *deep learning*, emergiu como uma ferramenta poderosa para automatizar a identificação e a classificação de estresses bióticos e abióticos em plantas a partir de imagens digitais no espectro visível (Vermelho, Verde e Azul (RGB)) (LIAKOS et al., 2018). CNNs, em particular, têm demonstrado acurácias expressivas na classificação de doenças folia-

res em diversas culturas, incluindo o café (FERENTINOS, 2018; MEHTA et al., 2025; UPADHYAY et al., 2025; SHOAIB et al., 2025). Mais recentemente, o surgimento dos ViTs e de arquiteturas híbridas que combinam convoluções e atenção expandiu o repertório de ferramentas disponíveis, o que introduz mecanismos de atenção global capazes de modelar dependências espaciais de longo alcance na imagem (DOSOVITSKIY et al., 2021; KHAN et al., 2022). Contudo, uma análise criteriosa da literatura revela que a grande maioria dos estudos aborda o diagnóstico de deficiências nutricionais como um problema de classificação estática, com *datasets* compostos predominantemente por imagens de folhas com sintomas já pronunciados, sem aproveitar a informação diagnóstica contida na dinâmica temporal da evolução desses sinais visuais. Essa abordagem estática é particularmente problemática porque desconsidera que a capacidade discriminativa dos sintomas visuais não é constante ao longo do tempo, o que tende a restringir o reconhecimento a estágios já avançados da deficiência, momento em que o valor prático da detecção precoce é substancialmente reduzido. Nos estágios iniciais, os sinais de diferentes deficiências podem ser quase indistinguíveis, enquanto nos estágios tardios, cada deficiência desenvolve um perfil fenotípico mais característico e diferenciado (MALAVOLTA, 2006). A incorporação da dimensão temporal no *pipeline* de classificação representa, portanto, não apenas um refinamento metodológico, mas uma mudança de paradigma na forma como o problema é formulado e abordado.

1.1 Motivação

No domínio específico do café, esse avanço se materializa em sistemas automatizados que oferecem vantagens concretas em relação à inspeção manual, tais como a capacidade de processar grandes volumes de imagens em tempo reduzido, a padronização dos critérios diagnósticos e a possibilidade de implantação em dispositivos embarcados ou plataformas móveis para uso direto no campo. Conjuntos de dados especificamente voltados para o café, como o conjunto de imagens de folhas de Esgario, Krohling e Ventura (2020), difundido na comunidade sob a denominação BRACOL, e o CoLeaf-DB (TUESTA-MONTEZA; MEJIA-CABRERA; ARCILA-DIAZ, 2023), têm contribuído para o avanço dessa linha de pesquisa ao disponibilizar imagens anotadas de folhas com diferentes tipos de estresse. No entanto, esses *datasets* compartilham uma limitação fundamental, em que as imagens constituem registros transversais, com cada folha capturada uma única vez, sem o acompanhamento sistemático da evolução temporal dos sintomas, o que restringe os modelos treinados a operar como classificadores instantâneos, incapazes de considerar a progressão do estresse ao longo do tempo. Trabalhos recentes, como o de Randive, Sapkal e Bornare (2025), reforçam o interesse da comunidade na classificação de deficiências nutricionais em café por meio de *deep learning*, porém mantêm a abordagem estática, sem incorporar a dimensão temporal da evolução dos sintomas, o que evidencia que essa lacuna permanece

não endereçada na literatura atual.

A natureza progressiva das deficiências nutricionais constitui uma característica central que a abordagem predominante na literatura ignora. Uma deficiência não se manifesta de forma abrupta, ela evolui de um estado assintomático, no qual a planta mobiliza reservas internas para compensar a carência, para estágios intermediários com sintomas sutis, como leves alterações de tonalidade e redução no crescimento, e, finalmente, para sinais fenotípicos claros e frequentemente irreversíveis, como clorose generalizada e necrose tecidual (MARSCHNER, 2012; MALAVOLTA, 2006). A prática comum de treinar modelos exclusivamente com imagens de estágios avançados descarta a informação diagnóstica contida nessa progressão e, mais criticamente, produz classificadores que só são eficazes quando as perdas de rendimento já se consolidaram. A questão central que permanece em aberto na literatura consiste em determinar a partir de que momento os sinais visuais se tornam suficientemente distintos para que um classificador baseado em *deep learning* consiga diferenciar múltiplas deficiências nutricionais com confiança. Responder a essa pergunta exige um *dataset* com anotação temporal explícita e um protocolo experimental que permita isolar o efeito do estágio de desenvolvimento dos sintomas sobre o desempenho classificatório.

Paralelamente à lacuna temporal, o campo da visão computacional tem testemunhado uma transformação arquitetural significativa com o surgimento dos ViTs. Introduzidos por Dosovitskiy et al. (2021), os ViTs processam a imagem como uma sequência de *patches* por meio de mecanismos de autoatenção que permitem a cada elemento da sequência interagir com todos os demais, o que modela dependências globais na imagem. Essa capacidade é teoricamente relevante para o diagnóstico de deficiências nutricionais, uma vez que sintomas como a clorose internerval se manifestam como padrões de descoloração espacialmente difusos que se estendem por grandes áreas da folha. Ademais, foram propostas arquiteturas híbridas que unem convoluções e mecanismos de atenção, com o intuito de conciliar o viés indutivo de localidade das convoluções, eficaz para capturar texturas e padrões de borda, com a modelagem de contexto global oferecida pelos *Transformers* (DAI et al., 2021; GUO et al., 2022). Essa sinergia arquitetural é particularmente promissora para sintomas que evoluem de alterações de cor difusas nos estágios iniciais para lesões morfológicas localizadas nos estágios tardios da deficiência.

A dimensão da interpretabilidade dos modelos constitui uma frente adicional de motivação. Em domínios de aplicação como a agricultura, onde as decisões baseadas nas predições de um modelo podem ter consequências econômicas significativas, a confiança dos usuários finais, como agrônomos, técnicos e produtores, depende não apenas da acurácia numérica, mas também da capacidade de explicar por que o modelo chegou a determinada conclusão. Técnicas de interpretabilidade como o Grad-CAM (SELVARAJU et al., 2017), permitem gerar mapas de calor que destacam as regiões da imagem mais relevantes para a decisão classificatória. No contexto do diagnóstico de deficiências nutricionais, essa

ferramenta possibilita verificar se o modelo foca em regiões agronomicamente coerentes, como margens foliares para K, espaços internervais para Mg, ou se baseia em artefatos irrelevantes, como o fundo da imagem ou variações de iluminação.

A motivação desta pesquisa reside, portanto, na convergência de quatro frentes complementares, quais sejam, a lacuna metodológica representada pela ausência de análise temporal sistemática no diagnóstico de deficiências nutricionais em cafeeiros; a oportunidade de comparar empiricamente o desempenho de arquiteturas convolucionais, baseadas em atenção e híbridas nesse domínio específico; a necessidade de investigar a interpretabilidade dos modelos à luz do conhecimento agrônomo; e a disponibilidade de um *dataset* adequado para essa investigação. O conjunto de dados utilizado neste trabalho, composto por 28.086 registros, correspondentes individualmente a imagens de folhas de *Coffea arabica* com seus metadados, coletadas em ambiente de estufa controlada ao longo de aproximadamente cinco meses (outubro de 2016 a março de 2017), com anotação temporal via *timestamps* e com sete classes, sendo elas o controle com nutrição completa e seis deficiências individuais de macronutrientes (N, P, K, Ca, Mg e S), oferece uma oportunidade singular para essa investigação. A estrutura temporal do *dataset* permite reconstruir a trajetória de desenvolvimento dos sintomas e avaliar, de forma controlada, o impacto do estágio temporal sobre a capacidade discriminativa dos modelos.

1.2 Objetivos

1.2.1 Objetivo Geral

O presente trabalho tem como objetivo geral analisar e comparar o desempenho de arquiteturas de *deep learning*, especificamente CNNs, ViTs e modelos híbridos CNN-ViT, na tarefa de classificação de múltiplas deficiências nutricionais em folhas de *Coffea arabica* a partir de imagens RGB, com foco na evolução temporal dos sintomas. Busca-se, de forma complementar, quantificar a viabilidade da detecção em diferentes estágios de desenvolvimento dos sintomas, o que permite identificar a janela temporal em que os sinais visuais se tornam suficientemente discriminativos para permitir uma classificação confiável. Cabe esclarecer que a dimensão temporal deste trabalho reside no janelamento do período de coleta dos dados de treinamento segundo a evolução dos sintomas, e não em uma classificação de sequências, uma vez que cada imagem é classificada individualmente. Essa investigação visa contribuir para o avanço do conhecimento sobre a aplicabilidade de técnicas de visão computacional no diagnóstico precoce de estresses nutricionais em cafeeiros, o que orienta o desenvolvimento de sistemas de monitoramento mais eficazes para a agricultura de precisão.

1.2.2 Objetivos Específicos

Para alcançar o objetivo geral, a presente pesquisa desdobra-se nos seguintes objetivos específicos:

1. Organizar e particionar temporalmente o *dataset*, com o conjunto de imagens de folhas de café coletadas ao longo de cinco meses em ambiente de estufa controlada distribuído em diferentes frações temporais, definidas como janelas de largura crescente do período experimental.
2. Implementar e avaliar três arquiteturas de *deep learning*, com o treinamento e a avaliação sistemáticos de uma CNN pura (ResNet50), um ViT (ViT-B/16) e um modelo híbrido CNN-ViT customizado, todos inicializados com pesos pré-treinados no ImageNet, sob um protocolo padronizado de treinamento que garanta uma comparação justa entre as arquiteturas. A padronização inclui o mesmo otimizador, taxa de aprendizado, critério de parada antecipada e estratégia de *data augmentation*.
3. Quantificar o impacto da dimensão temporal na classificação, por meio da mensuração sistemática e quantitativa de como a acurácia, o *F1-Score* macro e as métricas por classe variam em função do estágio temporal dos sintomas, o que estabelece a curva de desempenho temporal para cada arquitetura e identifica a janela a partir da qual a classificação se torna viável com confiança aceitável.
4. Investigar o efeito da resolução da imagem, por meio da avaliação do impacto de diferentes resoluções de entrada (314×314 e 512×512 pixels) sobre o desempenho classificatório, o que permite quantificar o *trade-off* entre ganho de acurácia e custo computacional adicional, de modo a orientar decisões práticas de implantação.
5. Aplicar técnicas de interpretabilidade para validação agrônômica, por meio do método Grad-CAM (SELVARAJU et al., 2017) para gerar mapas de ativação que evidenciem as regiões da folha mais relevantes para as decisões classificatórias dos modelos. A análise desses mapas, confrontada com o conhecimento agrônômico sobre a localização típica dos sintomas de cada deficiência, visa avaliar se os modelos focam em regiões compatíveis com a sintomatologia esperada ou se baseiam em artefatos espúrios.

1.3 Hipóteses

A investigação proposta é guiada por três hipóteses centrais, formuladas a partir da revisão da literatura em nutrição vegetal, visão computacional e *deep learning* aplicado à agricultura. Cada hipótese aborda uma dimensão específica do problema e orienta o desenho experimental adotado.

Hipótese H1: Dependência temporal do desempenho classificatório. A primeira hipótese postula que existe relação dependente do estágio temporal dos sintomas entre a janela de dados utilizada para treinamento e a acurácia de classificação de deficiências nutricionais, com expectativa *a priori* de melhoria à medida que as manifestações visuais se tornam mais pronunciadas, mas com a possibilidade de não monotonicidade caso janelas muito amplas incluam dados predominantemente assintomáticos. Essa hipótese fundamenta-se no conhecimento agrônomo consolidado de que as deficiências nutricionais constituem processos graduais, em que, nos estágios iniciais, as alterações visuais são sutis e frequentemente indistinguíveis entre diferentes deficiências, uma vez que a planta ainda dispõe de mecanismos de compensação e redistribuição de nutrientes (MARSCHNER, 2012). O N, por exemplo, sendo um nutriente altamente móvel na planta, é redistribuído das folhas mais velhas para as mais jovens antes que sintomas claros se manifestem, enquanto o Ca, por ser imóvel no floema, acumula sintomas preferencialmente nas folhas mais novas (MALAVOLTA, 2006). Essas diferenças na mobilidade dos nutrientes implicam que o tempo necessário para que os sintomas visuais se tornem discriminativos varia entre as diferentes deficiências, o que adiciona complexidade ao problema de classificação. Espera-se, portanto, que a expansão da janela temporal a partir das configurações mais restritas, limitadas pela escassez de dados, eleve a acurácia até uma faixa intermediária de melhor desempenho, a partir da qual a inclusão de dados de períodos iniciais, predominantemente assintomáticos, tenda a reduzi-la, o que configura uma relação não monotônica entre a largura da janela temporal e o desempenho classificatório. A confirmação dessa hipótese teria implicações diretas para o planejamento de sistemas de monitoramento, ao estabelecer a janela temporal mínima necessária para um diagnóstico confiável.

Hipótese H2: Efeito de arquiteturas com atenção global sobre o desempenho classificatório. A segunda hipótese trata, de forma exploratória, da comparação entre arquiteturas dotadas de mecanismos de atenção global, especificamente os ViTs e os modelos híbridos CNN-ViT, e as CNNs puras na classificação de deficiências nutricionais, sem assumir *a priori* a superioridade de um paradigma sobre o outro. A motivação teórica para investigar essa comparação reside no fato de que os sintomas de deficiências nutricionais frequentemente se manifestam como padrões espacialmente difusos, que se estendem por áreas amplas da folha, o que poderia favorecer a integração de informação de contexto global que os mecanismos de autoatenção dos *Transformers* são projetados para capturar (DOSOVITSKIY et al., 2021; VASWANI et al., 2017). As CNNs, em contrapartida, possuem um viés indutivo de localidade que as torna eficazes para padrões texturais de escala reduzida, mas que poderia limitar sua capacidade de integrar informações distribuídas ao longo de toda a imagem (DOSOVITSKIY et al., 2021). Os modelos híbridos, ao combinarem ambas as abordagens, poderiam capturar tanto padrões locais quanto globais. Cabe registrar que a literatura recente em classificação de imagens naturais tem demonstrado que a superioridade dos *Transformers* sobre as CNNs não é universal e depende

fortemente do volume de dados disponível, do viés indutivo das arquiteturas e do grau de pré-treinamento (TOUVRON et al., 2021). A comparação empírica dessas famílias arquiteturais no domínio específico de deficiências nutricionais em café é, portanto, conduzida com caráter exploratório, com o objetivo de verificar se as vantagens teóricas dos mecanismos de atenção se traduzem, ou não, em ganhos práticos de desempenho neste contexto.

Hipótese H3: Interpretabilidade e coerência agrônômica dos mapas de ativação. A terceira hipótese propõe que a análise de interpretabilidade, conduzida por meio do Grad-CAM, revelará que os modelos focam em regiões da folha que são agronomicamente consistentes com a sintomatologia conhecida de cada deficiência nutricional. Especificamente, espera-se que os mapas de ativação destaquem regiões como margens foliares, espaços internervais e o limbo foliar de forma generalizada (MALAVOLTA, 2006). Ademais, formula-se a expectativa de que a distribuição espacial da atenção varie em função do estágio temporal, em que, nos estágios iniciais, quando os sintomas são mais difusos, os mapas de ativação tendem a cobrir áreas mais amplas da folha, ao passo que nos estágios tardios, com lesões mais localizadas, a atenção se concentraria em regiões específicas. Essa hipótese é relevante não apenas para validar a confiabilidade dos modelos, mas também para avaliar seu potencial como ferramentas de apoio à decisão agrônômica, em que a transparência do processo decisório é fundamental para a adoção em contextos produtivos reais. A confirmação de que os modelos focam em regiões compatíveis com a sintomatologia esperada fortaleceria a argumentação em favor da sua utilização prática, enquanto a detecção de padrões de atenção inconsistentes com o conhecimento agrônômico sinalizaria a necessidade de refinamentos no *pipeline* de treinamento, como técnicas de regularização ou estratégias de *data augmentation* direcionadas. Nesse sentido, a análise de interpretabilidade desempenha um papel duplo, de validação científica dos modelos e de orientação para o seu aprimoramento.

1.4 Contribuições Esperadas

As contribuições esperadas deste trabalho organizam-se em três eixos complementares, o metodológico, o empírico e o aplicado, que, em conjunto, visam avançar o estado da arte na interseção entre visão computacional e o monitoramento de deficiências nutricionais.

Contribuição metodológica. No eixo metodológico, a principal contribuição reside no desenvolvimento e na formalização de um *pipeline* experimental para a análise temporal da classificação de estresses em plantas com o uso de *deep learning*. A estratégia de particionamento temporal, que permite variar sistematicamente a fração dos dados temporais incluída no treinamento e na avaliação, constitui uma abordagem reprodutível e parametrizável que pode ser aplicada a outros *datasets* agrícolas com anotação temporal, independentemente da cultura ou do tipo de estresse investigado. Diferentemente

de abordagens que categorizam manualmente os estágios de evolução dos sintomas em categorias discretas, como leve, moderado e severo, a estratégia proposta opera sobre a dimensão temporal contínua, o que preserva a granularidade da informação e permite uma análise mais fina da detectabilidade ao longo do tempo. Ademais, o presente trabalho documenta de forma detalhada as decisões metodológicas relativas à estratégia de validação, bem como a análise da aplicabilidade do protocolo *Leave-One-Group-Out* (LOGO) por vaso, que se revelou inadequado devido à alta variância temporal entre plantas do mesmo vaso, e a justificativa fundamentada para a adoção de *hold-out* estratificado 80/20. Esse registro transparente dos desafios e das limitações encontrados no processo de avaliação de modelos em dados agrícolas com estrutura hierárquica constitui, por si só, uma contribuição relevante, dado que muitos trabalhos na literatura omitem as tentativas malsucedidas de validação, e privam a comunidade de informações valiosas para o desenho de futuros experimentos.

Contribuição empírica. No eixo empírico, este trabalho oferece uma comparação sistemática e quantitativa entre três famílias arquiteturais de *deep learning* (CNNs, ViTs e modelos híbridos CNN-ViT) aplicadas à classificação de deficiências nutricionais em café ao longo de diferentes estágios temporais. A relevância de comparações rigorosas entre arquiteturas nesse domínio é ilustrada pelo estudo de Pushpa, Akash e Akash (2025), que confronta sistematicamente técnicas de aprendizado de máquina e de *deep learning* para a detecção de deficiências nutricionais em plantas e evidencia a superioridade das abordagens convolucionais sobre os métodos clássicos. Essa comparação preenche uma lacuna importante na literatura, em que os estudos existentes tipicamente avaliam uma única arquitetura em um único ponto temporal, o que impede conclusões sobre a adequação relativa de diferentes abordagens a diferentes condições diagnósticas. Os resultados empíricos gerados incluem curvas de desempenho temporal para cada arquitetura, o que permite visualizar como a acurácia evolui em função do estágio de desenvolvimento dos sintomas; análises detalhadas por classe que revelam a hierarquia de dificuldade entre as diferentes deficiências e identificam quais nutrientes apresentam sintomas mais facilmente discrimináveis, e uma avaliação quantitativa do impacto da resolução da imagem de entrada sobre a acurácia classificatória. Em conjunto, essas evidências empíricas constituem um corpo de conhecimento que pode orientar tanto futuras pesquisas acadêmicas, ao identificar as questões que permanecem em aberto, quanto aplicações práticas, ao informar a seleção de arquiteturas e configurações para sistemas de monitoramento em campo.

Contribuição aplicada. No eixo aplicado, o trabalho gera conhecimento potencialmente relevante para o desenvolvimento futuro de sistemas automatizados de monitoramento nutricional em lavouras de café. Ao indicar, sob as condições controladas do experimento, a faixa em que os sinais visuais se tornam discriminativos, os resultados oferecem subsídios preliminares para o planejamento de protocolos de monitoramento, sem extrapolação imediata para condições de campo, cuja validação é registrada como pré-requisito para

a tradução das observações deste trabalho em recomendações operacionais. A análise de interpretabilidade via Grad-CAM, ao validar que os modelos focam em regiões agronomicamente relevantes, reforça a confiabilidade das predições e facilita a adoção dessas tecnologias por profissionais do setor agropecuário, que necessitam de ferramentas não apenas acuradas, mas também explicáveis. Dessa forma, o presente trabalho contribui para a transição de um paradigma reativo de gestão nutricional, baseado na observação tardia de sintomas avançados, para um paradigma proativo, fundamentado na detecção precoce assistida por inteligência artificial.

1.5 Organização do Trabalho

A presente dissertação está estruturada em cinco capítulos, organizados de forma a conduzir progressivamente os fundamentos teóricos até os resultados experimentais e as conclusões da pesquisa. No Capítulo 1, ora apresentado, é contextualizado o problema de pesquisa no cenário da cafeicultura brasileira, é explicitada a motivação fundamentada na lacuna temporal identificada na literatura, são formalizados os objetivos gerais e específicos e são enunciadas as três hipóteses que orientam a investigação, bem como as contribuições esperadas nos eixos metodológico, empírico e aplicado.

No Capítulo 2 é apresentada a fundamentação teórica que sustenta a investigação, organizada em três eixos temáticos principais. O primeiro eixo aborda a nutrição mineral de cafeeiros ao detalhar o papel fisiológico dos macronutrientes, os mecanismos de manifestação e progressão dos sintomas de deficiência, e os métodos tradicionais de diagnóstico, desde a análise laboratorial até a inspeção visual. O segundo eixo cobre os fundamentos de *deep learning* aplicados à visão computacional, tais como a arquitetura de CNNs e seus componentes, como camadas convolucionais, *pooling*, normalização em lote, o mecanismo de autoatenção dos *Transformers* e sua adaptação para processamento de imagens nos ViTs, e as abordagens híbridas que combinam ambos os paradigmas. O terceiro eixo apresenta as técnicas de interpretabilidade de modelos, com destaque para o Grad-CAM e sua aplicabilidade na validação de predições em domínios agrícolas. Além disso, é apresentada uma revisão dos trabalhos relacionados que empregam aprendizado profundo para a classificação de estresses em plantas, ao situar a presente pesquisa em relação ao estado da arte e identificar explicitamente as lacunas que este trabalho se propõe a preencher.

No Capítulo 4 é descrita em detalhes a metodologia empregada, como a caracterização completa do *dataset*, com suas sete classes, organização temporal, o protocolo de pré-processamento, recorte central, redimensionamento, normalização e *data augmentation*, a estratégia de particionamento temporal e a configuração detalhada das três arquiteturas avaliadas (ResNet50, ViT-B/16 e o modelo híbrido CNN-ViT customizado) e o protocolo padronizado de treinamento e avaliação adotado.

No Capítulo 5 são apresentados e discutidos os resultados experimentais obtidos ao confrontar cada uma das três hipóteses formuladas com as evidências empíricas, comparar o desempenho das diferentes arquiteturas em cada fração temporal, analisar as métricas por classe para identificar as deficiências de maior e menor dificuldade classificatória, avaliar o efeito da resolução da imagem e examinar os mapas de interpretabilidade gerados pelo Grad-CAM à luz do conhecimento agrônômico.

Por fim, no Capítulo 6 são sintetizadas as principais descobertas da pesquisa, com a retomada das hipóteses e dos objetivos para verificar seu grau de atendimento, são discutidas as limitações do estudo e são apontadas direções promissoras para trabalhos futuros, bem como a possibilidade de extensão da abordagem temporal para outras culturas e tipos de estresse.

Fundamentação Teórica

Neste capítulo são apresentados os conceitos fundamentais que sustentam a presente pesquisa. A estrutura parte dos fundamentos agronômicos do problema, isto é, a nutrição mineral em plantas de café e a sintomatologia associada às deficiências, avança para as tecnologias de aquisição e processamento de imagens digitais na agricultura, e aprofunda-se nas arquiteturas de aprendizado profundo empregadas, nomeadamente as Redes Neurais Convolucionais, os *Vision Transformers* e os modelos híbridos. Em seguida, são discutidas as técnicas de interpretabilidade e explicabilidade que conferem transparência aos modelos, e, por fim, são apresentadas as análises dos trabalhos relacionados, com a identificação das lacunas que esta pesquisa visa preencher. A construção deste referencial teórico busca fornecer ao leitor uma compreensão abrangente e autocontida dos pilares sobre os quais o trabalho se sustenta.

2.1 Nutrição Mineral em Plantas de Café

A nutrição mineral constitui um dos pilares da fisiologia vegetal e da produção agrícola, determinante para o crescimento, o desenvolvimento e a produtividade das culturas. As plantas necessitam de elementos minerais essenciais, classificados em macronutrientes, Nitrogênio (N), Fósforo (P), Potássio (K), Cálcio (Ca), Magnésio (Mg) e Enxofre (S), e micronutrientes, Ferro (Fe), Manganês (Mn), Zinco (Zn), Cobre (Cu), Boro (B), Molibdênio (Mo) e Cloro (Cl), cada qual com funções metabólicas específicas e insubstituíveis. No caso do cafeeiro (*Coffea arabica*), a demanda nutricional é particularmente elevada devido ao ciclo fenológico complexo dessa espécie perene, que envolve fases simultâneas de crescimento vegetativo e reprodutivo, impondo uma competição interna por fotoassimilados e nutrientes minerais (MALAVOLTA, 2006). A deficiência de qualquer um desses elementos essenciais compromete funções vitais da planta, como a fotossíntese, a síntese proteica e a regulação osmótica, o que resulta em perdas de rendimento que podem ser irreversíveis quando o diagnóstico é tardio (MARSCHNER, 2012). Nesse contexto, a compreensão detalhada dos mecanismos de absorção, translocação e função de cada nutriente

é pré-requisito para qualquer abordagem de diagnóstico, seja ela tradicional ou baseada em técnicas computacionais.

A mobilidade dos nutrientes no floema é o princípio agrônômico mais relevante para a interpretação dos padrões visuais de deficiência, pois determina em quais órgãos da planta os sintomas se manifestam primeiro. Nutrientes com alta mobilidade no floema, como N, P, K, Mg e S, são redistribuídos pela planta a partir de tecidos maduros para suprir as demandas dos tecidos em crescimento ativo, conseqüentemente, os sintomas de deficiência desses elementos aparecem inicialmente nas folhas mais velhas, situadas na porção inferior da copa. Em contrapartida, nutrientes com mobilidade reduzida no floema apresentam sintomas preferencialmente nas folhas mais jovens, sendo que elementos pouco móveis, como Fe, Mn, Zn e B, são redistribuídos de forma limitada, enquanto elementos quase imóveis, como Ca, praticamente não se redistribuem, de modo que os sintomas surgem nas folhas mais novas, no ápice da planta (MALAVOLTA, 2006; MARSCHNER, 2012). Essa dicotomia entre sintomas em folhas velhas e folhas novas constitui o primeiro nível de discriminação no diagnóstico visual. A compreensão desse mecanismo agrônômico é fundamental para interpretar tanto os acertos quanto os erros dos classificadores avaliados neste trabalho, ainda que a posição da folha na copa não constitua um atributo fornecido aos modelos, que operam sobre recortes foliares individuais.

Cada deficiência nutricional produz uma sintomatologia visual característica que, em estágios avançados, permite a identificação por profissionais experientes. A deficiência de N manifesta-se como clorose generalizada e uniforme das folhas mais velhas, que adquirem coloração verde-amarelada pálida, com progressão para senescência precoce e abscisão foliar, dado o papel central do N na síntese de clorofila e proteínas (MALAVOLTA, 2006; MARSCHNER, 2012). A carência de P provoca escurecimento das folhas, que podem assumir tonalidade verde-escura, por vezes atingindo coloração arroxeada em algumas espécies, com redução do crescimento radicular e aéreo. A deficiência de K é reconhecida pela necrose marginal das folhas mais velhas, conhecida como “queima de bordas”, frequentemente acompanhada de clorose internerval. O Ca, por sua vez, é essencial para a estrutura da parede celular e, em sua ausência, as folhas jovens apresentam deformações, enrolamento e necrose apical. A deficiência de Mg, componente central da molécula de clorofila, gera clorose internerval nas folhas maduras, com as nervuras em tonalidade verde e o limbo amarelado ou avermelhado. Por fim, a carência de S produz clorose uniforme, cujos sintomas tendem a manifestar-se preferencialmente nas folhas mais novas quando o N está presente em níveis suficientes, com aspecto visual semelhante ao da deficiência de N em estágios iniciais, o que constitui uma das principais fontes de confusão diagnóstica (MALAVOLTA, 2006; MARSCHNER, 2012). Na Figura 1 são apresentados exemplos representativos dessas manifestações em folhas de *Coffea arabica*, que evidenciam a diversidade de padrões fenotípicos a serem discriminados pelos classificadores avaliados.

Um aspecto central desta pesquisa é a evolução temporal dos sintomas de deficiência.

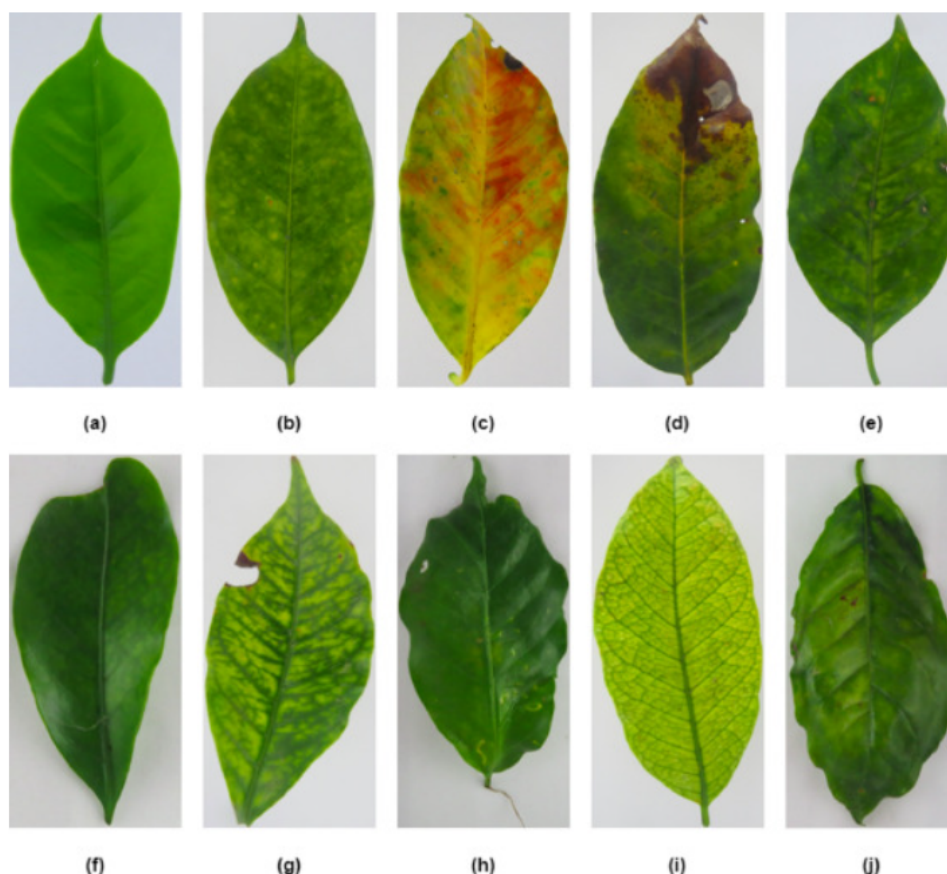


Figura 1 – Exemplos de folhas de café (*Coffea arabica*) com diferentes deficiências nutricionais: (a) folha saudável (Completo); (b) deficiência de N; (c) deficiência de P; (d) deficiência de K; (e) deficiência de Mg; (f) deficiência de B; (g) deficiência de Mn; (h) deficiência de Ca; (i) deficiência de Fe; (j) deficiências múltiplas. As alterações de coloração, textura e distribuição das lesões constituem padrões visuais característicos das deficiências nutricionais em cafeeiros. A figura é exibida com finalidade ilustrativa da diversidade de fenótipos foliares descritos na literatura; o conjunto de classes investigado neste trabalho corresponde a Completo, N, P, K, Ca, Mg e S, conforme detalhado no Capítulo 4, e não inclui as deficiências de B, Mn, Fe nem cenários de múltiplas deficiências aqui ilustrados. Adaptado de: Tuesta-Monteza, Mejia-Cabrera e Arcila-Diaz (2023).

Uma carência nutricional não se manifesta de forma abrupta, ao contrário, ela evolui progressivamente de um estado subclínico, onde as alterações bioquímicas já ocorrem, mas ainda não são visíveis, para sintomas iniciais sutis, como leves alterações de tonalidade, até atingir padrões fenotípicos claros e altamente discriminativos, como necroses extensas e deformações severas. Essa dinâmica temporal cria um desafio fundamental para sistemas de classificação automatizada, pois nos estágios iniciais diferentes deficiências produzem fenótipos visualmente similares, como a clorose difusa causada pela falta de N, que pode ser confundida com a causada pela falta de S ou até mesmo com o envelhecimento natural da folha, ao passo que, em estágios avançados, os sintomas se tornam progressivamente mais específicos e distinguíveis (MARSCHNER, 2012). As revisões abrangentes

de Upadhyay et al. (2025) e Mehta et al. (2025) documentam o predomínio do *deep learning* no diagnóstico de estresses em plantas. Não obstante, a abordagem dominante na literatura trata o diagnóstico como um problema de classificação estática, com modelos treinados e avaliados a partir de imagens de sintomas já estabelecidos, o que descarta a informação diagnóstica contida na evolução temporal dos sintomas. A presente pesquisa aborda diretamente essa lacuna ao investigar como o desempenho de classificação se relaciona com o estágio temporal de desenvolvimento dos sintomas.

A interação entre deficiências múltiplas e a influência de fatores ambientais adicionam camadas de complexidade ao diagnóstico visual. Na prática agrícola, é raro que uma planta sofra de uma única deficiência isolada, frequentemente, o desequilíbrio de um nutriente afeta a absorção ou a utilização de outros, o que gera quadros sintomatológicos compostos que desafiam até os profissionais mais experientes. Por exemplo, a deficiência de Ca pode ser exacerbada por excesso de K, que, em concentrações elevadas, exerce antagonismo iônico e reduz a quantidade de Ca absorvida (MALAVOLTA, 2006), e a carência de Mg é frequentemente observada em solos ácidos onde o alumínio tóxico inibe a absorção de cátions divalentes. Ademais, fatores como estresse hídrico, temperatura extrema e ataque de pragas podem mimetizar ou mascarar sintomas nutricionais, o que aumenta a ambiguidade do diagnóstico visual (MARSCHNER, 2012). No contexto experimental da presente pesquisa, o cultivo em solução nutritiva com deficiências individuais controladas, cada planta submetida à omissão de um único nutriente, elimina essas interações confusas e cria um cenário ideal para avaliar a capacidade dos modelos de discriminar sintomas puros de cada deficiência. Contudo, a transferência dos modelos treinados nesse ambiente controlado para condições de campo permanece um desafio a ser abordado em trabalhos futuros.

Os métodos tradicionais de diagnóstico nutricional em cafeeiros incluem a inspeção visual por agrônomos, a análise foliar por via úmida e a espectrometria de emissão atômica com plasma indutivamente acoplado, Espectrometria de Plasma Indutivamente Acoplado (ICP) (HANSEN et al., 2013). A inspeção visual, embora amplamente praticada, é subjetiva e particularmente falha nos estágios iniciais de deficiência, ao passo que a análise laboratorial e a ICP, apesar de constituírem o padrão-ouro quantitativo, são custosas, demoradas e destrutivas, o que inviabiliza o monitoramento contínuo em larga escala. Essas limitações motivam o desenvolvimento de abordagens não invasivas e automatizadas, baseadas em imagens digitais e aprendizado profundo, capazes de fornecer diagnósticos rápidos e escaláveis que complementem as técnicas tradicionais.

2.2 Diagnóstico por Imagens Digitais na Agricultura

A utilização de imagens digitais para o diagnóstico agrícola representa uma mudança de paradigma em relação aos métodos laboratoriais tradicionais, com a promessa de di-

agnósticos não destrutivos, rápidos e potencialmente escaláveis. O histórico dessa abordagem remonta às décadas de 1980 e 1990, com os primeiros sistemas de processamento digital de imagens aplicados à classificação de cobertura do solo via sensoriamento remoto por satélite. Contudo, a popularização em larga escala das redes neurais profundas e a disponibilização de grandes conjuntos de dados de imagens marcaram a transição de abordagens baseadas em engenharia manual de características, as *handcrafted features*, para a extração automática de representações por meio de *deep learning* (LIAKOS et al., 2018). Desde então, a comunidade científica tem explorado intensivamente o uso de imagens capturadas por câmeras convencionais, drones e *smartphones* para tarefas que vão desde a detecção de doenças foliares até a estimativa de produtividade e a avaliação do estado hídrico das culturas (LIAKOS et al., 2018).

A escolha da modalidade de imagem é um fator crítico que influencia diretamente o desempenho e a aplicabilidade de qualquer sistema de visão computacional agrícola. Imagens no espectro visível (RGB) são a modalidade mais acessível, capturadas por câmeras convencionais, drones comerciais e smartphones. Seu baixo custo e facilidade de aquisição as tornam ideais para aplicações em larga escala e para a disseminação de ferramentas diagnósticas em regiões com infraestrutura limitada. No entanto, imagens RGB, por operarem apenas no espectro visível, carecem de sensibilidade direta aos marcadores bioquímicos associados ao estresse fisiológico em estágios muito precoces, quando as alterações ainda não se traduziram em mudanças de coloração perceptíveis (SHOAIB et al., 2026). Em contrapartida, sensores hiperespectrais capturam centenas de bandas do espectro eletromagnético, enquanto sensores multiespectrais operam em um número menor de bandas selecionadas, tipicamente entre quatro e doze, abrangendo regiões como o infravermelho próximo e o infravermelho de ondas curtas, o que permite a detecção de alterações sutis na reflectância foliar associadas a variações no teor de clorofila, água e pigmentos carotenoides antes que sintomas sejam visíveis a olho nu. Contudo, o custo elevado dos sensores, a necessidade de calibração radiométrica rigorosa e a complexidade do processamento de dados hiperespectrais de alta dimensionalidade constituem barreiras significativas para a adoção massiva dessas tecnologias, especialmente por pequenos e médios produtores (SHOAIB et al., 2026).

A plataforma de aquisição constitui outro eixo de variabilidade que impacta a qualidade dos sistemas de diagnóstico por imagem. Plataformas terrestres, como câmeras montadas em estações fixas ou operadas manualmente, oferecem resolução espacial elevada e controle sobre as condições de iluminação, o que as torna particularmente adequadas para a captura de detalhes finos da superfície foliar, como alterações de textura associadas a estágios iniciais de deficiência. Em plataformas menos controladas, como drones e *smartphones*, a variabilidade nas condições de captura, como iluminação, ângulo, distância ao objeto e fundo da imagem, introduz viés de captura que pode comprometer a generalização dos modelos e exige estratégias robustas de pré-processamento e aumento

de dados (NOYAN, 2022).

Os desafios associados à padronização de dados constituem uma preocupação transversal a todas as modalidades e plataformas de aquisição. A variabilidade nas condições ambientais, como intensidade e direção da iluminação solar, presença de sombras e temperatura de cor, afeta diretamente as propriedades cromáticas das imagens, o que pode introduzir vieses que comprometem a generalização dos modelos. Estratégias de mitigação incluem a captura em condições controladas, como estufas e câmaras de imageamento, e técnicas computacionais de normalização de cor (WANG et al., 2025). Além disso, a variabilidade biológica intrínseca, como diferenças de genótipo, idade da folha, posição na copa e estado hídrico, adiciona uma camada de complexidade (UBBENS; STAVNESS, 2017) que motiva o uso de conjuntos de dados diversificados e de estratégias de aumento de dados, de modo a reduzir a influência de confundidores e favorecer a generalização (SHOAIB et al., 2025).

Este trabalho adota imagens RGB capturadas em ambiente de estufa controlada, uma decisão metodológica que minimiza a variabilidade ambiental e permite isolar o efeito das deficiências nutricionais sobre a aparência foliar. A captura foi realizada ao longo de aproximadamente cinco meses (outubro de 2016 a março de 2017), com registro temporal via *timestamps* nos diretórios de armazenamento, o que gera um dataset longitudinal de 28.086 registros distribuídos em sete classes (Completo, N, P, K, Ca, Mg, S) e quatro vasos por classe. A contribuição diferencial reside não na modalidade de imagem empregada, mas na exploração sistemática da dimensão temporal dos dados. Em vez de tratar o diagnóstico como um problema de classificação estática, abordagem predominante na literatura, a presente pesquisa investiga como o desempenho de classificação evolui à medida que os sintomas se desenvolvem ao longo do tempo, e quantifica a partir de que momento os sinais visuais se tornam suficientemente distintos para permitir a discriminação automatizada entre deficiências.

2.3 CNNs

2.3.1 Fundamentos

As CNNs constituem a classe de modelos que dominou o campo da visão computacional durante a última década, com fundamentos estabelecidos por LeCun et al. (1998) com a arquitetura LeNet-5, originalmente projetada para o reconhecimento de dígitos manuscritos. A arquitetura de uma CNN é inspirada na organização hierárquica do córtex visual de mamíferos e se baseia em três operações fundamentais, sendo elas a convolução, o agrupamento, *pooling*, e a ativação não linear. A operação de convolução aplica filtros aprendíveis, denominados *kernels*, que deslizam sobre o volume de entrada para produzir mapas de características, os *feature maps*, cada filtro responsável por detectar um padrão

especial específico (DUMOULIN; VISIN, 2018). Em camadas iniciais, esses filtros tipicamente respondem a bordas, gradientes e texturas simples; em camadas mais profundas, combinações de filtros capturam padrões progressivamente mais abstratos e semanticamente ricos, como formas geométricas, partes de objetos e, no contexto desta pesquisa, padrões de clorose e necrose em folhas (GOODFELLOW; BENGIO; COURVILLE, 2016).

A operação de *pooling*, tipicamente *max pooling* ou *average pooling*, reduz a dimensionalidade espacial dos mapas de características, o que promove invariância local a pequenas translações e deformações (SCHERER; MÜLLER; BEHNKE, 2010), além de diminuir o custo computacional das camadas subsequentes (GOODFELLOW; BENGIO; COURVILLE, 2016). Ao condensar a informação de regiões vizinhas em um único valor representativo, o *pooling* permite que a rede mantenha as informações mais salientes enquanto descarta redundâncias espaciais. Essa operação, combinada com a aplicação sucessiva de convoluções, confere às CNNs dois vieses indutivos fundamentais, sendo eles a localidade, em que cada unidade responde apenas a uma região restrita da entrada, denominada campo receptivo, *receptive field*, e a invariância translacional, em que o compartilhamento de pesos garante que o mesmo filtro detecta o mesmo padrão independentemente de sua posição na imagem (GOODFELLOW; BENGIO; COURVILLE, 2016). Esses vieses indutivos tornam as CNNs extremamente eficientes em termos de parâmetros, o que permite que modelos relativamente compactos alcancem desempenho elevado mesmo com conjuntos de dados moderados. Na Figura 2 é ilustrado um exemplo de arquitetura CNN para classificação de imagens, com as etapas de convolução, *pooling* e classificação por camadas totalmente conectadas.

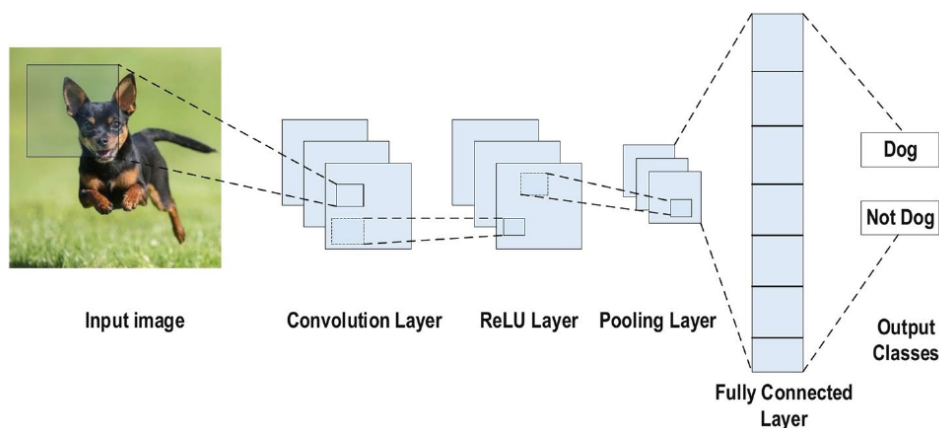


Figura 2 – Exemplo de arquitetura CNN para classificação de imagens: a imagem de entrada é processada por camadas de convolução com ativação ReLU, seguidas de camadas de *pooling* para redução espacial, e por fim por camadas totalmente conectadas, *fully connected*, que produzem a classificação final. Adaptado de: Alzubaidi et al. (2021).

As funções de ativação não linear são componentes essenciais que introduzem capacidade de aproximação de funções complexas nas redes neurais. A ReLU, definida como

$f(x) = \max(0, x)$, tornou-se a função de ativação padrão em CNNs modernas devido à sua simplicidade computacional e à mitigação do problema do gradiente desvanecente, o *vanishing gradient*, que dificultava o treinamento de redes profundas com funções sigmoideais e tangente hiperbólica (GOODFELLOW; BENGIO; COURVILLE, 2016). A ReLU introduz esparsidade nas ativações, pois neurônios com entrada negativa são desativados, o que contribui para representações mais eficientes e interpretáveis. Variantes como a Leaky ReLU e a ELU atribuem valores não nulos a entradas negativas, em vez de desativá-las completamente como a ReLU, o que reduz o deslocamento de viés nas ativações (CLEVERT; UNTERTHINER; HOCHREITER, 2016), enquanto a GELU pondera as entradas por seu valor em vez de apenas pelo sinal (HENDRYCKS; GIMPEL, 2016), sendo a ativação empregada nas camadas *feed-forward* da arquitetura ViT (DOSOVITSKIY et al., 2021). A ReLU original, contudo, permanece predominante em arquiteturas residuais como a ResNet, empregada neste trabalho.

A técnica de *Batch Normalization* (BN), proposta por Ioffe e Szegedy (2015), representou um avanço no treinamento de redes profundas ao normalizar as ativações de cada camada para ter média zero e variância unitária dentro de cada *mini-batch*. Essa normalização reduz o fenômeno de deslocamento interno de covariáveis, o *internal covariate shift*, o que permite o uso de taxas de aprendizado mais elevadas, acelera a convergência e atua como regularizador implícito. Em arquiteturas modernas, camadas de BN são inseridas sistematicamente após cada convolução e antes da função de ativação, e constituem um componente padrão do bloco convolucional. Ademais, a técnica de *Dropout*, proposta por Srivastava et al. (2014), complementa a regularização ao desativar aleatoriamente uma fração dos neurônios durante o treinamento, o que força a rede a aprender representações mais robustas e distribuídas, menos dependentes de neurônios individuais. A combinação de BN e *Dropout* constitui a estratégia de regularização predominante em CNNs contemporâneas e é empregada nas arquiteturas avaliadas nesta dissertação.

O treinamento de CNNs é realizado por meio do algoritmo de retropropagação (*back-propagation*), que calcula o gradiente da função de perda em relação a cada parâmetro da rede por meio da regra da cadeia, e de algoritmos de otimização baseados em gradiente, como o *Stochastic Gradient Descent* (SGD) e suas variantes adaptativas (Adam, AdaGrad, RMSProp). O otimizador Adam, utilizado neste trabalho, combina as vantagens do momentum e da adaptação por parâmetro da taxa de aprendizado, particularmente eficaz para paisagens de perda ruidosas e não convexas típicas de redes profundas (GOODFELLOW; BENGIO; COURVILLE, 2016). A escolha da função de perda, neste caso a entropia cruzada categórica, *cross-entropy loss*, e a estratégia de regularização influenciam diretamente a capacidade do modelo de generalizar para dados não vistos durante o treinamento, um aspecto crítico em aplicações agrícolas em que a variabilidade natural das amostras é elevada.

A função de perda *Cross-Entropy*, definida pela Equação 1, penaliza proporcional-

mente a confiança do modelo em predições incorretas, o que incentiva distribuições de probabilidade concentradas na classe correta.

$$\mathcal{L} = - \sum_{c=1}^C y_c \log(\hat{y}_c) \quad (1)$$

Nessa formulação, y_c é o rótulo verdadeiro em codificação *one-hot* e $\hat{y}_c$ é a probabilidade predita para a classe c . Em problemas de classificação multiclasse como o desta pesquisa, com sete classes de deficiência nutricional, a saída da rede passa por uma camada *softmax* que normaliza os *logits* em uma distribuição de probabilidade, e a *Cross-Entropy* é calculada sobre essa distribuição. A combinação de *softmax* com *Cross-Entropy* possui propriedades numéricas favoráveis, com gradientes simples e estáveis, o que contribui para a convergência eficiente do treinamento. Ademais, estratégias como o ajuste adaptativo da taxa de aprendizado por estagnação da perda de validação, conhecido como *learning rate scheduling*, e a parada antecipada, *early stopping*, que interrompe o treinamento quando a perda de validação não melhora por um número definido de épocas, são mecanismos adicionais de regularização que previnem o sobreajuste e garantem que o modelo retido seja aquele com melhor capacidade de generalização (GOODFELLOW; BENGIO; COURVILLE, 2016).

2.3.2 Arquiteturas Residuais e a ResNet

O treinamento de CNNs muito profundas era, até 2015, limitado pelo chamado problema da degradação, *degradation problem*, em que, ao aumentar o número de camadas além de um certo limiar, a acurácia de treinamento, e não apenas a de validação, começava a diminuir, o que contrariava a expectativa de que redes mais profundas, por possuírem maior capacidade representacional, deveriam aprender pelo menos tão bem quanto redes mais rasas. Esse fenômeno, distinto do sobreajuste, *overfitting*, indicava uma dificuldade de otimização, e não de capacidade, o que sugere que o algoritmo de retropropagação não conseguia encontrar soluções eficientes em paisagens de perda extremamente profundas. A arquitetura ResNet, proposta por He et al. (2016), abordou esse problema de forma elegante por meio da introdução de conexões residuais, *skip connections*, que permitem que o sinal de entrada de um bloco seja somado diretamente à sua saída, de modo que a rede aprende o resíduo em vez da transformação completa $H(x)$, conforme a Equação 2.

$$\mathcal{F}(x) = H(x) - x \quad (2)$$

A intuição por trás das conexões residuais é que, se a transformação identidade $H(x) = x$ for próxima da solução ótima, é mais fácil para a rede aprender a zerar os pesos da função residual $\mathcal{F}(x) \approx 0$ do que aprender a transformação identidade por meio de uma sequência de camadas convolucionais não lineares. Formalmente, a saída de um bloco residual é definida pela Equação 3, em que $\mathcal{F}$ representa as operações do bloco, como

convoluções, BN e ReLU, e x é a entrada que “atalha” essas operações.

$$y = \mathcal{F}(x, \{W_i\}) + x \quad (3)$$

Quando as dimensões de $\mathcal{F}(x)$ e x diferem, por exemplo ao mudar o número de canais ou a resolução espacial, uma projeção linear W_s é aplicada à entrada, conforme a Equação 4 (HE et al., 2016).

$$y = \mathcal{F}(x, \{W_i\}) + W_s x \quad (4)$$

Essas conexões de atalho garantem que os gradientes possam fluir diretamente da função de perda até as camadas iniciais da rede, o que mitiga o desvanecimento e a explosão de gradientes e permite o treinamento eficaz de redes com centenas de camadas. Na Figura 3 é ilustrada a estrutura do bloco residual, em que a conexão de atalho soma a entrada $\mathbf{x}$ diretamente à saída da transformação residual $\mathcal{F}(\mathbf{x})$.

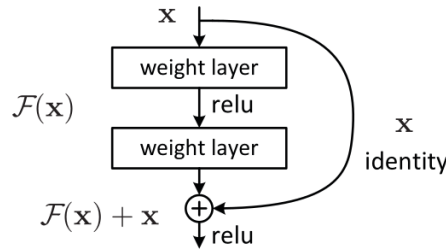


Figura 3 – Bloco residual, *building block*, da ResNet: a conexão de atalho, *skip connection*, soma a entrada $\mathbf{x}$ diretamente à saída da transformação residual $\mathcal{F}(\mathbf{x})$, o que permite que o gradiente flua sem atenuação pelas camadas. Adaptado de: He et al. (2016).

A ResNet50, variante com 50 camadas utilizada neste trabalho, emprega blocos residuais do tipo *bottleneck*, que consistem em três convoluções empilhadas, sendo elas uma convolução 1×1 para redução de dimensionalidade, uma convolução 3×3 para extração de características espaciais e uma convolução 1×1 para expansão dos canais. Essa arquitetura em gargalo, *bottleneck*, reduz significativamente o número de parâmetros e o custo computacional em comparação com blocos residuais simples de duas convoluções 3×3 , ao mesmo tempo em que mantém, ou até melhora, a capacidade representacional. A família ResNet, na variante de 152 camadas, venceu a competição ImageNet Large Scale Visual Recognition Challenge (ILSVRC) de 2015 (HE et al., 2016), e a ResNet50 consolidou-se como a variante mais amplamente adotada na literatura pelo equilíbrio entre capacidade representacional e custo computacional, empregada tanto como modelo final quanto como *backbone* para extração de características em arquiteturas mais complexas.

A família ResNet compreende variantes com diferentes profundidades (ResNet18, ResNet34, ResNet50, ResNet101, ResNet152), cada uma com um compromisso distinto entre capacidade representacional e custo computacional. Estudos empíricos demonstraram que

a acurácia melhora consistentemente com o aumento da profundidade até certo ponto, mas com retornos decrescentes, visto que a ResNet152 oferece ganhos marginais sobre a ResNet101, que por sua vez oferece ganhos modestos sobre a ResNet50, enquanto o custo computacional cresce linearmente com o número de camadas (HE et al., 2016). A ResNet50 figura entre as variantes de ResNet mais adotadas na literatura (ALZUBAIDI et al., 2021), com capacidade representacional suficiente para a maioria das tarefas de classificação a um custo computacional gerenciável. Além disso, a ampla disponibilidade de pesos pré-treinados de alta qualidade para a ResNet50, obtidos com técnicas modernas de aumento de dados e regularização, torna essa arquitetura uma escolha pragmática para pesquisas em domínios especializados.

A estratégia de transferência de aprendizado é particularmente relevante para o emprego de ResNets em domínios agrícolas, onde conjuntos de dados rotulados são tipicamente menores do que os utilizados em *benchmarks* de visão computacional geral. Nessa abordagem, a rede é inicializada com pesos pré-treinados em um grande conjunto de dados genérico, neste caso o ImageNet, com mais de um milhão de imagens em mil categorias, e subsequentemente ajustada, num processo de *fine-tuning*, para a tarefa-alvo. As camadas iniciais capturam características de baixo nível, como bordas e texturas, enquanto as camadas finais, mais específicas da tarefa, são as preferencialmente ajustadas durante o *fine-tuning* para discriminar as classes do novo problema (ALZUBAIDI et al., 2021). A configuração de transferência de aprendizado adotada para a ResNet50 neste trabalho, incluindo a variante de pesos pré-treinados e a adaptação da camada de classificação, é detalhada no Capítulo 4.

2.3.3 Extração Hierárquica de Características

Uma propriedade fundamental das CNNs, e em particular das arquiteturas profundas como a ResNet, é a extração hierárquica de características visuais, na qual cada camada da rede opera em um nível de abstração progressivamente mais elevado. Zeiler e Fergus (2014) demonstraram, por meio de redes de deconvolução, que as camadas iniciais de CNNs profundas respondem a características de baixo nível, como bordas, gradientes de intensidade, texturas simples e padrões de cor, enquanto camadas intermediárias combinam essas primitivas para formar representações de nível médio, como texturas complexas, partes de objetos e padrões repetitivos, e camadas finais codificam características de alto nível, como formas globais, relações semânticas entre partes e categorias de objetos. Essa hierarquia composicional é uma consequência direta da arquitetura empilhada de convoluções com campos receptivos crescentes, de modo que cada camada “vê” uma porção maior da imagem original, ao integrar informações de regiões progressivamente mais amplas.

O conceito de campo receptivo é central para compreender como CNNs processam informação visual em diferentes escalas. O campo receptivo teórico de um neurônio em uma dada camada corresponde à região da imagem de entrada que influencia a ativação

desse neurônio. Nas camadas iniciais, os campos receptivos são pequenos e capturam detalhes locais; à medida que as camadas se empilham, os campos receptivos crescem por composição, o que permite que camadas profundas integrem contexto espacial mais amplo. Na ResNet50, os campos receptivos efetivos crescem de dezenas de píxeis nas primeiras camadas para centenas de píxeis nas últimas, e cobrem porções significativas da imagem de entrada (GOODFELLOW; BENGIO; COURVILLE, 2016). No contexto da classificação de deficiências nutricionais em folhas de café, essa hierarquia é particularmente relevante, pois camadas iniciais podem capturar mudanças finas de textura e coloração associadas ao início da clorose, enquanto camadas mais profundas podem integrar padrões mais globais, como a distribuição espacial da clorose em relação às nervuras (clorose internerval no caso de Mg) ou a concentração de necrose nas margens foliares (no caso de K). Na Figura 4 é apresentada essa progressão hierárquica, com a exibição dos padrões que ativam unidades em diferentes camadas de uma CNN profunda.

A natureza hierárquica da extração de características em CNNs tem implicações diretas para a interpretabilidade dos modelos e para a compreensão de quais informações visuais sustentam a discriminação entre classes. Técnicas de visualização, como o Grad-CAM, discutido na Seção 2.7, exploram precisamente essa hierarquia ao produzir mapas de ativação que revelam quais regiões espaciais, e indiretamente quais padrões visuais, são mais relevantes para a decisão do classificador. No caso específico da presente pesquisa, a análise dos mapas de ativação da ResNet50 permite verificar se o modelo identifica sintomas agronomicamente consistentes, como bordas necrosadas para deficiência de K ou clorose internerval para deficiência de Mg, ou se apoia em artefatos irrelevantes, como o fundo da imagem ou variações de iluminação. Essa capacidade de introspecção constitui uma ponte essencial entre a modelagem computacional e o conhecimento agrônomo especializado, o que reforça a validade e a utilidade prática dos classificadores desenvolvidos.

2.4 ViTs

2.4.1 Mecanismo de Atenção

O mecanismo de atenção, *attention*, originalmente concebido para melhorar o desempenho de modelos de tradução automática, tornou-se o componente central da arquitetura *Transformer*, proposta por Vaswani et al. (2017). A ideia fundamental é permitir que o modelo pondere dinamicamente a importância relativa de diferentes partes da entrada ao produzir cada elemento da saída, o que supera as limitações dos modelos recorrentes, que processavam sequências de forma estritamente serial e sofriam de dificuldades em capturar dependências de longo alcance. No contexto de visão computacional, o mecanismo de atenção permite que cada região da imagem consulte todas as demais regiões para determinar quais informações contextuais são mais relevantes para sua representação, uma



Figura 4 – Visualização das características aprendidas por uma CNN profunda em diferentes camadas: as camadas iniciais (Layer 1–2) capturam primitivas visuais como bordas, gradientes e padrões de textura simples; camadas intermediárias (Layer 3) combinam essas primitivas em representações mais complexas; camadas profundas (Layer 4–5) codificam partes de objetos e categorias semânticas. Para cada mapa de características, são exibidos os padrões que maximizam a ativação correspondente e os *patches* da imagem de validação que os ativam com maior intensidade. Adaptado de: Zeiler e Fergus (2014).

capacidade que vai além da localidade imposta pelas convoluções.

O mecanismo de autoatenção opera por meio de três projeções lineares da entrada, denominadas matrizes de *Query* (Q), *Key* (K) e *Value* (V). Para cada elemento da sequência, o vetor de *Query* representa o que este elemento está buscando, o vetor de *Key* representa o que este elemento pode oferecer e o vetor de *Value* contém a informação efetiva a ser transmitida. O peso de atenção entre dois elementos é calculado como o produto escalar de seus vetores Q e K, normalizado pela raiz quadrada da dimensão d_k , seguido de uma função *softmax* para obter uma distribuição de probabilidade. A saída é a soma

ponderada dos vetores V , cuja fórmula canônica é dada pela Equação 5 (VASWANI et al., 2017).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

Essa operação tem complexidade quadrática $\mathcal{O}(n^2d)$ em relação ao comprimento da sequência n e à dimensão d , o que impõe restrições computacionais significativas para sequências longas, como as geradas pela divisão de imagens de alta resolução em *patches* pequenos. Na Figura 5 são ilustrados os componentes do mecanismo de atenção escalada e da atenção multicabeça.

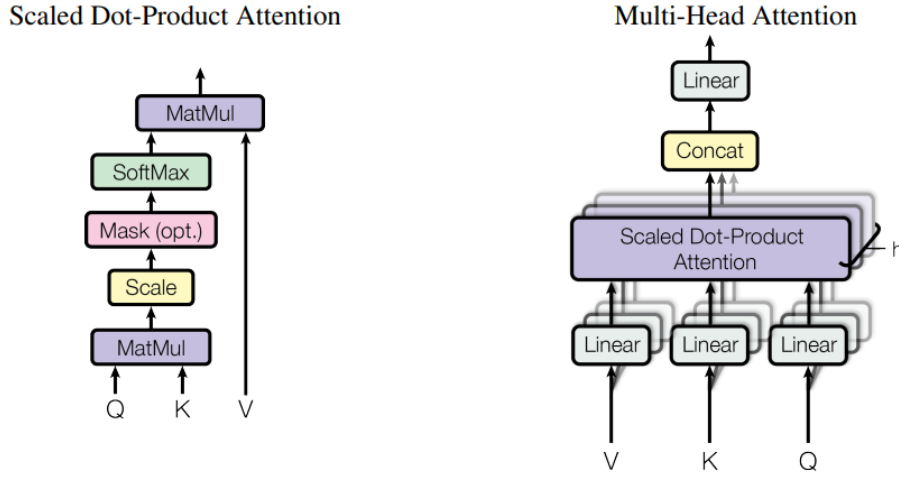


Figura 5 – *Scaled Dot-Product Attention* (esquerda) e *Multi-Head Attention* (direita): o mecanismo de atenção escalada computa pesos a partir do produto escalar entre *Query* (Q) e *Key* (K), normalizado por $\sqrt{d_k}$ e aplicado ao *softmax*, para ponderar os vetores *Value* (V). A atenção multicabeça projeta Q , K e V em h subespaços independentes, executa a atenção em paralelo e concatena os resultados, o que permite ao modelo capturar padrões de atenção distintos simultaneamente. Adaptado de: Vaswani et al. (2017).

A extensão para atenção multicabeça (*Multi-Head Attention*, MHA) permite que o modelo aprenda múltiplos padrões de atenção simultaneamente, cada cabeça em um subespaço de projeção distinto. Formalmente, dado h cabeças de atenção, cada uma projeta Q , K e V com matrizes de pesos independentes W_i^Q , W_i^K , W_i^V , computa a atenção separadamente e concatena os resultados, conforme a Equação 6, em que W^O é uma projeção de saída (VASWANI et al., 2017).

$$\text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (6)$$

Essa decomposição em múltiplas cabeças permite que o modelo capture diferentes tipos de relações simultaneamente, pois, por exemplo, uma cabeça pode especializar-se em relações espaciais locais (regiões vizinhas de uma folha), enquanto outra pode capturar correspondências globais (comparação entre o padrão da borda e o centro da folha). No

contexto da análise de folhas de café, essa capacidade é teoricamente vantajosa para reconhecer padrões de deficiência que envolvem relações espaciais complexas, como a clorose internerval, que requer a comparação entre a coloração das nervuras e dos espaços entre elas.

A complexidade computacional do mecanismo de autoatenção é um aspecto que merece consideração detalhada, especialmente para aplicações em visão computacional. Para uma sequência de n *tokens* com dimensão d , o cálculo da atenção requer a computação dos produtos matriciais QK^T e a multiplicação pela matriz V , o que resulta em complexidade temporal e espacial de $\mathcal{O}(n^2d)$. Enquanto em processamento de linguagem natural as sequências típicas possuem centenas de *tokens*, uma imagem de 224×224 dividida em *patches* de 16×16 gera 196 *tokens*, e uma imagem de 512×512 com os mesmos *patches* gera 1024 *tokens*, o que torna o custo quadrático significativo. Essa restrição tem motivado o desenvolvimento de variantes com atenção linear ou local, como o Swin Transformer (LIU et al., 2021), que aplica atenção apenas dentro de janelas, *windows*, de tamanho fixo, o que reduz a complexidade para $\mathcal{O}(n)$ ao custo de perder parte da capacidade de modelagem global. Na presente pesquisa, a resolução das imagens e o número resultante de *tokens* são suficientemente moderados para permitir o uso da atenção global completa sem restrições computacionais proibitivas.

A arquitetura *Transformer* organiza o mecanismo de MHA em blocos codificadores, denominados *encoder blocks*, que incluem, além da autoatenção, uma rede *feed-forward*, ou *FFN*, com duas camadas lineares e uma ativação não linear (tipicamente GELU), conexões residuais ao redor de cada sublayer e normalização por camada. A composição desses blocos em sequência produz representações progressivamente mais refinadas, onde cada bloco integra informações contextuais de toda a sequência de entrada. Diferentemente das CNNs, que constroem contexto global apenas implicitamente por meio do empilhamento de convoluções com campos receptivos locais, cada bloco *Transformer* tem acesso explícito ao contexto global desde a primeira camada, uma propriedade que tem implicações profundas para a modelagem de padrões visuais que se distribuem por toda a extensão da imagem (KHAN et al., 2022).

2.4.2 Arquitetura do Vision Transformer

O ViT, proposto por Dosovitskiy et al. (2021), adaptou a arquitetura *Transformer*, originalmente concebida para sequências textuais, ao domínio da visão computacional. A ideia central é tratar uma imagem como uma sequência de palavras, onde a imagem de entrada, com dimensões $H \times W \times C$, é dividida em uma grade de *patches* não sobrepostos de tamanho $P \times P$, o que resulta em $N = \frac{HW}{P^2}$ *patches*. Cada *patch* é achatado em um vetor unidimensional e projetado linearmente para um espaço de dimensão D por meio de uma camada de *patch embedding*, e produz a sequência de *tokens* visuais que será processada pelo codificador *Transformer*. Na variante ViT-B/16 utilizada neste trabalho,

a imagem é dividida em *patches* de 16×16 píxeis, cada um projetado para um espaço de 768 dimensões (DOSOVITSKIY et al., 2021).

Um aspecto crucial da arquitetura ViT é a incorporação de informação posicional, necessária porque o mecanismo de autoatenção é, por construção, invariante à permutação dos *tokens* de entrada, ou seja, não possui noção intrínseca de ordem ou posição espacial. Para suprir essa deficiência, *embeddings* posicionais aprendíveis são somados aos *patch embeddings* antes da entrada no codificador, o que fornece ao modelo informações sobre a localização espacial de cada *patch* na imagem original. Diferentemente do *Transformer* original para processamento de linguagem natural, que utilizava *embeddings* posicionais baseados em funções sinusoidais fixas (VASWANI et al., 2017), o ViT emprega *embeddings* posicionais totalmente aprendíveis, o que permite que o modelo descubra automaticamente a estrutura espacial mais informativa para a tarefa (DOSOVITSKIY et al., 2021). Quando a resolução de entrada difere daquela utilizada no pré-treinamento, os *embeddings* posicionais são interpolados bidimensionalmente, uma técnica que preserva razoavelmente a informação posicional, embora com alguma degradação de desempenho. Neste trabalho, o ViT-B/16 opera em resolução 512×512 , o que resulta em 1.024 *tokens* de *patch* mais o *token* [Token de Classificação (CLS)], frente aos 196 *tokens* do pré-treinamento em 224×224 . Essa expansão, de aproximadamente cinco vezes no número de *tokens*, implica a interpolação dos *embeddings* posicionais de uma grade 14×14 para 32×32 , regime significativamente mais distante do pré-treinamento original. Conjectura-se que essa degradação posicional possa contribuir para a menor eficiência relativa do ViT-B/16 observada nos experimentos, embora o efeito não tenha sido isolado experimentalmente neste trabalho.

Além dos *tokens* correspondentes aos *patches* da imagem, um *token* especial denominado [CLS] é inserido no início da sequência de entrada. Esse *token*, inicializado aleatoriamente e atualizado durante o treinamento, não corresponde a nenhuma região específica da imagem, mas funciona como um agregador global de informação, pois, por meio do mecanismo de autoatenção, ele atende a todos os *patches* e, ao final do codificador, sua representação é utilizada como entrada para a cabeça de classificação, tipicamente uma camada linear que mapeia o espaço de representação para o número de classes (DOSOVITSKIY et al., 2021). Essa estratégia é análoga ao uso de *tokens* [CLS] em modelos de linguagem como o BERT. Alternativamente, a média global, *global average pooling*, dos *tokens* de saída pode ser utilizada, sem necessidade do *token* [CLS], embora a abordagem original com [CLS] tenha se mostrado eficaz em diversas tarefas. Na Figura 6 é apresentado um diagrama da arquitetura ViT, com destaque para o processo de divisão em *patches*, a incorporação de *embeddings* posicionais e a estrutura do codificador *Transformer*.

O codificador *Transformer* do ViT consiste em L blocos idênticos empilhados, cada um composto por duas sublayers, sendo elas uma camada de *Multi-Head Self-Attention* (MSA)

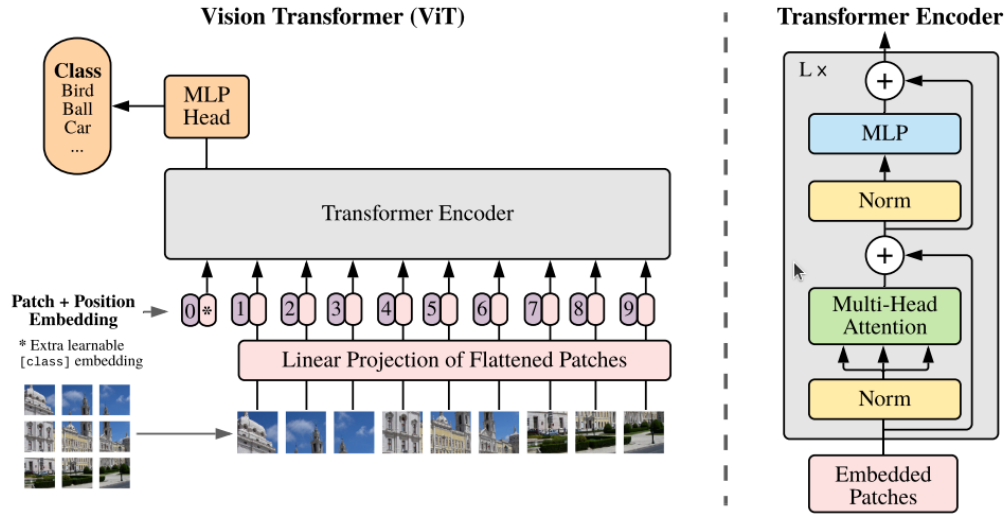


Figura 6 – Visão geral da arquitetura do ViT: a imagem é dividida em *patches* de tamanho fixo, cada um projetado linearmente em um *token* visual com *embedding* posicional. A sequência de *tokens*, acrescida de um *token* [CLS], é processada por um codificador *Transformer* (detalhe à direita), cuja saída do [CLS] alimenta a cabeça de classificação. Adaptado de: Dosovitskiy et al. (2021).

e uma rede *feed-forward* (FFN) com duas camadas lineares e ativação GELU (*Gaussian Error Linear Unit*). Cada sublayer é envolvida por uma conexão residual e seguida de *Layer Normalization* (LN), segundo o esquema *Pre-Norm* descrito nas Equações 7 e 8.

$$z'_l = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1} \quad (7)$$

$$z_l = \text{FFN}(\text{LN}(z'_l)) + z'_l \quad (8)$$

Na variante ViT-B/16, denominada *Base*, o codificador possui $L = 12$ camadas, dimensão de embedding $D = 768$, 12 cabeças de atenção e uma dimensão interna da FFN de 3072, o que totaliza aproximadamente 86 milhões de parâmetros (DOSOVITSKIY et al., 2021). Essa arquitetura é significativamente maior do que a ResNet50 (aproximadamente 25 milhões de parâmetros), o que reflete a necessidade de maior capacidade para compensar a ausência de vieses indutivos de localidade. A regularização é realizada por meio de *Dropout* nas camadas de atenção e na FFN, além de técnicas de aumento de dados durante o treinamento.

O treinamento de ViTs apresenta características distintas em relação às CNNs. Devido à ausência de vieses indutivos de localidade e invariância translacional, os *Transformers* necessitam de conjuntos de dados substancialmente maiores para aprender a estrutura espacial das imagens a partir dos dados brutos. Dosovitskiy et al. (2021) demonstraram que, quando treinados com o ImageNet-1K (1,28 milhões de imagens), os ViTs apresentam desempenho inferior ao de CNNs de capacidade comparável; contudo, quando pré-treinados em conjuntos massivos como o JFT-300M (300 milhões de imagens) ou o ImageNet-21K

(14 milhões de imagens), os ViTs superam as CNNs estado da arte, o que sugere que a atenção global compensa a falta de vieses indutivos quando dados suficientes estão disponíveis. Touvron et al. (2021) propuseram o DeiT (*Data-efficient Image Transformers*), que introduziu estratégias de destilação de conhecimento e aumento de dados para tornar o treinamento de ViTs viável com conjuntos menores, o que democratizou o acesso a essa classe de modelos. Para aplicações em domínios especializados como a agricultura, a transferência de aprendizado a partir de pesos pré-treinados no ImageNet é, portanto, uma estratégia essencial, adotada também neste trabalho. As condições específicas de pré-treinamento empregadas para cada arquitetura, bem como suas implicações para a interpretação comparativa dos resultados, são detalhadas no Capítulo 4.

2.4.3 Comparação entre CNNs e ViTs

A distinção fundamental entre CNNs e ViTs reside em seus vieses indutivos, pois enquanto CNNs incorporam localidade, em que cada filtro opera em uma vizinhança restrita, e invariância translacional, com filtros compartilhados em todas as posições, o *Transformer* é, em princípio, agnóstico à estrutura espacial dos dados e aprende relações posicionais exclusivamente a partir dos dados de treinamento (CORDONNIER; LOUKAS; JAGGI, 2020). Essa diferença tem consequências práticas profundas. As CNNs são mais eficientes em termos de dados, *data-efficient*, capazes de alcançar desempenho elevado com conjuntos de treinamento moderados, pois seus vieses indutivos atuam como um forte regularizador implícito. Os ViTs, por outro lado, necessitam de mais dados para compensar a ausência desses vieses, mas, em contrapartida, possuem maior flexibilidade para modelar relações de longo alcance e padrões globais que CNNs capturam apenas indiretamente por meio do empilhamento de camadas (KHAN et al., 2022).

Em termos de escalabilidade, os ViTs apresentam vantagens notáveis, visto que seu desempenho melhora de forma mais consistente com o aumento da escala do modelo e dos dados de treinamento, enquanto as CNNs tendem a saturar mais rapidamente. Contudo, a complexidade quadrática da autoatenção em relação ao número de *tokens* impõe limitações computacionais significativas para imagens de alta resolução, um desafio que tem motivado o desenvolvimento de variantes eficientes como o Swin Transformer (LIU et al., 2021) e variantes com atenção linear. No contexto específico desta pesquisa, a comparação entre ResNet50, uma CNN, e ViT-B/16, um *Transformer*, é particularmente pertinente, uma vez que o viés indutivo de localidade das CNNs tende a ser vantajoso em tarefas nas quais as diferenças discriminativas são predominantemente texturais e locais, como alterações de cor e textura na superfície foliar, ao passo que relações semânticas globais favoreceriam os ViTs. Essa expectativa decorre do princípio teórico de que o viés indutivo adequado para um problema depende da estrutura intrínseca dos dados (DOSOVITSKIY et al., 2021; KHAN et al., 2022). A verificação empírica dessa relação no domínio da classificação de deficiências nutricionais em folhas de café é apresentada no Capítulo 5.

Uma distinção relevante para a compreensão dos resultados desta pesquisa diz respeito à forma como cada paradigma lida com o aumento de dados. As CNNs, graças à invariância translacional conferida pelo compartilhamento de pesos, são naturalmente robustas a deslocamentos espaciais e se beneficiam de aumentos geométricos simples, como espelhamento horizontal, rotação e recorte aleatório. Os ViTs, por outro lado, por serem agnósticos à estrutura espacial, requerem estratégias de aumento mais intensivas para aprender invariâncias que as CNNs incorporam por design. Tuvron et al. (2021) demonstraram que o desempenho dos ViTs é particularmente sensível à política de aumento de dados, com técnicas como RandAugment, CutMix e Mixup essenciais para alcançar resultados competitivos. A política de aumento de dados efetivamente adotada nesta pesquisa é detalhada no Capítulo 4.

A complementaridade entre os paradigmas convolucional e de atenção motivou o desenvolvimento de arquiteturas híbridas que buscam combinar o melhor de ambos os mundos, unindo a eficiência local das CNNs à capacidade de modelagem global dos *Transformers*. Essa direção de pesquisa, discutida em detalhe na seção seguinte, representa uma das fronteiras mais ativas da visão computacional contemporânea e constitui o terceiro modelo avaliado nesta dissertação.

2.5 Arquiteturas Híbridas entre Convoluções e Atenção

A motivação central para o desenvolvimento de arquiteturas híbridas que combinam convoluções e atenção reside na observação de que CNNs e *Transformers* possuem capacidades complementares, visto que as convoluções são particularmente eficazes na extração de características locais de baixo nível, como bordas, texturas e padrões de cor, enquanto o mecanismo de autoatenção se destaca na modelagem de dependências globais e relações semânticas de alto nível entre regiões distantes da imagem (KHAN et al., 2022). A hipótese subjacente é que, ao combinar ambos os paradigmas em uma única arquitetura, é possível obter modelos que herdem a eficiência de dados das CNNs (graças a seus vieses indutivos de localidade) e a capacidade de captura de contexto global dos *Transformers*, o que resulta em representações mais ricas e robustas do que aquelas produzidas por qualquer paradigma isoladamente.

Diversas estratégias de hibridização foram propostas na literatura, e diferem na forma como os componentes convolucionais e de atenção são integrados. A abordagem mais comum emprega um *backbone* convolucional como extrator de características, cujos mapas de saída são reorganizados em uma sequência de *tokens* e processados por um codificador *Transformer*. Essa estratégia é adotada no modelo híbrido avaliado neste trabalho, em que a ResNet50 (sem as camadas finais de classificação) extrai mapas de características de alta dimensionalidade (2048 canais), que são reduzidos espacialmente por *adaptive average*

pooling para uma grade de 5×5 regiões, o que resulta em 25 *tokens* visuais, configuração adotada no benchmark principal, sendo que o tamanho dessa grade é sistematicamente variado no estudo de ablação descrito no Capítulo 4. Uma projeção convolucional 1×1 reduz a dimensionalidade dos canais de 2048 para 256, e os *tokens* resultantes, complementados por um *token* [CLS] e *embeddings* posicionais aprendíveis, são processados por um codificador *Transformer* com seis camadas. Essa arquitetura permite que as convoluções da ResNet capturem detalhes texturais finos das folhas, enquanto o *Transformer* integra essas informações em uma representação global contextualizada. Na Figura 7 é apresentado o diagrama completo dessa arquitetura, com os componentes e as dimensões dos tensores em cada estágio do processamento.

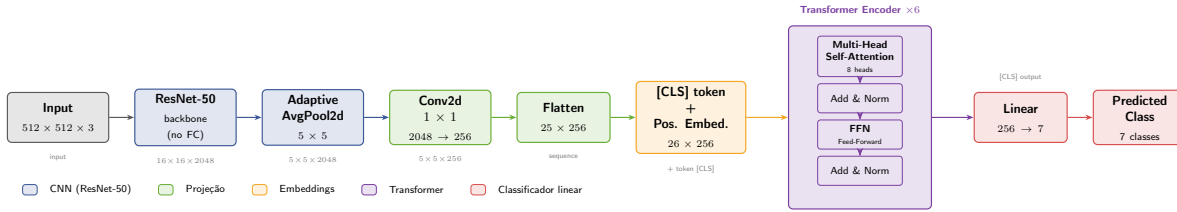


Figura 7 – Arquitetura do modelo Híbrido CNN-ViT avaliado nesta dissertação: o *backbone* ResNet50 extrai mapas de características de 2048 canais, com dimensão espacial dependente da resolução de entrada, sendo 16×16 para 512×512 e 10×10 para 314×314 , reduzidos por *adaptive average pooling* para uma grade uniforme de 5×5 *tokens*. Uma projeção convolucional 1×1 reduz os canais para 256, e os 25 *tokens* resultantes, complementados por *token* [CLS] e *embeddings* posicionais aprendíveis (26 *tokens* no total), são processados por um codificador *Transformer* de 6 camadas (5 padrão + 1 customizada) com 8 cabeças de atenção para a classificação final em 7 classes. Fonte: elaborado pelo autor, inspirado em Li et al. (2025).

Trabalhos relevantes na literatura ilustram a diversidade de estratégias de hibridização. O CoAtNet, proposto por Dai et al. (2021), combina convoluções depthwise-separable com atenção relativa em um design hierárquico de cinco estágios (S0–S4), o que demonstra que a integração cuidadosa de ambos os componentes pode superar tanto CNNs quanto ViTs puros em tarefas de classificação de imagens. O CMT, sigla de *Convolutional Neural Networks Meet Vision Transformers*, proposto por Guo et al. (2022), introduz um módulo de convolução local antes de cada bloco de atenção, o que enriquece a representação de *tokens* com informações de vizinhança antes da modelagem global. Mais recentemente, a CMTNet (GUO; FENG; GUO, 2025) demonstrou a eficácia de abordagens híbridas para a classificação de imagens hiperespectrais capturadas por Veículo Aéreo Não Tripulado (UAV)s, com blocos convolucionais para extração espacial-espectral seguidos de codificadores *Transformer* para modelagem de dependências de longo alcance. No domínio específico da fitopatologia, o PlantXViT (THAKUR et al., 2022) propôs uma arquitetura

híbrida leve e interpretável que combina um *backbone* convolucional com camadas de atenção para a detecção de doenças em plantas, o que demonstra que modelos híbridos podem ser eficientes tanto em termos de acurácia quanto de custo computacional, adequados para implantação em dispositivos de borda. No mesmo período, Li et al. (2025) propuseram uma abordagem híbrida entre convoluções e atenção com mecanismo de atenção cruzada, *cross-attention*, o que demonstra ganhos de classificação ao integrar características locais e globais de forma mais eficiente.

No contexto desta dissertação, a arquitetura híbrida combina o viés indutivo de localidade das convoluções da ResNet50 com a modelagem de contexto global do codificador *Transformer*, em uma divisão de trabalho na qual as convoluções extraem características locais e o mecanismo de atenção integra relações entre regiões distantes da folha. Os detalhes de implementação dessa arquitetura são apresentados na Seção 4.4.3, e sua avaliação empírica comparativa frente às demais arquiteturas é discutida no Capítulo 5.

É importante notar que a arquitetura híbrida difere do ViT puro na forma como os *tokens* visuais são construídos. No ViT original, cada *patch* bruto da imagem é projetado diretamente para o espaço de *embedding*, de modo que o codificador *Transformer* recebe representações de baixo nível e deve aprender tanto características locais quanto relações globais. No modelo híbrido, os *tokens* são derivados dos mapas de características da ResNet50, que já codificam informações ricas sobre texturas, bordas e padrões cromáticos locais. Dessa forma, o codificador *Transformer* opera sobre representações de nível intermediário a alto, com sua capacidade concentrada na modelagem de relações globais entre regiões pré-processadas. Essa divisão de trabalho é teoricamente vantajosa, pois as convoluções fornecem um pré-processamento com viés indutivo de localidade, enquanto a atenção global complementa com a integração de informações entre regiões distantes. A hipótese subjacente é que essa combinação possa resultar em desempenho superior ao de cada paradigma isoladamente.

2.6 Avaliação de Desempenho em Classificação

A avaliação de modelos de classificação multiclasse apoia-se em um conjunto de métricas derivadas da matriz de confusão, cada uma sensível a aspectos distintos do comportamento do classificador (GRANDINI; BAGLI; VISANI, 2020). A acurácia global, *accuracy*, definida como a proporção de amostras corretamente classificadas em relação ao total, é a métrica mais intuitiva e amplamente reportada na literatura de classificação de imagens. Contudo, a acurácia global pode mascarar deficiências severas de desempenho em classes específicas, especialmente sob desbalanceamento, pois um modelo que classifica corretamente as amostras das classes majoritárias mas falha sistematicamente nas minoritárias ainda pode apresentar acurácia global aparentemente satisfatória.

Para uma análise em nível de classe, três métricas complementares são empregadas.

A precisão, *precision*, de uma classe indica a proporção de predições corretas dentre todas as predições realizadas para aquela classe, ao passo que a revocação, *recall*, indica a proporção de amostras daquela classe que foram corretamente identificadas pelo modelo. O *F1-score*, definido como a média harmônica entre precisão e revocação, fornece uma medida equilibrada que penaliza discrepâncias entre as duas. Para agregar o desempenho por classe em um único valor robusto ao desbalanceamento, o *macro F1-score* é calculado como a média aritmética simples dos *F1-scores* individuais, o que atribui igual peso a todas as classes independentemente de sua frequência, em contraste com a média ponderada, *weighted*, dominada pelas classes de maior volume.

A matriz de confusão, por sua vez, constitui um instrumento complementar de análise qualitativa. Quando normalizada por linha, cada célula representa a proporção de amostras de uma classe verdadeira atribuída a cada classe predita, o que elimina o efeito do desbalanceamento na visualização e permite a identificação imediata de padrões de confusão entre classes específicas (GRANDINI; BAGLI; VISANI, 2020).

2.7 Interpretabilidade e Explicabilidade (XAI)

2.7.1 Grad-CAM

A complexidade dos modelos de *deep learning*, frequentemente dotados de milhões de parâmetros e múltiplas camadas de transformações não lineares, os torna inerentemente caixas-pretas, o que dificulta a compreensão do processo de decisão subjacente às suas predições (ADADI; BERRADA, 2018). Essa opacidade levanta preocupações legítimas em domínios de aplicação críticos, como a agricultura, onde decisões baseadas em diagnósticos incorretos podem resultar em perdas econômicas significativas. As técnicas de XAI surgiram para abordar essa limitação, com o objetivo de tornar as decisões dos modelos compreensíveis para seres humanos (MOHAN; RAYANOOTHALA; SREE, 2025). No espectro das técnicas de XAI, destacam-se as abordagens baseadas em atribuição, que identificam quais regiões ou características da entrada foram mais influentes para uma determinada predição (SELVARAJU et al., 2017).

O Grad-CAM, proposto por Selvaraju et al. (2017), é uma técnica de visualização post-hoc amplamente utilizada que gera mapas de calor e destaca as regiões da imagem mais relevantes para a predição de uma classe específica. O princípio do Grad-CAM baseia-se no cálculo dos gradientes do *logit* da classe-alvo y^c em relação aos mapas de ativação A^k da última camada convolucional. Esses gradientes são agregados espacialmente por *global average pooling*, conforme descrito na Equação 9, para obter os pesos de importância que ponderam os mapas de ativação correspondentes. O mapa de calor final é obtido pela combinação linear ponderada dos mapas de ativação, seguida de uma função ReLU, conforme a Equação 10, para reter apenas as ativações com influência positiva (SELVARAJU

et al., 2017). O resultado é um mapa de calor com a resolução espacial da última camada convolucional, que é redimensionado para a resolução da imagem original por interpolação bilinear. O *pipeline* completo do Grad-CAM é ilustrado na Figura 8 desde o cálculo dos gradientes até a geração do mapa de calor sobreposto à imagem original.

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (9)$$

$$L_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right) \quad (10)$$

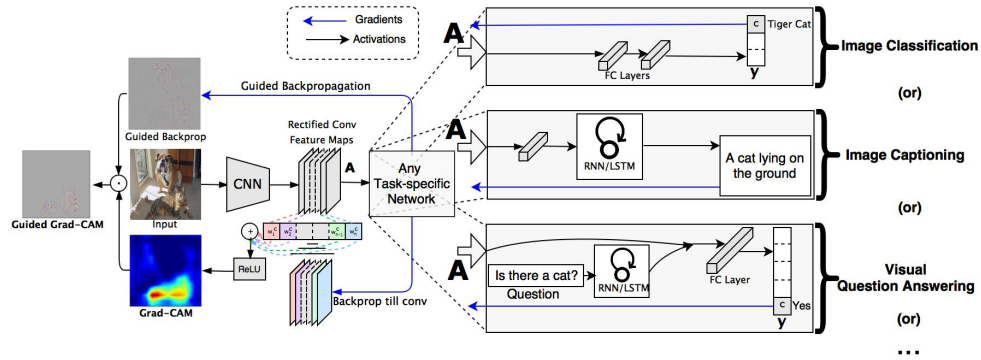


Figura 8 – *Pipeline* do Grad-CAM: os gradientes do *logit* da classe-alvo em relação aos mapas de ativação da última camada convolucional são agregados por *global average pooling* para obter pesos de importância por canal; a combinação linear ponderada desses mapas, seguida de ReLU, produz o mapa de calor que é redimensionado e sobreposto à imagem original. Regiões em vermelho indicam alta relevância para a predição da classe, enquanto regiões em azul indicam baixa relevância. Adaptado de: Selvaraju et al. (2017).

A versatilidade do Grad-CAM reside em sua aplicabilidade a qualquer arquitetura que contenha camadas convolucionais, sem necessidade de modificações na rede ou de retreinamento, e em sua capacidade de gerar explicações específicas por classe, isto é, mapas de calor distintos para diferentes classes, o que revela quais regiões sustentam cada predição particular. No contexto da análise de imagens agrícolas, essa especificidade por classe é particularmente valiosa, pois ao gerar mapas Grad-CAM para a predição de deficiência de K, por exemplo, espera-se que as regiões de maior ativação correspondam às margens foliares onde a necrose é agronomicamente esperada, para a deficiência de Mg, espera-se destaque nas regiões internervais onde a clorose se manifesta. Essa correspondência entre as regiões de atenção do modelo e o conhecimento agrônomo especializado serve como uma forma de verificação qualitativa, que aumenta a confiança de que o modelo aprende características discriminativas, e não artefatos espúrios dos dados, sem substituir métricas formais de interpretabilidade (SELVARAJU et al., 2017).

A aplicação concreta do Grad-CAM aos modelos avaliados nesta pesquisa, com a análise da correspondência entre as regiões destacadas e a sintomatologia descrita na Seção 2.1, é apresentada e discutida no Capítulo 5.

2.7.2 Mapas de Atenção em Transformers

Nos modelos baseados em *Transformers*, os próprios pesos de atenção constituem uma ferramenta de interpretabilidade intrínseca, uma vez que os valores α_{ij} da matriz de atenção indicam explicitamente o grau em que o *token* i “atende” ao *token* j ao compor sua representação. Ao visualizar esses pesos de atenção como mapas bidimensionais sobrepostos à imagem original, obtém-se uma indicação das regiões que o modelo considera mais informativas para cada predição. Essa forma de interpretabilidade é considerada *nativa*, pois emerge diretamente do mecanismo de computação do modelo, sem necessidade de técnicas post hoc adicionais (KHAN et al., 2022). Na arquitetura híbrida avaliada neste trabalho, a última camada do codificador *Transformer* foi implementada com a capacidade de capturar e armazenar os pesos de atenção, o que possibilita a visualização de quais *tokens* (regiões da imagem) o modelo prioriza ao realizar a classificação. Na Figura 9 esse princípio é ilustrado com exemplos do ViT original treinado em imagens naturais; no modelo híbrido desta pesquisa, mapas análogos são obtidos e analisados no Capítulo 5.

Contudo, a interpretação dos mapas de atenção requer cautela. É reconhecido que os pesos de atenção nem sempre correspondem a explicações causais do comportamento do modelo, pois correlações elevadas entre atenção e importância de gradiente não são garantidas, e diferentes cabeças de atenção podem exibir padrões contraditórios. Ademais, em arquiteturas com múltiplas camadas de atenção, os pesos das camadas finais refletem a atenção sobre representações já transformadas por camadas anteriores, e não diretamente sobre os dados de entrada, o que torna a interpretação menos direta do que nos modelos com uma única camada de atenção. Neste trabalho, essa limitação aplica-se diretamente à análise de interpretabilidade do modelo híbrido, pois os mapas de atenção correspondem exclusivamente à última das seis camadas do codificador *Transformer*, de modo que refletem a atenção sobre representações abstratas já processadas pelas cinco camadas anteriores, e não sobre os *tokens* visuais originais. Essa restrição metodológica é reconhecida como uma limitação da análise de interpretabilidade conduzida, e seus resultados devem ser interpretados com essa ressalva. Apesar dessas considerações, os mapas de atenção permanecem uma ferramenta valiosa de análise exploratória, especialmente quando combinados com técnicas complementares como o Grad-CAM, o que permite uma avaliação multifacetada dos mecanismos de decisão dos modelos (KHAN et al., 2022).

Em termos comparativos, a disponibilidade de mecanismos de interpretabilidade tanto para CNNs (Grad-CAM) quanto para *Transformers* (mapas de atenção) permite que a presente pesquisa conduza uma análise de explicabilidade que transcende arquiteturas individuais. Ao contrastar as regiões de foco da ResNet50, obtidas por meio do Grad-CAM, com os padrões de atenção do modelo híbrido, obtém-se perspectivas complementares sobre como diferentes paradigmas computacionais percebem os sintomas de deficiência nutricional. Essa comparação não apenas enriquece a compreensão dos modelos avaliados, mas também contribui para o corpo crescente de literatura sobre XAI na agricultura, o

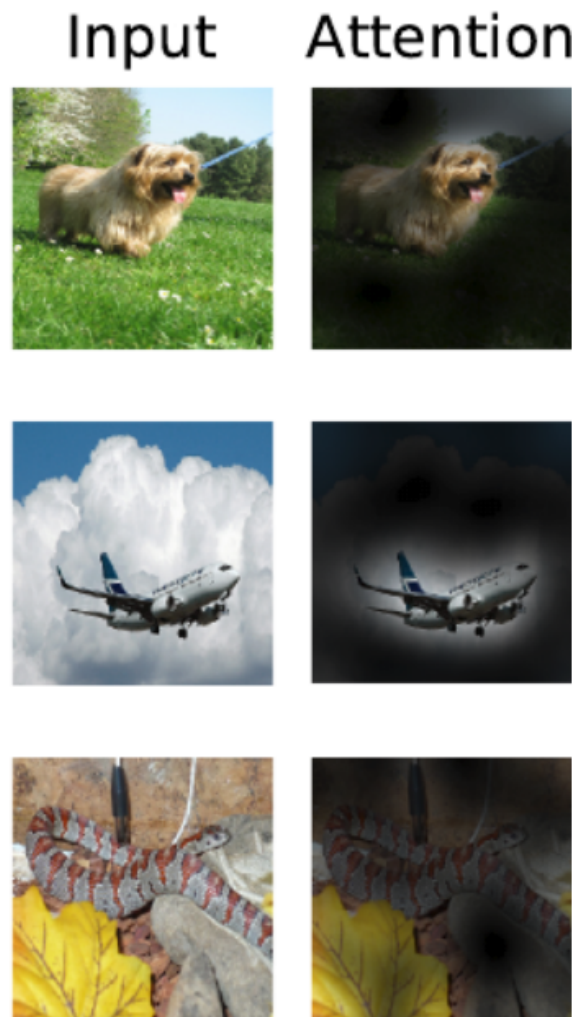


Figura 9 – Mapas de atenção do ViT original treinado em imagens naturais: para cada imagem de entrada, o *token* [CLS] atende seletivamente às regiões semanticamente relevantes, ilustrando a capacidade do mecanismo de autoatenção de focar em objetos de interesse sem supervisão espacial explícita. Adaptado de: Dosovitskiy et al. (2021).

que demonstra como técnicas de interpretabilidade podem ser utilizadas para validar a relevância das representações aprendidas por modelos de *deep learning*.

2.7.3 Importância da Interpretabilidade na Agricultura

A adoção de sistemas de inteligência artificial na agricultura depende não apenas de sua acurácia, mas, de forma crítica, da confiança que os usuários finais, agrônomos, extensionistas rurais, produtores, depositam nas suas recomendações. Modelos de *deep learning* que operam como caixas-pretas ao fornecer predições sem justificativas compreensíveis enfrentam resistência significativa para adoção prática, especialmente em contextos onde decisões equivocadas podem acarretar consequências econômicas severas, como a aplicação desnecessária de fertilizantes ou a negligência de uma deficiência real. A inter-

pretabilidade, nesse sentido, desempenha um papel dual, pois, por um lado, permite que especialistas de domínio validem se o modelo opera com base em informações agronomicamente relevantes, por outro, fornece explicações que tornam as predições mais acionáveis e auxilia o processo de tomada de decisão (MOHAN; RAYANOOTHALA; SREE, 2025; THAKUR et al., 2022).

Ademais, a interpretabilidade desempenha um papel fundamental na detecção de vieses e artefatos nos dados de treinamento. Modelos de *deep learning* são notoriamente suscetíveis a aprender atalhos espúrios, o *shortcut learning*, por exemplo ao basear suas predições no fundo da imagem, em marcações de anotação ou em características visuais irreconhecíveis para observadores humanos, ao invés de tomarem por base sintomas reais de interesse (GEIRHOS et al., 2020). Sem ferramentas de interpretabilidade, esses vieses podem passar despercebidos, o que resulta em modelos com acurácia artificialmente elevada nos dados de treinamento, mas que falham catastroficamente quando aplicados em condições reais. A análise sistemática dos mapas de ativação permite identificar esses problemas, o que oferece uma forma de auditoria que complementa as métricas quantitativas tradicionais e contribui para a construção de modelos mais confiáveis e robustos.

No contexto específico do diagnóstico de deficiências nutricionais em cafeeiros, a interpretabilidade assume relevância adicional pela possibilidade de integração entre o conhecimento agrônomo consolidado e os padrões aprendidos computacionalmente. A sintomatologia de cada deficiência é bem documentada na literatura agrônoma (MALAVOLTA, 2006; MARSCHNER, 2012), a exemplo da clorose internerval para Mg, da necrose marginal para K e da clorose uniforme para N, o que fornece um referencial externo contra o qual as explicações dos modelos podem ser confrontadas. Quando os mapas de ativação do Grad-CAM ou os pesos de atenção do *Transformer* destacam precisamente as regiões foliares onde os sintomas são agronomicamente esperados, obtém-se uma validação cruzada que reforça tanto a confiança no modelo quanto a compreensão dos padrões visuais que discriminam as deficiências. Essa sinergia entre *machine learning* e conhecimento de domínio representa uma das contribuições mais promissoras da XAI para a agricultura de precisão, ao transformar modelos preditivos em ferramentas de descoberta e confirmação científica.

Trabalhos Relacionados

A literatura sobre visão computacional aplicada à agricultura experimentou um crescimento exponencial na última década, impulsionada pelos avanços em *deep learning* e pela crescente disponibilidade de dados de imagem em larga escala. No domínio específico do cafeeiro, diversos conjuntos de dados públicos e estudos aplicados foram desenvolvidos para fomentar a pesquisa em detecção de doenças e deficiências. O BRACOL (*Brazilian Coffee Leaf*), proposto por Esgario, Krohling e Ventura (2020), é um dos *datasets* mais utilizados, com imagens de folhas de café exibindo sintomas de doenças bióticas, entre elas ferrugem, cercosporiose e bicho-mineiro, além de folhas saudáveis. O JMuBEN, descrito por Jepkoech et al. (2021), expandiu o escopo ao reunir imagens de folhas de café do Quênia com múltiplas categorias de doenças, ampliando a diversidade geográfica e clínica dos dados disponíveis. Mais recentemente, o CoLeaf-DB, proposto por Tuesta-Monteza, Mejia-Cabrera e Arcila-Diaz (2023), voltou-se especificamente para deficiências nutricionais em folhas de café, constituindo o *dataset* tematicamente mais próximo ao utilizado nesta pesquisa. No plano dos modelos, Randive, Sapkal e Bornare (2025) aplicaram arquiteturas convolucionais à detecção e classificação de deficiências nutricionais em folhas de café, com resultados promissores, o que reforça o interesse crescente da comunidade científica nessa aplicação. Apesar da diversidade de enfoques, todos esses trabalhos tratam o diagnóstico como um problema de classificação estática, com imagens de sintomas já estabelecidos, descartando a informação diagnóstica contida na evolução temporal do quadro nutricional.

A aplicação de *deep learning* para a classificação de doenças e estresses em plantas tem sido dominada por arquiteturas convolucionais, com resultados expressivos relatados para diversas culturas e patologias. Ferentinos (2018) conduziram um dos primeiros estudos abrangentes com o uso de CNNs para identificação de 58 combinações de planta-doença, com acurácia de 99,53% no dataset PlantVillage, embora esse resultado tenha sido criticado por viés de captura, *capture bias*, documentado por Noyan (2022), e pela ausência de validação com dados de campo. Upadhyay et al. (2025), em revisão de mais de 278 artigos, documenta o progresso contínuo nesse campo e destaca a migração de arquiteturas

como VGG e AlexNet para modelos mais sofisticados como ResNets, EfficientNets e, mais recentemente, *Vision Transformers*, ao passo que Mehta et al. (2025), por meio de revisão sistemática com diretrizes PRISMA, sistematiza o panorama de aplicações de CNNs na área. Shoaib et al. (2025) ampliam o escopo ao incluir também a detecção de pragas, com ênfase nos desafios de implantação em condições de campo. Estudos específicos para o cafeeiro reportaram acurácias elevadas na classificação de doenças foliares com CNNs pré-treinadas em *transfer learning*, o que confirma a viabilidade da abordagem para essa cultura (ESGARIO; KROHLING; VENTURA, 2020). No contexto de deficiências nutricionais, Amudha e Brindha (2024) propuseram uma rede baseada em cápsulas para a classificação de deficiências em folhas de arroz, o que demonstra que arquiteturas de *deep learning* podem capturar padrões cromáticos e texturais sutis associados a carências minerais. Ademais, Pushpa, Akash e Akash (2025) conduziram um estudo comparativo entre técnicas de *machine learning* e *deep learning* para a detecção de deficiências nutricionais por análise de imagens foliares, o que evidencia a superioridade consistente dos modelos profundos nessa tarefa. Contudo, a grande maioria desses trabalhos trata o diagnóstico como um problema de classificação estática, com imagens de sintomas já estabelecidos, sem considerar o estágio de desenvolvimento do estresse.

A dimensão temporal do desenvolvimento de estresses em plantas é reconhecida como relevante na literatura de fenotipagem vegetal, mas permanece pouco explorada no contexto de classificação automatizada com *deep learning*. Ubbens e Stavness (2017), ao apresentarem a plataforma *Deep Plant Phenomics*, demonstraram a aplicabilidade de *deep learning* a tarefas de fenotipagem vegetal de *Arabidopsis thaliana*, incluindo contagem de folhas, classificação de mutantes e regressão de idade a partir de imagens, o que estabelece precedente para a incorporação de variáveis temporais em *pipelines* visuais de fenotipagem. Rußwurm e Körner (2017) demonstraram a viabilidade da classificação multitemporal de cobertura do solo com o uso de redes LSTM aplicadas a séries temporais de imagens de satélite, o que ilustra o potencial de abordagens temporais em sensoriamento remoto agrícola. No entanto, a transição dessa metodologia para o nível de parcela e de planta individual, onde a resolução temporal e espacial deve ser muito maior, permanece um desafio aberto. Trabalhos como o de Tan et al. (2022), que utilizaram *machine learning* para detectar estágios de crescimento de mudas de arroz ao longo do tempo, representam passos nessa direção, mas com escopo e metodologia distintos da classificação de deficiências nutricionais a partir da evolução temporal dos sintomas abordada neste trabalho.

A emergência de arquiteturas baseadas em *Transformers* e de modelos híbridos que combinam convoluções e atenção tem gerado interesse crescente na comunidade de visão computacional agrícola. Sakka et al. (2025) conduziram uma revisão abrangente sobre aplicações de CNNs em agricultura inteligente com dados multimodais, que abrange tarefas como detecção de doenças, estimativa de rendimento e classificação de culturas,

e apontam as arquiteturas baseadas em *Transformers* e os modelos híbridos como uma direção promissora para investigações futuras. O PlantXViT (THAKUR et al., 2022) demonstrou que arquiteturas híbridas podem alcançar desempenho competitivo com modelos consideravelmente maiores, enquanto mantêm custos computacionais baixos, adequados para implantação em dispositivos de borda, *edge computing*. A CMTNet (GUO; FENG; GUO, 2025) aplicou uma abordagem híbrida entre convoluções e atenção à classificação de imagens hiperespectrais capturadas por UAVs, o que demonstra ganhos em relação a CNNs e ViTs puros. O TAC-UNet (HE et al., 2024) explorou a combinação de convoluções com atenção em tarefas de segmentação, o que reforça a versatilidade dos modelos híbridos para diferentes paradigmas de visão computacional. Apesar desses avanços, a aplicação de arquiteturas híbridas ao problema específico de classificação de deficiências nutricionais em cafeeiros a partir da evolução temporal dos sintomas permanece inexplorada. Cabe ressaltar, contudo, que os ganhos relatados para arquiteturas baseadas em atenção e para modelos híbridos dependem fortemente do volume de dados e do domínio de aplicação, de modo que sua superioridade sobre as CNNs não é garantida em regimes de dados moderados, hipótese que a presente pesquisa avalia empiricamente.

A questão da transferência de aprendizado no domínio agrícola merece destaque pela sua importância prática. A maioria dos datasets agrícolas é significativamente menor do que os *benchmarks* de visão computacional geral (ImageNet, COCO, JFT-300M), o que torna o treinamento de modelos profundos a partir do zero (*from scratch*) propenso ao sobreajuste. A estratégia de transferência de aprendizado, na qual modelos pré-treinados em grandes conjuntos de dados genéricos como o ImageNet são ajustados para a tarefa-alvo, tem sido amplamente adotada no domínio agrícola como forma de mitigar a escassez de dados rotulados (ALZUBAIDI et al., 2021; SAKKA et al., 2025). A premissa subjacente é que as características de baixo nível aprendidas em grandes datasets, como bordas, texturas e gradientes, são suficientemente genéricas para transferir-se entre domínios visuais distintos, enquanto as camadas finais se especializam nas discriminações específicas da tarefa-alvo. No *dataset* utilizado nesta pesquisa, com 28.086 registros distribuídos em sete classes, a transferência de aprendizado a partir do ImageNet constitui uma estratégia essencial para alcançar desempenho competitivo, aplicada às três arquiteturas avaliadas, sendo elas ResNet50, ViT-B/16 e modelo híbrido.

A interpretabilidade dos modelos de *deep learning* aplicados à agricultura tem recebido atenção crescente na literatura recente. Selvaraju et al. (2017) introduziram o Grad-CAM, técnica amplamente adotada para visualização de regiões discriminativas em CNNs, e sua aplicação ao domínio agrícola tem permitido a validação qualitativa de modelos contra o conhecimento de domínio. O PlantXViT (THAKUR et al., 2022) avaliou a interpretabilidade de suas predições com o uso do Grad-CAM e LIME como ferramentas de análise post hoc, o que reforça que a explicabilidade é um aspecto relevante para a adoção prática de sistemas de diagnóstico automatizado na agricultura. Essa perspectiva

alinha-se com a abordagem da presente pesquisa, com o uso tanto do Grad-CAM (para a ResNet50) quanto dos mapas de atenção (para o modelo híbrido) como ferramentas de cotejo qualitativo entre as predições computacionais e o conhecimento agrônomo sobre a sintomatologia de deficiências nutricionais. A interpretabilidade dos modelos que processem esses dados é um componente essencial para a transição de protótipos laboratoriais para sistemas de campo.

Na Tabela 1 é apresentada uma síntese dos principais trabalhos relacionados e suas características, o que evidencia as lacunas que a presente pesquisa visa preencher.

Tabela 1 – Comparação dos principais trabalhos relacionados com a presente pesquisa. “Temp.” indica análise temporal; “Def. Nutr.” indica foco em deficiências nutricionais; “XAI” indica uso de técnicas de interpretabilidade.

Trabalho	Cultura	Tarefa	Arquitetura	Temp.	Def. Nutr.	XAI
BRACOL (ESGARIO; KROHLING; VENTURA, 2020)	Café	Doenças	CNN	Não	Não	Não
JMuBEN (JEPKOECH et al., 2021)	Café	Doenças	Dataset	Não	Não	Não
CoLeaf-DB (TUESTA-MONTEZA; MEJIA-CABRERA; ARCILA-DIAZ, 2023)	Café	Deficiências	CNN	Não	Sim	Não
Ferentinos (2018)	Múltiplas	Doenças	CNN	Não	Não	Não
Rußwurm e Körner (2017)	Cobertura solo	Classificação	LSTM	Sim	Não	Não
PlantViT (THAKUR et al., 2022)	Múltiplas	Doenças	Híbrida	Não	Não	Sim
CMTNet (GUO; FENG; GUO, 2025)	Múltiplas	Classificação	Híbrida	Não	Não	Não
Amudha e Brindha (2024)	Arroz	Deficiências	CNN (Cápsulas)	Não	Sim	Não
Randive, Sapkal e Bornare (2025)	Café	Deficiências	CNN	Não	Sim	Não
Tan et al. (2022)	Arroz	Crescimento	ML	Sim	Não	Não
Este trabalho	Café	Deficiências	CNN, ViT, Híbrida	Sim	Sim	Sim

Conforme fica evidenciado na Tabela 1, a presente pesquisa se diferencia da literatura existente em três eixos complementares que, até onde se pôde identificar, não foram simultaneamente abordados em nenhum trabalho anterior. Primeiramente, a análise temporal sistemática. Em vez de uma classificação estática que desconsidera o estágio de desenvolvimento dos sintomas, este trabalho investiga explicitamente como o desempenho diagnóstico evolui à medida que as deficiências se manifestam ao longo do tempo, com a quantificação da viabilidade da detecção em diferentes estágios. Em segundo lugar, a comparação abrangente de arquiteturas no contexto temporal, em que se avalia diretamente o desempenho de uma CNN (ResNet50), de um ViT (ViT-B/16) e de um modelo híbrido CNN-ViT, o que fornece *insights* empíricos sobre qual paradigma se adapta melhor à natureza evolutiva dos sintomas de deficiência nutricional. Em terceiro lugar, a análise de interpretabilidade como validação agrônoma, sendo que o uso conjunto de Grad-CAM e mapas de atenção permite verificar se os padrões aprendidos pelos modelos correspondem à sintomatologia documentada na literatura agrônoma, o que confere uma dimensão de cotejo qualitativo que vai além das métricas quantitativas tradicionais. Essa combinação de análise temporal, comparação arquitetural e interpretabilidade, aplicada ao problema de deficiências nutricionais em cafeeiros, constitui a contribuição original desta dissertação.

Cabe ressaltar que a dimensão temporal do presente trabalho não se confunde com abordagens de séries temporais que utilizam arquiteturas recorrentes ou convoluções temporais para modelar a evolução dos dados ao longo do tempo. A estratégia adotada é distinta, pois diferentes subconjuntos temporais do *dataset* são utilizados para treinar e

avaliar modelos de classificação estática, com o intuito de investigar como o desempenho varia em função da largura da janela temporal dos dados incluídos no treinamento. Essa abordagem permite responder à pergunta central da pesquisa, a partir de que momento os sinais visuais se tornam suficientemente discriminativos, sem a necessidade de arquiteturas temporais complexas, ao focar na relação entre maturidade dos sintomas e classificabilidade. Cabe observar que não há um conjunto de teste fixo e compartilhado entre as configurações temporais, visto que cada configuração temporal define um subconjunto distinto do *dataset* global, que é então dividido em 80% para treinamento e 20% para avaliação, com estratificação por classe. Dessa forma, as métricas de cada configuração são calculadas sobre conjuntos de avaliação distintos, de modo que diferenças de acurácia entre frações refletem tanto a variação no conjunto de treinamento quanto nos dados de avaliação correspondentes; essa característica do protocolo é detalhada no Capítulo 4. A contribuição é, portanto, tanto empírica, com evidências quantitativas sobre a dependência temporal da classificação, quanto metodológica, ao propor um protocolo experimental replicável para a avaliação temporal de modelos de classificação de estresses em plantas.

Com os fundamentos teóricos e o posicionamento em relação à literatura estabelecidos, o Capítulo 4 a seguir detalha a proposta metodológica desenvolvida para operacionalizar essa investigação, com a descrição do conjunto de dados utilizado, do protocolo de particionamento temporal, das arquiteturas implementadas e das estratégias de treinamento e avaliação adotadas.

Proposta Metodológica

A investigação sobre a classificação de deficiências nutricionais em folhas de café a partir da evolução temporal dos sintomas demanda uma abordagem metodológica que contemple não apenas a escolha de arquiteturas de aprendizado profundo, mas também o tratamento cuidadoso da dimensão temporal inerente ao problema. Conforme discutido nos capítulos anteriores, a literatura de *deep learning* aplicada à agricultura frequentemente ignora a progressão dos sintomas ao longo do tempo, ao tratar a classificação como uma tarefa estática (UPADHYAY et al., 2025; MEHTA et al., 2025). Essa simplificação, embora conveniente do ponto de vista experimental, descarta informação diagnóstica valiosa, pois a evolução temporal dos sintomas, desde um estado assintomático até manifestações fenotípicas pronunciadas, contém padrões discriminativos que podem ser explorados por modelos de aprendizado profundo. A presente pesquisa propõe um *framework* experimental que coloca essa evolução temporal no centro da análise, o que permite investigar a partir de que momento os sinais visuais se tornam suficientemente discriminativos para que um classificador diferencie múltiplas deficiências com confiança.

Neste capítulo são detalhadas todas as etapas do processo experimental, desde a caracterização do conjunto de dados até as métricas de avaliação e os mecanismos de interpretabilidade empregados. A metodologia foi estruturada para garantir a reprodutibilidade dos resultados e permitir uma análise sistemática do impacto tanto da dimensão temporal quanto da escolha arquitetural. Cada decisão de *design* é apresentada com sua respectiva justificativa, tais como os pivôs metodológicos realizados durante a pesquisa, a exemplo da transição da validação LOGO para o *hold-out* estratificado, de modo a oferecer uma visão transparente do processo investigativo. A organização das seções segue a ordem lógica do *pipeline* experimental, que abrange a caracterização dos dados (Seção 4.1), o pré-processamento (Seção 4.2), o particionamento temporal (Seção 4.3), as arquiteturas de classificação (Seção 4.4), o protocolo de treinamento (Seção 4.5), a estratégia de validação (Seção 4.6), as métricas de avaliação (Seção 4.7) e a interpretabilidade (Seção 4.8).

O fluxograma completo do *pipeline* experimental é apresentado na Figura 10, na qual são apresentadas, de forma organizada, as etapas descritas nas seções seguintes em três

fases sequenciais, sendo elas a preparação e o particionamento temporal dos dados, conforme a Seção 4.3; treinamento dos modelos com divisão *hold-out* estratificada, conforme as Seções 4.4–4.6; e avaliação quantitativa e qualitativa dos resultados, conforme as Seções 4.7–4.8.

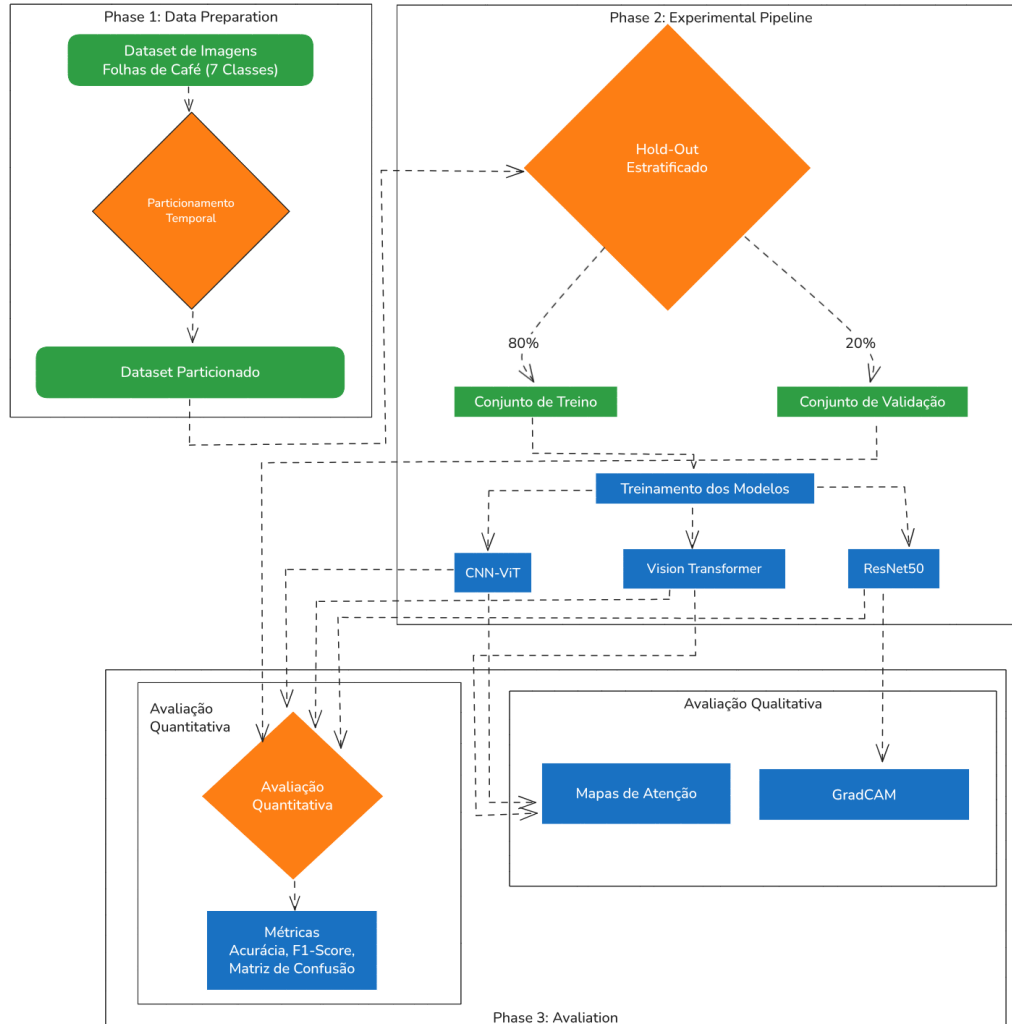


Figura 10 – Fluxograma do *pipeline* experimental adotado neste trabalho, organizado em três fases: preparação dos dados com particionamento temporal, treinamento dos modelos com divisão *hold-out* estratificada, e avaliação quantitativa (métricas) e qualitativa (mapas de atenção e Grad-CAM).

4.1 Conjunto de Dados

O conjunto de dados utilizado nesta pesquisa foi construído a partir de um experimento controlado de cultivo de cafeeiros da espécie *Coffea arabica*, conduzido em ambiente de estufa, pelos professores Thiago Pirola Ribeiro e Enio Tarso de Souza Costa. A escolha por um experimento em condições controladas, em detrimento de coletas em campo aberto, foi deliberada e fundamenta-se no fato de o ambiente de estufa permitir isolar o efeito de cada

deficiência nutricional individualmente, o que elimina variáveis de confusão como pragas, doenças, variações climáticas extremas, interações nutricionais múltiplas e diferenças de solo que inevitavelmente estariam presentes em condições de campo (MARSCHNER, 2012). Dessa forma, cada sintoma visual observado nas folhas pode ser atribuído com maior confiança à deficiência específica induzida no tratamento correspondente, o que confere ao *dataset* uma qualidade de anotação que seria difícil de alcançar com dados coletados em lavouras comerciais. Essa pureza na relação causa-efeito é essencial para que as conclusões sobre a classificabilidade temporal dos sintomas sejam válidas e não contaminadas por fatores exógenos. Na Figura 11 é apresentado o registro fotográfico desse arranjo experimental, que evidencia tanto a identificação individual dos vasos por tratamento quanto o protocolo de aquisição das imagens.



(a) Vaso identificado por tratamento e réplica.



(b) Aquisição de imagem sobre suporte branco.

Figura 11 – Registro fotográfico do experimento controlado conduzido em casa de vegetação. Em (a), um dos vasos do experimento, embrulhado para isolamento do substrato e identificado por etiqueta que informa o número do vaso, a réplica biológica R4 e o tratamento nutricional aplicado, neste caso a omissão de enxofre, sinalizada por –S. Em (b), uma folha de *Coffea arabica* posicionada sobre suporte branco, o que ilustra o protocolo de aquisição das imagens no espectro RGB por câmera digital convencional. Fonte: o autor.

As plantas foram submetidas a sete tratamentos nutricionais distintos, cada um como uma classe no problema de classificação. O grupo de controle, denominado “Completo”, recebeu solução nutritiva balanceada com todos os macronutrientes essenciais, e serve como referência para o estado saudável das plantas. Os seis grupos experimentais foram constituídos por soluções nutritivas com a omissão individual de um macronutriente essencial, sendo eles Nitrogênio (N), Fósforo (P), Potássio (K), Cálcio (Ca), Magnésio (Mg) e Enxofre (S). A seleção desses seis macronutrientes fundamenta-se na literatura agrônômica clássica, que os identifica como os nutrientes cujas deficiências produzem sintomas

visuais mais característicos e distinguíveis em cafeeiros (MALAVOLTA, 2006). Cada macronutriente desempenha funções fisiológicas distintas na planta, sendo o N constituinte de aminoácidos e clorofila, o P participante de reações energéticas e transferência de fosfatos, o K regulador da pressão osmótica e da ativação enzimática, o Ca componente estrutural das paredes celulares, o Mg o átomo central da molécula de clorofila e o S integrador de aminoácidos sulfurados e coenzimas (MARSCHNER, 2012). Essas diferenças funcionais resultam em manifestações visuais distintas quando cada nutriente é omitido, embora com graus variados de diferenciação visual, o que torna o problema de classificação particularmente desafiador.

O período de coleta das imagens estendeu-se de outubro de 2016 a março de 2017, o que compreende aproximadamente cinco meses de acompanhamento contínuo. Essa escolha de intervalo se deu para que fosse possível capturar a progressão completa dos sintomas, desde os estágios assintomáticos iniciais, quando as reservas nutricionais da planta ainda são suficientes para manter a aparência saudável, até as manifestações avançadas de cada deficiência, com sinais fenotípicos claramente visíveis como clorose generalizada, necrose marginal e descoloração internerval. As imagens foram adquiridas com câmera digital convencional no espectro visível (RGB), uma escolha que, conforme discutido no Capítulo 2, favorece a escalabilidade e a aplicabilidade prática do sistema, uma vez que sensores RGB são onipresentes em *smartphones* e câmeras de baixo custo (LIAKOS et al., 2018). Cada registro no conjunto de dados contém, além da imagem propriamente dita, metadados essenciais para a análise temporal, nomeadamente a classe de deficiência correspondente, a identificação da planta individual (vaso) e um *timestamp* que indica a data de captura. Essa tripla de informações (classe, vaso e data) constitui a base para o particionamento temporal descrito na Seção 4.3.

O conjunto de dados totaliza 28.086 registros no arquivo CSV de metadados, dos quais 2.724 correspondem a imagens de referência e 25.362 a amostras experimentais propriamente ditas. As imagens de referência são empregadas exclusivamente na etapa de calibração do pré-processamento, descrita na Seção 4.2.1, ao passo que as 25.362 amostras experimentais constituem o conjunto efetivamente submetido à classificação. A organização dos dados segue uma estrutura hierárquica de diretórios no formato `classe/vaso/timestamp/imagem.JPG`, onde cada classe contém quatro réplicas biológicas identificadas como R1, R2, R3 e R4, e cada vaso contém subdiretórios organizados por data de captura no formato YYYYMMDD. As quatro réplicas biológicas por tratamento correspondem a um número típico de experimentos em casa de vegetação, suficiente para inferência sobre os tratamentos nutricionais avaliados, mas que limita a generalização sobre variabilidade genotípica e fenológica em condições naturais. Essa estrutura é também relevante para a investigação de estratégias de validação baseadas em agrupamento por planta, conforme discutido na Seção 4.6. Cada vaso contém aproximadamente mil imagens distribuídas ao longo do período experimental, e a ordenação cronológica dos diretórios de *timestamp*

corresponde diretamente à linha temporal do experimento, o que facilita a implementação do particionamento temporal de forma programática.

A distribuição de amostras por classe é apresentada na Tabela 2 e, graficamente ilustrada na Figura 12. A análise revela um desbalanceamento moderado entre as classes, pois a classe Completo, correspondente ao grupo de controle, possui o menor número de amostras (3.528), enquanto a classe Mg apresenta o maior volume (4.339), o que resulta em uma razão de desbalanceamento de aproximadamente 1:1,23. Essa diferença, embora existente, não é severa o suficiente para exigir técnicas agressivas de balanceamento como *oversampling* sintético, a exemplo do SMOTE, ou *undersampling* aleatório, sendo adequadamente tratada pela adoção de métricas que consideram igualmente todas as classes na avaliação, conforme será detalhado na Seção 4.7. O menor número de amostras da classe Completo é esperado, pois plantas com nutrição completa tendem a apresentar menor variabilidade visual ao longo do tempo, o que resultou em menos capturas consideradas relevantes pelo protocolo de aquisição.

Tabela 2 – Distribuição de amostras por classe no conjunto de dados. A coluna “Amostras” indica o número total de registros, além das referências e amostras experimentais.

Classe	Descrição	Amostras
Completo	Controle (nutrição completa)	3.528
N	Deficiência de N	3.987
P	Deficiência de P	4.007
K	Deficiência de K	4.028
Ca	Deficiência de Ca	4.118
Mg	Deficiência de Mg	4.339
S	Deficiência de S	4.079
Total		28.086

Na Figura 13 são apresentados os exemplos representativos de folhas para cada uma das sete classes, o que permite uma apreciação visual das diferenças fenotípicas entre os tratamentos em estágios avançados de manifestação. Observa-se que, enquanto algumas deficiências produzem sintomas facilmente distinguíveis a olho nu, como a clorose generalizada e uniforme característica da deficiência de N ou a necrose marginal típica da deficiência de K, outras apresentam manifestações visuais mais sutis e potencialmente confundíveis entre si. Em particular, as deficiências de P e S, embora se manifestem em compartimentos foliares distintos, com sintomas predominantes em folhas mais velhas no caso de P, pela elevada mobilidade do nutriente no floema, e em folhas mais novas no caso de S, de mobilidade restrita, podem produzir alterações de coloração com tonalidades amareladas pouco distintivas em estágios iniciais, o que representa um desafio significativo para a classificação automatizada (MALAVOLTA, 2006). A classe Completo, por sua vez, apresenta folhas com aparência saudável que, nos estágios iniciais de outras defi-

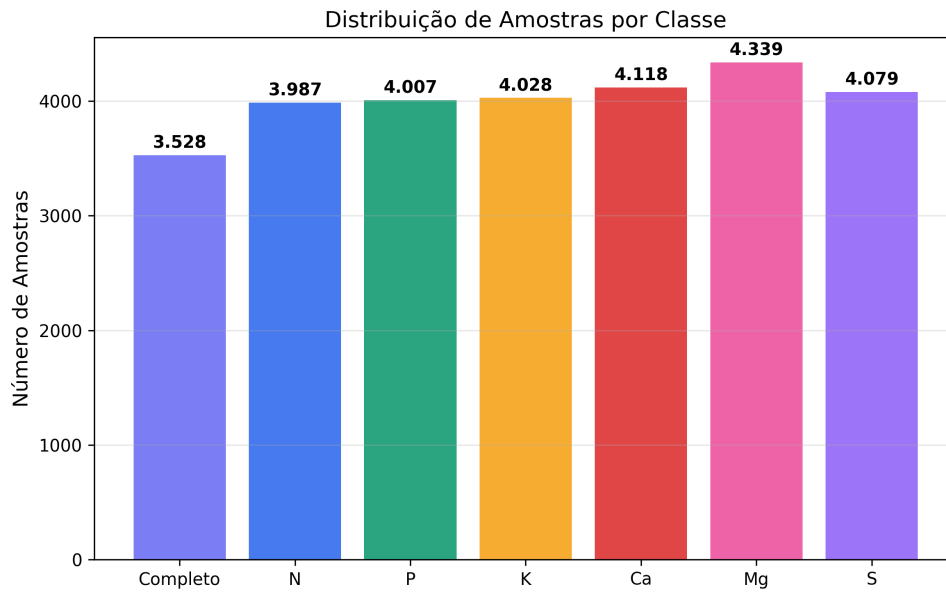


Figura 12 – Distribuição do número de amostras por classe no conjunto de dados. Observa-se um desbalanceamento moderado, com a classe Completo com o menor número de amostras (3.528) e a classe Mg com o maior (4.339).

ciências, podem ser visualmente indistinguíveis das folhas com nutrição deficiente, o que ajuda a explicar a dificuldade de classificação observada para esta classe nos resultados experimentais. Essas observações qualitativas motivam a necessidade de modelos de classificação com capacidade de capturar padrões texturais finos e reforçam a importância da análise temporal proposta neste trabalho.

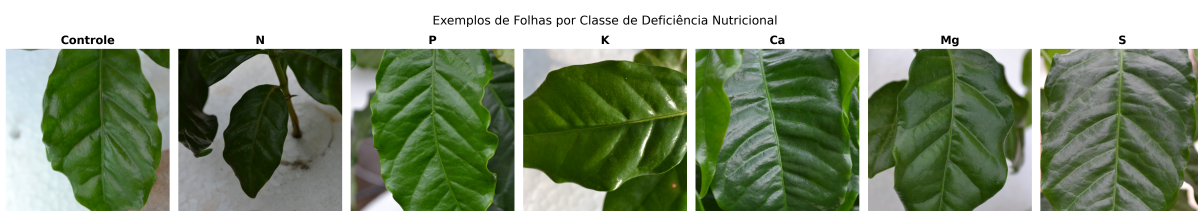


Figura 13 – Exemplos representativos de folhas para cada uma das sete classes do conjunto de dados em estágios avançados de manifestação. As imagens exemplares foram selecionadas a partir das frações temporais avançadas do dataset, sem critério algorítmico formal, com a finalidade exclusiva de ilustrar a diversidade fenotípica entre classes; não constituem amostra estatisticamente representativa. As diferenças fenotípicas variam de sintomas pronunciados, como a clorose uniforme em N e a necrose marginal em K, a manifestações mais sutis, como em Ca e S, o que evidencia a complexidade do problema de classificação multiclasse.

4.2 Pré-processamento

O pré-processamento das imagens constitui uma etapa fundamental para a padronização da entrada dos modelos de *deep learning* e para a maximização do aproveitamento da informação visual contida nas folhas. O *pipeline* de pré-processamento foi implementado integralmente com a biblioteca `torchvision.transforms` do *framework* PyTorch (PASZKE et al., 2019), o que garante integração nativa com o fluxo de treinamento, reprodutibilidade das transformações aplicadas e compatibilidade com a aceleração por Unidade de Processamento Gráfico (GPU). As subseções a seguir detalham cada componente do *pipeline*.

4.2.1 Recorte Central Quadrado

As imagens originais do conjunto de dados possuem dimensões variadas, uma vez que as folhas foram fotografadas em diferentes ângulos e distâncias ao longo do período experimental, e folhas de tamanhos distintos resultam naturalmente em enquadramentos com proporções diversas. Para criar uma entrada uniforme sem distorcer a proporção natural das folhas, foi implementada uma transformação personalizada denominada `SquareCenterCrop`. Essa transformação opera em duas etapas logicamente distintas. Primeiramente, uma etapa de calibração percorre todas as imagens de referência do *dataset* para identificar a menor dimensão, seja a altura ou a largura, dentre todas elas, e estabelece esse valor como parâmetro global de recorte. Em seguida, durante o carregamento de cada imagem, um recorte quadrado centralizado é aplicado com essa dimensão, o que garante que a região de maior interesse, a folha posicionada no centro do enquadramento, seja preservada integralmente em todas as amostras.

A escolha pelo recorte central quadrado, em detrimento de alternativas como o redimensionamento direto com alteração de proporção, *aspect ratio*, ou o recorte aleatório, *random crop*, fundamenta-se na necessidade de preservar as características geométricas e texturais das folhas sem introduzir variabilidade artificial indesejada. Um redimensionamento não proporcional poderia distorcer a geometria das folhas e os padrões de clorose e necrose, sintomas reconhecidos como os principais indicadores visuais de deficiência nutricional (MARSCHNER, 2012). Já o recorte aleatório, embora útil como técnica de aumento de dados, poderia excluir regiões relevantes da folha de forma não controlada, o que compromete a consistência entre amostras. O recorte central opera sob a premissa razoável de que a folha ocupa predominantemente a região central da imagem, o que é consistente com o protocolo padronizado de captura utilizado no experimento controlado em estufa. Reconhece-se, contudo, que essa premissa de centralização não foi verificada quantitativamente e que o recorte pelo menor lado global pode, em folhas de maiores dimensões, suprimir porções marginais da lâmina, justamente onde se manifestam sintomas como a necrose marginal característica da deficiência de K, o que constitui uma ameaça

à validade a ser considerada na interpretação dos resultados.

Após o recorte central quadrado, as imagens são redimensionadas para a resolução de entrada de cada modelo avaliado por meio da transformação `Resize` da `torchvision`¹. Duas resoluções foram investigadas neste trabalho, sendo elas 314×314 e 512×512 pixels. A resolução de 314×314 foi adotada como padrão para os experimentos comparativos entre as três arquiteturas, por oferecer um equilíbrio adequado entre a quantidade de informação visual preservada e o custo computacional do treinamento, que inclui tanto o tempo de processamento quanto o consumo de memória GPU. A resolução superior de 512×512 foi empregada em experimentos adicionais com a arquitetura ResNet50 e, parcialmente, com o modelo híbrido, para avaliar quantitativamente o impacto da resolução de entrada na acurácia de classificação. A análise desse impacto é apresentada no Capítulo 5.

4.2.2 Aumento de Dados e Normalização

Técnicas de aumento de dados foram aplicadas exclusivamente ao conjunto de treinamento, com o objetivo de enriquecer artificialmente a variabilidade das amostras e, conseqüentemente, melhorar a capacidade de generalização dos modelos para dados não vistos. É fundamental ressaltar que o conjunto de validação não recebe nenhuma transformação de aumento, sendo submetido apenas ao recorte central, ao redimensionamento e à normalização. Essa distinção entre os *pipelines* de treinamento e validação é uma prática essencial em aprendizado de máquina (GOODFELLOW; BENGIO; COURVILLE, 2016), pois garante que as métricas de avaliação reflitam o desempenho real do modelo em dados não perturbados, o que evita estimativas artificialmente otimistas de acurácia que comprometeriam a validade das comparações entre modelos e configurações temporais.

As transformações de aumento de dados aplicadas ao conjunto de treinamento foram selecionadas com base em critérios de plausibilidade biológica e relevância prática e implementadas por meio da biblioteca `torchvision.transforms`, conforme descrito a seguir.

- ❑ **RandomHorizontalFlip**: espelhamento horizontal aleatório da imagem com probabilidade de 50%. Essa transformação é biologicamente plausível, uma vez que a orientação horizontal de uma folha não altera suas características de deficiência nutricional, pois os padrões de clorose, necrose e descoloração são simétricos ou independentes da orientação horizontal. A transformação contribui para que o modelo aprenda representações invariantes à orientação, o que aumenta efetivamente o tamanho do conjunto de treinamento sem introduzir artefatos que possam confundir a classificação.
- ❑ **ColorJitter**: variações aleatórias nos parâmetros de brilho (*brightness*=0,2), contraste (*contrast*=0,2) e saturação (*saturation*=0,2). Cada parâmetro define uma

¹ <<https://pytorch.org/vision/stable/>>

faixa de variação de $\pm 20\%$ em torno do valor original, ao simular diferentes condições de iluminação que poderiam ocorrer durante a captura das imagens em uma aplicação real. Essa transformação é particularmente relevante para o contexto agrícola, em que a iluminação natural varia ao longo do dia e entre sessões de captura, e torna o modelo mais robusto a variações fotométricas que não estão intrinsecamente relacionadas à deficiência nutricional em si.

A opção por não incluir transformações geométricas adicionais, como rotações arbitrárias, distorções de perspectiva ou *random erasing*, foi deliberada. Rotações de ângulos grandes poderiam alterar a orientação relativa das nervuras e margens foliares, que são referências anatômicas importantes para a manifestação de certos sintomas, como a *chlorotic intervein* característica da deficiência de Mg, que depende da relação espacial entre nervuras e lâmina foliar (MALAVOLTA, 2006). Já técnicas de oclusão aleatória poderiam mascarar justamente as regiões foliares que contêm a informação diagnóstica mais relevante, o que prejudica o aprendizado das representações discriminativas.

A etapa final do pré-processamento consiste na conversão das imagens para o formato de tensor e na normalização dos valores de pixel com o uso das médias ($\mu = [0, 485; 0, 456; 0, 406]$) e os desvios padrão ($\sigma = [0, 229; 0, 224; 0, 225]$) calculados sobre o conjunto ImageNet. Essa normalização é uma prática padrão quando se utiliza *transfer learning* a partir de modelos pré-treinados na ImageNet, pois garante que a distribuição estatística das entradas seja compatível com aquela sobre a qual os pesos pré-treinados foram otimizados. Uma normalização baseada nas estatísticas do próprio conjunto de dados de folhas de café poderia ser uma alternativa, porém a adoção das estatísticas do ImageNet preserva a compatibilidade com os filtros aprendidos na fase de pré-treinamento, que constituem o ponto de partida para o *fine-tuning*. A aplicação da normalização ImageNet tanto ao conjunto de treinamento quanto ao de validação assegura a consistência do pré-processamento em ambas as fases do processo experimental.

4.3 Particionamento Temporal

O particionamento temporal constitui o elemento central e mais distintivo da metodologia proposta neste trabalho, o que representa a principal contribuição metodológica em relação à abordagem convencional da literatura. Enquanto trabalhos anteriores sobre classificação de doenças e deficiências em plantas utilizam todas as imagens disponíveis indiscriminadamente para treinar modelos de classificação, sem considerar o estágio temporal de desenvolvimento dos sintomas (FERENTINOS, 2018; ESGARIO; KROHLING; VENTURA, 2020; AMUDHA; BRINDHA, 2024), a presente pesquisa introduz uma dimensão temporal explícita ao processo experimental. Essa abordagem é motivada pela natureza inerentemente progressiva dos estresses nutricionais, uma vez que uma deficiência

não se manifesta instantaneamente, mas evolui gradualmente de um estado assintomático, quando as reservas internas da planta ainda são suficientes para manter as funções metabólicas, para sintomas sutis e, finalmente, para sinais fenotípicos pronunciados e facilmente reconhecíveis (MALAVOLTA, 2006; MARSCHNER, 2012). Compreender em que ponto dessa progressão os sintomas se tornam visualmente discriminativos, e se modelos de *deep learning* conseguem capturar sinais de deficiência mesmo nos estágios iniciais, constitui uma contribuição metodológica complementar à maximização de acurácia.

O mecanismo criado, neste trabalho, para controlar o particionamento temporal foi adotar um parâmetro denominado `DATA_START_FRAC`, que define a fração temporal a partir da qual os dados são selecionados para o treinamento e a avaliação de cada experimento. Concretamente, o procedimento opera da seguinte forma. Todas as imagens do *dataset* são ordenadas cronologicamente com base no *timestamp* de captura e, em seguida, apenas as que se encontram na fração final do período total são selecionadas para compor o conjunto de dados daquele experimento, conforme formaliza a Equação 11.

$$\mathcal{D}_f = \{ x_{(i)} : \lfloor f \cdot N \rfloor < i \leq N \} \quad (11)$$

Nessa formulação, $x_{(i)}$ denota a i -ésima imagem do *dataset* ordenado por *timestamp* crescente, N é o número total de imagens e $f = \text{DATA_START_FRAC}$ é a fração inicial descartada, de modo que são retidas as últimas $N - \lfloor f \cdot N \rfloor$ imagens, correspondentes aproximadamente à fração final $(1 - f)$ do período. A título de ilustração, a fração retida corresponde aos últimos 40% para `frac60` e aos últimos 5% para `frac95`. Dessa forma, valores menores de `DATA_START_FRAC` abrangem janelas temporais mais amplas, o que engloba tanto os estágios de sintomas mais pronunciados quanto períodos anteriores em que as manifestações visuais ainda eram incipientes ou ausentes, enquanto valores maiores restringem a seleção aos *timestamps* mais recentes do período experimental, nos quais os sintomas se encontram em estágios avançados de desenvolvimento, porém com um volume de amostras substancialmente menor.

Foram investigadas dez configurações temporais, espaçadas em intervalos de 5 pontos percentuais, cada uma correspondente a uma largura distinta de janela temporal:

- `frac50`, que cobre os últimos 50% do período experimental: Configuração mais abrangente, que inclui a maior janela temporal. Embora abranja os estágios com sintomas mais pronunciados, incorpora também um volume significativo de imagens de períodos assintomáticos ou de sintomas incipientes, o que introduz ruído no conjunto de treinamento.
- `frac55` a `frac65`, que cobrem os últimos 45% a 35% do período experimental: Configurações intermediárias que progressivamente excluem os dados mais antigos, restringindo o conjunto majoritariamente a períodos em que os sintomas já se encontram desenvolvidos, com exclusão progressiva das imagens assintomáticas mais antigas.

- ❑ **frac70 a frac75**, que cobrem os últimos 30% a 25% do período experimental: Configurações que restringem os dados ao terço final do período experimental, em que os sintomas são visualmente pronunciados, porém com menor volume de treinamento.
- ❑ **frac80 a frac90**, que cobrem os últimos 20% a 10% do período experimental: Janelas progressivamente mais estreitas, com volume de dados significativamente reduzido. Embora os *timestamps* retidos correspondam a estágios avançados dos sintomas, a escassez de amostras limita a capacidade de treinamento dos modelos, o que resulta em queda de desempenho.
- ❑ **frac95**, que retém apenas os últimos 5% do período experimental: Configuração mais restritiva. Os sintomas encontram-se em estágio avançado de desenvolvimento; contudo, o volume de amostras é extremamente reduzido, o que torna a diferenciação entre classes particularmente desafiadora devido à escassez de dados de treinamento.

Cabe esclarecer que valores numericamente maiores de `DATA_START_FRAC` correspondem a janelas temporais mais estreitas e, portanto, a menores volumes de dados. Como cada janela mais estreita é um subconjunto da janela mais ampla ($\text{frac95} \subset \text{frac75} \subset \text{frac60} \subset \text{frac50}$), as configurações mais restritivas retêm apenas os *timestamps* mais recentes do período experimental, nos quais os sintomas se encontram em estágio avançado de desenvolvimento, contudo, o volume de amostras resultante pode ser insuficiente para o treinamento eficaz dos modelos.

Cabe registrar que as dez frações investigadas resultam de um processo de busca exploratória sobre o espaço de janelas temporais, em que parte das configurações foi incluída à luz de tendências preliminares observadas nas frações inicialmente avaliadas. É reconhecido que esse desenho adaptativo limita a inferência estatística formal sobre a posição exata do ponto ótimo identificado, e que a confirmação rigorosa dessa posição exigiria experimento confirmatório com frações pré-registradas, conduzido independentemente dos dados que motivaram a expansão da grade temporal.

Na Figura 14 é possível visualizar a evolução temporal dos sintomas de deficiência de cálcio (Ca) ao longo de três momentos representativos do período experimental. No estágio inicial, a folha apresenta coloração verde uniforme, sem alterações visuais perceptíveis. No estágio intermediário, surgem lesões necróticas punctiformes de coloração alaranjada distribuídas pela lâmina foliar, o que indica o início do comprometimento celular. No estágio tardio, observa-se amarelecimento difuso da lâmina foliar com manchas amarelo-esverdeadas, característica do estágio avançado da deficiência. Essa progressão visual é representativa do fenômeno que o particionamento temporal busca capturar, especificamente, a informação diagnóstica contida na evolução dos sintomas, que é inevitavelmente perdida quando todos os estágios temporais são misturados indiscriminadamente em um único conjunto de treinamento. Ao segregar os dados por janela temporal, cada modelo é desafiado a aprender representações discriminativas apenas com o nível de informação

visual disponível naquele estágio específico, o que permite uma avaliação mais precisa e informativa da capacidade real dos algoritmos.



Figura 14 – Evolução temporal dos sintomas de deficiência de Ca em três estágios representativos do período experimental. Da esquerda para a direita: estágio inicial (28/10/2016), com folha de coloração verde uniforme e sem alterações visíveis; estágio intermediário (10/02/2017), com lesões necróticas punctiformes distribuídas pela lâmina; e estágio tardio (15/03/2017), com amarelecimento difuso e manchas amarelo-esverdeadas, características do estágio avançado da deficiência de Ca.

É importante destacar que a variação no número total de amostras entre as configurações temporais é uma consequência natural da estratégia de particionamento, visto que a configuração **frac50**, que inclui os últimos 50% dos dados temporais, contém substancialmente mais imagens do que a configuração **frac95**, que inclui apenas os últimos 5%. As implicações dessa diferença de volume como potencial fator de confusão na interpretação dos resultados, bem como as estratégias para isolá-lo, são discutidas no Capítulo 5.

Após a seleção temporal das imagens, o conjunto resultante para cada configuração temporal é dividido em dois subconjuntos, treinamento e validação, na proporção de 80% e 20%, respectivamente. Essa divisão é realizada de forma estratificada por classe, por meio da função **train_test_split** da biblioteca **scikit-learn** (PEDREGOSA et al., 2011), o que garante que a proporção relativa de cada classe seja mantida tanto no conjunto de treinamento quanto no de validação. A estratificação é particularmente importante neste contexto, dado o desbalanceamento moderado entre as classes identificado na Seção 4.1, de modo que nenhuma classe fique sub-representada em um dos subconjuntos e comprometa a confiabilidade da avaliação. A semente aleatória fixa (**SEED=42**) assegura que a divisão seja determinística e reproduzível em todas as execuções, o que permite que qualquer pesquisador replique exatamente os mesmos conjuntos de treinamento e validação ao utilizar os mesmos parâmetros.

4.4 Arquiteturas de Classificação

A escolha das arquiteturas de classificação foi orientada por dois critérios complementares, a representatividade em relação ao estado da arte em visão computacional e a diversidade de mecanismos de extração de características. Foram avaliadas três arquiteturas distintas, uma CNN pura, um ViT puro e um modelo híbrido CNN-ViT de concepção original, todas com pesos pré-treinados no conjunto ImageNet como ponto de partida, em uma estratégia de *transfer learning* (HE et al., 2016; DOSOVITSKIY et al., 2021). Essa estratégia permite que os modelos iniciem o treinamento com representações visuais genéricas já aprendidas, como detecção de bordas, texturas e formas básicas, que são então refinadas para o domínio específico de folhas de café com deficiências nutricionais. A diversidade arquitetural permite investigar uma questão fundamental, qual seja, se o viés indutivo de localidade das convoluções ou a capacidade de modelagem de dependências globais dos mecanismos de atenção é mais adequado para capturar os padrões visuais característicos de deficiências nutricionais em folhas de *Coffea arabica*. Na Figura 15 é apresentada uma visão geral comparativa das três arquiteturas investigadas de modo a ficar evidenciada as diferentes estratégias de extração de características adotadas por cada uma, pela extração local via convoluções, pela global via mecanismos de atenção e pela combinação de ambas, cujos detalhes são descritos nas subseções 4.4.1, 4.4.2 e 4.4.3.

4.4.1 ResNet50

A arquitetura ResNet50, proposta por He et al. (2016), foi adotada como a principal linha de base convolucional deste trabalho. A contribuição fundamental da família ResNet reside na introdução das conexões residuais, *skip connections*, que permitem o treinamento eficaz de redes profundas ao criar caminhos diretos para o fluxo do gradiente, o que mitiga o problema de degradação que afeta redes convolucionais com muitas camadas (HE et al., 2016). A variante com 50 camadas oferece um equilíbrio favorável entre profundidade representacional, suficiente para capturar hierarquias complexas de características visuais, e custo computacional, e figura entre as arquiteturas CNN amplamente empregadas em tarefas agrícolas, entre elas a detecção de doenças foliares (SAKKA et al., 2025). Neste trabalho, o modelo foi inicializado com os pesos pré-treinados `IMAGENET1K_V2`, disponibilizados pela biblioteca `torchvision`, que correspondem a uma versão aprimorada do treinamento original, com técnicas modernas de regularização e aumento de dados que resultam em representações mais robustas.

A adaptação da ResNet50 para o problema de classificação de deficiências nutricionais consistiu na substituição da camada totalmente conectada, *fully connected*, originalmente projetada para as 1.000 classes do ImageNet, por uma nova camada linear com saída para sete classes, correspondentes aos sete tratamentos nutricionais do experimento. Todas as demais camadas do modelo foram mantidas com seus pesos pré-treinados como inici-

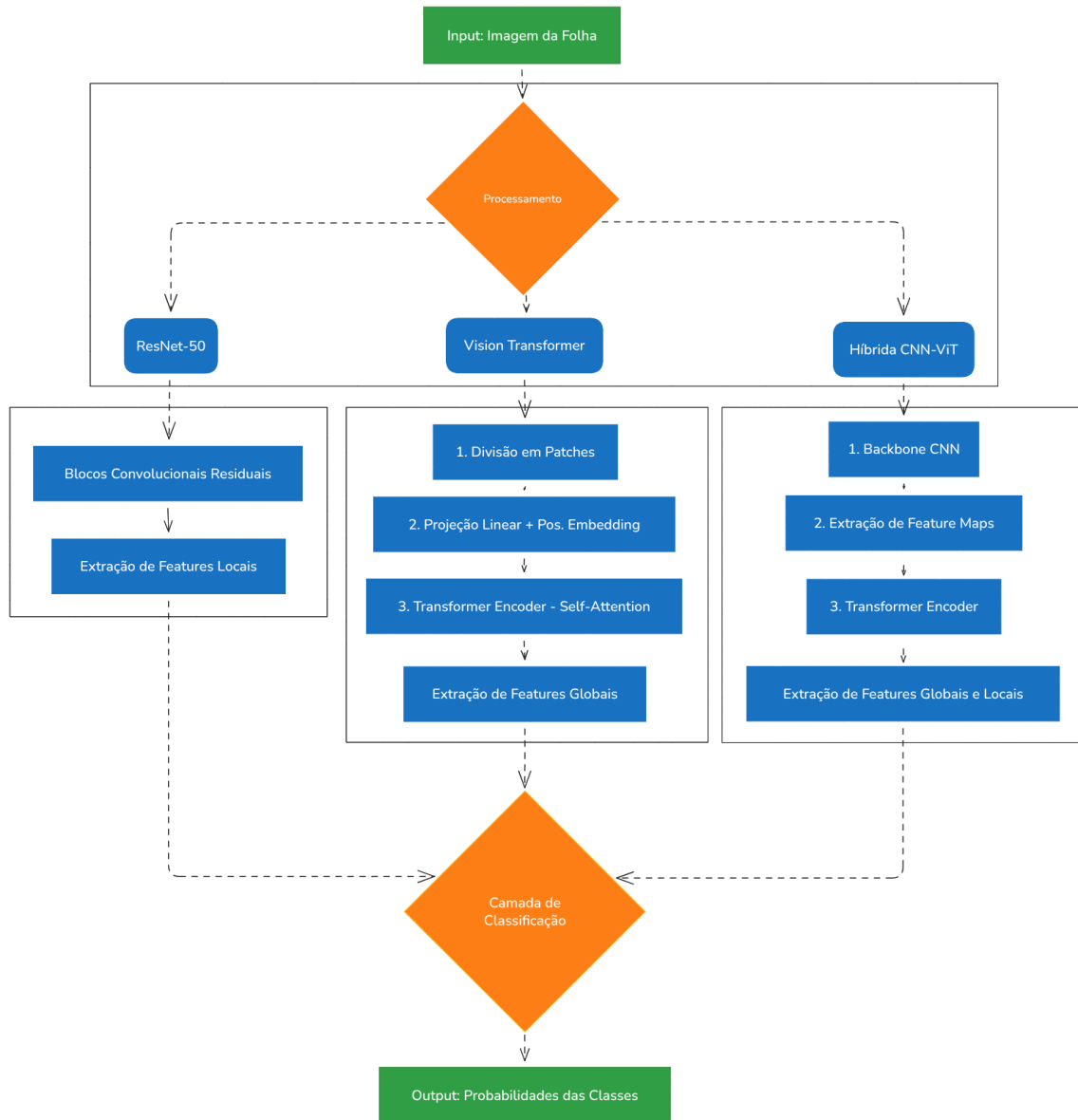


Figura 15 – Visão geral comparativa das três arquiteturas investigadas: ResNet50 (esquerda), ViT (centro) e Híbrido CNN-ViT (direita). O diagrama ilustra o fluxo de processamento desde a imagem de entrada até a saída de probabilidades das classes e destaca as diferenças nos mecanismos de extração de características de cada arquitetura.

alização, porém configuradas como ajustáveis durante o treinamento, o que caracteriza uma estratégia de *fine-tuning* completo. Essa decisão, em contraste com o congelamento parcial de camadas, *feature extraction*, permite que todas as representações aprendidas sejam refinadas para o domínio específico de folhas de café, uma vez que as características visuais de deficiências nutricionais, como clorose, necrose e alterações de textura, diferem substancialmente dos objetos e cenas presentes nas categorias genéricas do ImageNet.

O viés indutivo de localidade inerente às operações convolucionais constitui uma vantagem teórica significativa para a tarefa em questão. As deficiências nutricionais

manifestam-se por meio de alterações texturais frequentemente localizadas nas folhas, como clorose internerval, necrose marginal, descoloração de pontas e manchas difusas, que são naturalmente capturadas pelos filtros convolucionais em múltiplas escalas (ZEILER; FERGUS, 2014). Nas camadas iniciais da rede, os filtros aprendem a detectar bordas, gradientes de cor e texturas de baixo nível; nas camadas intermediárias, padrões mais complexos como manchas, transições de cor e contornos foliares são representados; e nas camadas mais profundas, especificamente a **layer4**, representações semânticas de alto nível encapsulam os padrões composicionais que distinguem as diferentes deficiências. Essa extração hierárquica progressiva é particularmente relevante para distinguir entre deficiências que produzem padrões visuais semelhantes mas em regiões foliares distintas, como a clorose internerval do Mg versus a clorose uniforme do N.

A ResNet50 foi avaliada em duas resoluções de entrada, 314×314 e 512×512 pixels, e em todas as dez configurações temporais, de frac50 a frac95, o que constitui o conjunto mais abrangente de experimentos dentre as três arquiteturas. Essa extensividade justifica-se pelo desempenho consistentemente superior observado nos experimentos preliminares, que motivou uma investigação mais detalhada de suas capacidades em diversas condições. A comparação entre resoluções, em particular, permite avaliar o *trade-off* entre informação visual e custo computacional no contexto específico da classificação de deficiências nutricionais.

4.4.2 ViT-B/16

O ViT, proposto por Dosovitskiy et al. (2021), representa uma abordagem fundamentalmente distinta para a classificação de imagens, baseada exclusivamente em mecanismos de autoatenção sem qualquer operação convolucional. Na variante ViT-B/16 utilizada neste trabalho, a imagem de entrada é dividida em *patches* não sobrepostos de 16×16 pixels, que são projetados linearmente para um espaço de representação com dimensão fixa de 768 na variante *Base* e processados como uma sequência de *tokens*, de maneira análoga às palavras em um modelo de processamento de linguagem natural. *Embeddings* posicionais aprendíveis são adicionados a cada *token* para preservar a informação de localização espacial, e um *token* especial de classificação (CLS) é preposto à sequência para agregar a representação global da imagem ao longo das camadas do codificador *Transformer*. O modelo foi inicializado com pesos pré-treinados IMAGENET1K_V1, disponibilizados pela biblioteca `torchvision`, correspondentes ao cenário de pré-treinamento exclusivo no ImageNet-1K, sem os dados adicionais que propiciariam vantagem de escala. A ResNet50, em contrapartida, utiliza a variante IMAGENET1K_V2. Essa assimetria nas condições de pré-treinamento decorre de uma restrição técnica da biblioteca `torchvision`, que disponibiliza a designação IMAGENET1K_V2 exclusivamente para a família ResNet a que foi aplicada a chamada *improved training recipe*, não havendo variante equivalente publicada para o ViT-B/16. As demais opções de pré-treinamento disponíveis para essa arquite-

tura, baseadas no paradigma SWAG, envolvem aprendizado fracamente supervisionado em larga escala, com supervisão derivada de *hashtags*, e não permitiriam comparação direta com o pré-treinamento supervisionado da ResNet50. A adoção da IMAGENET1K_V1 para o ViT-B/16, portanto, não é uma escolha discricionária, mas a única alternativa de pré-treinamento supervisionado em ImageNet-1K disponibilizada pela biblioteca para essa arquitetura, assimetria que, ainda que tecnicamente inevitável, constitui um fator relevante para a interpretação comparativa dos resultados entre as arquiteturas.

A adaptação do ViT-B/16 para resoluções de entrada diferentes daquelas utilizadas no pré-treinamento original (224×224) exige um procedimento de interpolação dos *embeddings* posicionais para a nova dimensão da grade de *patches*. Especificamente, os *embeddings* posicionais originais, que formam uma grade correspondente à resolução de pré-treinamento, são reamostrados por interpolação bicúbica para a grade correspondente à nova resolução de entrada. Essa técnica, implementada no código de treinamento, permite que o modelo opere com resoluções arbitrárias sem restrição estrutural, embora a eficácia da interpolação dependa da proximidade entre a resolução original e a nova, pois resoluções muito distantes podem degradar a qualidade da informação posicional. A cabeça de classificação do ViT foi substituída por uma camada linear com saída para sete classes, segundo a mesma estratégia de adaptação utilizada na ResNet50.

O ViT-B/16 foi avaliado nas mesmas dez configurações temporais (frac50 a frac95) aplicadas às demais arquiteturas, porém restrito à resolução de 512×512 pixels. Essa restrição de resolução justifica-se pelo custo computacional substancialmente maior dos *Transformers* puros em resoluções elevadas, aliado ao desempenho consistentemente inferior observado nos experimentos preliminares, que não justificou a extensão a ambas as resoluções. Essa limitação é consistente com observações amplamente reportadas na literatura, que indicam que *Transformers* puros tendem a requerer volumes de dados substancialmente maiores para superar CNNs em tarefas de escala moderada (TOUVRON et al., 2021; KHAN et al., 2022).

4.4.3 Modelo Híbrido CNN-ViT

A arquitetura híbrida CNN-ViT, implementada na classe `HybridCnnViT`, foi desenvolvida como uma contribuição original deste trabalho, com o objetivo de combinar as capacidades complementares de extração de características locais das CNNs com a modelagem de dependências globais dos mecanismos de atenção. A motivação teórica reside na hipótese de que os sintomas de deficiências nutricionais podem se manifestar tanto por padrões texturais localizados, como manchas de clorose, pontos necróticos e alterações de textura em regiões específicas da folha, adequadamente capturados por convoluções, quanto por relações espaciais entre regiões distintas, como a distribuição espacial da clorose em relação às nervuras ou a progressão da necrose das margens para o centro da folha, que são melhor modeladas por mecanismos de atenção global (DAI et al., 2021;

GUO et al., 2022). Trabalhos recentes têm demonstrado que modelos híbridos que integram convoluções e atenção alcançam resultados competitivos em diversas tarefas de visão computacional, tais como classificação de imagens agrícolas e sensoriamento remoto (GUO; FENG; GUO, 2025; THAKUR et al., 2022).

O *backbone* do modelo híbrido consiste na arquitetura ResNet50 com a remoção da camada de classificação final (`fc`) e da camada de *pooling* global médio (`avgpool1`). Em seu lugar, uma camada de *pooling* adaptativo (`AdaptiveAvgPool2d`) com saída de dimensão fixa 5×5 é aplicada aos mapas de características da última camada convolucional, a `layer4`, o que resulta em 25 regiões espaciais, cada uma representada por um vetor de 2.048 dimensões, correspondente à dimensionalidade dos canais de saída da `layer4` da ResNet50. Essas 25 regiões são conceitualmente análogas aos *patches* do ViT, porém com a diferença fundamental de que já passaram por todos os estágios de extração de características convolucionais da ResNet50, com representações semânticas ricas e abstratas em vez de informações de pixels brutos. Uma camada convolucional de projeção (`Conv2d` com *kernel* 1×1 e sem viés) reduz a dimensionalidade de cada região de 2.048 para 256, o que define o espaço de representação compatível com o codificador *Transformer* subsequente e proporciona uma redução significativa na quantidade de parâmetros da etapa de atenção.

A sequência de entrada do codificador *Transformer* é composta por 26 *tokens*, sendo os 25 *patches* provenientes do *backbone* convolucional acrescidos de um *token* de classificação (CLS), inicializado com valores aleatórios e aprendido conjuntamente com os demais parâmetros durante o treinamento. *Embeddings* posicionais aprendíveis, também inicializados aleatoriamente, são adicionados a todos os 26 *tokens* para codificar a informação de posição espacial relativa de cada região na grade 5×5 . O codificador *Transformer* é composto por seis camadas no total, sendo cinco delas camadas padrão de `TransformerEncoderLayer` do PyTorch, configuradas com oito cabeças de atenção, dimensão de modelo `d_model` de 256 e dimensão da camada *feedforward* de 512, seguidas por uma sexta camada personalizada denominada `CustomTransformerEncoderLayer`. Esta última camada implementa a mesma computação padrão de autoatenção multicabeça, *Multi-Head Self-Attention*, seguida por uma rede *feedforward* com normalização de camada e conexões residuais, conforme a formulação original dos *Transformers* (VASWANI et al., 2017), porém com uma particularidade, pois ela armazena explicitamente os pesos de atenção computados durante a passagem *forward*, expondo-os como um atributo acessível para análise de interpretabilidade posterior.

A classificação final é realizada por meio da extração da representação do *token* CLS após o processamento sequencial pelas seis camadas do codificador, seguida por uma camada linear que projeta o vetor de 256 dimensões para as sete classes do problema. A escolha do *token* CLS como representação agregada segue a prática padrão estabelecida pelo ViT (DOSOVITSKIY et al., 2021), sob a premissa de que o mecanismo de autoatenção permite que esse *token* especial acumule progressivamente informações de todas as 25

regiões da imagem ao longo das camadas do codificador. A camada personalizada final, ao expor os pesos de atenção associados ao *token* CLS, possibilita a geração de mapas de atenção que indicam quais das 25 regiões espaciais foram mais relevantes para a decisão de classificação, conforme detalhado na Seção 4.8.

O modelo híbrido foi avaliado nas resoluções de 314×314 e 512×512 pixels e em todas as dez configurações temporais (frac50 a frac95). A motivação para a investigação desse modelo é dupla. Por um lado, busca-se avaliar se a combinação de características locais (extraídas pela CNN) com relações globais (modeladas pelo *Transformer*) oferece vantagem de desempenho sobre a CNN pura, por outro, verificar se o componente de atenção global é capaz de capturar relações espaciais entre regiões foliares que sejam indicativas de deficiências nutricionais específicas e que escapem à capacidade de modelagem das convoluções.

4.5 Protocolo de Treinamento

Todos os modelos foram treinados com o uso de um protocolo rigorosamente padronizado, de modo a garantir que as diferenças de desempenho observadas entre arquiteturas e configurações temporais possam ser atribuídas exclusivamente às variáveis sob investigação e não a variações no procedimento de otimização. Essa padronização é essencial para a validade das comparações e para a reprodutibilidade dos experimentos. O protocolo foi implementado em um repositório público no GitHub e inserido inicialmente em um *notebook*² Foi executado inicialmente no ambiente **Google Colab**; contudo, devido às restrições de disponibilidade de GPU, com limite de aproximadamente 4 horas por sessão, todos os experimentos finais foram conduzidos em ambiente local equipado com GPU **NVIDIA RTX 5060 Ti** (16 GB VRAM) e 32 GB de RAM. A seguir, cada componente do protocolo é detalhado com suas respectivas justificativas.

O otimizador selecionado foi o Adam, sigla de *Adaptive Moment Estimation*, com uma taxa de aprendizado, *learning rate*, inicial de 1×10^{-4} . O Adam combina as vantagens da adaptação individual da taxa de aprendizado por parâmetro, derivada do RMSprop, com a utilização de momentos de primeira ordem, derivada do SGD com momento, o que resulta em atualizações que se adaptam automaticamente à magnitude e à frequência dos gradientes de cada parâmetro (GOODFELLOW; BENGIO; COURVILLE, 2016). Essa propriedade o torna particularmente adequada para o *fine-tuning* de modelos pré-treinados, em que diferentes camadas podem requerer magnitudes de atualização distintas, visto que as camadas iniciais, cujas representações genéricas, como bordas e texturas, são largamente transferíveis, necessitam de atualizações menores, enquanto as camadas mais profundas e a camada de classificação requerem adaptações mais substanciais ao domínio-alvo. A taxa de aprendizado de 1×10^{-4} é conservadora em relação ao valor padrão do

² disponível em: <<https://github.com/hericlesferraz/coffee-soil-nutrients-dissertation>>.

Adam (1×10^{-3}), uma escolha deliberada para evitar que atualizações excessivamente grandes nos primeiros passos de treinamento destruam as representações pré-treinadas que servem como ponto de partida.

A função de perda utilizada foi a **CrossEntropyLoss**, implementação padrão do PyTorch que combina internamente uma camada *LogSoftmax* com a função de perda de entropia cruzada negativa, também conhecida como *Negative Log-Likelihood Loss*. Essa combinação é o padrão de fato para problemas de classificação multiclasse, e sua implementação unificada é numericamente mais estável do que a aplicação sequencial das duas operações, pois evita o cálculo explícito da exponencial na *Softmax* seguido do logaritmo. A **CrossEntropyLoss** foi aplicada sem pesos de classe, de forma uniforme a todas as sete classes. Embora a atribuição de pesos inversamente proporcionais à frequência de cada classe seja uma técnica comum para lidar com desbalanceamento, o desbalanceamento moderado observado no conjunto de dados (razão máxima de 1:1,23) foi considerado insuficiente para justificar essa intervenção, sendo adequadamente tratado pela adoção de métricas que consideram igualmente todas as classes na avaliação, conforme detalhado na Seção 4.7.

Para ajustar dinamicamente a taxa de aprendizado durante o treinamento, foi empregado o agendador **ReduceLROnPlateau**, configurado com um fator de redução de 0,5 e uma paciência de 3 épocas. Esse agendador monitora continuamente a perda de validação ao final de cada época e, quando não detecta melhoria por três épocas consecutivas, reduz a taxa de aprendizado pela metade, o que permite que o otimizador explore regiões mais finas do espaço de parâmetros em fases avançadas do treinamento, quando a trajetória de otimização se aproxima de mínimos locais. Complementarmente, um mecanismo de parada antecipada com uma paciência de 5 épocas foi implementado para interromper o treinamento quando a perda de validação deixa de apresentar melhoria por 5 épocas consecutivas. Esse mecanismo cumpre dupla função. Por um lado, evita o *overfitting* ao interromper o treinamento antes que o modelo comece a memorizar idiosincrasias do conjunto de treinamento em detrimento da capacidade de generalização. Por outro lado, em conjunto com um mecanismo de *checkpoint*, garante que os pesos retidos ao final do treinamento correspondam ao modelo com a menor perda de validação observada, e não ao estado do último passo de treinamento. O número máximo de épocas foi definido em 20, um valor que, combinado com a parada antecipada, mostrou-se suficiente para a convergência de todos os modelos avaliados em todas as configurações experimentais.

O tamanho de lote, *batch size*, foi fixado em 16 amostras por iteração de treinamento, uma escolha conservadora que se mostrou adequada às restrições de memória GPU durante os experimentos preliminares e foi mantida para garantir consistência em toda a investigação. Embora lotes maiores possam oferecer estimativas de gradiente mais estáveis e favorecer a convergência em termos de épocas, pesquisas recentes demonstram que lotes menores introduzem um efeito de regularização implícita através da variância do

gradiente estocástico, que pode ser benéfico para a generalização do modelo (GOODFELLOW; BENGIO; COURVILLE, 2016). Finalmente, uma semente aleatória (`SEED=42`) foi fixada em todas as bibliotecas relevantes, sendo elas `PyTorch`, com suas operações CUDA em GPU, `NumPy` e o módulo `random` do Python, para garantir a reprodutibilidade completa dos experimentos. Essa semente controla tanto a inicialização dos parâmetros aleatórios (como o *token CLS* e os *embeddings* posicionais do modelo híbrido, e a camada de classificação de todos os modelos) quanto a ordem de apresentação dos dados nos lotes de treinamento e as transformações aleatórias de aumento de dados.

4.6 Estratégia de Validação

A definição da estratégia de validação passou por um processo iterativo motivado pelos desafios inerentes à estrutura do conjunto de dados, composto por imagens de plantas individuais identificáveis ao longo do tempo. A abordagem inicialmente concebida foi a validação cruzada LOGO, na qual cada grupo correspondia a um vaso individual, com quatro vasos por classe, de modo que cada *fold* reservava as imagens de um vaso para validação e treinava o modelo com as imagens dos três vasos restantes. Embora essa premissa fosse metodologicamente rigorosa, ao simular o cenário de aplicação com plantas não vistas durante o treinamento, os experimentos preliminares produziram resultados altamente instáveis, com variância entre *folds* muito superior ao esperado. A análise por *fold* revelou que a causa fundamental não era uma deficiência dos modelos, mas sim a alta variância temporal na progressão dos sintomas entre plantas do mesmo tratamento, dado que fatores individuais, como o estado nutricional prévio e a eficiência de absorção radicular, fazem com que diferentes vasos atinjam estágios de manifestação sintomática em velocidades distintas, o que torna as características visuais da planta reservada para validação substancialmente diferentes daquelas aprendidas a partir das plantas de treinamento.

Diante dessa limitação empírica, a estratégia foi ajustada para um *hold-out* estratificado, no qual as imagens de todas as plantas são agrupadas em um único conjunto e divididas na proporção de 80% para treinamento e 20% para validação, com estratificação por classe via função `train_test_split` da biblioteca `scikit-learn`. É necessário enfatizar, desde já, a principal limitação dessa estratégia, qual seja, a possibilidade de que imagens de uma mesma planta, ainda que distintas entre si, estejam presentes simultaneamente nos conjuntos de treinamento e de validação. Como cada vaso contém aproximadamente mil imagens distribuídas ao longo do período experimental e há apenas quatro réplicas biológicas por classe, o *split* aleatório 80/20 estratificado por classe e não por vaso torna virtualmente certa a presença de cada planta em ambos os subconjuntos. Cabe distinguir essa situação de uma duplicação direta de amostras, dado que a função `train_test_split` opera no nível de imagem individual e, portanto, nenhuma imagem

aparece simultaneamente em treino e validação; o que se compartilha entre os subconjuntos é a identidade da planta fotografada em *timestamps* distintos, e não a mesma imagem em ambos os lados. Em consequência, as métricas reportadas neste trabalho devem ser interpretadas como estimativas de teto da capacidade discriminativa dos modelos sob a estrutura de dados disponível, e não como estimativas de generalização para plantas não observadas.

4.7 Métricas de Avaliação

A avaliação do desempenho dos modelos empregou o conjunto de métricas de classificação descrito na Seção 2.6, selecionado para oferecer uma visão abrangente do comportamento de cada modelo tanto em termos globais quanto em nível de classe individual. A acurácia global foi adotada como métrica primária de comparação, por permitir o confronto direto com trabalhos anteriores que utilizam a mesma métrica. Dado o desbalanceamento moderado entre as classes identificado na Seção 4.1, foram reportados adicionalmente o *macro F1-score*, que atribui igual importância a todas as classes independentemente de sua prevalência, e as métricas de precisão, revocação e *F1-score* por classe, calculadas por meio da função `classification_report` da biblioteca `scikit-learn` (PEDREGOSA et al., 2011). Como instrumento complementar de análise qualitativa, foram geradas matrizes de confusão normalizadas por linha para cada configuração avaliada, o que permite verificar se as confusões do modelo são agronomicamente plausíveis, a exemplo da sobreposição visual esperada entre as deficiências de P e S em estágios iniciais, com P predominante em folhas mais velhas e S em folhas mais novas (MALAVOLTA, 2006), ou entre a classe Completo e os estágios incipientes de outras deficiências.

4.8 Interpretabilidade

A interpretabilidade dos modelos, cujos fundamentos teóricos são apresentados na Seção 2.7, foi operacionalizada neste trabalho por meio de duas técnicas complementares, cada uma adequada a uma das arquiteturas avaliadas. Para a ResNet50, o Grad-CAM (SELVARAJU et al., 2017) foi aplicado a camadas convolucionais profundas do modelo, especificamente a `layer3` e a `layer4`. Embora a `layer4`, última camada convolucional, encapsule as representações semânticas de mais alto nível, a `layer3` preserva maior resolução espacial; ambas são comparadas no Capítulo 5, onde se avalia qual delas produz mapas mais coerentes com a escala dos sintomas foliares.

Para o modelo híbrido CNN-ViT, a interpretabilidade foi obtida a partir dos pesos de atenção do *token* CLS na última camada do codificador, implementada como a `CustomTransformerEncoderLayer` justamente para expor esses pesos sem alterar o fluxo computacional. Os 25 valores de atenção, correspondentes à grade 5×5 produzida pelo

`AdaptiveAvgPool2d`, são reorganizados em um mapa bidimensional sobreposto à imagem original. Cabe registrar que essa grade é consideravelmente mais grosseira do que a resolução dos mapas Grad-CAM, limitação que restringe a precisão espacial dos mapas de atenção do híbrido e é retomada na análise do Capítulo 5.

Experimentos e Resultados

Neste capítulo são apresentados os resultados experimentais obtidos a partir da avaliação sistemática das arquiteturas de *deep learning* propostas para a classificação de deficiências nutricionais em folhas de *Coffea arabica*. A investigação compreende um total de 98 experimentos organizados em dois blocos complementares, sendo um benchmark principal com 50 experimentos, no qual as arquiteturas ResNet50, ViT-B/16 e Híbrido CNN-ViT foram avaliadas em diferentes resoluções e dez frações temporais distintas; e um estudo de ablação com 48 experimentos, que investiga o impacto da granularidade dos *patches* no desempenho do modelo Híbrido ao variar a grade de atenção entre 5×5 e 14×14 *patches*. A análise está organizada de modo a responder às três hipóteses formuladas no Capítulo 1, sendo elas a dependência temporal do desempenho de classificação (H1), a comparação entre arquiteturas convolucionais e baseadas em atenção (H2), e a capacidade de interpretabilidade dos modelos treinados (H3). A escala experimental adotada, com 98 configurações avaliadas de forma controlada, permite uma investigação abrangente dos fatores que influenciam o diagnóstico automatizado de estresses nutricionais, o que oferece evidências empíricas em profundidade para cada uma das hipóteses investigadas.

A estrutura do capítulo segue uma lógica progressiva e cumulativa. Inicialmente, descreve-se a configuração experimental adotada, o ambiente de execução, as métricas de avaliação e o espaço de configurações explorado. Em seguida, são apresentados os resultados gerais consolidados em tabelas comparativas que abrangem todas as configurações do *benchmark*, organizadas por arquitetura. Posteriormente, cada hipótese é analisada com base nas evidências experimentais coletadas, onde a análise temporal (H1) examina a curva de desempenho ao longo de dez frações; na comparação arquitetural (H2) são quantificadas as diferenças entre as três arquiteturas avaliadas; e na análise de interpretabilidade (H3) são avaliados qualitativamente os mapas Grad-CAM e de atenção. Na seção de análise por classe é detalhado, pelo uso de matriz de confusão e métricas individuais, o desempenho do melhor modelo para cada uma das sete deficiências. O estudo de ablação é utilizado para investigar o impacto da resolução do *grid* de *patches* no modelo Híbrido. Por fim, uma discussão integrada sintetiza os achados, contextualiza-os em rela-

ção à literatura existente e explicita as limitações do trabalho. Todas as figuras e tabelas apresentadas neste capítulo foram geradas a partir dos dados experimentais reais, e os valores reportados refletem o desempenho no conjunto de validação, conforme o protocolo de avaliação descrito no Capítulo 4.

A análise apresentada é tanto quantitativa, baseada em métricas numéricas de classificação, quanto qualitativa, fundamentada na interpretação visual dos mapas de ativação, matrizes de confusão e mapas de atenção dos *Transformers*. A combinação dessas duas perspectivas permite uma avaliação mais completa e robusta da capacidade dos modelos propostos para o problema de classificação de deficiências nutricionais em folhas de café, e conecta os resultados numéricos ao conhecimento agrônomo de domínio e às implicações práticas para o diagnóstico automatizado em campo.

5.1 Configuração Experimental

Os experimentos foram conduzidos em um ambiente local equipado com uma GPU NVIDIA RTX 5060 Ti, com o uso do sistema de rastreamento *MLflow*¹ para o registro sistemático de todos os hiperparâmetros, métricas e artefatos de cada execução. Para a implementação e treinamento dos modelos foi utilizado o *framework* *PyTorch* em conjunto com a biblioteca *torchvision* para o carregamento de pesos pré-treinados e aplicação de transformações nas imagens. A biblioteca *scikit-learn* foi utilizada para a divisão estratificada dos dados e para o cálculo das métricas de avaliação, enquanto a biblioteca *Pandas* foi empregada para auxiliar na manipulação dos metadados temporais do *dataset*. A semente aleatória, denominada *seed*, foi fixada em 42 para todas as execuções, a fim de garantir a reprodutibilidade dos resultados e a comparabilidade entre diferentes configurações experimentais. O uso do *MLflow* permitiu não apenas o registro automatizado de cada experimento, mas também a comparação visual entre configurações e a recuperação dos melhores modelos para análises subsequentes de interpretabilidade.

Todos os hiperparâmetros foram mantidos constantes em todos os 98 experimentos para garantir uma comparação justa entre as configurações, conforme detalhado no Capítulo 4. Especificamente, utilizou-se o otimizador Adam com taxa de aprendizado de 1×10^{-4} , a função de perda *CrossEntropyLoss* e o *scheduler* *ReduceLROnPlateau* com fator de redução de 0,5 e paciência de 3 épocas. O mecanismo de parada antecipada foi configurado com paciência de 5 épocas para monitoramento da perda de validação, o que previne o sobreajuste ao interromper o treinamento quando o modelo deixa de apresentar melhorias. O tamanho do *batch* foi definido em 16 amostras, e o número máximo de épocas foi limitado a 20. A divisão dos dados seguiu a estratégia *hold-out* estratificada na proporção de 80% para treinamento e 20% para teste, o que preserva a distribuição proporcional de cada classe em ambos os subconjuntos. Essa abordagem foi adotada após

¹ <<https://mlflow.org>>

uma tentativa inicial com LOGO por vaso, que apresentou alta variância devido às diferenças temporais entre plantas de um mesmo vaso, conforme documentado no Capítulo 4. A manutenção rigorosa desses hiperparâmetros ao longo de todos os experimentos é uma decisão metodológica intencional que privilegia a comparabilidade direta entre configurações em detrimento da otimização individual de cada modelo.

As métricas de avaliação adotadas incluem a acurácia global, que mede a proporção de predições corretas sobre o total de amostras, e métricas por classe, sendo elas *Precision*, *Recall* e *F1-Score*, que capturam a capacidade do modelo de distinguir cada deficiência individualmente. A *Precision* quantifica a proporção de predições positivas que são verdadeiramente positivas, o *Recall* mede a fração de exemplos positivos que foram corretamente identificados, e o F1-Score constitui a média harmônica entre ambas, o que penaliza desequilíbrios entre as duas métricas. O Macro F1-Score, calculado como a média aritmética dos F1-Scores de todas as classes, é particularmente relevante neste contexto, pois oferece uma visão equilibrada do desempenho que não é enviesada por eventuais desbalanceamentos na distribuição das amostras (GRANDINI; BAGLI; VISANI, 2020). Para a análise de interpretabilidade, empregou-se a técnica Grad-CAM (*Gradient-weighted Class Activation Mapping*) (SELVARAJU et al., 2017), aplicada a diferentes camadas convolucionais do ResNet50, especificamente `layer3` e `layer4`, para gerar mapas de ativação que evidenciam as regiões da imagem mais relevantes para cada predição. A comparação entre camadas permite avaliar o efeito do campo receptivo na qualidade dos mapas de interpretabilidade.

O espaço experimental foi definido pela combinação de três fatores no *benchmark* principal, sendo eles a arquitetura do modelo, entre ResNet50, ViT-B/16 e Híbrido CNN-ViT; a resolução das imagens de entrada; e a fração temporal dos dados utilizada para treinamento e avaliação, controlada via parâmetro. Esse parâmetro determina a porcentagem inicial dos dados que é descartada, de modo que `frac50` utiliza os últimos 50% do período temporal e `frac95` utiliza apenas os últimos 5%. As dez frações avaliadas (50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90% e 95%) oferecem uma granularidade temporal mais fina do que aquela identificada nos trabalhos revisados, o que permite a identificação de limiares a partir dos quais os sintomas se tornam suficientemente discriminativos para a classificação automatizada. As arquiteturas ResNet50 e Híbrido CNN-ViT foram avaliadas em duas resoluções (314×314 e 512×512 pixels), num total de 40 configurações. O ViT-B/16, por ser projetado para *patches* fixos de 16×16 pixels, foi avaliado exclusivamente na resolução 512×512 , o que resulta em 10 configurações adicionais. O total do *benchmark* é de 50 configurações. Adicionalmente, o estudo de ablação compreende 4 configurações de *patches* (5×5 , 7×7 , 10×10 e 14×14) $\times$ 2 resoluções $\times$ 6 frações temporais selecionadas, num total de 48 experimentos adicionais. Ao todo, portanto, foram avaliadas 98 configurações experimentais distintas, cujos resultados são apresentados nas seções subsequentes deste capítulo (5.2 à 5.8) .

5.2 Resultados Gerais

As informações apresentadas nas Tabelas 3 e 4 consolidam os resultados de todas as 40 configurações do *benchmark*, organizadas por arquitetura. Essa apresentação separada permite identificar imediatamente as tendências internas de cada modelo antes da comparação cruzada, e o uso de dez frações temporais oferece uma resolução fina, mais granular do que aquela identificada nos trabalhos revisados, para a análise da relação entre estágio de manifestação dos sintomas e capacidade discriminativa dos modelos. O melhor resultado absoluto foi alcançado pela ResNet50 com resolução 512×512 e fração temporal de 60%, que atingiu uma acurácia de 77,97% e um Macro F1-Score de 0,774.

A acurácia de 77,97% obtida pelo melhor modelo deve ser interpretada como estimativa de teto sob protocolo *hold-out* estratificado por classe, comparável a estudos da literatura que adotam a mesma estratégia (ESGARIO; KROHLING; VENTURA, 2020; RANDIVE; SAPKAL; BORNARE, 2025). Como cada vaso contém aproximadamente mil imagens distribuídas ao longo do período experimental, o *split* aleatório 80/20 não isola plantas individuais entre treino e validação, o que implica que esse valor mede a capacidade discriminativa máxima dos modelos sob a estrutura de dados disponível, e não a generalização para indivíduos não observados. Cabe distinguir essa situação de uma duplicação direta de amostras, uma vez que nenhuma imagem aparece simultaneamente em treino e validação, pois a função `train_test_split` opera no nível de imagem individual, o que se repete entre os subconjuntos é a identidade da planta fotografada em *timestamps* distintos, e não a mesma imagem em ambos os lados.

Tabela 3 – Acurácia (%) da arquitetura ResNet50 para todas as combinações de resolução e fração temporal.

Resolução (pixels)	frac50	frac55	frac60	frac65	frac70	frac75	frac80	frac85	frac90	frac95
314×314	74,73	76,22	76,34	76,86	75,10	73,52	73,50	69,78	70,47	59,84
512×512	75,92	75,95	77,97	77,25	76,41	73,84	73,10	72,93	71,46	64,57

Tabela 4 – Acurácia (%) da arquitetura Híbrido CNN-ViT para todas as combinações de resolução e fração temporal.

Resolução (pixels)	frac50	frac55	frac60	frac65	frac70	frac75	frac80	frac85	frac90	frac95
314×314	67,28	66,01	70,82	70,83	71,68	69,82	67,09	65,05	64,96	50,39
512×512	69,69	67,76	71,27	67,23	65,57	67,77	66,40	65,05	61,61	53,54

Tabela 5 – Acurácia (%) da arquitetura ViT-B/16 na resolução 512×512 pixels para todas as frações temporais avaliadas.

Resolução (pixels)	frac50	frac55	frac60	frac65	frac70	frac75	frac80	frac85	frac90	frac95
512×512	47,65	67,54	66,54	67,74	65,37	63,12	61,48	59,92	57,09	46,46

A análise dos dados apresentados nas Tabelas 3, 4 e 5 revela três tendências fundamentais, detalhadas nas Seções 5.3, 5.4 e 5.5. Primeiramente, observa-se uma curva em

formato de sino na relação entre fração temporal e acurácia, onde o desempenho não melhora monotonicamente com a inclusão de mais dados temporais, mas atinge um pico nas frações intermediárias, `frac60` a `frac65`, que declinam quando dados de períodos mais antigos, nos quais os sintomas ainda não se manifestaram, são incluídos, e também quando apenas os dados mais recentes são utilizados. Esse padrão é consistente em todas as combinações de arquitetura e resolução avaliadas, o que confere robustez à observação e constitui uma das contribuições deste trabalho. Em segundo lugar, a superioridade da ResNet50 sobre as demais arquiteturas é consistente em todas as comparações, o que indica uma vantagem estrutural dessa rede para o problema em questão. Em terceiro lugar, a faixa de variação da acurácia é substancial entre o melhor resultado global (77,97%, ResNet50 512 × 512 `frac60`) e o pior resultado global (46,46%, ViT-B/16 `frac95`), há uma diferença de 31,51 pontos percentuais (pp), o que reflete a interação complexa entre arquitetura, resolução e fração temporal.

A amplitude dessa variação reforça a relevância de cada fator para o desempenho final do sistema e evidencia que a tarefa de classificação de deficiências nutricionais a partir de imagens RGB permanece desafiadora, o que é consistente com a natureza sutil e frequentemente sobreposta dos sintomas visuais dessas deficiências (MALAVOLTA, 2006). Conforme discutido no Capítulo 2, muitas deficiências compartilham manifestações fenotípicas similares, como a clorose, que pode ser indicativa de deficiências de N, Mg, S ou Fe, o que torna a classificação multiclasse um problema inerentemente complexo, mesmo para observadores humanos experientes. O Macro F1-Score de 0,774 do melhor modelo, próximo à acurácia de 77,97%, sugere que o desempenho é relativamente equilibrado entre as classes, sem uma dominância excessiva de classes majoritárias na composição da acurácia global. Esse equilíbrio é desejável em aplicações de diagnóstico, nas quais a capacidade de identificar corretamente cada tipo de deficiência é tão importante quanto o desempenho agregado, uma vez que a ação corretiva varia substancialmente conforme o nutriente deficiente (MARSCHNER, 2012).

5.3 Análise do Impacto Temporal (H1)

A primeira hipótese desta pesquisa (H1) postula que o desempenho de classificação de deficiências nutricionais é temporalmente dependente, o que melhora de forma significativa à medida que os sintomas visuais progridem ao longo do tempo. Para avaliar essa hipótese com a granularidade necessária, os modelos foram treinados e testados com dez frações temporais progressivamente mais restritivas, de `frac50` (últimos 50% do período experimental) até `frac95` (últimos 5%, janela mais restrita com menor volume de dados). Na Figura 16 são apresentadas as curvas de acurácia em função da fração temporal para todas as quatro combinações de arquitetura e resolução, o que permite uma visualização direta da tendência de desempenho e, crucialmente, a identificação do formato de sino

que caracteriza a relação temporal.

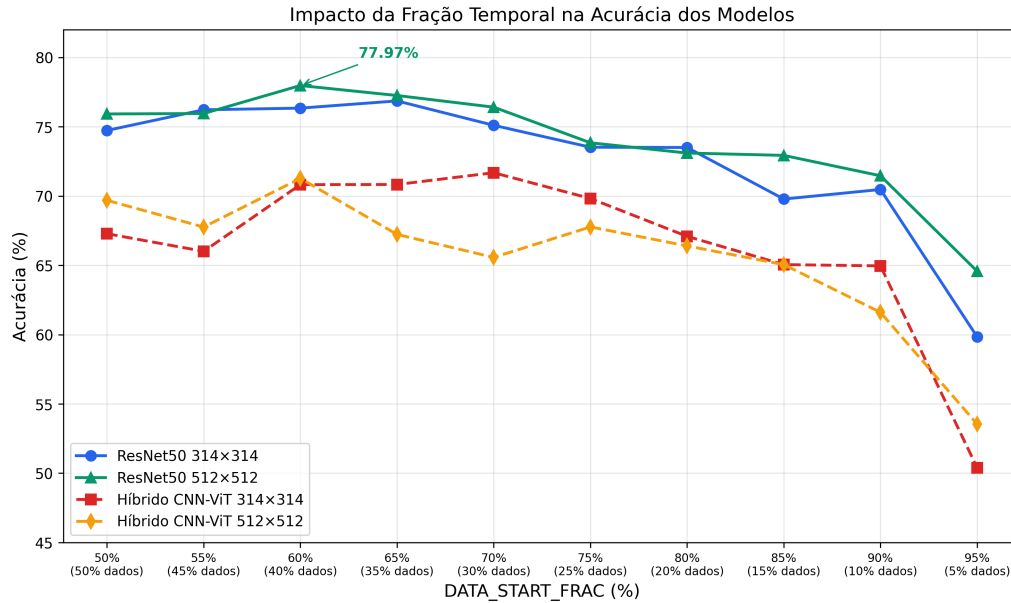


Figura 16 – Acurácia dos modelos ResNet50 e Híbrido CNN-ViT em função da fração temporal dos dados, para ambas as resoluções. A curva em formato de sino é visível em todas as configurações, com pico entre frac60 e frac65. Frações mais altas, à direita, correspondem a janelas temporais mais estreitas com menor volume de dados, e frações mais baixas, à esquerda, correspondem a janelas mais amplas que incluem dados de períodos assintomáticos.

Os resultados sustentam a hipótese H1 e refinam sua formulação inicial, uma vez que, em vez da relação monotônica originalmente antecipada, observa-se uma curva em formato de sino que evidencia a interação entre dois efeitos opostos decorrentes do alargamento da janela temporal, sendo eles o ganho de volume de dados ainda majoritariamente sintomáticos, predominante nas janelas intermediárias, e a diluição por dados assintomáticos, predominante nas janelas mais amplas. Para a arquitetura ResNet50 na resolução 512×512 , a acurácia sobe de 64,57% na fração 95%, correspondente à janela mais restrita com menor volume de dados, até o pico de 77,97% na fração 60%, o que representa um ganho de 13,40 pp. No entanto, ao incluir dados ainda mais antigos, nas frações 55% e 50%, a acurácia recua levemente para 75,95% e 75,92%, respectivamente. Esse comportamento configura uma curva de sino assimétrica, na qual o ramo ascendente, de frac95 a frac60, é mais íngreme que o ramo descendente, de frac60 a frac50. Analogamente, o ResNet50 na resolução 314×314 apresenta pico em frac65 (76,86%) com declínio para 74,73% em frac50, e o modelo Híbrido CNN-ViT na resolução 314×314 atinge seu máximo em frac70 (71,68%) antes de recuar para 67,28% em frac50. Esses valores constituem evidência empírica forte de que existe uma janela temporal de melhor desempenho médio para a classificação e que a inclusão indiscriminada de todos os dados disponíveis não é a estratégia ideal.

A existência da curva de sino pode ser explicada pela interação de dois mecanismos opostos associados ao alargamento da janela temporal. Cabe observar, preliminarmente, que a adoção de frações menores expande a janela retida sempre em direção aos *timestamps* mais antigos, de modo que as imagens com sintomas mais desenvolvidos, correspondentes aos registros mais recentes, estão presentes em todas as configurações, ao passo que as janelas mais amplas apenas acrescentam dados de períodos progressivamente anteriores. O primeiro mecanismo, dominante no ramo ascendente da curva, é o ganho de volume de dados de treinamento, uma vez que a janela mais restrita, frac95, dispõe de um número de amostras insuficiente para o ajuste adequado das redes, e a expansão até frac60 eleva substancialmente esse volume com dados ainda majoritariamente provenientes de períodos em que os sintomas já se encontram desenvolvidos. O segundo mecanismo, responsável pelo ramo descendente, é a diluição por dados assintomáticos, que se manifesta quando frações muito abrangentes, como frac55 e frac50, passam a incorporar imagens dos períodos iniciais do experimento, nos quais as folhas deficientes eram visualmente semelhantes às saudáveis, o que adiciona ruído ao conjunto de treinamento e dificulta a aprendizagem de fronteiras de decisão nítidas entre as classes. Deficiências como a de K, que se manifesta por necrose marginal, ou a de N, caracterizada por clorose generalizada, somente produzem padrões visuais distintivos com o avanço do estresse (MARSCHNER, 2012), razão pela qual os registros mais antigos contribuem pouco para a discriminação. O argumento mais robusto em favor da dependência temporal reside na própria localização do ótimo no interior da grade de frações, e não no extremo de maior volume, frac50, visto que, sob um regime governado exclusivamente pelo volume de dados, a acurácia cresceria de forma monotônica até frac50, comportamento que não se observa em nenhuma das arquiteturas ou resoluções avaliadas. Cabe ressaltar, contudo, que o ramo ascendente permanece confundido com o efeito de volume, dado que janelas mais amplas contêm simultaneamente mais amostras e dados de períodos mais desenvolvidos, de modo que a separação rigorosa desses fatores exigiria um experimento confirmatório por subamostragem das janelas amplas ao volume das janelas estreitas, registrado como trabalho futuro no Capítulo 6. Sob essa ressalva, o ponto numericamente ótimo localiza-se entre frac60 e frac65 para a maioria das configurações.

A análise detalhada do ramo ascendente da curva revela que os maiores ganhos de acurácia coincidem com os maiores incrementos relativos de volume de dados. Para o ResNet50 512×512 , o salto mais acentuado ocorre entre frac95 (64,57%) e frac90 (71,46%), com um ganho de 6,89 pp ao se dobrar a fração de dados retida, dos últimos 5% para os últimos 10% do período. Os ganhos subsequentes são mais graduais, sendo 1,47 pp de frac90 para frac85, 0,17 pp de frac85 para frac80, e 0,74 pp de frac80 para frac75, à medida que cada nova fração representa um acréscimo relativo de amostras progressivamente menor. Esse padrão é consistente com a hipótese de que a janela mais restrita opera em um regime de escassez de dados, no qual a adição de novas amostras de períodos ainda

sintomáticos produz ganhos expressivos, que se atenuam conforme o volume se torna suficiente. A partir da fração 75% observa-se uma segunda fase de ganhos até o pico em frac60–frac65, possivelmente associada à incorporação de exemplos que reforçam a distinção entre classes visualmente próximas, como P e S, antes que a inclusão de dados dos períodos iniciais passe a predominar e reverta a tendência.

A progressão temporal observada no modelo Híbrido CNN-ViT exhibe um comportamento qualitativo semelhante, porém com maior variabilidade entre frações consecutivas. Na resolução 314×314 , a acurácia apresenta oscilações não monotônicas nas frações intermediárias (por exemplo, 66,01% em frac55 seguido de 70,82% em frac60 e 70,83% em frac65), o que sugere que o modelo Híbrido é mais sensível às variações na composição dos dados de treinamento do que o ResNet50. Essa maior variabilidade pode ser atribuída à complexidade adicional do mecanismo de atenção do *Transformer*, que, por não possuir o viés indutivo de localidade das convoluções, necessita de distribuições de treinamento mais estáveis para convergir a soluções consistentes (TOUVRON et al., 2021). Na resolução 512×512 , o Híbrido apresenta um padrão ainda mais irregular, com pico em frac60 (71,27%) mas oscilações significativas nas frações adjacentes, o que reforça a hipótese de que essa arquitetura é mais sensível à composição temporal dos dados.

Do ponto de vista prático, esses resultados têm implicações importantes para o diagnóstico automatizado em campo. A acurácia de aproximadamente 73% obtida pelo ResNet50 nas frações intermediárias, frac75 a frac80, indica que, mesmo em janelas temporais mais estreitas, com menor volume de dados, os modelos já são capazes de fornecer classificações com uma taxa de acerto consideravelmente superior ao acaso, que é de 14,3% para sete classes. Essa constatação é relevante porque sugere que um sistema de monitoramento contínuo poderia detectar tendências de deficiência nutricional antes que os sintomas se tornem visualmente óbvios para um observador humano, o que potencialmente permite intervenções corretivas mais precoces (SHOAIB et al., 2026). Não obstante, a queda acentuada de desempenho nas janelas mais restritas (64,57% em frac95 para o ResNet50 512×512 e apenas 50,39% para o Híbrido 314×314) evidencia que existe um volume mínimo de dados abaixo do qual a classificação se torna pouco confiável, o que reforça a necessidade de estratégias de coleta que garantam amostras suficientes, bem como de abordagens complementares, como sensoramento multiespectral ou análises laboratoriais (SOUSSI et al., 2024). A identificação da janela temporal de melhor desempenho médio, situada entre frac60 e frac65 sob as condições controladas deste experimento, fornece uma indicação preliminar, condicionada a futura validação em campo, sobre faixa temporal candidata a maximizar a relação esforço-acurácia. A tradução dessa observação em diretriz operacional para o produtor depende de validação externa, conforme detalhado no Capítulo 6.

Cabe ainda ressaltar que o refinamento da hipótese H1, com a observação adicional do formato de sino, tem implicações metodológicas potencialmente aplicáveis a outros

problemas de classificação de estresses vegetais. A dependência temporal do desempenho sugere que qualquer estudo de classificação de estresses vegetais baseado em imagens deveria reportar não apenas a acurácia máxima obtida, mas também a curva temporal completa, o que fornece assim uma avaliação mais honesta e completa da capacidade diagnóstica real do sistema. A omissão dessa dimensão em grande parte da literatura, a exemplo de Ferentinos (2018), pode levar a expectativas inflacionadas sobre o desempenho de modelos em condições reais de campo, onde os sintomas podem estar em estágios variados de desenvolvimento.

5.4 Análise Comparativa das Arquiteturas (H2)

A segunda hipótese (H2) postulava que arquiteturas com mecanismos de atenção global superariam as CNNs puras na tarefa de classificação de deficiências nutricionais, em virtude da capacidade dos *Transformers* de modelar dependências de longo alcance na imagem. Os resultados experimentais não sustentam essa hipótese nas condições controladas adotadas neste trabalho, em que a arquitetura ResNet50 apresentou desempenho superior ao modelo Híbrido CNN-ViT em todas as 20 comparações pareadas (10 frações $\times$ 2 resoluções), e também superou o ViT-B/16 em todas as 10 frações avaliadas na resolução 512×512 . Ao se observar a Figura 17 é possível visualizar essa hierarquia consistente entre as três arquiteturas.

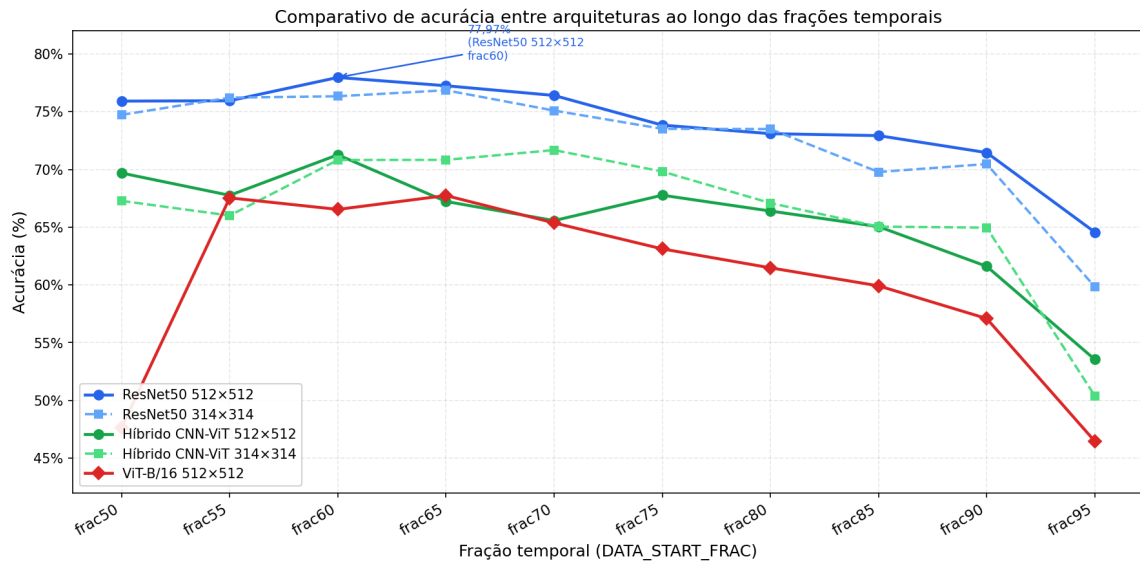


Figura 17 – Comparação da acurácia entre as três arquiteturas avaliadas (ResNet50, Híbrido CNN-ViT e ViT-B/16) ao longo das frações temporais. A hierarquia ResNet50 > Híbrido > ViT-B/16 é consistente em todo o eixo temporal. O ViT-B/16 apresenta comportamento de colapso nas frações extremas (frac50 e frac95), em contraste com a degradação gradual das demais arquiteturas.

Na configuração de melhor desempenho para cada arquitetura, o ResNet50 alcançou

77,97% de acurácia (resolução 512×512 , frac60), o Híbrido CNN-ViT atingiu 71,68% (resolução 314×314 , frac70) e o ViT-B/16 obteve 67,74% (resolução 512×512 , frac65), o que estabelece uma hierarquia clara em que o ResNet50 supera o Híbrido CNN-ViT, que por sua vez supera o ViT-B/16. As diferenças numéricas observadas, de 6,29 pp entre ResNet50 e Híbrido e de 10,23 pp entre ResNet50 e ViT-B/16 nos respectivos picos, excedem a faixa típica de variabilidade entre sementes reportada em outros estudos do mesmo porte, ainda que a confirmação formal dessa significância demande repetição com múltiplas sementes, o que não foi conduzido neste trabalho. Quando comparados diretamente na mesma resolução de 512×512 pixels, o ResNet50 obteve valor numericamente superior ao ViT-B/16 em todas as 10 frações temporais, com diferença média de aproximadamente 13 pp.

Um comportamento particularmente revelador no desempenho do ViT-B/16 é a queda acentuada de desempenho na fração **frac50** (47,65%), que situa os resultados obtidos com esse modelo próximo ao nível de acurácia da fração mais restritiva (frac95: 46,46%). Esse comportamento diverge do padrão de sino suave observado nas demais arquiteturas, nas quais a inclusão de frac50 resulta em apenas uma queda marginal em relação ao pico. O comportamento observado é compatível com pelo menos duas interpretações: a primeira é a hipersensibilidade do ViT-B/16 à inclusão de dados assintomáticos, dada a ausência do viés indutivo de localidade que permitiria às convoluções isolar padrões discriminativos mesmo com proporção elevada de amostras sem sintomas visíveis; a segunda é a falha de convergência em uma única realização de treinamento, possibilidade não descartável dada a ausência de repetição com múltiplas sementes neste trabalho. Ambas as interpretações são consistentes com os dados disponíveis e são reportadas como hipóteses concorrentes, cuja distinção exigiria experimentos adicionais com sementes variadas.

5.4.1 Limitações da comparação arquitetural

Antes de discutir os fatores técnicos que explicam a hierarquia observada, é necessário registrar que a comparação entre arquiteturas foi conduzida sob um protocolo padronizado que, embora favoreça a comparabilidade, não foi otimizado individualmente para cada modelo. Quatro fatores assimétricos merecem destaque. Em primeiro lugar, os pesos pré-treinados utilizados não são equivalentes entre arquiteturas, sendo a ResNet50 inicializada com a variante **IMAGENET1K_V2**, treinada com técnicas modernas de aumento de dados e regularização, enquanto o ViT-B/16 utiliza a variante **IMAGENET1K_V1**, sem essas técnicas. Cabe esclarecer que essa assimetria não é uma escolha discricionária, mas decorre da disponibilidade restrita de pesos na biblioteca **torchvision**, que publica a designação **IMAGENET1K_V2** exclusivamente para a família ResNet, sem variante equivalente para o ViT-B/16, conforme detalhado no Capítulo 2. Em segundo lugar, a taxa de aprendizado foi mantida fixa em 1×10^{-4} para todas as arquiteturas, valor adequado para o *fine-tuning* de CNNs, mas não necessariamente ótimo para *Transformers*, que tipicamente

requerem taxas distintas. Em terceiro lugar, a política de aumento de dados restringe-se a *RandomHorizontalFlip* e *ColorJitter*, sem técnicas mais agressivas como *RandAugment* ou *Mixup* apontadas por Touvron et al. (2021) como críticas para o desempenho de ViTs em regimes de dados moderados. Por fim, a operação do ViT-B/16 em resolução 512×512 exige interpolação bicúbica dos *embeddings* posicionais a partir da grade 14×14 de *patches* do pré-treinamento original em 224×224 pixels, regime que pode degradar a qualidade da informação posicional.

Sob configurações otimizadas individualmente para cada arquitetura, a hierarquia observada poderia ser atenuada. A não sustentação de H2 deve, portanto, ser interpretada como contextual ao protocolo padronizado deste trabalho, e não como afirmação geral sobre a inadequação de *Transformers* ao domínio de classificação de deficiências nutricionais.

5.4.2 Fatores técnicos que sustentam a hierarquia observada

A hierarquia observada nesse trabalho pode ser atribuída a uma conjunção de fatores técnicos complementares. O primeiro e mais fundamental é o viés indutivo de localidade inerente às operações de convolução, que assume que pixels vizinhos tendem a ser semanticamente relacionados (LECUN et al., 1998). Esse viés é particularmente vantajoso para a análise de texturas foliares, onde os sintomas de deficiência se manifestam como alterações locais de cor, textura e padrão nas nervuras, margens e superfície das folhas. A necrose marginal do K, a clorose internerval do Mg e as pontuações necróticas do Ca são exemplos de padrões espacialmente localizados que são naturalmente capturados por filtros convolucionais que operam em vizinhanças espaciais restritas (MALAVOLTA, 2006). Conforme demonstrado por Cordonnier, Loukas e Jaggi (2020), a atenção global dos *Transformers* pode, em princípio, expressar qualquer operação convolucional e, na prática, aprende padrões locais similares aos das CNNs; contudo, essa convergência exige quantidades substanciais de dados, ao passo que as convoluções alcançam o mesmo efeito por construção.

O segundo fator é o tamanho relativamente modesto do *dataset*. Embora as 25.362 imagens classificadas, distribuídas em 7 classes, constituam um volume significativo para padrões de agricultura computacional, essa quantidade está muito aquém dos milhões de exemplos tipicamente necessários para que *Vision Transformers* superem CNNs (DO-SOVITSKIY et al., 2021; TOUVRON et al., 2021). A literatura demonstra de forma consistente que a eficiência amostral dos *Transformers* é inferior à das CNNs em regimes de dados limitados, o que é um cenário comum em aplicações agrícolas (SAKKA et al., 2025). O fato de que o ViT-B/16, que depende ainda mais de grandes volumes de dados do que o Híbrido por não possuir o *backbone* convolucional pré-treinado como extrator de *features* locais, apresentou o pior desempenho entre as três arquiteturas é consistente com essa explicação. O terceiro fator é a eficiência do *transfer learning* a partir dos pesos

pré-treinados no ImageNet, que é notadamente superior para a ResNet50, cuja arquitetura foi extensivamente validada nesse domínio ao longo de quase uma década de uso pela comunidade (HE et al., 2016). O quarto fator, exclusivo do Híbrido, é a resolução limitada dos *patches*, investigada em profundidade na Seção 5.8.

O custo computacional constitui uma dimensão adicional relevante para a análise comparativa. Na Tabela 6 são apresentados os tempos de treinamento registrados para cada configuração na resolução 512×512 , o detalhamento por fração temporal pode ser observado na Figura 24 do Apêndice B. O ViT-B/16 demandou tempo de treinamento substancialmente superior às demais arquiteturas em todas as frações. Na configuração frac50, correspondente ao maior volume de dados, o ViT-B/16 levou 297,1 minutos, contra 69,4 minutos para o ResNet50 e 56,9 minutos para o Híbrido, o que resulta em um fator de aproximadamente 4,3 vezes em relação ao ResNet50. Na fração ótima de desempenho, correspondente ao frac60, os tempos foram 157,4 minutos no ViT, 55,0 minutos na ResNet50 e 48,1 minutos no Híbrido, o que mantém a proporção de aproximadamente 2,9 vezes entre ViT e ResNet50. Esse custo adicional está associado à complexidade quadrática do mecanismo de autoatenção em relação ao número de *tokens* processados (VASWANI et al., 2017), uma vez que, para imagens 512×512 divididas em *patches* de 16×16 pixels, o ViT-B/16 opera com 1.024 *tokens* por imagem, o que gera matrizes de atenção de dimensão 1.024×1.024 em cada cabeça de atenção. A combinação de pior desempenho de classificação (67,74% contra 77,97% do ResNet50) e custo computacional 3 a 4 vezes superior torna o ViT-B/16 puro a opção menos favorável entre as três arquiteturas avaliadas para este domínio.

Tabela 6 – Resumo do custo computacional por arquitetura na resolução 512×512 : tempo médio por experimento, tempo total do conjunto de 10 frações e razão em relação ao ResNet50. Tempos registrados pelo MLflow.

Arquitetura	Tempo médio (min)	Tempo total (h)	Razão vs. ResNet50
ResNet50	35,6	5,9	1,0×
Híbrido CNN-ViT	26,7	4,4	0,75×
ViT-B/16	108,6	18,1	3,1×

Uma análise mais granular dos dados de tempo revela uma distinção importante entre custo por época e custo total. O tempo médio por época é praticamente idêntico entre ResNet50 e Híbrido (1,89 min/época e 1,94 min/época, respectivamente), o que indica que as duas arquiteturas possuem custo computacional por passagem equivalente para as configurações avaliadas. A diferença no tempo total em favor do Híbrido decorre exclusivamente do menor número médio de épocas treinadas, sendo 12,9 épocas contra 18,8 do ResNet50. Esse padrão reflete o acionamento mais precoce do mecanismo de parada antecipada no modelo Híbrido, que atinge o platô de melhoria na perda de validação em menos iterações. A interpretação mais consistente com os resultados de acurácia é que o Híbrido converge mais rapidamente para seu teto de capacidade neste domínio,

e não que possui maior eficiência de aprendizado por época. Em outras palavras, o menor tempo total não é um indicativo de eficiência arquitetural, mas sim um reflexo da limitação do modelo, que esgota sua capacidade de generalização antes de atingir o limite máximo de épocas. O ViT-B/16, por sua vez, apresenta um custo por época de 8,23 minutos em média, 4,3 vezes superior ao das demais arquiteturas, em decorrência da complexidade quadrática do mecanismo de autoatenção, e treina um número de épocas similar ao Híbrido (12,3 em média), o que resulta em um custo total substancialmente mais elevado, como pode ser observado na Tabela 6.

É relevante contextualizar esses resultados em relação à literatura de arquiteturas híbridas. Guo et al. (2022) demonstraram que arquiteturas CMT, sigla de *Convolutional meets Transformer*, podem alcançar desempenho competitivo em *benchmarks* como ImageNet com custos computacionais significativamente menores que ViTs puros, com o conjunto padrão ImageNet-1K (1,28 milhão de imagens). Da mesma forma, Thakur et al. (2022) obtiveram resultados promissores com a arquitetura PlantXViT para identificação de doenças em plantas, com *datasets* curados de imagens de alta qualidade e em maior quantidade. Essas comparações sugerem que a superioridade do ResNet50 observada neste trabalho não é necessariamente uma limitação intrínseca das arquiteturas baseadas em atenção em geral, mas uma consequência da escala de dados disponível. Li et al. (2025) demonstraram recentemente, em classificação de imagens de sensoriamento remoto, que abordagens híbridas podem ser competitivas quando a integração entre convoluções e atenção é cuidadosamente projetada, o que reforça a perspectiva de que o espaço de configurações arquiteturais híbridas permanece promissor para investigações futuras com *datasets* maiores.

A análise da interação entre arquitetura e fração temporal revela padrões distintos para cada modelo. Para o ResNet50 e o Híbrido, o *gap* de desempenho tende a ser mais pronunciado nas janelas mais amplas, especialmente de $\text{frac}50$ a $\text{frac}65$, nas quais o ResNet50 explora com maior eficiência o maior volume e a riqueza visual disponíveis. O ViT-B/16, por sua vez, apresenta comportamento mais irregular ao longo do eixo temporal, com oscilações mais acentuadas entre frações consecutivas, o que sugere maior sensibilidade à composição temporal dos dados de treinamento. Esse padrão reforça a interpretação de que o viés indutivo convolucional proporciona uma extração de características mais estável e robusta ao longo das diferentes condições temporais avaliadas.

5.5 Impacto da Resolução

A influência da resolução de entrada no desempenho da classificação foi avaliada por meio da comparação entre os resultados obtidos em ambas as arquiteturas nas resoluções 512×512 e 314×314 pixels, ao longo de todas as dez frações temporais. Na Figura 18 é ilustrada essa comparação para o ResNet50, o que evidencia o ganho associado à resolução

mais alta.

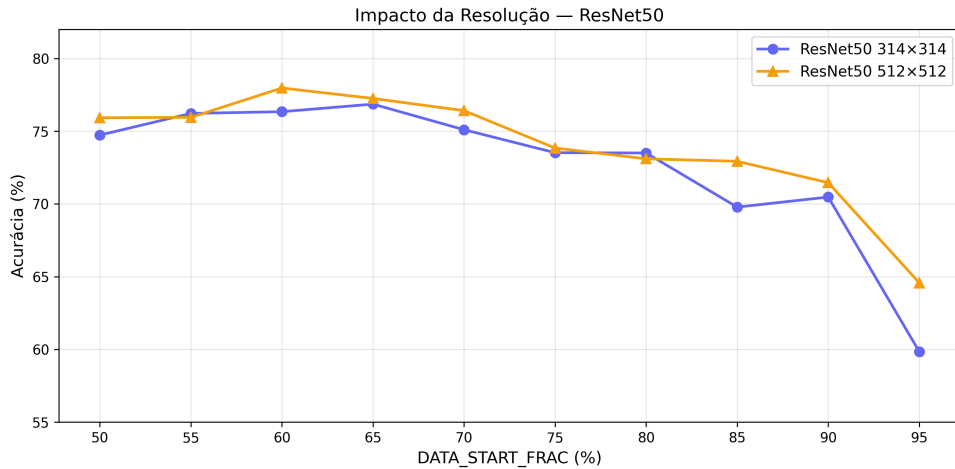


Figura 18 – Impacto da resolução de entrada na acurácia do modelo ResNet50 ao longo das frações temporais. A resolução 512×512 apresenta vantagem consistente, especialmente nas frações extremas (frac90–frac95).

Os resultados indicam que a resolução 512×512 pixels proporcionou uma diferença numericamente modesta e recorrente a favor do ResNet50 na maioria das frações temporais, dentro da faixa de variabilidade em que múltiplas sementes seriam necessárias para qualificar como estatisticamente significativa. Na configuração de melhor desempenho, a diferença é de 1,63 pp (77,97% em frac60 utilizando imagens de 512×512 pixels, enquanto imagens de 314×314 pixels alcançaram 76,34%). Embora essa diferença seja modesta em comparação com o impacto da fração temporal (13,40 pp) ou da escolha da arquitetura (cerca de 6 pp), ela é notavelmente mais acentuada nas frações com menor volume de dados, como em frac95, em que a resolução 512×512 pixels alcança 64,57% contra 59,84% para 314×314 pixels, uma diferença de 4,73 pp. Esse resultado sugere que os detalhes visuais adicionais preservados pela resolução mais alta são particularmente informativos quando o volume de dados de treinamento é reduzido, condição na qual a preservação de nuances cromáticas e texturais finas pode compensar parcialmente a escassez de amostras e fazer a diferença entre uma classificação correta e uma confusão entre classes.

A comparação com os resultados do modelo Híbrido CNN-ViT oferece um contraponto revelador em relação ao impacto da resolução. Diferentemente do ResNet50, o Híbrido apresenta uma relação inconsistente com a resolução, pois em algumas frações a resolução 512×512 pixels é superior (por exemplo, frac50: 69,69% contra 67,28%), mas em outras é inferior (frac70: 65,57% contra 71,68%). Esse comportamento divergente sugere que o aumento da resolução de entrada, embora benéfico para o ResNet50, pode ter efeito adverso no modelo Híbrido em determinadas configurações. Uma explicação plausível é que a redução para 25 *patches* via *pooling* adaptativo descarta proporcionalmente mais informação quando a entrada tem maior resolução, o que anula os benefícios potenciais da imagem com maior resolução espacial e introduz um gargalo de informação que se

agrava com o aumento dimensional. No caso do ResNet50, cujas operações convolucionais processam diretamente a grade de pixels em múltiplas escalas hierárquicas, a resolução superior se traduz em mapas de características mais ricos em cada camada, o que permite a detecção de padrões mais finos.

Contudo, o ganho marginal proporcionado pela resolução 512×512 pixels para o ResNet50 deve ser ponderado em relação ao custo computacional associado. A resolução superior implica um aumento de aproximadamente 2,66 vezes no número de pixels em relação à 314×314 pixels, o que se traduz em um maior consumo de memória da GPU e em um maior tempo de treinamento por época. Em cenários de implantação prática, nos quais os recursos computacionais podem ser limitados, como em dispositivos embarcados ou aplicações em tempo real no campo, a resolução 314×314 pode representar um compromisso mais adequado entre desempenho e eficiência, oferecendo acurácias competitivas (76,86% no pico) com custo computacional significativamente menor. Em contrapartida, em cenários nos quais a resolução mais alta é viável, como em estações de monitoramento fixas equipadas com câmeras de alta definição, o ganho adicional pode se traduzir em diagnósticos mais precisos, especialmente para a detecção precoce de sintomas. A escolha da resolução deve, portanto, ser guiada pelo contexto específico de implantação, ao considerar tanto as restrições de *hardware* e conectividade quanto o estágio temporal esperado dos sintomas a serem classificados.

5.6 Análise por Classe

Para compreender as nuances do desempenho do melhor modelo (ResNet50 com resolução 512×512 e fração temporal de 60%), esta seção apresenta uma análise detalhada das métricas por classe individual. Na Tabela 7 são apresentados, de forma consolidada, os valores de *Precision*, *Recall* e F1-Score para cada uma das sete classes de deficiência nutricional; a visualização gráfica correspondente encontra-se na Figura 25 do Apêndice B.

Tabela 7 – Métricas detalhadas por classe para o melhor modelo (ResNet50, 512×512 , frac60). As métricas macro representam a média aritmética sobre todas as classes.

Classe	<i>Precision</i>	<i>Recall</i>	F1-Score	Suporte
Ca	0,885	0,740	0,806	312
Completo	0,702	0,635	0,667	211
K	0,710	0,768	0,738	297
Mg	0,744	0,859	0,797	348
N	0,906	0,964	0,934	280
P	0,701	0,794	0,745	281
S	0,828	0,657	0,732	300
Macro	0,782	0,774	0,774	—

O resultado mais expressivo da análise por classe é o desempenho excepcional na

classificação da deficiência de N, que alcançou um *Recall* de 0,964, ou seja, 96,4% das amostras dessa classe foram corretamente identificadas, combinado com uma *Precision* de 0,906, o que resulta em um F1-Score de 0,934. Esse desempenho pode ser explicado pela natureza visualmente distinta dos sintomas de deficiência de N, que se manifestam como uma clorose generalizada e uniforme das folhas, frequentemente acompanhada de amarelecimento progressivo que se inicia nas folhas mais velhas (MARSCHNER, 2012). Diferentemente de outras deficiências, cujos sintomas podem ser localizados em regiões específicas da folha, a deficiência de N produz uma alteração cromática difusa e homogênea que é relativamente fácil de distinguir tanto para observadores humanos quanto para algoritmos de classificação visual. A alta *Precision*, de 90,6%, indica que poucas amostras de outras classes são erroneamente classificadas como N, o que confirma que os padrões visuais dessa deficiência são genuinamente únicos no contexto deste *dataset*.

Em contrapartida, a classe Completo, que representa o grupo de controle, apresentou o menor *Recall* entre todas as classes (0,635), o que indica que aproximadamente 36,5% das folhas saudáveis foram erroneamente classificadas como deficientes em algum nutriente. Essa dificuldade de classificação é compreensível do ponto de vista agrônomo, pois folhas saudáveis podem apresentar variações naturais de coloração e textura que se assemelham visualmente aos estágios iniciais de deficiência, e a fronteira visual entre saudável e levemente deficiente é inerentemente difusa (MALAVOLTA, 2006). A *Precision* de 0,702 para essa classe indica que, quando o modelo prediz uma folha como saudável, essa predição é correta em cerca de 70% dos casos, o que sugere que o modelo tende a ser conservador na atribuição do *status* saudável. Esse padrão é particularmente relevante para aplicações práticas, pois, em um contexto de triagem, a tendência a classificar folhas saudáveis como deficientes, ou seja, a gerar falsos positivos de deficiência, é preferível ao oposto, pois gera alertas que podem ser verificados, em vez de omitir deficiências que necessitariam de intervenção. O F1-Score de 0,667 para esta classe, o mais baixo entre todas, reflete justamente esse desequilíbrio entre a cautela do modelo e a similaridade visual entre folhas saudáveis e levemente deficientes.

A classe Ca destacou-se por apresentar o segundo maior valor da métrica *Precision* entre todas as deficiências avaliadas (0,885), o que significa que, quando o modelo identifica uma amostra como deficiente em Ca, essa predição é correta em 88,5% dos casos. Esse resultado está associado aos sintomas característicos dessa deficiência em cafeeiros, como deformação das folhas jovens, redução do crescimento apical e necrose dos meristemas, que constituem padrões visuais distintivos e relativamente raros em outras deficiências (MARSCHNER, 2012). O valor da métrica *Recall* mais modesto (0,740) indica que cerca de 26% das amostras de Ca são confundidas com outras classes, possivelmente nos estágios iniciais em que os sintomas ainda não se distinguiram completamente. A classe Mg apresentou o segundo maior valor para a métrica *Recall* (0,859), o que reflete a clorose internerval característica dessa deficiência, na qual as nervuras permanecem verdes en-

quanto o tecido ao redor amarelece, um padrão visual com geometria bem definida que favorece a detecção por CNNs. O valor da métrica *Precision* de 0,744 sugere que a clorose internerval pode ser confundida com padrões similares de outras classes, o que resulta em um F1-Score de 0,797 que reflete esse compromisso entre sensibilidade e especificidade.

A análise da confusão entre as classes P e S merece atenção particular, pois esse foi o par de classes com maior sobreposição de erros. A classe P obteve valores de 0,701 para a métrica *Precision* e 0,794 para a métrica *Recall*, enquanto a classe S apresentou valor de 0,828 para *Precision*, porém *Recall* de apenas 0,657, o segundo menor entre todas as classes. Essa assimetria sugere que o modelo tende a classificar amostras de S como P com frequência significativa, pois S tem alta *Precision*, o que significa que, quando prediz S, geralmente acerta, mas baixo *Recall*, pois muitas amostras da classe S são incorretamente classificadas em outras classes, predominantemente como classe P. A sobreposição entre essas duas deficiências é consistente com a literatura agrônômica, que documenta que ambas podem causar alterações de pigmentação e coloração nas folhas, como escurecimento e arroxamento no caso de P e clorose uniforme no caso de S, e que a distinção frequentemente requer a observação de características sutis, como a localização preferencial dos sintomas e a progressão temporal específica (MARSCHNER, 2012; MALAVOLTA, 2006). A classe K, com F1-Score de 0,738, apresentou um perfil equilibrado entre *Precision* (0,710) e *Recall* (0,768), o que é coerente com a necrose marginal visualmente distinta que caracteriza essa deficiência, embora com alguma confusão residual com Ca em estágios iniciais, quando ambas afetam as bordas foliares.

Na Figura 19 é apresentada a matriz de confusão completa do melhor modelo, o que permite uma análise detalhada dos padrões de erro. A diagonal principal concentra a maioria das predições, o que confirma o desempenho global satisfatório, mas as regiões fora da diagonal revelam as confusões sistemáticas discutidas acima. A dominância visual da classe N na diagonal é evidente, enquanto a dispersão na linha da classe Completo confirma a dificuldade de classificação de folhas saudáveis. A confusão bidirecional entre as classes P e S é visível nos elementos correspondentes fora da diagonal, o que corrobora a análise quantitativa das métricas por classe.

Para contextualizar o impacto da dimensão temporal na capacidade de discriminação por classe, a comparação entre o cenário ótimo (frac60, 77,97%) e o de pior desempenho temporal (frac95, 64,57%) revela uma degradação qualitativa nos padrões de confusão. No cenário mais restrito, a dispersão fora da diagonal é marcadamente maior, especialmente para as classes Completo, P e S, cujas características visuais em estágios iniciais são particularmente ambíguas. As matrizes de confusão correspondentes encontram-se no Apêndice B (Figuras 26 e 27), bem como a do melhor modelo Híbrido CNN-ViT (314×314 pixels, frac70, 71,68%), cujas confusões mais acentuadas nas classes Completo e S são coerentes com a limitação espacial do mecanismo de atenção com 25 *patches*.

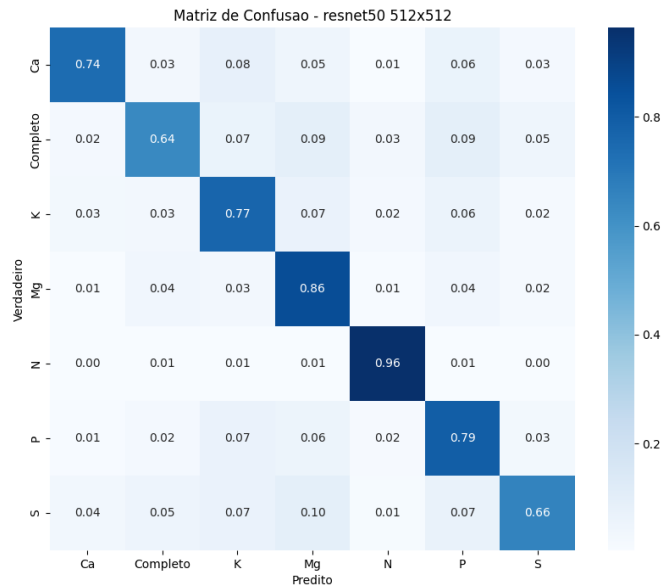


Figura 19 – Matriz de confusão do melhor modelo (ResNet50, 512×512 , frac60, acurácia de 77,97%). A diagonal principal evidencia o bom desempenho geral, com destaque para a classe N. Confusões entre P e S refletem a sobreposição de sintomas visuais dessas deficiências.

5.7 Análise de Interpretabilidade (H3)

A análise apresentada nesta seção é de natureza qualitativa e exploratória. Métricas formais de avaliação de interpretabilidade, como *Pointing Game*, *Intersection over Union* contra máscaras anatômicas, ou *deletion/insertion*, não foram aplicadas; sua incorporação fica registrada como uma sugestão de trabalho futuro. As observações a seguir devem ser interpretadas como evidências exploratórias compatíveis com a hipótese, e não como sua confirmação formal.

A terceira hipótese (H3) investiga se os modelos treinados são capazes de localizar, nas imagens das folhas, regiões agronomicamente relevantes para a classificação das deficiências nutricionais. A verificação dessa hipótese é fundamental não apenas para avaliar a confiabilidade das predições, mas também para aferir o potencial dos modelos como ferramentas de apoio à decisão para profissionais do setor, que necessitam compreender o raciocínio subjacente às classificações automatizadas. Para essa análise, foram utilizadas duas abordagens complementares, sendo elas a técnica Grad-CAM aplicada ao ResNet50 em duas camadas distintas (**layer3** e **layer4**), e a visualização dos mapas de atenção do modelo Híbrido CNN-ViT. A comparação entre camadas do ResNet50 permite avaliar o efeito do campo receptivo na qualidade e na especificidade espacial dos mapas de interpretabilidade, enquanto a análise dos mapas de atenção do Híbrido investiga se o mecanismo de *self-attention* dos *Transformers* oferece uma via alternativa de explicabilidade.

Uma contribuição metodológica desta análise é a comparação sistemática entre os mapas Grad-CAM gerados a partir de **layer3** e **layer4** do ResNet50. Na Figura 20 é apresentada essa comparação para amostras representativas de diferentes classes de deficiência. A **layer3**, localizada mais próxima do início da rede, possui um campo receptivo menor e produz mapas de ativação com maior resolução espacial, o que resulta em regiões de ativação mais localizadas e específicas. Em contrapartida, a **layer4**, como última camada convolucional, possui um campo receptivo que abrange praticamente toda a imagem, o que gera mapas de ativação mais difusos e menos informativos do ponto de vista da localização anatômica dos sintomas. A superioridade dos mapas da **layer3** é particularmente evidente para classes com sintomas espacialmente localizados, como K, caracterizado por necrose marginal, e Mg, caracterizado por clorose internerval, nas quais a maior resolução espacial permite a identificação precisa das regiões anatômicas afetadas. Para classes com sintomas mais difusos, como N, a diferença entre as duas camadas é menos pronunciada, o que é esperado, dado que a clorose generalizada não possui uma localização preferencial na folha.

A razão pela qual a **layer3** produz mapas superiores para este problema específico está relacionada à escala dos padrões visuais de interesse. Os sintomas de deficiências nutricionais, como clorose internerval, necrose marginal, pontuações necróticas e alterações de pigmentação, são padrões de escala intermediária, maiores que texturas de baixo nível, capturadas nas primeiras camadas, mas menores que a folha inteira, representada nas últimas camadas. A **layer3**, cujo campo receptivo corresponde aproximadamente a essa escala intermediária, está, portanto, naturalmente sintonizada com a granularidade dos sintomas foliares. Essa observação tem implicações práticas para a aplicação de Grad-CAM em problemas de classificação de estresses vegetais, pois a escolha da camada-alvo não deve ser automática (última camada convolucional, como é a prática comum), mas deve considerar a escala espacial dos padrões de interesse, o que pode variar conforme a espécie, o tipo de estresse e a resolução da imagem.

A análise qualitativa dos mapas Grad-CAM da **layer3** do ResNet50 revela padrões de ativação que são notavelmente consistentes com a sintomatologia agrônômica conhecida para cada deficiência. Para a classe Ca, os mapas de ativação concentram-se nas regiões amareladas e danificadas da folha, particularmente nas áreas de necrose apical e deformação dos tecidos jovens, o que corresponde ao padrão esperado de manifestação dessa deficiência (MARSCHNER, 2012). O alto valor obtido na métrica *Precision* dessa classe (88,5%) é, portanto, coerente com a capacidade do modelo de focalizar em regiões anatomicamente relevantes e visualmente distintivas. Para a classe Mg, os mapas de ativação destacam as áreas internervais da folha, onde a clorose se desenvolve enquanto as nervuras permanecem verdes. Esse padrão de ativação reproduz com fidelidade a sintomatologia clássica da deficiência de Mg, na qual o tecido entre as nervuras perde progressivamente a coloração verde, o que cria um contraste visual com as nervuras preservadas que serve

como assinatura diagnóstica (MALAVOLTA, 2006). O alto *Recall* dessa classe (85,9%) é consistente com a presença de um padrão visual geometricamente bem definido que favorece a detecção automatizada. Para a classe K, os mapas Grad-CAM concentram a ativação predominantemente nas margens e bordas das folhas, o que reflete a necrose marginal característica dessa deficiência.

Na Figura 21 é apresentado um *grid* ampliado com mapas Grad-CAM da `layer3` para todas as sete classes, o que oferece uma visão completa da capacidade de localização do modelo.

Para a deficiência de N, os mapas de ativação apresentam uma distribuição mais ampla e sem padrão geométrico preferencial em comparação com classes de sintomas localizados, como K e Mg, o que é coerente com a natureza da clorose generalizada que caracteriza essa deficiência. Esse comportamento é semanticamente correto, visto que o sintoma relevante é uma alteração cromática global da folha e não um padrão restrito a uma região anatômica específica, de modo que o gradiente do modelo tende a se dispersar em vez de convergir para um único foco. Para as deficiências de P e S, os mapas revelam ativações igualmente distribuídas e menos focalizadas, o que é consistente com a natureza mais difusa e inespecífica dos sintomas dessas deficiências. Os sintomas de P incluem escurecimento generalizado e pigmentação arroxeada, enquanto S se manifesta por uma clorose que se inicia nas folhas mais jovens, ambos padrões que não possuem uma localização anatômica tão definida quanto a necrose marginal do K. Para a classe Completo, que representa o grupo de controle sem sintomas, o Grad-CAM é fundamentalmente não-informativo, e a ausência de correspondência nas informações apresentadas na Tabela 8 reflete essa limitação. Por definição matemática, o método sempre produz alguma ativação, visto que o gradiente do *score* da classe em relação às ativações da camada é sempre não-nulo. No entanto, para uma classe definida pela ausência de padrões patológicos, não existe região anatômica que discrimine positivamente a classe, de modo que a ativação resultante é difusa e não deve ser interpretada como evidência de que o modelo identificou uma característica diagnóstica significativa. Essa limitação intrínseca do Grad-CAM para classes de controle é consistente com o menor *Recall* da classe Completo e reforça a necessidade de interpretar mapas de ativação sempre em conjunto com o conhecimento da natureza do problema.

Ao se observarem os dados contidos na Tabela 8, nota-se a aderência aparente entre os padrões de ativação dos mapas Grad-CAM da `layer3` do ResNet50 e a sintomatologia documentada na literatura agrônômica clássica de nutrição mineral de plantas. Para cada classe de deficiência, são confrontadas as regiões de ativação identificadas pelo modelo com os sintomas visuais esperados segundo Malavolta (2006), o que permite uma avaliação estruturada da coerência agrônômica aparente das representações aprendidas, sem substituir métrica formal.

Na Tabela 8 é possível observar que as três deficiências com padrões espaciais mais

Tabela 8 – Aderência qualitativa entre os padrões de ativação do Grad-CAM (**layer3**, ResNet50) e a sintomatologia agrônômica conhecida. A coluna “Aderência qualitativa” indica o grau aparente de alinhamento entre o comportamento do modelo e o conhecimento agrônômico, conforme avaliação subjetiva conduzida pelo autor por inspeção visual; não substitui métrica formal de interpretabilidade.

Classe	Sintoma esperado	Localização	Ativação Grad-CAM	Aderência qualitativa
N	Clorose uniforme	Folhas velhas	Difusa sobre toda a superfície	Alta
K	Necrose das pontas e margens	Folhas velhas	Concentrada nas bordas e extremidades	Alta
Mg	Clorose internerval	Folhas velhas	Destaque nas áreas internervais	Alta
Ca	Amarelecimento marginal, deformações	Folhas novas	Regiões amareladas e danificadas	Média
P	Escurecimento, pigmentação arroxeada	Folhas velhas	Distribuída, pouco focalizada	Média
S	Clorose uniforme	Folhas novas	Distribuída, inespecífica	Média
Completo	Ausência de sintomas	—	Difusa, não-informativa (limitação do método)	—

definidos na literatura agrônômica, nomeadamente N, caracterizado por clorose uniforme, K, por necrose marginal, e Mg, por clorose internerval, são precisamente aquelas para as quais o Grad-CAM produz mapas de ativação com maior correspondência anatômica. Essa convergência é particularmente significativa porque o modelo não recebeu qualquer supervisão espacial durante o treinamento, sendo que as regiões de interesse foram aprendidas exclusivamente a partir dos rótulos de classe, o que sugere que o ResNet50 aprendeu, de forma emergente e sem supervisão espacial, representações compatíveis com os critérios visuais reconhecidos na literatura agrônômica para a identificação de deficiências nutricionais. Para as classes P, S e Ca, cujos sintomas são mais difusos ou menos geometricamente definidos, os mapas de ativação são correspondentemente menos localizados, o que é coerente com a natureza dos sintomas descritos na literatura. A classe Completo, por sua vez, apresenta um padrão de ativação que sugere classificação por exclusão e foca na região central e na nervura principal em busca da ausência de sintomas patológicos.

Com base nessa análise, observa-se, nos mapas Grad-CAM da **layer3** do ResNet50, que o modelo é capaz de localizar regiões agronomicamente relevantes nas folhas, com padrões de ativação consistentes com a sintomatologia conhecida de cada deficiência. A superioridade da **layer3** sobre a **layer4** constitui uma contribuição metodológica relevante para a aplicação de técnicas de XAI em classificação de estresses vegetais, pois evidencia que a escolha da camada-alvo deve considerar a escala espacial dos sintomas de interesse e não seguir automaticamente a prática de selecionar a última camada convolucional (SELVARAJU et al., 2017). A análise comparativa entre o Grad-CAM e os mapas de atenção do modelo Híbrido CNN-ViT, bem como o impacto da granularidade do *grid* de atenção, é desenvolvida na seção seguinte.

5.8 Estudo de Ablação: Granularidade dos *Patches*

Os resultados do *benchmark*, ao evidenciar que o modelo Híbrido CNN-ViT permaneceu consistentemente abaixo da ResNet50 em todas as configurações avaliadas, motivaram a investigação de um possível fator limitante na própria arquitetura híbrida, qual seja, o

tamanho do *pool* de regiões gerado pelo *backbone* convolucional antes da etapa de atenção. Esse estudo de ablação foi concebido *post hoc*, como resposta direta aos achados do *benchmark*.

Os resultados da Seção 5.4 identificaram a baixa resolução do *grid* de *patches* (5×5 , 25 *patches*) como uma possível causa da inferioridade do modelo Híbrido CNN-ViT em relação ao ResNet50. Para investigar essa hipótese de forma controlada, foi conduzido um estudo de ablação no qual a grade de *patches* do modelo Híbrido foi sistematicamente variada entre quatro configurações, sendo elas 5×5 (25 *patches*), 7×7 (49 *patches*), 10×10 (100 *patches*) e 14×14 (196 *patches*). Cada configuração foi avaliada em duas resoluções (314×314 e 512×512 pixels) e seis frações temporais selecionadas, num total de 48 experimentos adicionais. O objetivo primário dessa ablação é quantificar o impacto da granularidade espacial no desempenho do modelo, e o objetivo secundário é avaliar se o aumento do número de *patches* melhora a qualidade dos mapas de atenção para fins de interpretabilidade. A configuração 14×14 (196 *patches*) é particularmente relevante, pois se aproxima da resolução utilizada pelo ViT-B/16 padrão, o que permite uma comparação indireta com a granularidade adotada naquela arquitetura (DOSOVITSKIY et al., 2021).

Na Tabela 9 e na Figura 22 são apresentados os melhores desempenhos de cada configuração de *patches*. Os resultados revelam um ganho marginal e saturante com o aumento do número de *patches*, uma vez que a melhor configuração, 10×10 , alcança 71,51%, apenas 1,43 pp acima da configuração original 5×5 . A configuração 14×14 apresenta desempenho ligeiramente inferior à 10×10 , o que indica saturação e possível início de *overfitting* com o aumento excessivo da dimensionalidade do *grid*. Embora a magnitude do ganho esteja abaixo do limiar que exigiria múltiplas sementes para confirmação formal, o padrão de saturação é consistente entre as quatro configurações avaliadas, o que sugere, sem comprovar formalmente, que o número de *patches* não é o gargalo dominante. A diferença observada permanece insuficiente para fechar a distância de cerca de 6 pp em relação ao ResNet50 (77,97%).

Tabela 9 – Melhor desempenho do modelo Híbrido CNN-ViT por configuração da grade de *patches* no estudo de ablação. A acurácia satura a partir de 10×10 *patches*, com ganho marginal sobre a configuração original.

Grade de <i>patches</i>	Acurácia máx. (%)	Fração	Resolução (pixels)
5×5 (25 <i>patches</i>)	70,08	frac60	314×314
7×7 (49 <i>patches</i>)	70,18	frac60	512×512
10×10 (100 <i>patches</i>)	71,51	frac60	314×314
14×14 (196 <i>patches</i>)	70,55	frac65	314×314

A saturação do desempenho observada a partir de 10×10 *patches* é um resultado significativo, pois indica que o gargalo principal do modelo Híbrido não reside na resolução do *grid* de *patches*. Se a baixa granularidade fosse a causa primária da diferença de desempenho em relação ao ResNet50, seria esperado um ganho substancial e contí-

nuo com o aumento progressivo do número de *patches*. A ausência desse ganho sugere que outros fatores, como a limitação do volume de dados para o treinamento eficaz do componente *Transformer*, a profundidade do módulo de atenção, com apenas seis camadas, ou a eficiência do *transfer learning* para o *backbone* convolucional compartilhado, desempenham um papel mais determinante na limitação do desempenho do modelo Híbrido. Essa conclusão tem implicações importantes para o *design* de futuras arquiteturas híbridas para este domínio. Em vez de investir exclusivamente no aumento da resolução espacial do mecanismo de atenção, seria mais produtivo explorar estratégias como o pré-treinamento auto-supervisionado do componente *Transformer*, o aumento do volume de dados de treinamento, ou a utilização de mecanismos de atenção mais sofisticados que combinem atenção local e global (DAI et al., 2021).

Do ponto de vista da interpretabilidade, na Figura 23 é apresentada a comparação direta entre os mapas Grad-CAM da **layer3** do ResNet50 e os mapas de atenção do Híbrido nas quatro configurações de *patches*, para as classes Ca, Mg e K, cujos sintomas apresentam localização anatômica mais definida na literatura agrônômica. Observa-se uma melhoria progressiva na resolução espacial dos mapas do Híbrido com o aumento do número de *patches*, sendo que a configuração 5×5 , com 25 regiões de atenção, produz mapas grosseiros, enquanto a configuração 14×14 , com 196 regiões, permite uma localização mais refinada das áreas de interesse. No entanto, mesmo na configuração mais granular, os mapas de atenção do Híbrido não alcançam a precisão e a coerência anatômica dos mapas Grad-CAM da **layer3** do ResNet50, pois ainda carecem da capacidade de distinguir com precisão nervuras, áreas internervais e margens foliares que o Grad-CAM evidencia com clareza. Essa observação reforça a conclusão de que, para aplicações de interpretabilidade em problemas de classificação de estresses foliares, a técnica Grad-CAM aplicada a CNNs permanece a abordagem mais eficaz entre as investigadas neste trabalho. Uma visualização abrangente para todas as sete classes encontra-se na Figura 28 do Apêndice C.

A análise conjunta dos resultados quantitativos e visuais da ablação conduz a três conclusões principais. Primeiro, o número de *patches* tem impacto limitado no desempenho classificatório do modelo Híbrido, com ganho máximo de cerca de 1,5 pp e saturação a partir de 10×10 *patches*. Segundo, o aumento de *patches* melhora a resolução visual dos mapas de atenção, mas essa melhoria é insuficiente para equiparar a qualidade do Grad-CAM. Terceiro, o gargalo principal do modelo Híbrido reside em fatores que transcendem a resolução do *grid*, o que redireciona as perspectivas de melhoria para aspectos como volume de dados, estratégias de pré-treinamento e arquitetura do módulo de atenção. Essas conclusões fundamentam as recomendações para trabalhos futuros apresentadas no Capítulo 6 e contribuem para a compreensão mais ampla das condições sob as quais arquiteturas híbridas que combinam convoluções e atenção podem ser competitivas em domínios especializados com *datasets* de tamanho moderado.

5.9 Discussão

Os resultados experimentais apresentados neste capítulo, com 98 configurações experimentais distintas, permitem uma avaliação abrangente das três hipóteses formuladas nesta pesquisa e proporcionam *insights* relevantes para o campo da classificação automatizada de deficiências nutricionais em plantas. Esta seção sintetiza os achados principais, contextualiza-os em relação à literatura existente, discute suas implicações e explicita as limitações do trabalho.

A hipótese H1, que postulava a dependência temporal do desempenho de classificação, foi confirmada de forma robusta, o que revela adicionalmente um padrão de curva de sino que não havia sido antecipado. A amplitude de variação do desempenho ao longo das frações temporais, da ordem de 13,40 pp entre frac95 (64,57%) e o pico em frac60 (77,97%) para o ResNet50 em 512×512 pixels, supera tanto o efeito da escolha da arquitetura, de cerca de 6 pp, quanto o da resolução de entrada, de cerca de 1 a 2 pp, o que evidencia ser o particionamento temporal dos dados o fator de maior impacto sobre a acurácia. A evidência mais direta da dependência temporal, todavia, não reside na magnitude desse ganho, parcialmente confundida com o aumento de volume das janelas mais amplas, mas na localização do ótimo no interior da grade de frações, dado que a acurácia declina em direção a frac50 apesar do maior volume de dados disponível nesse extremo, padrão consistente em todas as arquiteturas e resoluções avaliadas. A revelação do formato de sino, com pico entre frac60 e frac65 e declínio em frações mais abrangentes, é uma contribuição original desta pesquisa que não encontra precedentes diretos na literatura de *deep learning* aplicada à agricultura. A existência de uma janela temporal de melhor desempenho médio, explicada pela interação entre o ganho de volume de dados e a diluição por dados assintomáticos, tem implicações práticas diretas para o projeto de sistemas de monitoramento e para o planejamento de campanhas de coleta de dados. A literatura de *deep learning* aplicada à agricultura tipicamente trata o problema de classificação como uma tarefa estática, a exemplo de Ferentinos (2018), e a presente pesquisa demonstra que essa simplificação pode levar a conclusões excessivamente otimistas sobre a capacidade diagnóstica dos modelos.

A hipótese H2, relativa à suposta superioridade das arquiteturas com atenção global, não foi sustentada sob o protocolo padronizado adotado, com evidências convergentes em múltiplas camadas de análise. A superioridade do ResNet50 sobre o Híbrido CNN-ViT e, por conseguinte, deste sobre o ViT-B/16 puro, foi observada de forma inequívoca, visto que o ResNet50 superou o Híbrido em todas as 20 comparações pareadas (diferença média de cerca de 6 pp) e o ViT-B/16 em todas as 10 frações avaliadas (diferença média de cerca de 13 pp). Notavelmente, o ViT-B/16, a arquitetura que mais depende de grandes volumes de dados e que carece do viés indutivo de localidade das convoluções, apresentou o pior desempenho entre as três, com colapso de acurácia tanto na janela mais restrita, limitada pela escassez de dados (frac95: 46,46%), quanto na inclusão de dados assintomá-

ticos (frac50: 47,65%). O estudo de ablação complementa esse quadro ao demonstrar que o aumento do número de *patches* do Híbrido de 5×5 para 14×14 produziu um ganho de apenas cerca de 1,5 pp, o que confirma que o gargalo do modelo Híbrido não reside na resolução do *grid*. A convergência de todas essas evidências fortalece substancialmente a conclusão de que o viés indutivo convolucional é estruturalmente mais adequado para este domínio. Essa conclusão é corroborada pelo estudo comparativo de Pushpa, Akash e Akash (2025), que demonstrou a superioridade de modelos convolucionais, especificamente o VGG16, sobre técnicas clássicas de aprendizado de máquina, como o Random Forest e o SVC, na classificação de deficiências nutricionais em folhas de café. No entanto, é importante ressaltar que a refutação de H2 é contextual, uma vez que se aplica às condições específicas desta pesquisa (25.362 imagens classificadas, 7 classes, 20 épocas máximas), e não deve ser generalizada para cenários com *datasets* significativamente maiores ou configurações arquiteturais distintas (LI et al., 2025; GUO et al., 2022). Soma-se a esses fatores o protocolo de aumento de dados padronizado, deliberadamente conservador e restrito ao espelhamento horizontal e a variações cromáticas, *ColorJitter*. Como os ViTs são particularmente sensíveis à política de aumento de dados e dependem de estratégias mais intensivas, como RandAugment, CutMix e Mixup, para atingir desempenho competitivo (TOUVRON et al., 2021), é plausível que a padronização adotada tenha favorecido as arquiteturas convolucionais em detrimento das baseadas em atenção, o que constitui uma limitação adicional a ser considerada na interpretação da refutação de H2.

A hipótese H3 encontrou evidências exploratórias compatíveis com sua formulação, com resultados que contribuem para a discussão crescente sobre a importância da XAI na agricultura computacional (SHOAIB et al., 2025). A confirmação formal requer avaliação quantitativa, mediante métricas como *Pointing Game* ou *deletion/insertion*, não conduzida neste trabalho. Os mapas Grad-CAM da **layer3** do ResNet50 demonstraram capacidade de localizar regiões agronomicamente relevantes, como regiões amareladas para Ca, áreas internervais para Mg e bordas foliares para K, com coerência fisiológica notável. A comparação entre **layer3** e **layer4** revelou que a escolha da camada-alvo é um fator determinante para a qualidade dos mapas, com a **layer3** produzindo visualizações superiores devido à correspondência entre seu campo receptivo e a escala dos sintomas foliares. Em contrapartida, os mapas de atenção do Híbrido apresentaram limitações significativas, mesmo com o aumento de *patches* (ablação), e não alcançaram a precisão do Grad-CAM. A capacidade de gerar explicações visuais compreensíveis é um requisito funcional, e não apenas estético, para a adoção de sistemas de diagnóstico automatizado por profissionais do campo.

Ao comparar os resultados obtidos com trabalhos anteriores que utilizaram *datasets* de folhas de café, a acurácia de 77,97% obtida pelo melhor modelo situa-se em uma faixa competitiva, dada a complexidade do problema. Esgario, Krohling e Ventura (2020), ao trabalharem com o *dataset* BRACOL para classificação de severidade de doenças bióticas

em café (ferrugem, cercosporiose, bicho-mineiro e folhas saudáveis), alcançaram acurácias superiores a 95% com a arquitetura ResNet50 em abordagem *multi-task*, embora em um problema de natureza distinta (doenças bióticas versus deficiências nutricionais). Tuesta-Monteza, Mejia-Cabrera e Arcila-Diaz (2023), ao disponibilizarem o *dataset* CoLeaf-DB para classificação de deficiências nutricionais em café, contribuíram com um recurso valioso para a comunidade, embora sem relatar resultados de classificação extensivos no artigo de descrição do *dataset*. Trabalhos mais recentes reforçam a relevância da linha de investigação, como demonstrado por Randive, Sapkal e Bornare (2025), que aplicaram redes de *deep learning* para detecção de deficiências nutricionais em folhas de café com o CoLeaf-DB, e reportaram acurácias superiores a 97% com a arquitetura YOLOv8n-cls, porém em um problema com apenas três classes (N, P, K) e sem a incorporação da variável temporal. É fundamental ressaltar que comparações diretas entre trabalhos são limitadas pelas diferenças nos *datasets*, nos protocolos de avaliação e nas configurações experimentais, visto que nenhum dos trabalhos citados incorporou a dimensão temporal de forma controlada, e o *dataset* utilizado nesta pesquisa é composto exclusivamente por imagens de *Coffea arabica* cultivada em estufa controlada. A principal contribuição diferencial desta pesquisa reside, portanto, não na acurácia absoluta obtida, mas na demonstração empírica da dependência temporal e na análise de interpretabilidade que conecta as predições do modelo ao conhecimento agrônomo de domínio.

A análise conjunta dos resultados temporais e por classe revela uma interação importante entre essas duas dimensões. O alto *Recall* da classe N (96,4%) é observado no cenário com sintomas avançados, mas a clorose generalizada que caracteriza essa deficiência provavelmente emerge de forma mais precoce e uniforme do que os sintomas localizados de outras deficiências. Essa hipótese é corroborada pelo fato de que, mesmo nos cenários temporais mais restritivos, as confusões relacionadas ao N são relativamente menores em comparação com as confusões entre as demais classes. Isso sugere que diferentes deficiências possuem “velocidades” distintas de manifestação visual e que futuras pesquisas poderiam explorar estratégias de classificação adaptativas, com o estágio temporal estimado dos sintomas como variável auxiliar na predição. A observação de que a confusão P-S é mais pronunciada que outras confusões reforça a necessidade de incorporar informações temporais complementares, uma vez que essas duas deficiências podem ser mais facilmente distinguidas pela progressão de seus sintomas ao longo do tempo do que pela aparência em um único instante.

Não obstante os resultados significativos, algumas limitações dos experimentos devem ser explicitadas para permitir uma avaliação informada do escopo e da generalidade dos achados. Em primeiro lugar, o *dataset* é composto por uma única variedade de *Coffea arabica* cultivada em condições de estufa controlada, o que limita a generalização dos resultados para outras variedades, condições de campo aberto ou ambientes climáticos distintos. As condições controladas de estufa eliminam efeitos de confusão, como vari-

ações de iluminação, presença de pragas e danos mecânicos, que estariam presentes em imagens de campo e poderiam alterar significativamente o desempenho dos modelos. Em segundo lugar, o treinamento foi limitado a 20 épocas com parada antecipada, o que, embora previna o sobreajuste, pode não ter permitido que o modelo Híbrido CNN-ViT, que possui mais parâmetros a otimizar no componente *Transformer*, convergisse plenamente. Em terceiro lugar, a busca de hiperparâmetros foi restrita a uma configuração fixa, sem exploração sistemática de taxas de aprendizado, tamanhos de *batch* ou estratégias de *data augmentation* mais agressivas. Em quarto lugar, a validação foi baseada em *hold-out* simples, sem a realização de validação cruzada *k-fold* ou repetições com diferentes sementes aleatórias, o que impede a estimativa de intervalos de confiança e a realização de testes estatísticos formais para confirmar a significância das diferenças observadas. Em consequência, comparações cujo *gap* é inferior a aproximadamente 3 pp devem ser interpretadas com cautela, pois podem refletir variabilidade entre realizações da mesma configuração e não diferenças efetivas de capacidade. Por fim, a avaliação da interpretabilidade foi conduzida de forma qualitativa, sem métricas quantitativas formais como *Pointing Game* ou *Insertion/Deletion*, embora a concordância observada com a sintomatologia agrônômica forneça uma validação indireta substancial.

Do ponto de vista das implicações práticas, é necessário enfatizar que as observações a seguir são apresentadas sob a condição expressa de que dependem de validação em condições de campo não realizada neste trabalho. Sob essa ressalva, os resultados apontam para a viabilidade técnica, em condições controladas de estufa, de sistemas automatizados de diagnóstico de deficiências nutricionais baseados em imagens RGB, particularmente quando combinados com monitoramento temporal contínuo. A acurácia de 77,97% obtida pelo melhor modelo, embora insuficiente para substituir completamente a análise laboratorial, é potencialmente promissora como ferramenta de triagem que, após validação externa, poderia direcionar as intervenções de manejo nutricional e reduzir a dependência de análises químicas tradicionais como alternativas de referência. A identificação da janela temporal de melhor desempenho médio (frac60–frac65) oferece indicação preliminar para a definição de protocolos de coleta de imagens. A incorporação da técnica Grad-CAM oferece uma via de validação cruzada na qual o agrônomo pode não apenas receber a predição da deficiência, mas também verificar se as regiões indicadas pelo modelo correspondem às áreas onde se espera encontrar os sintomas específicos, o que aumenta a confiança na recomendação automatizada. A escolha da ResNet50 como arquitetura preferencial, aliada à sua menor complexidade computacional em comparação com modelos baseados em *Transformers*, facilita a implantação em dispositivos com recursos limitados, o que é relevante para o contexto da agricultura de precisão em regiões com infraestrutura tecnológica restrita (SOUSSI et al., 2024).

Em síntese, os 98 experimentos conduzidos nesta pesquisa demonstram que a classificação de deficiências nutricionais em cafeeiros é viável com acurácia de 77,97%, com o uso

de CNNs pré-treinadas e imagens RGB, com a janela temporal de melhor desempenho médio identificada na fração 60%. A dimensão temporal é o fator de maior impacto no desempenho, com uma curva de sino que revela a interação entre discriminabilidade visual e diluição por dados assintomáticos. A hierarquia de desempenho entre as arquiteturas evidencia que o viés indutivo de localidade das CNNs é mais adequado do que mecanismos de atenção global para capturar padrões texturais finos em folhas, vantagem que não é eliminada pelo aumento de *patches* na ablação. Já a técnica Grad-CAM, aplicada à **layer3** do ResNet50, fornece explicações visuais agronomicamente consistentes que superam os mapas de atenção do modelo Híbrido em especificidade e coerência fisiológica. As limitações identificadas apontam direções claras para trabalhos futuros, conforme discutido no Capítulo 6.

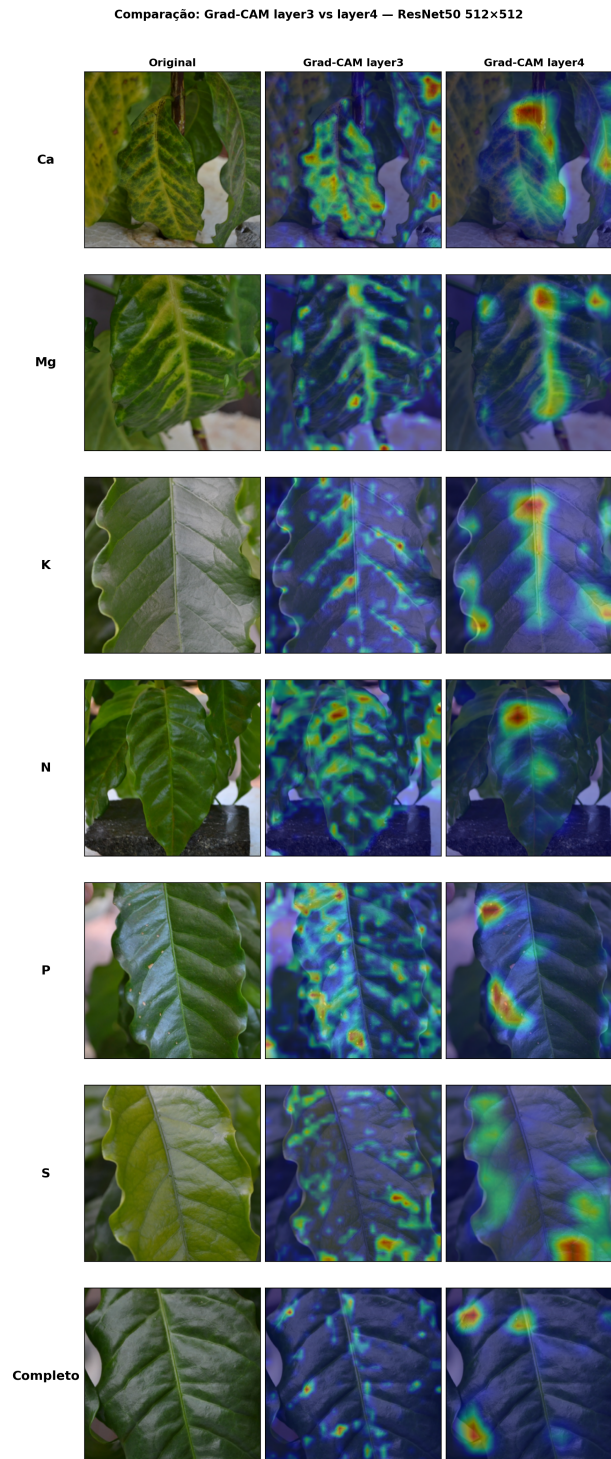


Figura 20 – Comparação entre os mapas Grad-CAM gerados a partir de `layer3` e `layer4` do ResNet50 para amostras representativas das sete classes. Cada linha apresenta uma classe, com a imagem original (coluna da esquerda), o mapa Grad-CAM extraído da `layer3` (coluna do centro) e o mapa extraído da `layer4` (coluna da direita). A `layer3` produz mapas mais localizados e fisiologicamente coerentes, enquanto a `layer4` gera ativações mais difusas que cobrem áreas extensas da folha.

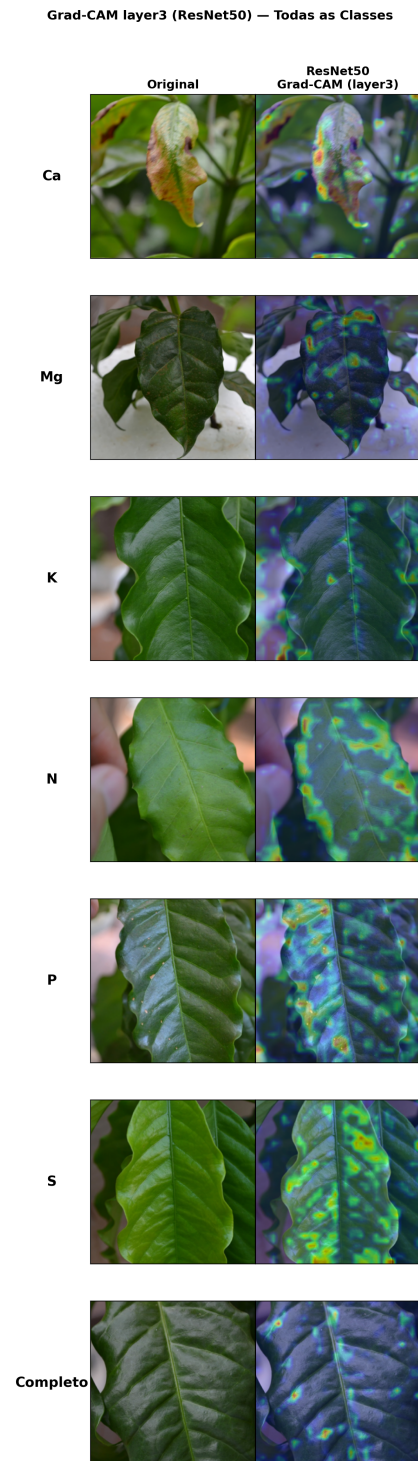


Figura 21 – *Grid* completo dos mapas Grad-CAM da `layer3` do ResNet50 para todas as sete classes. Observam-se padrões de ativação consistentes com a sintomatologia agrônômica: N apresenta ativação difusa (clorose generalizada), K foca nas bordas (necrose marginal), Mg destaca áreas internervais (clorose internerval), e Ca ativa regiões de dano localizado.

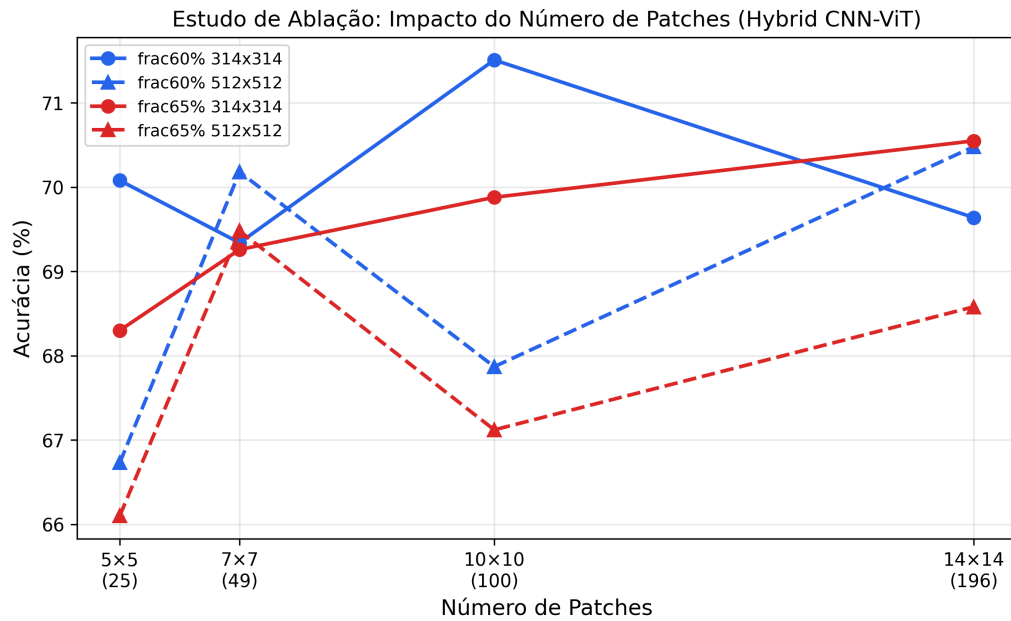


Figura 22 – Resultado do estudo de ablação: acurácia do modelo Híbrido CNN-ViT em função da configuração de *patches* (5×5 a 14×14). O ganho com o aumento de *patches* é marginal ($\sim 1,5$ pp), o que indica que a resolução do *grid* não é o principal gargalo de desempenho.

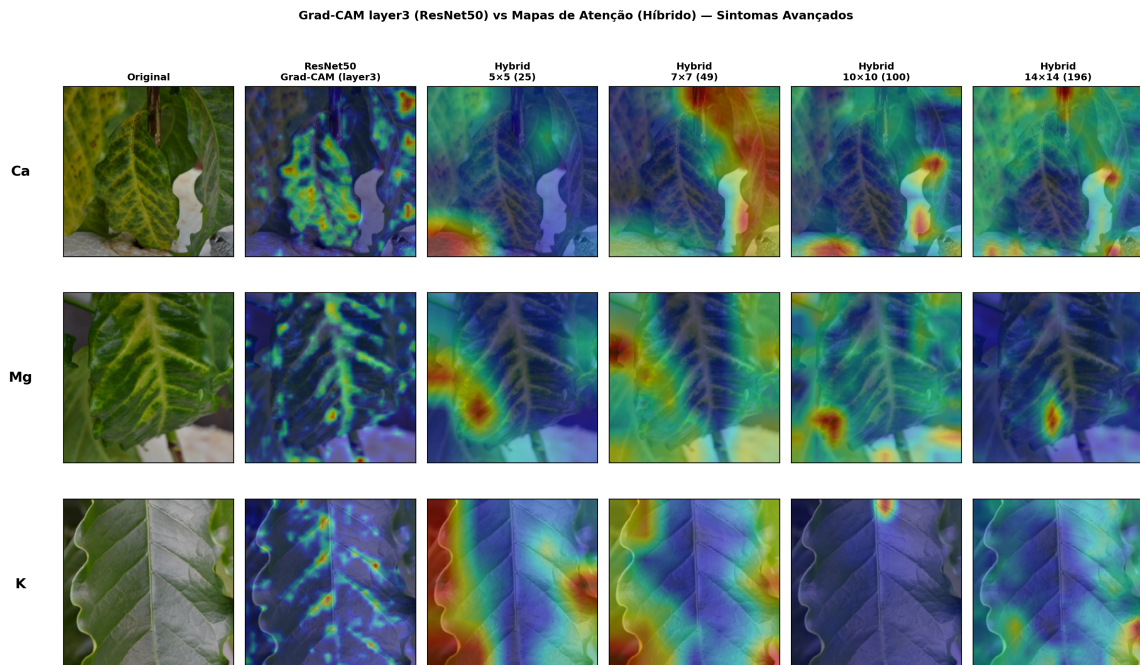


Figura 23 – Comparação entre os mapas Grad-CAM da *layer3* do ResNet50 e os mapas de atenção do Híbrido CNN-ViT nas quatro configurações de *patches* (5×5 , 7×7 , 10×10 , 14×14) para as classes Ca, Mg e K. O Grad-CAM localiza com precisão as regiões de sintomas, enquanto os mapas de atenção do Híbrido apresentam resolução espacial progressivamente maior com o aumento dos *patches*, sem alcançar a especificidade do Grad-CAM.

Conclusão

A presente dissertação investigou um problema central na interseção entre agricultura de precisão e visão computacional, por meio da classificação automatizada de deficiências nutricionais em folhas de *Coffea arabica* com o uso de técnicas de *deep learning*, com ênfase na dimensão temporal da manifestação dos sintomas. O diagnóstico convencional dessas deficiências depende da inspeção visual realizada por agrônomos, um processo inerentemente subjetivo e reativo, uma vez que os sintomas fenotípicos, como clorose, necrose e deformações foliares, só se tornam visualmente evidentes em estágios avançados, quando perdas de produtividade já podem ser irreversíveis (MARSCHNER, 2012; MALAVOLTA, 2006). A literatura de *deep learning* aplicada à detecção de estresses em plantas, embora crescente, trata predominantemente o problema como tarefa de classificação estática a partir de imagens capturadas em um único momento temporal, sem incorporar sistematicamente a natureza progressiva dos sintomas (FERENTINOS, 2018; MEHTA et al., 2025). Mesmo trabalhos recentes voltados especificamente à classificação de deficiências nutricionais em café, como o de Randive, Sapkal e Bornare (2025), não incorporam a dimensão temporal da evolução dos sintomas, o que confirma que a lacuna abordada por esta pesquisa permanece relevante e não endereçada.

Diante dessa lacuna, a pesquisa desenvolveu um *pipeline* experimental que permitiu avaliar sistematicamente três arquiteturas de *deep learning* (ResNet50, ViT-B/16 e um modelo Híbrido CNN-ViT) em diferentes janelas temporais do desenvolvimento dos sintomas, com um conjunto de dados de aproximadamente 28 mil imagens de sete classes nutricionais. O controle de variação temporal, possibilitou quantificar no contexto de deficiências nutricionais em café, a relação entre o estágio de evolução dos sintomas e a capacidade discriminativa dos classificadores. Os resultados, apresentados e discutidos no Capítulo 5, forneceram evidências quantitativas robustas para a avaliação das três hipóteses formuladas na pesquisa, além de revelar achados inesperados sobre a superioridade das arquiteturas convolucionais neste domínio.

O conjunto dos experimentos conduzidos permitiu não apenas atingir os objetivos propostos na Seção 1.2, mas também gerar *insights* relevantes para a comunidade científica

nas áreas de visão computacional aplicada à agricultura e de nutrição vegetal. As seções seguintes sintetizam os principais resultados obtidos, detalham as contribuições da pesquisa em diferentes dimensões, reconhecem as limitações do estudo e delineiam direções promissoras para trabalhos futuros.

6.1 Síntese dos Resultados

A primeira hipótese da pesquisa (H1), que postulava a existência de uma dependência temporal significativa no desempenho de classificação, foi sustentada e refinada pelos dados, que substituíram a expectativa monotônica original por uma relação não monotônica em formato de sino. A acurácia do modelo ResNet50 aplicado à imagens com resolução de 512×512 pixels revelou uma curva de sino ao longo das janelas temporais, com valor de 64,57% na fração mais restrita (últimos 5% dos dados temporais, **frac95**), com pico de 77,97% na fração **frac60** (últimos 40% dos dados) e declínio para 75,92% na fração mais ampla (**frac50**, últimos 50%). Essa curva evidencia a existência de uma janela temporal de melhor desempenho médio, pois nas janelas mais restritas, o volume de dados é insuficiente para o treinamento eficaz; nas fases intermediárias (**frac60-frac65**), os sinais visuais atingem seu máximo de distinguibilidade; e ao incluir dados excessivamente antigos (assintomáticos), o modelo é parcialmente confundido. O melhor desempenho global do estudo (77,97% de acurácia e *macro* F1 de 0,774) foi obtido nessa janela, o que consolida a tendência observada em todas as configurações testadas. Esse valor deve ser interpretado como estimativa de teto sob protocolo *hold-out* estratificado por classe, e não como estimativa de generalização para plantas não observadas, dado que a estrutura do conjunto de dados, com quatro vasos por classe e milhares de imagens por vaso, faz com que cada planta esteja presente em ambos os subconjuntos de treino e validação. Esses achados possuem implicações diretas para a prática agronômica, ao indicar que existe uma janela temporal associada ao melhor desempenho médio observado para o diagnóstico automatizado.

A comparação exploratória associada à segunda hipótese (H2), que avaliava se arquiteturas baseadas em mecanismos de atenção global superariam as arquiteturas puramente convolucionais, revelou o padrão oposto ao dos ganhos teoricamente esperados, sob o protocolo padronizado adotado neste trabalho. A hierarquia de desempenho observada foi inequívoca, uma vez que o ResNet50 superou o Híbrido CNN-ViT, que por sua vez superou o ViT-B/16 puro, com diferenças de 6,29 pp e 10,23 pp entre o ResNet50 e cada uma das arquiteturas baseadas em atenção nos respectivos picos de desempenho (77,97%, 71,68% e 67,74%). O ResNet50 superou o Híbrido em todas as 20 comparações pareadas e o ViT-B/16 em todas as 10 frações avaliadas na resolução 512×512 . Notavelmente, o ViT-B/16 apresentou comportamento de colapso nas frações extremas, tanto em **frac95** (46,46%) quanto em **frac50** (47,65%), o que evidencia que a ausência do viés indutivo

de localidade o torna hipersensível tanto à escassez de sintomas quanto à inclusão de dados assintomáticos. Para investigar se o gargalo do modelo Híbrido residia na resolução espacial limitada de sua grade de *patches*, foi conduzido um estudo de ablação com 48 experimentos adicionais ao variar o tamanho da grade de 5×5 (25 *patches*) para 7×7 , 10×10 e 14×14 (196 *patches*). Os resultados revelaram um ganho marginal de apenas 1,5 pontos percentuais, o que indica que o gargalo não reside na granularidade espacial da atenção. Conforme discutido no Capítulo 5, a interpretação mais plausível para a hierarquia observada reside no viés indutivo de localidade inerente às operações convolucionais, que se mostrou particularmente adequado para a captura de padrões texturais finos, como variações sutis de coloração, manchas internervais e alterações nas margens foliares, que caracterizam as deficiências nutricionais (KHAN et al., 2022; CORDONNIER; LOUKAS; JAGGI, 2020). Investigar configurações otimizadas individualmente para cada arquitetura, com pesos pré-treinados, taxas de aprendizado e estratégias de aumento de dados ajustadas a cada paradigma, permanece como direção de trabalho futuro.

A terceira hipótese (H3), relativa à interpretabilidade dos modelos por meio de mapas de ativação, encontrou evidências exploratórias compatíveis com sua formulação; a confirmação formal requer avaliação quantitativa não conduzida neste trabalho. A análise qualitativa com Grad-CAM (SELVARAJU et al., 2017) revelou que o melhor modelo (ResNet50 512×512) concentra suas ativações em regiões agronomicamente relevantes. A investigação comparativa entre as camadas **layer3** e **layer4** da ResNet50 demonstrou que a **layer3**, com *receptive field* menor, produz mapas de ativação mais localizados e fisiologicamente coerentes, sendo que, para a deficiência de Ca, a ativação concentra-se nas regiões amareladas; para Mg, nas áreas internervais em que a clorose é esperada; e para K, nas bordas e extremidades foliares (MARSCHNER, 2012). Os mapas de atenção do modelo Híbrido, embora tenham melhorado com o aumento do número de *patches* (de 5×5 para 14×14), não alcançaram a precisão espacial do Grad-CAM, o que reforça a superioridade interpretativa da abordagem convolucional. No entanto, essa validação permanece de natureza qualitativa e carece de uma avaliação quantitativa mais rigorosa para que a hipótese possa ser considerada plenamente confirmada.

Além das três hipóteses, dois achados adicionais merecem destaque. Primeiramente, a análise por classe do melhor modelo revelou uma assimetria importante no desempenho, pois a classe N apresentou o maior *recall* entre todas as classes (96,4%), o que indica que seus sintomas, tipicamente clorose generalizada e amarelecimento uniforme, são visualmente altamente discriminativos. Em contrapartida, a classe Completo (controle) obteve o menor *recall* (63,5%), o que pode ser explicado pela similaridade visual entre folhas saudáveis e folhas nos estágios iniciais de algumas deficiências, que ainda não apresentam sintomas evidentes. Em segundo lugar, o impacto da resolução das imagens mostrou-se positivo, porém modesto, visto que a passagem de 314×314 para 512×512 proporcionou um ganho de aproximadamente 1 ponto percentual na acurácia, um resultado que deve

ser ponderado frente ao custo computacional adicional envolvido no processamento de imagens de maior resolução.

6.2 Contribuições

6.2.1 Contribuições Metodológicas

A principal contribuição metodológica desta dissertação reside na proposição e validação de um protocolo de particionamento temporal para a avaliação de classificadores de estresses visuais em plantas. O protocolo de particionamento temporal introduzido neste trabalho permitiu simular cenários de diagnóstico em diferentes estágios de evolução dos sintomas, desde janelas muito restritas, como **frac95**, com os últimos 5% dos dados, até janelas com ampla cobertura temporal, como **frac50**, com os últimos 50%, o que revelou a existência de um ponto ótimo em **frac60–frac65**. Essa abordagem transcende o cenário específico da cafeicultura e pode ser adaptada para qualquer estudo longitudinal de estresses bióticos ou abióticos em culturas agrícolas, desde que os dados possuam anotação temporal adequada.

Ademais, o protocolo comparativo entre arquiteturas de naturezas distintas (convolucionais, baseadas em atenção e híbridas) sob condições experimentais rigorosamente padronizadas, com o mesmo otimizador, a mesma taxa de aprendizado, o mesmo *scheduler* e o mesmo critério de parada antecipada, constitui uma contribuição metodológica relevante. A decisão de manter todos os hiperparâmetros fixos entre as arquiteturas, embora possa não extrair o desempenho máximo de cada modelo individualmente, garante que as diferenças observadas sejam atribuíveis exclusivamente às características arquiteturais, e não a escolhas de otimização. A documentação da tentativa inicial com validação cruzada *Leave-One-Group-Out* (LOGO) por vaso, que foi abandonada em favor do *hold-out* estratificado devido à alta variância temporal entre plantas, também representa uma contribuição ao estabelecer um registro transparente de decisões metodológicas que pode orientar pesquisas futuras.

6.2.2 Contribuições Empíricas

No plano empírico, a pesquisa forneceu evidência quantitativa no escopo da literatura revisada, no contexto de deficiências nutricionais em café, da magnitude do impacto temporal sobre o desempenho de classificação. A identificação de uma curva de sino no desempenho temporal, com acurácia que subiu de 64,57% (**frac95**) para 77,97% (**frac60**) e recuou para 75,92% (**frac50**), constitui um dado, no melhor de nosso conhecimento, ainda não reportado, que permite dimensionar tanto o desafio da detecção precoce quanto o efeito deletério da inclusão excessiva de dados assintomáticos. Esse achado, documen-

tado em 50 experimentos com *pipeline* consistente, é relevante para calibrar expectativas realistas sobre o potencial e os limites de sistemas automatizados de diagnóstico visual.

A comparação entre arquiteturas produziu uma hierarquia empiricamente robusta, pois a superioridade da ResNet50 sobre o Híbrido CNN-ViT e o ViT-B/16 puro foi confirmada em 50 experimentos de *benchmark* com *pipeline* padronizado. O ViT-B/16, avaliado na resolução 512×512 , obteve acurácia máxima de apenas 67,74%, inferior ao Híbrido (71,68%) e muito abaixo da ResNet50 (77,97%). O estudo de ablação com 48 experimentos adicionais, ao variar o número de *patches* do modelo Híbrido de 25 (5×5) a 196 (14×14), demonstrou que o gargalo de desempenho não reside na granularidade espacial da atenção, mas na própria adequação do mecanismo de atenção global para este domínio. Esses achados contribuem para o debate na literatura sobre a adequação de *Transformers* para domínios com dados limitados e padrões predominantemente texturais (DOSOVITSKIY et al., 2021; TOUVRON et al., 2021). O detalhamento por classe, com destaque para os extremos de N (96,4% de *recall*) e Completo (63,5% de *recall*), oferece ainda um mapeamento granular da dificuldade relativa de cada categoria, informação que pode direcionar esforços futuros de coleta de dados e engenharia de *features*.

6.2.3 Contribuições Aplicadas

Do ponto de vista aplicado, a pesquisa demonstrou a viabilidade técnica em condições controladas de estufa da classificação automatizada de deficiências nutricionais em folhas de café a partir de imagens RGB, com acurácia de 77,97% obtida pelo modelo ResNet50 na resolução 512×512 na fração temporal ótima. Embora esse resultado não seja ainda suficiente para substituir integralmente o diagnóstico por um especialista, ele representa uma prova de conceito, em condições controladas, para o desenvolvimento de ferramentas de triagem e apoio à decisão. Em cenários práticos, um sistema com esse nível de acurácia poderia servir como uma primeira etapa de diagnóstico, o que alerta o produtor sobre possíveis deficiências e reduz a dependência de análises laboratoriais custosas e demoradas, como a espectrometria de massa com plasma indutivamente acoplado (ICP-MS) (HANSEN et al., 2013).

A identificação de que a deficiência de N é altamente discriminativa (96,4% de *recall*) possui relevância direta para a prática agrônômica, dado que o N é um dos macronutrientes mais demandados pela cultura do café e cuja deficiência causa impactos significativos na produtividade (MALAVOLTA, 2006). Esse resultado sugere que, mesmo com as limitações atuais dos modelos, um sistema de diagnóstico automatizado já poderia oferecer valor concreto para a detecção dessa deficiência específica. Adicionalmente, os mapas de ativação gerados pelo Grad-CAM oferecem uma camada de explicabilidade que pode aumentar a confiança dos agrônomos na utilização de tais ferramentas, ao demonstrar que o modelo fundamenta suas decisões em regiões foliares coerentes com o conhecimento agrônômico estabelecido.

6.3 Limitações

A primeira e mais relevante limitação do estudo refere-se ao ambiente de coleta dos dados. Todas as imagens foram obtidas em condições de estufa controlada, com iluminação padronizada, fundo uniforme e ausência de fatores de confusão como sombreamento, sobreposição foliar, presença de pragas e variações climáticas. Essas condições, embora ideais para isolar o efeito das deficiências nutricionais e garantir a validade interna do experimento, não refletem a complexidade do ambiente de campo, onde os modelos seriam efetivamente utilizados. A generalização dos resultados obtidos para cenários reais de produção permanece, portanto, uma questão em aberto que necessita de investigação específica. De forma complementar, o estudo foi conduzido exclusivamente com a espécie *Coffea arabica*, o que limita a extensibilidade das conclusões para outras espécies de café, como *Coffea canephora*, ou para outras culturas agrícolas que também sofrem de deficiências nutricionais similares.

No que tange à escala dos dados, o *dataset* utilizado, com aproximadamente 28 mil imagens distribuídas em sete classes e quatro vasos, pode ser considerado de porte moderado para os padrões atuais de *deep learning*. Essa limitação de escala pode ter impactado de forma desproporcional as arquiteturas baseadas em *Transformers*, reconhecidamente mais dependentes de grandes volumes de dados para superar o desempenho de CNNs pré-treinadas (DOSOVITSKIY et al., 2021; TOUVRON et al., 2021). Dessa forma, a inferência de que as CNNs são intrinsecamente superiores para este tipo de problema deve ser interpretada com cautela, pois pode refletir, ao menos parcialmente, uma limitação do volume de dados disponível e não apenas uma vantagem arquitetural intrínseca. Adicionalmente, o número de réplicas biológicas por tratamento, de quatro vasos por classe, embora típico de experimentos em casa de vegetação, é insuficiente para isolar a variabilidade individual entre plantas; em consequência, as conclusões deste trabalho referem-se aos tratamentos nutricionais avaliados sob condições controladas, e não à variabilidade genotípica e fenológica que se espera em campo.

Uma limitação adicional, de natureza conceitual, decorre da estratégia de rotulagem e de particionamento temporal. As classes foram definidas pelo tratamento nutricional aplicado a cada vaso, o que assegura a origem controlada da deficiência, mas não corresponde a um diagnóstico de estágio conduzido por especialista. Ademais, o particionamento temporal opera sobre um único corte cronológico global do conjunto ordenado por data, e não de forma independente por classe, de modo que se assume, de forma implícita, um tempo de manifestação de sintomas semelhante entre as diferentes deficiências. Como o período efetivamente assintomático de cada deficiência não foi estabelecido de forma independente, as janelas temporais devem ser interpretadas como faixas do período experimental, e não como estágios sintomáticos calibrados por nutriente.

Do ponto de vista do treinamento, o número máximo de 20 épocas, embora complementado por mecanismos de *early stopping* e redução adaptativa da taxa de aprendi-

zado, pode não ter sido suficiente para que os modelos, especialmente os baseados em *Transformers*, convergissem plenamente. A estratégia de validação adotada, o *hold-out* estratificado 80/20, embora estável e adequada para o escopo deste trabalho, não oferece as garantias estatísticas da validação cruzada *k-fold* completa, o que limita a robustez das estimativas de desempenho reportadas. Finalmente, a ausência de uma exploração sistemática de hiperparâmetros, como diferentes taxas de aprendizado, tamanhos de *batch* e estratégias de *data augmentation* específicas para cada arquitetura, constitui uma limitação adicional, dado que cada família de modelos pode responder de maneira distinta a diferentes configurações de otimização.

6.4 Trabalhos Futuros

A direção mais imediata e crítica para a continuidade desta pesquisa consiste na validação dos modelos em condições de campo, que constitui pré-requisito para a tradução das observações deste trabalho em recomendações operacionais para o produtor. Um estudo de validação externa, com imagens coletadas em lavouras comerciais com variações naturais de iluminação, ângulo de captura, estágio fenológico e fundo complexo, permitiria avaliar a real capacidade de generalização dos classificadores treinados em ambiente controlado. Essa etapa poderia incluir a construção de um novo *dataset* anotado por agrônomos especialistas, com imagens capturadas por *smartphones* e drones, de forma a simular cenários realistas de utilização. A comparação entre o desempenho em estufa e em campo forneceria uma medida objetiva do *domain gap*, o que orienta o desenvolvimento de estratégias de adaptação de domínio caso a degradação de desempenho se mostre significativa.

Uma direção complementar, diretamente ligada à motivação de detecção precoce que orientou este trabalho, consiste em quantificar a detectabilidade em termos de tempo de calendário. Em vez de expressar a viabilidade do diagnóstico apenas em função da largura da janela de dados, tal análise buscaria estimar a partir de quantas semanas ou meses após a indução da deficiência os modelos passam a discriminá-la de forma confiável, e confrontar esse marco com o momento em que os sintomas se tornam perceptíveis à inspeção humana. Essa caracterização permitiria mensurar de forma direta a antecipação diagnóstica proporcionada pela abordagem automatizada, aspecto que o presente desenho experimental, centrado na largura da janela temporal, não isolou.

Outra vertente promissora é a adoção de uma formulação hierárquica do problema de classificação, na qual um primeiro estágio distinguiria folhas saudáveis de folhas deficientes e um segundo estágio identificaria a deficiência específica. Essa decomposição poderia mitigar a confusão observada entre a classe de controle e os estágios iniciais das deficiências, evidenciada pelo menor *recall* da classe Completo, além de alinhar a estrutura do classificador ao fluxo de decisão agrônomo, que tipicamente parte da detecção

de anormalidade para o diagnóstico do nutriente específico.

Uma segunda linha de investigação promissora envolve a exploração de arquiteturas mais recentes que buscam reconciliar as vantagens dos mecanismos convolucionais e de atenção de formas distintas das exploradas nesta dissertação. Modelos como *Swin Transformer* (LIU et al., 2021) e variantes hierárquicas baseadas em atenção representam avanços arquiteturais que podem alterar o equilíbrio observado entre CNNs e *Transformers*. O *Swin Transformer*, em particular, introduz atenção em janelas locais com deslocamento, e combina a eficiência computacional das convoluções com a flexibilidade dos mecanismos de atenção. Avaliar essas arquiteturas sob o mesmo protocolo experimental estabelecido nesta pesquisa permitiria expandir o *benchmark* e verificar se a superioridade das CNNs se mantém frente a *designs* híbridos mais sofisticados.

A integração de modalidades de dados complementares constitui uma terceira direção de trabalho futuro com elevado potencial, alinhada às tendências identificadas em revisões recentes da literatura (UPADHYAY et al., 2025; SHOAIB et al., 2025), que apontam a fusão multimodal e a implantação em dispositivos móveis como fronteiras promissoras para a visão computacional aplicada à agricultura. A combinação de imagens RGB com dados de sensoriamento remoto hiperespectral ou multiespectral poderia capturar alterações bioquímicas nas folhas que precedem a manifestação de sintomas visíveis, o que antecipa o diagnóstico para estágios ainda mais precoces do estresse nutricional (LIAKOS et al., 2018; SOUSSI et al., 2024). Abordagens de fusão multimodal, que combinam *features* extraídas de diferentes fontes de dados em uma representação unificada, têm demonstrado resultados promissores em outros domínios (ALBAHLI, 2025) e poderiam ser adaptadas para este contexto. Adicionalmente, a incorporação de dados meteorológicos, edáficos e fenológicos em modelos multimodais poderia enriquecer o poder diagnóstico dos classificadores, o que fornece contexto ambiental que auxilie na diferenciação entre deficiências nutricionais e outros estresses abióticos.

O desenvolvimento e a implantação de soluções em dispositivos móveis e de borda, *edge devices*, representa uma quarta vertente de trabalho futuro de grande relevância aplicada. A otimização dos modelos por meio de técnicas como destilação de conhecimento, *knowledge distillation*, quantização e poda, *pruning*, permitiria sua execução em *smartphones* e dispositivos embarcados de baixo custo, o que viabiliza o uso em campo por produtores que não dispõem de conectividade permanente com a internet. Um aplicativo móvel que permita ao agricultor capturar uma imagem de folha e obter, em tempo real, uma estimativa da deficiência nutricional poderia democratizar o acesso a diagnósticos que hoje dependem de infraestrutura laboratorial ou de consultoria especializada.

Por fim, a expansão do *dataset* constitui uma necessidade transversal a todas as direções mencionadas. A inclusão de outras espécies do gênero *Coffea*, como *C. canephora*, e de variedades híbridas, de deficiências adicionais, como micronutrientes, entre eles ferro, manganês e zinco, e de condições ambientais diversificadas, como diferentes altitudes, ti-

pos de solo e regimes hídricos, permitiria treinar modelos mais robustos e generalizáveis. A integração desses classificadores com sistemas de recomendação de fertilizantes, capazes de converter o diagnóstico visual em prescrições agronômicas específicas, entre elas o tipo de fertilizante, a dosagem e o momento de aplicação, representaria o passo final para transformar a pesquisa em uma ferramenta de impacto efetivo na cafeicultura de precisão.

6.5 Considerações Finais

Esta dissertação avançou na compreensão de um problema relevante e pouco explorado na literatura, qual seja, a relação entre a dimensão temporal dos sintomas de deficiências nutricionais e a capacidade de classificação de modelos de *deep learning*. Ao demonstrar quantitativamente que a acurácia de classificação segue uma curva de sino temporal, de 64,57% na janela mais restrita, correspondente ao **frac95**, para 77,97% na janela de melhor desempenho, correspondente ao **frac60**, e 75,92% com inclusão excessiva de dados assintomáticos, a pesquisa forneceu uma resposta concreta e matizada à pergunta central que a motivou. A existência de uma janela temporal de melhor desempenho indica que não basta “mais dados”; é necessário selecionar a janela temporal adequada para maximizar a capacidade discriminativa do classificador. A constatação de que a ResNet50 superou consistentemente abordagens híbridas CNN-ViT, confirmada em 50 experimentos de *benchmark* e corroborada por 48 experimentos de ablação, reforça a importância de avaliar empiricamente as premissas teóricas sobre a adequação de arquiteturas para domínios específicos, em vez de assumir que modelos mais recentes são universalmente superiores.

Os resultados obtidos, embora circunscritos a um cenário experimental controlado, lançam bases sólidas para o desenvolvimento de ferramentas de diagnóstico automatizado que possam, no futuro, auxiliar produtores e agrônomos na identificação precoce de estresses nutricionais em cafeeiros. A intersecção entre visão computacional e agronomia, explorada neste trabalho, permanece um campo fértil para inovações que combinem rigor metodológico com impacto prático, o que contribui para uma cafeicultura mais sustentável, eficiente e baseada em evidências.

Referências

ADADI, A.; BERRADA, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). **IEEE Access**, IEEE, v. 6, p. 52138–52160, 2018. Disponível em: <<https://doi.org/10.1109/ACCESS.2018.2870052>>.

ALBAHLI, S. Agrifusionnet: A lightweight deep learning model for multisource plant disease diagnosis. **Agriculture**, v. 15, n. 14, p. 1523, 2025. Disponível em: <<https://doi.org/10.3390/agriculture15141523>>.

ALZUBAIDI, L. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. **Journal of Big Data**, v. 8, n. 1, p. 53, 2021. Disponível em: <<https://doi.org/10.1186/s40537-021-00444-8>>.

AMUDHA, M.; BRINDHA, K. Rice leaf nutrient deficiency classification system using car-capsule network. **IEEE Access**, v. 12, p. 169518–169532, 2024. Disponível em: <<https://doi.org/10.1109/ACCESS.2024.3498606>>.

CLEVERT, D.-A.; UNTERTHINER, T.; HOCHREITER, S. Fast and accurate deep network learning by exponential linear units (elus). In: **International Conference on Learning Representations (ICLR)**. [s.n.], 2016. Disponível em: <<https://doi.org/10.48550/arXiv.1511.07289>>.

CORDONNIER, J.-B.; LOUKAS, A.; JAGGI, M. On the relationship between self-attention and convolutional layers. In: **International conference on learning representations**. [s.n.], 2020. Disponível em: <<https://doi.org/10.48550/arXiv.1911.03584>>.

DAI, Z. et al. Coatnet: Marrying convolution and attention for all data sizes. **arXiv preprint arXiv:2106.04803**, 2021. Disponível em: <<https://doi.org/10.48550/arXiv.2106.04803>>.

DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: **International Conference on Learning Representations (ICLR)**. [s.n.], 2021. Disponível em: <<https://doi.org/10.48550/arXiv.2010.11929>>.

DUMOULIN, V.; VISIN, F. **A guide to convolution arithmetic for deep learning**. 2018. Disponível em: <<https://doi.org/10.48550/arXiv.1603.07285>>.

- ESGARIO, J. G. M.; KROHLING, R. A.; VENTURA, J. A. Deep learning for classification and severity estimation of coffee leaf biotic stress. **Computers and Electronics in Agriculture**, v. 169, p. 105162, 2020. Disponível em: <<https://doi.org/10.1016/j.compag.2019.105162>>.
- FERENTINOS, K. P. Deep learning models for plant disease detection and diagnosis. **Computers and Electronics in Agriculture**, v. 145, p. 311–318, 2018. Disponível em: <<https://doi.org/10.1016/j.compag.2018.01.009>>.
- FERREIRA, M. L.; ANDRADE, A. M. d.; SANTIAGO, W. E. Geoprocessamento aplicado à identificação de áreas propícias ao cultivo de café. **Research, Society and Development**, v. 10, n. 4, p. e17410414050, 2021. Disponível em: <<https://doi.org/10.33448/rsd-v10i4.14050>>.
- GANDOLFI, M. R. C.; JESUS, C. M.; GANDOLFI, P. E. Qualidade do emprego nas culturas de café, cana-de-açúcar e milho: estudo comparativo 2000/2010 entre mesorregiões de minas gerais. **Revista de Economia e Sociologia Rural**, v. 57, n. 1, p. 63–80, 2019. Disponível em: <<https://doi.org/10.1590/1234-56781806-94790570104>>.
- GEIRHOS, R. et al. Shortcut learning in deep neural networks. **Nature Machine Intelligence**, Nature Publishing Group, v. 2, n. 11, p. 665–673, 2020. Disponível em: <<https://doi.org/10.1038/s42256-020-00257-z>>.
- GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. <<http://www.deeplearningbook.org>>.
- GRANDINI, M.; BAGLI, E.; VISANI, G. Metrics for multi-class classification: an overview. **arXiv preprint arXiv:2008.05756**, 2020. Disponível em: <<https://doi.org/10.48550/arXiv.2008.05756>>.
- GUO, J. et al. Cmt: Convolutional neural networks meet vision transformers. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [s.n.], 2022. p. 12175–12185. Disponível em: <<https://doi.org/10.1109/CVPR52688.2022.01186>>.
- GUO, X.; FENG, Q.; GUO, F. Cmtnet: a hybrid cnn-transformer network for uav-based hyperspectral crop classification in precision agriculture. **Scientific Reports**, Nature Publishing Group, v. 15, n. 1, p. 1–15, 2025. Disponível em: <<https://doi.org/10.1038/s41598-025-97052-w>>.
- HANSEN, T. H. et al. Multielement plant tissue analysis using icp spectrometry. In: **Plant Mineral Nutrients: Methods and Protocols**. Humana Press, 2013, (Methods in Molecular Biology, v. 953). p. 121–141. Disponível em: <https://doi.org/10.1007/978-1-62703-152-3_8>.
- HE, J. et al. TAC-UNet: Transformer-assisted convolutional neural network for medical image segmentation. **Quantitative Imaging in Medicine and Surgery**, v. 14, n. 12, p. 8824–8839, 2024. Disponível em: <<https://doi.org/10.21037/qims-24-1229>>.
- HE, K. et al. Deep residual learning for image recognition. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [s.n.], 2016. p. 770–778. Disponível em: <<https://doi.org/10.1109/CVPR.2016.90>>.

- HENDRYCKS, D.; GIMPEL, K. Gaussian error linear units (gelus). **arXiv preprint arXiv:1606.08415**, 2016. Disponível em: <<https://doi.org/10.48550/arXiv.1606.08415>>.
- IBGE. **A Geografia do Café**. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística, 2016. (Dinâmica Territorial da Produção Agropecuária). ISBN 978-85-240-4401-4.
- IOFFE, S.; SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: PMLR. **International conference on machine learning**. 2015. p. 448–456. Disponível em: <<https://doi.org/10.48550/arXiv.1502.03167>>.
- JEPKOECH, J. et al. Arabica coffee leaf images dataset for coffee leaf disease detection and classification. **Data in Brief**, v. 36, p. 107142, 2021. Disponível em: <<https://doi.org/10.1016/j.dib.2021.107142>>.
- KHAN, S. et al. Transformers in vision: A survey. **ACM Computing Surveys**, v. 54, n. 10s, p. 1–41, 2022. Disponível em: <<https://doi.org/10.1145/3505244>>.
- LECUN, Y. et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998. Disponível em: <<https://doi.org/10.1109/5.726791>>.
- LI, Z. et al. Enhanced hyperspectral image classification for coastal wetlands using a hybrid cnn-transformer approach with cross-attention mechanism. **Frontiers in Marine Science**, 2025. Disponível em: <<https://doi.org/10.3389/fmars.2025.1613565>>.
- LIAKOS, K. G. et al. Machine learning in agriculture: A review. **Sensors**, v. 18, n. 8, p. 2674, 2018. Disponível em: <<https://doi.org/10.3390/s18082674>>.
- LIU, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**. [s.n.], 2021. p. 10012–10022. Disponível em: <<https://doi.org/10.1109/ICCV48922.2021.00986>>.
- MALAVOLTA, E. **Manual de Nutrição Mineral de Plantas**. São Paulo: Editora Agronômica Ceres, 2006.
- MARSCHNER, P. (Ed.). **Marschner's Mineral Nutrition of Higher Plants**. 3rd. ed. Academic Press, 2012. Disponível em: <<https://doi.org/10.1016/C2009-0-63043-9>>.
- MEHTA, A. R. et al. Ai-powered innovations in agriculture: A systematic review on plant disease detection and classification. **Indian Journal of Agricultural Research**, v. 59, n. 9, p. 1321–1330, 2025. Disponível em: <<https://doi.org/10.18805/IJARE.A-6371>>.
- MOHAN, R. N. V. J.; RAYANOOTHALA, P. S.; SREE, R. P. Next-gen agriculture: integrating AI and XAI for precision crop yield predictions. **Frontiers in Plant Science**, v. 15, p. 1451607, 2025. Disponível em: <<https://doi.org/10.3389/fpls.2024.1451607>>.
- NOYAN, M. A. **Uncovering bias in the PlantVillage dataset**. 2022. Disponível em: <<https://doi.org/10.48550/arXiv.2206.04374>>.

PASZKE, A. et al. PyTorch: An imperative style, high-performance deep learning library. In: **Advances in Neural Information Processing Systems**. [s.n.], 2019. v. 32. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html>.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Disponível em: <<https://jmlr.org/papers/v12/pedregosa11a.html>>.

PUSHPA, B. R.; AKASH, V. V.; AKASH, V. P. A comparative study of machine learning and deep learning frameworks in identifying and detecting nutrient deficiencies in plants using leaf image analysis. In: **2025 6th International Conference for Emerging Technology (INCET)**. [s.n.], 2025. p. 1–6. Disponível em: <<https://doi.org/10.1109/INCET64471.2025.11140314>>.

RANDIVE, L.; SAPKAL, S.; BORNARE, D. Nutrient deficiency detection and classification in coffee leaves using deep learning models. In: **EPJ Web of Conferences**. [s.n.], 2025. v. 328, p. 01038. Disponível em: <<https://doi.org/10.1051/epjconf/202532801038>>.

RUBWURM, M.; KÖRNER, M. Multi-temporal land cover classification with long short-term memory neural networks. **The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, XLII-1/W1, p. 551–558, 2017. Disponível em: <<https://doi.org/10.5194/isprs-archives-XLII-1-W1-551-2017>>.

SAKKA, M. E. et al. A review of cnn applications in smart agriculture using multimodal data. **Sensors**, MDPI, v. 25, n. 2, p. 472, 2025. Disponível em: <<https://doi.org/10.3390/s25020472>>.

SCHERER, D.; MÜLLER, A.; BEHNKE, S. Evaluation of pooling operations in convolutional architectures for object recognition. In: SPRINGER. **Artificial Neural Networks–ICANN 2010**. 2010. p. 92–101. Disponível em: <https://doi.org/10.1007/978-3-642-15825-4_10>.

SELVARAJU, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: **Proceedings of the IEEE International Conference on Computer Vision (ICCV)**. [s.n.], 2017. Disponível em: <<https://doi.org/10.1109/ICCV.2017.74>>.

SESSO, P. P.; FILHO, U. A. S.; PEREIRA, L. F. P. Dimensionamento do agronegócio do café no brasil. **Cadernos de Ciência & Tecnologia**, v. 38, n. 2, p. e26901, 2021. Disponível em: <<https://doi.org/10.35977/0104-1096.cct2021.v38.26901>>.

SHOAIB, M. et al. Plant stress detection using multimodal imaging and machine learning: from leaf spectra to smartphone applications. **Frontiers in Plant Science**, v. 16, p. 1670593, 2026. Disponível em: <<https://doi.org/10.3389/fpls.2025.1670593>>.

_____. Leveraging deep learning for plant disease and pest detection: a comprehensive review and future directions. **Frontiers in Plant Science**, v. 16, 2025. Disponível em: <<https://doi.org/10.3389/fpls.2025.1538163>>.

- SOUSSI, A. et al. Smart sensors and smart data for precision agriculture: A review. **Sensors**, v. 24, n. 8, p. 2647, 2024. Disponível em: <<https://doi.org/10.3390/s24082647>>.
- SRIVASTAVA, N. et al. Dropout: a simple way to prevent neural networks from overfitting. **The journal of machine learning research**, JMLRORG, v. 15, n. 1, p. 1929–1958, 2014. Disponível em: <<https://jmlr.org/papers/v15/srivastava14a.html>>.
- TAN, S. et al. Machine learning approaches for rice seedling growth stages detection. **Frontiers in Plant Science**, v. 13, p. 914771, 2022. Disponível em: <<https://doi.org/10.3389/fpls.2022.914771>>.
- THAKUR, P. S. et al. **Explainable vision transformer enabled convolutional neural network for plant disease identification: PlantXViT**. 2022. ArXiv preprint arXiv:2207.07919. Disponível em: <<https://doi.org/10.48550/arXiv.2207.07919>>.
- TOUVRON, H. et al. Training data-efficient image transformers & distillation through attention. In: PMLR. **International Conference on Machine Learning**. 2021. p. 10347–10357. Disponível em: <<https://doi.org/10.48550/arXiv.2012.12877>>.
- TUESTA-MONTEZA, V. A.; MEJIA-CABRERA, H. I.; ARCILA-DIAZ, J. Coleaf-db: Peruvian coffee leaf images dataset for coffee leaf nutritional deficiencies detection and classification. **Data in Brief**, v. 48, p. 109226, 2023. Disponível em: <<https://doi.org/10.1016/j.dib.2023.109226>>.
- UBBENS, J. R.; STAVNESS, I. Deep plant phenomics: A deep learning platform for complex plant phenotyping tasks. **Frontiers in Plant Science**, Frontiers, v. 8, p. 1190, 2017. Disponível em: <<https://doi.org/10.3389/fpls.2017.01190>>.
- UPADHYAY, A. et al. Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture. **Artificial Intelligence Review**, v. 58, 2025. Disponível em: <<https://doi.org/10.1007/s10462-024-11100-x>>.
- VASWANI, A. et al. Attention is all you need. In: **Advances in Neural Information Processing Systems 30 (NIPS 2017)**. [s.n.], 2017. Disponível em: <<https://doi.org/10.48550/arXiv.1706.03762>>.
- WANG, S. et al. Advances in deep learning applications for plant disease and pest detection: A review. **Remote Sensing**, v. 17, n. 4, p. 698, 2025. Disponível em: <<https://doi.org/10.3390/rs17040698>>.
- ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. In: **Computer Vision – ECCV 2014**. Springer, 2014. p. 818–833. Disponível em: <https://doi.org/10.1007/978-3-319-10590-1_53>.

Apêndices

Resultados Completos dos Experimentos

Este apêndice consolida os resultados numéricos de todos os 98 experimentos conduzidos nesta pesquisa, organizados em duas tabelas: o benchmark principal (50 experimentos) e o estudo de ablação (48 experimentos). Para cada configuração, são reportados a acurácia (%) e o Macro F1-Score no conjunto de validação. Os resultados resumidos e as análises correspondentes encontram-se no Capítulo 5.

A.1 Benchmark Principal

A Tabela 10 apresenta os resultados de todas as 50 configurações do benchmark, organizadas por arquitetura e resolução. Cada célula reporta a acurácia (%) e o Macro F1-Score separados por barra. O melhor resultado de cada combinação de modelo e resolução está destacado em negrito.

Tabela 10 – Resultados completos do benchmark: acurácia (%) e Macro F1-Score para os 50 experimentos. O melhor resultado de cada configuração de modelo e resolução está destacado em negrito.

Modelo / Resolução	frac50	frac55	frac60	frac65	frac70	frac75	frac80	frac85	frac90	frac95
ResNet50 314 × 314	74,73 / 0,741	76,22 / 0,755	76,34 / 0,758	76,86 / 0,761	75,10 / 0,746	73,52 / 0,733	73,50 / 0,735	69,78 / 0,685	70,47 / 0,703	59,84 / 0,597
ResNet50 512 × 512	75,92 / 0,754	75,95 / 0,755	77,97 / 0,774	77,25 / 0,767	76,41 / 0,760	73,84 / 0,736	73,10 / 0,724	72,93 / 0,722	71,46 / 0,715	64,57 / 0,645
Híbrido 314 × 314	67,28 / 0,667	66,01 / 0,659	70,82 / 0,698	70,83 / 0,702	71,68 / 0,713	69,82 / 0,692	67,09 / 0,671	65,05 / 0,636	64,96 / 0,640	50,39 / 0,491
Híbrido 512 × 512	69,69 / 0,687	67,76 / 0,670	71,27 / 0,707	67,23 / 0,667	65,57 / 0,653	67,77 / 0,676	66,40 / 0,655	65,05 / 0,649	61,61 / 0,608	53,54 / 0,551
ViT-B/16 512 × 512	47,65 / 0,471	67,54 / 0,667	66,54 / 0,660	67,74 / 0,667	65,37 / 0,652	63,12 / 0,630	61,48 / 0,609	59,92 / 0,592	57,09 / 0,573	46,46 / 0,464

A.2 Estudo de Ablação

A Tabela 11 apresenta os resultados de todas as 48 configurações do estudo de ablação sobre a granularidade dos *patches* do modelo Híbrido CNN-ViT. As frações avaliadas foram frac50 a frac75, num total de seis valores por configuração. Cada célula reporta a acurácia (%) e o Macro F1-Score. O melhor resultado global está destacado em negrito.

Tabela 11 – Resultados completos do estudo de ablação: acurácia (%) e Macro F1-Score para os 48 experimentos do Híbrido CNN-ViT. O melhor resultado global está destacado em negrito.

Patch / Resolução	frac50	frac55	frac60	frac65	frac70	frac75
$5 \times 5 - 314 \times 314$	67,28 / 0,667	67,41 / 0,668	70,08 / 0,689	68,30 / 0,681	69,32 / 0,694	68,16 / 0,669
$5 \times 5 - 512 \times 512$	67,40 / 0,661	66,54 / 0,660	66,73 / 0,672	66,10 / 0,654	66,56 / 0,666	65,09 / 0,644
$7 \times 7 - 314 \times 314$	66,46 / 0,658	65,88 / 0,645	69,34 / 0,687	69,26 / 0,693	64,78 / 0,644	68,40 / 0,683
$7 \times 7 - 512 \times 512$	65,39 / 0,640	68,33 / 0,681	70,18 / 0,698	69,48 / 0,693	68,40 / 0,677	66,04 / 0,656
$10 \times 10 - 314 \times 314$	65,55 / 0,646	67,72 / 0,669	71,51 / 0,712	69,88 / 0,697	70,96 / 0,707	65,41 / 0,654
$10 \times 10 - 512 \times 512$	66,10 / 0,656	65,22 / 0,639	67,87 / 0,668	67,12 / 0,651	68,07 / 0,673	65,48 / 0,648
$14 \times 14 - 314 \times 314$	66,93 / 0,665	66,97 / 0,662	69,64 / 0,695	70,55 / 0,697	70,04 / 0,688	68,64 / 0,684
$14 \times 14 - 512 \times 512$	66,89 / 0,661	64,91 / 0,642	70,48 / 0,702	68,58 / 0,671	66,56 / 0,654	60,84 / 0,614

Análise Complementar de Desempenho

Este apêndice reúne figuras de suporte à análise de desempenho apresentada no Capítulo 5: o custo computacional detalhado por fração temporal, as métricas individuais por classe em formato gráfico e as matrizes de confusão dos cenários complementares ao melhor modelo.

B.1 Custo Computacional por Fração Temporal

A Figura 24 detalha os tempos de treinamento registrados pelo MLflow para cada arquitetura e fração temporal na resolução 512×512 . A diferença entre o ViT-B/16 e as demais arquiteturas é pronunciada nas frações de maior volume de dados (frac50–frac60), com o ViT a demandar até 297 minutos contra 70 minutos do ResNet50. A análise agregada encontra-se na Tabela 6 do Capítulo 5.

B.2 Métricas por Classe — Visualização Gráfica

A Figura 25 apresenta as métricas de *Precision*, *Recall* e F1-Score do melhor modelo (ResNet50, 512×512 , frac60) em formato gráfico de barras agrupadas. Os valores numéricos correspondentes encontram-se na Tabela 7 do Capítulo 5.

B.3 Matrizes de Confusão Complementares

As Figuras 26 e 27 complementam a análise da Seção 5.6 do Capítulo 5 e apresentam as matrizes de confusão do cenário de pior desempenho temporal e do melhor modelo Híbrido, respectivamente.

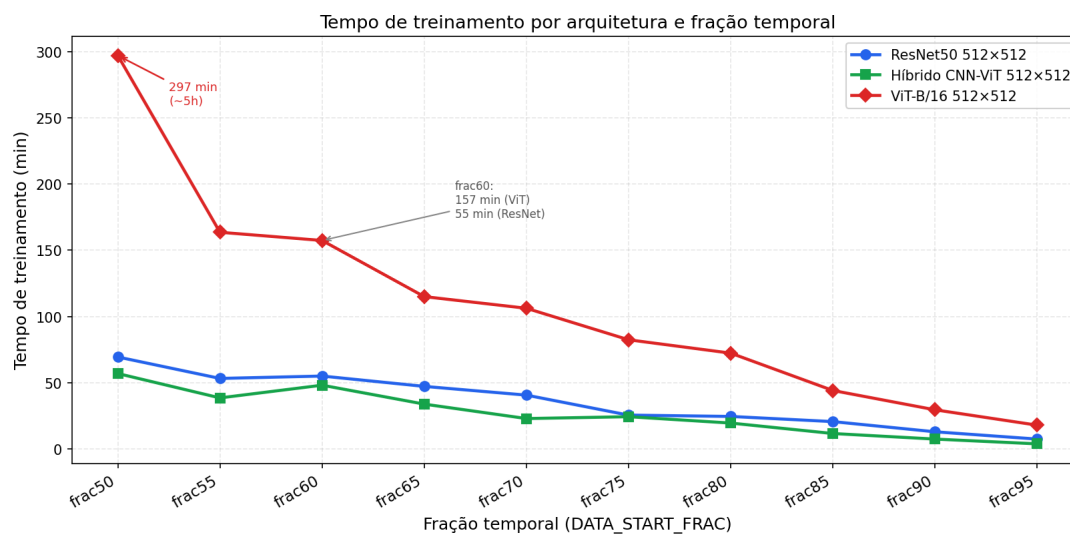


Figura 24 – Tempo de treinamento (em minutos) por arquitetura e fração temporal, registrado pelo **MLflow** para a resolução 512×512 . O ViT-B/16 apresenta custo computacional 3–4× superior ao ResNet50, com maior discrepância nas frações de maior volume de dados (frac50–frac60).



Figura 25 – Visualização das métricas de *Precision*, *Recall* e F1-Score por classe para o melhor modelo (ResNet50, 512×512 , frac60). A deficiência de Nitrogênio apresenta métricas significativamente superiores às demais, enquanto as classes Completo e Enxofre exibem os menores F1-Scores.

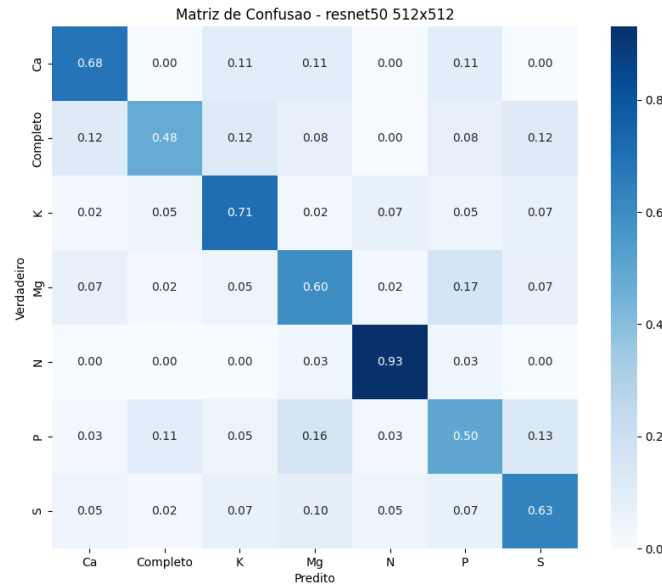


Figura 26 – Matriz de confusão do cenário com janela temporal mais restrita (ResNet50, 512×512 , frac95, acurácia de 64,57%). A dispersão significativa fora da diagonal principal evidencia a dificuldade de discriminação quando o volume de dados de treinamento é severamente limitado. Comparar com a Figura 19.

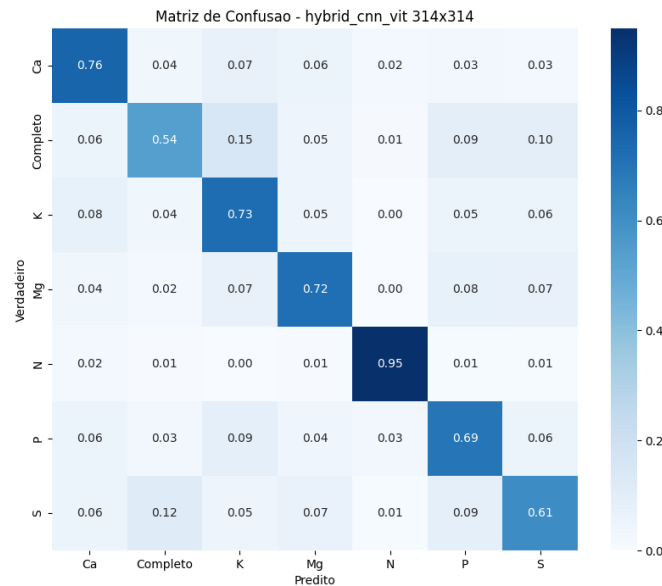


Figura 27 – Matriz de confusão do melhor modelo Híbrido CNN-ViT (314×314 , frac70, acurácia de 71,68%). A diagonal principal menos definida em relação ao ResNet50 evidencia maior confusão entre classes, especialmente para Completo e Enxofre.

Galeria de Mapas de Interpretabilidade

Este apêndice apresenta a visualização abrangente da análise de interpretabilidade desenvolvida nas Seções 5.7 e 5.8 do Capítulo 5, ampliando a cobertura para todas as sete classes do problema.

C.1 Grad-CAM versus Atenção por Tamanho de *Patch*

A Figura 28 constitui a visualização mais abrangente da análise de interpretabilidade. Para cada uma das sete classes, compara-se diretamente o mapa Grad-CAM da `layer3` do ResNet50 com os mapas de atenção do Híbrido nas quatro configurações de *patches*. Essa comparação unificada permite ao leitor avaliar simultaneamente a qualidade relativa de cada abordagem de interpretabilidade e o efeito da granularidade do *grid* de atenção em todo o espaço de classes do problema.

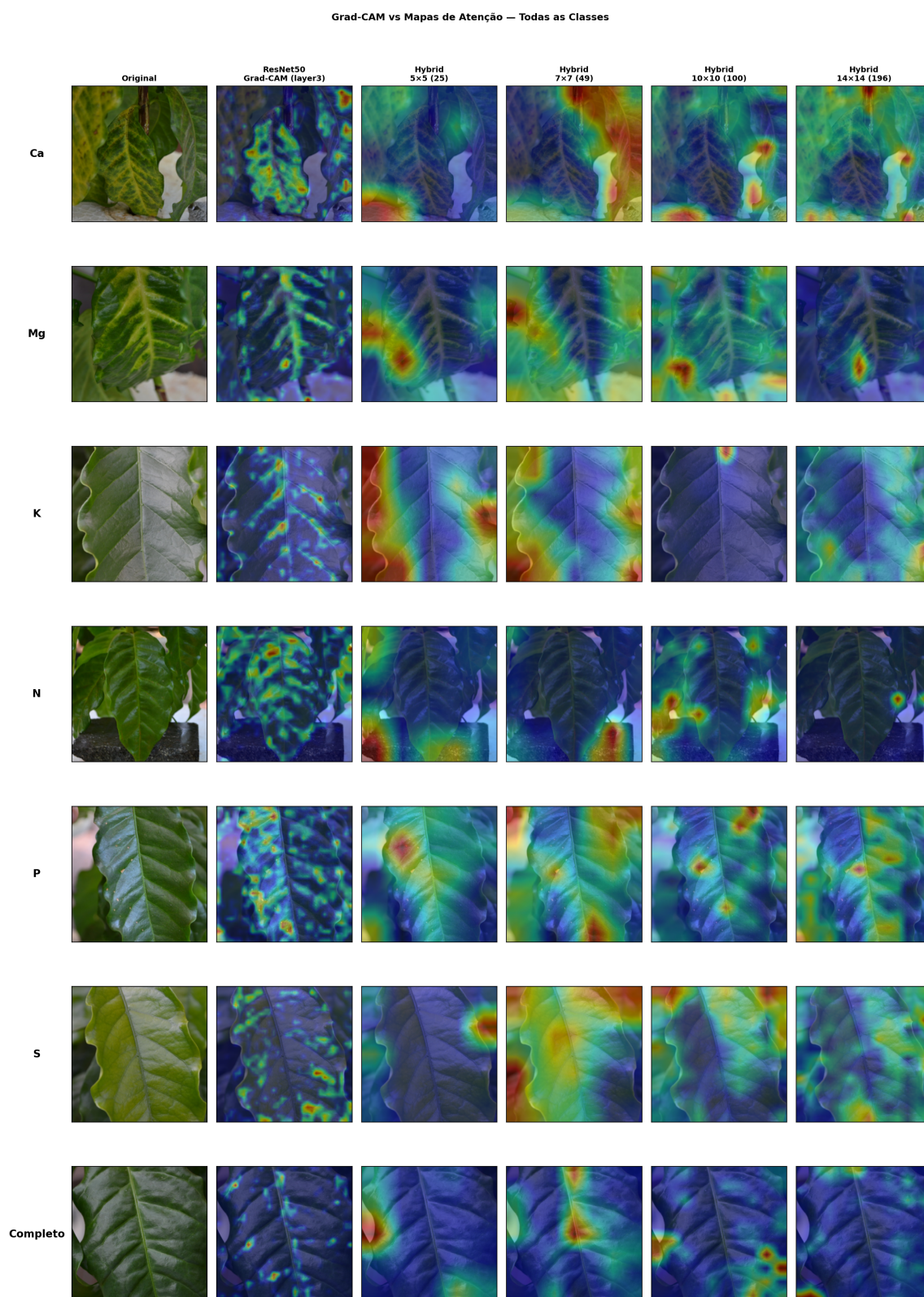


Figura 28 – Comparativo completo entre Grad-CAM (layer3, ResNet50) e mapas de atenção do Híbrido CNN-ViT nas quatro configurações de *patches* para todas as sete classes. O Grad-CAM apresenta consistentemente padrões de ativação mais coerentes com a sintomatologia agrônômica em todas as classes, independentemente do tamanho do *grid* de atenção utilizado pelo Híbrido.