Explicabilidade na Classificação de Imagens com VGG16: Um Estudo com Grad-CAM e Mapas de Calor

David Marques[1]

Marcos Hideki Sakaguchi[1]

João Ricardo Favan[2]

Mauricio Duarte[2]

RESUMO

Este trabalho tem como foco a utilização do *VGG16*, aliado à técnica de *Grad-CAM* para demonstrar uma das formas de explicar o caminho que o modelo percorre, e quais características são levadas em consideração para a decisão de sua resposta, sendo aplicado à identificação de doenças em tomateiros. O principal objetivo foi avaliar visualmente as decisões do modelo de classificação, gerando heatmaps que demonstram quais regiões das imagens foram mais determinantes para a predição. O foco na área de ativação é uma ótima pauta pois muito se utiliza desses modelos, mas não se sabe exatamente o que acontece internamente. Foi utilizado o *Grad-CAM* como uma técnica que permite identificar as áreas específicas de uma imagem que mais influenciam a decisão do modelo, oferecendo de forma visual sobre como a rede convolucional está processando as informações. Ao aplicar essa abordagem treinando o modelo, foi obtido uma acurácia de 70% em cima dos dados de teste. A partir da geração dos mapas de calor, foi possível observar que o modelo utilizou de características importantes como a superfície das folhas referindo-se aos danos do patógeno como principal enfoque, mas também, foi utilizado mais de outros elementos para fazer a classificação. Assim, a análise dos resultados visa demonstrar que modelos complexos como o VGG16 podem ser compreendidos de forma intuitiva, se for aliado com ferramentas auxiliares, como por exemplo a CAM, o que promove maior confiança no uso de inteligência artificial para fins práticos.

Palavras-chave: Mapa de calor, *Grad-CAM*, D*eep learning*.

1 INTRODUÇÃO

A Inteligência Artificial Explicável *(XAI*) tem se destacado como uma área crucial para melhorar a transparência e a interpretabilidade dos modelos de aprendizado profundo, especialmente em aplicações práticas como a identificação

[1] Discente em Big Data no Agronegócio na FATEC Pompeia, Pompeia-SP,
[2] Docentes do curso Big Data no Agronegócio, FATEC Pompeia, Pompeia-SP.

de doenças em plantas (Mersha et al., 2024). Este trabalho concentra-se na utilização do modelo *VGG16* aliado à técnica *Grad-CAM* (*Gradient-weighted Class Activation Mapping*), com o objetivo de demonstrar como os modelos convolucionais processam as informações das imagens e quais características são determinantes para suas decisões. Ao gerar mapas de calor sobre as imagens, é possível visualizar as áreas específicas que influenciam mais diretamente as predições do modelo, promovendo maior confiança e compreensão sobre o funcionamento interno da rede (Natarajan;Nambiar.,2024).

O tomate é uma das culturas agrícolas mais importantes no Brasil, posicionando o país entre os maiores produtores globais. Além de sua relevância econômica, o tomate é amplamente cultivado por pequenos agricultores, cuja subsistência frequentemente depende da saúde e produtividade dessa cultura. Doenças comuns em plantações brasileiras, podem causar perdas significativas, tornando a detecção precoce essencial para mitigar danos. Neste cenário, a inteligência artificial aplicada ao diagnóstico de doenças em tomateiros surge como uma ferramenta estratégica para aprimorar a gestão agrícola e reduzir prejuízos(Viana, 2023).

Modelos como o *VGG16*, que alcançou uma acurácia de 70% nos testes deste estudo, mostram-se eficazes, mas enfrentam desafios de interpretabilidade. A técnica *Grad-CAM* se torna fundamental para compreender se o modelo utiliza características relevantes, como danos na superfície das folhas, ou se inclui elementos irrelevantes no processo de classificação. Esta abordagem visual proporciona uma análise intuitiva, ampliando o potencial prático da inteligência artificial para agricultores e especialistas.

Além disso, a aplicação de ferramentas como Grad-CAM reforça a importância da aplicabilidade para validação das decisões em sistemas baseados em redes convolucionais (Barbosa, 2023). A análise apresentada neste trabalho destaca que, embora o caminho entre a análise da imagem pelo modelo, e o resultado final seja complexo à primeira vista, é possível compreender suas decisões e, assim, aumentar a confiança na utilização da inteligência artificial na

agricultura de precisão. Este avanço é essencial para superar a tradicional visão de "caixa-preta" associada a esses modelos, tornando-os mais acessíveis e éticos em aplicações práticas.

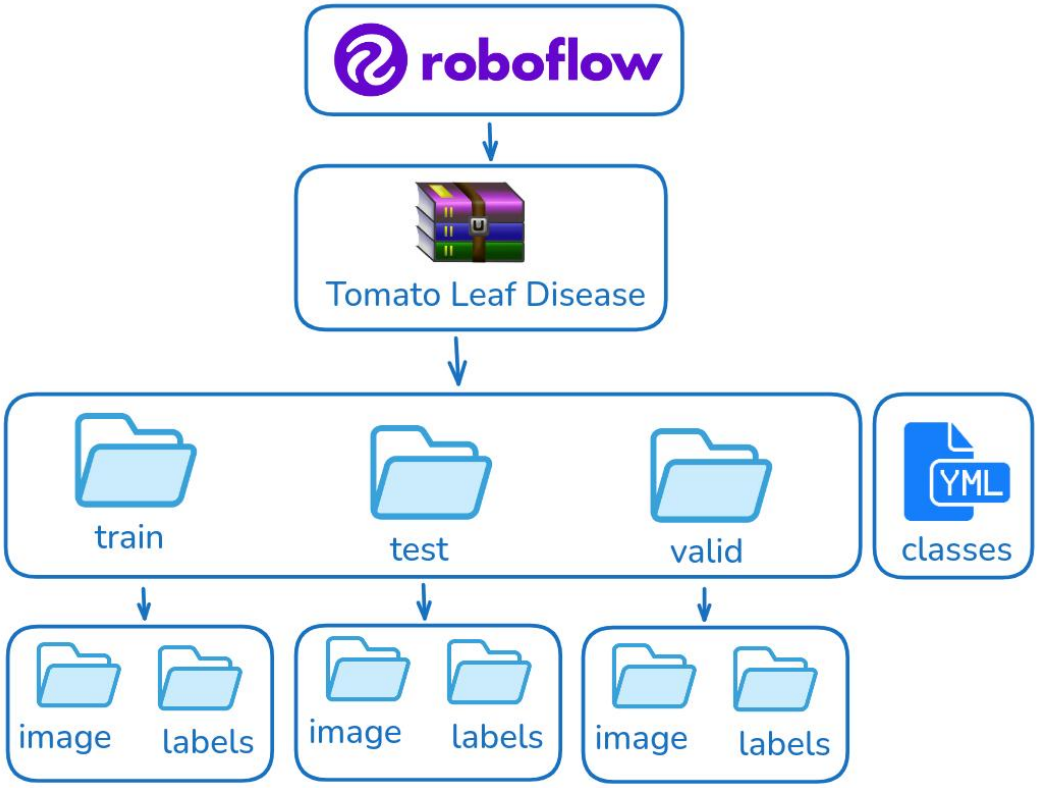
2 MATERIAL E MÉTODOS

O conjunto de dados utilizado foi o *Tomato Leaf Disease Dataset,* obtido do repositório do site Roboflow (Roboflow, 2024), de Bryan, nomeado *Tomato Leaf Disease Dataset*, Roboflow Universe, 2023. Contendo 9 mil imagens de folhas de tomateiro infectadas por diversas doenças, servindo como base de estudo e para o treinamento e validação do modelo de classificação. Após fazer o *download* e extrair o conteúdo, foram obtidos 6 itens, sendo 4 deles utilizados no projeto e os outros 2 sendo arquivos de documentação.

Para a etapa de tratamento de dados e preparação do dataset, inicialmente foi estruturada a arquitetura do dataset original, conforme ilustrado na Figura 1. Em seguida, por meio da aplicação de um padrão de organização, os dados foram reorganizados de maneira que permitisse melhor visualização, como detalhado na Figura 2.

Além disso, foi implementada uma regra para utilizar apenas imagens com uma única anotação, conforme evidenciado na Figura 3. Por fim, a Figura 4 apresenta a configuração final do dataset, onde somente as classes com mais de 450 imagens foram selecionadas, garantindo maior representatividade no treinamento do modelo.

Figura 1 — Estrutura do Dataset para Detecção de Doenças em Folhas de Tomate.



Fonte: Elaborada pelos autores (2024).

Como é visto na figura 1, o diagrama ilustra a estrutura de organização do conjunto de dados usado para a detecção de doenças em folhas de tomate. O dataset é dividido em três subconjuntos principais: treino, validação e teste, cada subconjunto contém diretórios separados para imagens e rótulos. Um arquivo de configuração no formato YML contém as classes presentes no dataset.

Figura 2 — Relação entre Rótulos e Classes no Dataset.

Fonte: Elaborada pelos autores (2024).

A figura 2 ilustra o funcionamento do arquivo de rótulos (labels), associado a cada imagem. A classe é definida pela primeira letra do rótulo e mapeada diretamente a uma das categorias listadas no names como por exemplo: 1 refere-se a EarlyBlight.

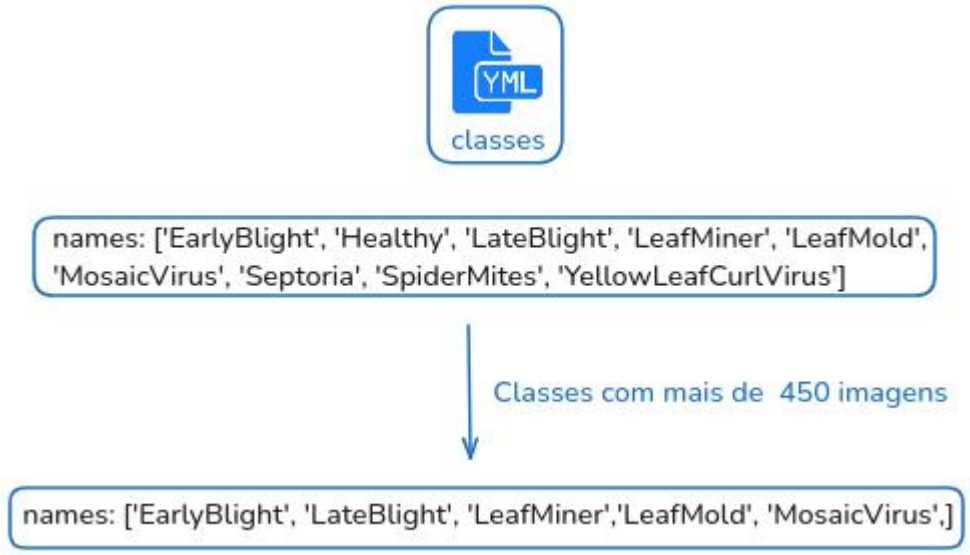
Figura 3 — Seleção de Imagens com Apenas uma Anotação.

Fonte: Elaborada pelos autores (2024).

Ilustrado na figura 3, foi realizado um processo de filtragem no dataset que reduziu os dados pela metade, cerca de 4 mil, mantendo somente imagens com uma única linha de anotação, pois dela se define a classe.

Utilizando-se desse padrão, foi criado o script que faz essa filtragem por classe, cria as pastas com o nome de cada classe e as copiam.

Figura 4 — Seleção de Classes com Base no Número de Imagens.

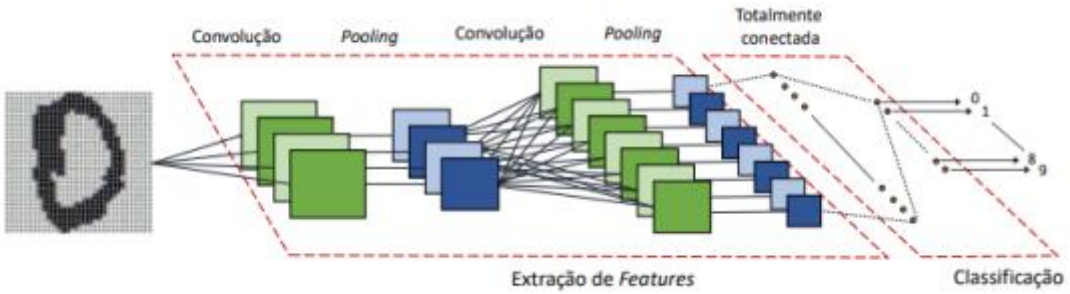


Fonte: Elaborada pelos autores (2024).

A figura 4 demonstra o processo de um novo requisito criado, que foi a de apenas utilizar classes com mais de 450 imagens, essa restrição foi criada para reduzir a grande variabilidade entre as classes, que apresentavam quantidades de imagens variando de 290 a 970.

O treinamento e avaliação foram realizados em um computador com processador Intel i5 de 11ª geração, 16 GB de RAM e gráficos integrados, utilizando Python 3.11.6 (Python, 2024) como linguagem principal. O modelo foi implementado na biblioteca Keras (Keras, 2024), adotando a arquitetura pré-treinada VGG 16 (Keras, 2024), como ilustrado na figura 5, ajustada via transfer learning para potencializar a classificação de doenças. A arquitetura opera com imagens no formato RGB de 224x224 pixels, utilizando filtros convolucionais 3x3 e a função de ativação ReLU para detecção de padrões relevantes.

Figura 5 — Esquema da Arquitetura *VGG16* para Classificação de Imagens com Camadas de Convolução, *Pooling e Softmax*



Fonte: Becker, 2017.

Na imagem é possível acompanhar a arquitetura do *VGG16*, e seu funcionamento. O nome *VGG16* vem da profundidade de sua arquitetura, que como o nome indica, contém 16 camadas de treinamento, compostas por 13 camadas de convolução e 3 camadas conectadas, conta também com camadas de *max pooling*, camadas utilizadas para reduzir a dimensionalidade das imagens, e manter as características mais importantes, e uma camada de *softmax,* camada que transforma os valores de saída da rede neural em probabilidades, para a classificação final (Tammina, 2019).

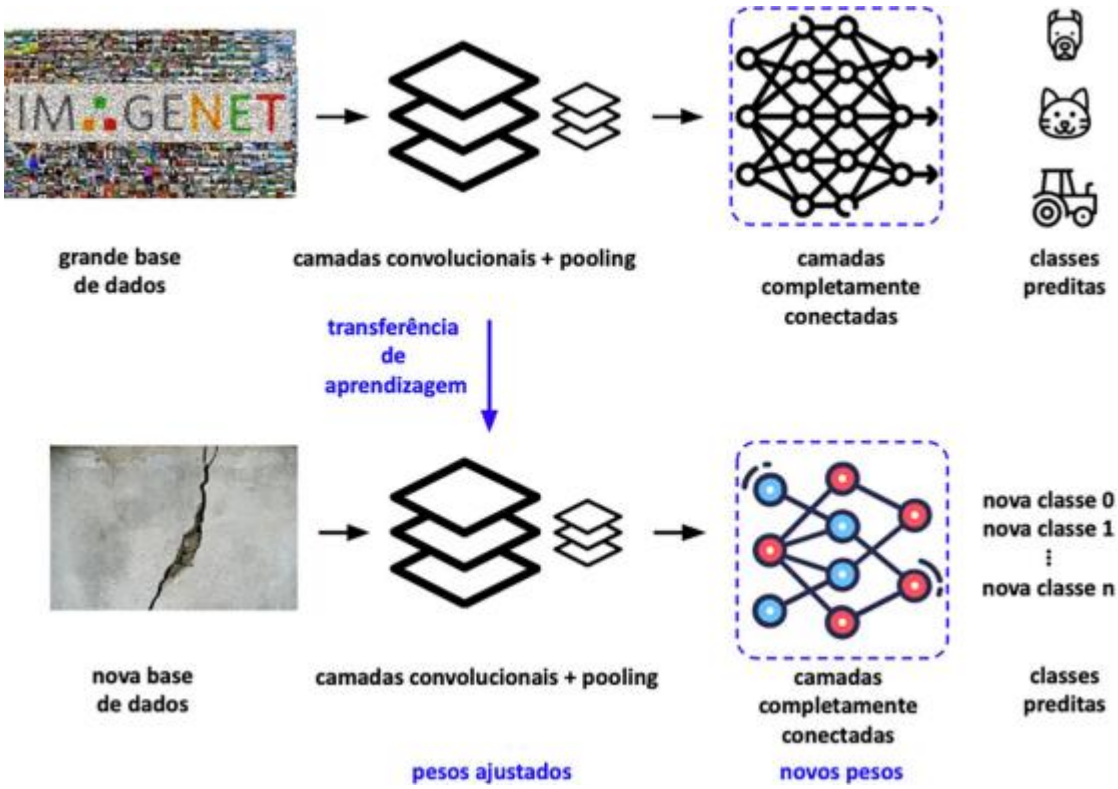
O modelo recebe imagens de entrada com um tamanho fixo de 224 pixels de largura por 224 pixels de altura, e com três canais de cor, que correspondem ao vermelho, verde e azul (canais *RGB*, que são as cores básicas usadas para formar imagens coloridas). Em cada camada de processamento (chamada camada convolucional), são utilizados pequenos blocos chamados "filtros" de tamanho 3x3 pixels, que percorrem a imagem e ajudam a identificar características importantes, como bordas ou texturas (Mody; Mathew; Jagannathan, 2019).

Cada camada utiliza uma função matemática chamada *ReLU (Rectified Linear Unit*), que ajuda o modelo a aprender padrões mais complexos. Essa função transforma os valores negativos em zero permitindo o modelo lidar melhor com diferentes tipos de informações e capturar padrões que não são simplesmente lineares (como curvas ou formas mais complexas) (Bhagat; Kumar, 2023).

Ao utilizar essa arquitetura, torna-se possível a utilização de uma técnica chamada transfer learning (ilustrada na Figura 6), que se resume a reutilizar o conhecimento já

adquirido de uma tarefa, para potencializar o desempenho de outra tarefa relacionada (Lysdahlgaard, 2023).

Figura 6 — Ilustração da Transferência de Aprendizado: Reaproveitamento de Pesos Treinados em Grandes Bases de Dados para Novas Classes.



Fonte: Pereira et al., 2024.

Na Figura 6, é representado o funcionamento do conceito de *Transfer Learning* (Aprendizado por Transferência), um conceito bastante difundido, frequentemente utilizado quando existe um modelo já pré-treinado, em uma tarefa e que tem a possibilidade de ser reutilizado para resolver uma nova tarefa, diminuindo a necessidade de treinar um modelo novo (Iman;Arabnia;Rasheed, 2023).

O pré-processamento das imagens incluiu aumento de dados com rotação, zoom e espelhamentos horizontais, aplicados com a classe ImageDataGenerator do Keras. Para manipulação e visualização, foram usadas as bibliotecas OpenCV

(OpenCV, 2024) e PIL (PIL, 2024), responsáveis por ajustes como redimensionamento e conversões de formato.

A explicabilidade do modelo foi aprimorada com a aplicação da técnica Grad-CAM, que destacou regiões da imagem mais relevantes para a classificação, contribuindo para identificar os padrões associados às doenças foliares. Os dados e operações foram gerenciados com NumPy (NumPy, 2024), enquanto os módulos os e glob automatizaram o carregamento das imagens.

3 RESULTADOS E DISCUSSÃO

Os resultados basearam-se na acurácia de predição (70% nos dados de teste) e na análise visual dos mapas de calor gerados pelo Grad-CAM. Avaliou-se se os pontos de maior intensidade concentravam-se na folha ou dispersaram além dela. Foram analisadas 48 imagens de teste, categorizando a concentração em três focos: folha, ambos e plano de fundo. O estudo destaca a utilidade dos mapas de calor para verificar se o modelo considera efetivamente a região da folha, com base nesses parâmetros principais (Aquino, Costa, Filho, 2022).

Após realizar o treino, foi obtido uma acurácia de 98%, foi observado uma variação de 0.6 a 0.9 na fidelidade por classe, o que resulta em uma média de acurácia final aproximada de 70% no conjunto de validação, o que indica uma boa capacidade em generalizar ao predizer novas imagens. Além disso, foi revelado pela acurácia, a dificuldade que certas classes apresentaram na classificação, como é o caso de requeima tardia e mofo nas folhas, indicando que as características visuais dessas classes provavelmente são mais complexas, devido a uma possível mescla entre a folha e o plano de fundo, que foi observada dentro dos dados de treino.

Em comparação com o trabalho de Pereira Júnior et al. (2024), que obteve uma acurácia de 99,27%, utilizando de *VGG16, VGG19* e *ResNet50* com *transfer learning* por meio de ajuste fino, e utilizando um banco de dados de 40.000 imagens. Pode se levar em consideração que a grande diferença na acurácia pode ser atribuída à variabilidade nos datasets e a discrepância de quantidade de imagens,

considerando que há uma diferença de 20 vezes no número de amostras utilizadas, assim escalando as capacidades do modelo poder aprender mais.

Após a aplicação do método Grad-CAM nas cinco classes e análise dos resultados, foi possível estruturar os dados em formato de tabela, como apresentado na Figura 7, contendo a distribuição das 48 imagens para cada classe. Observou-se que duas classes apresentaram maior enfoque na folha: Requeima Precoce, ilustrada na Figura 8, e Minador de Folhas, representada na Figura 9.

A classe Requeima Precoce demonstrou predominância no uso do plano de fundo nas análises, enquanto a classe Vírus do Mosaico, também evidenciada na Figura 11, apresentou maior foco no plano de fundo. Por fim, duas classes, Requeima Tardia e Mofo nas Folhas, mostraram distribuição entre folha e plano de fundo, conforme ilustrado na Figura 10. Esses resultados foram extraídos com base na interpretação dos heatmaps gerados, seguindo as diretrizes de Samek et al. (2019).

Figura 7 — Distribuição das imagens do conjunto de dados em relação ao foco das áreas de interesse para cada classe.

Classe	Quantidade de Imagens	Enfoque na Folha	Enfoque em ambas	Enfoque no Plano de Fundo
Requeima Precoce	48	46 (95.83%)	1 (2.08%)	2 (4.17%)
Requeima Tardia	48	19 (39.58%)	23 (47.92%)	6 (12.5%)
Minador de Folhas	48	44 (91.67%)	4 (8.33%)	0 (0.0%)
Mofo nas Folhas	48	12 (25.0%)	36 (75.0%)	0 (0.0%)
Vírus do Mosaico	48	0 (0.0%)	5 (10.42%)	43 (89.58%)

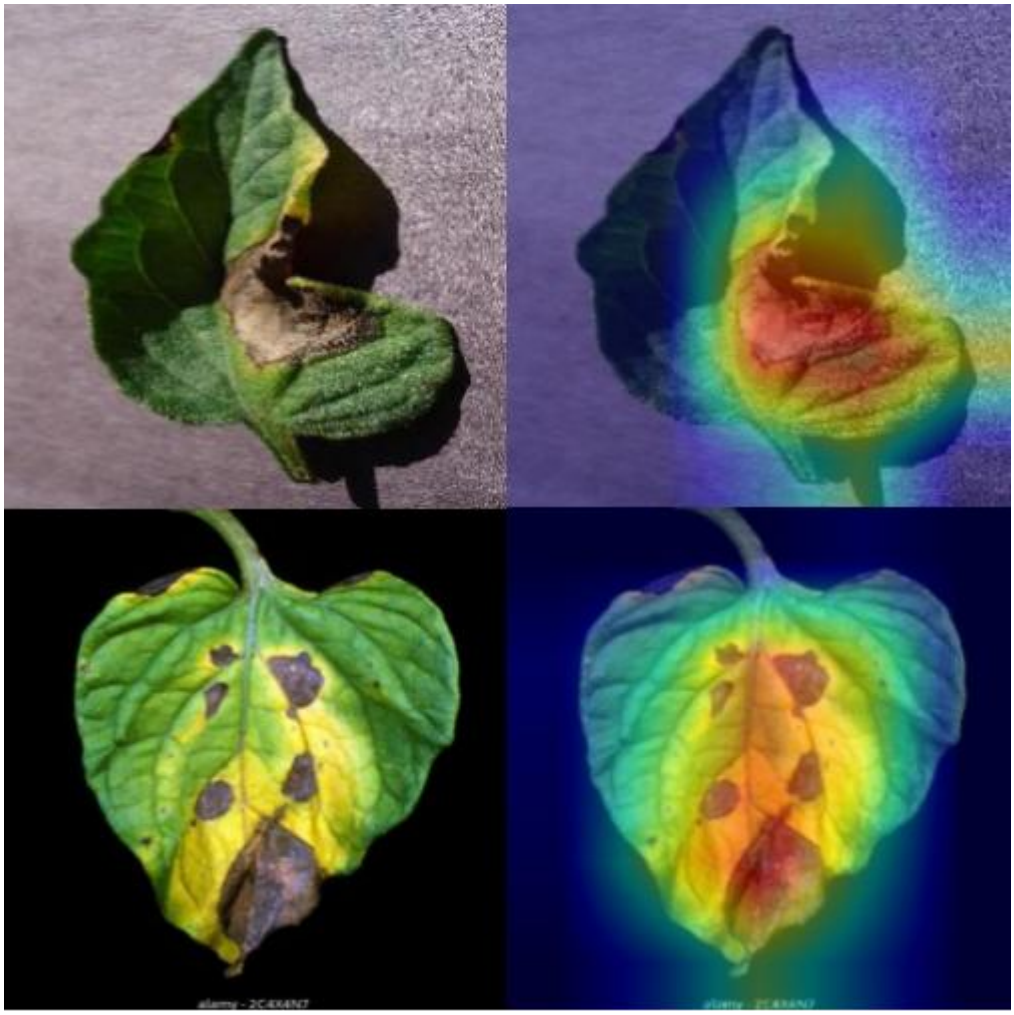
Fonte: Elaborada pelos autores (2024).

Com base nesta figura, é possível observar a divisão do número de imagens por classe e suas respectivas áreas de ativação. Cada uma das 48 imagens de cada classe foi categorizada entre "enfoque na folha", "enfoque em ambas" (folha e plano de fundo) e "enfoque no plano de fundo".

Os resultados destacam os padrões detectados por meio dos mapas de calor gerados pelo Grad-CAM, permitindo uma análise visual detalhada das áreas de maior relevância para a classificação, conforme ilustrado nas figuras subsequentes.

Figura 8 — Heatmap na folha de tomate com Requeima Precoce (Early Blight):

A Requeima Precoce, causada pelo fungo Alternaria solani, caracteriza-se pela formação de manchas escuras e circulares com bordas amareladas nas folhas.



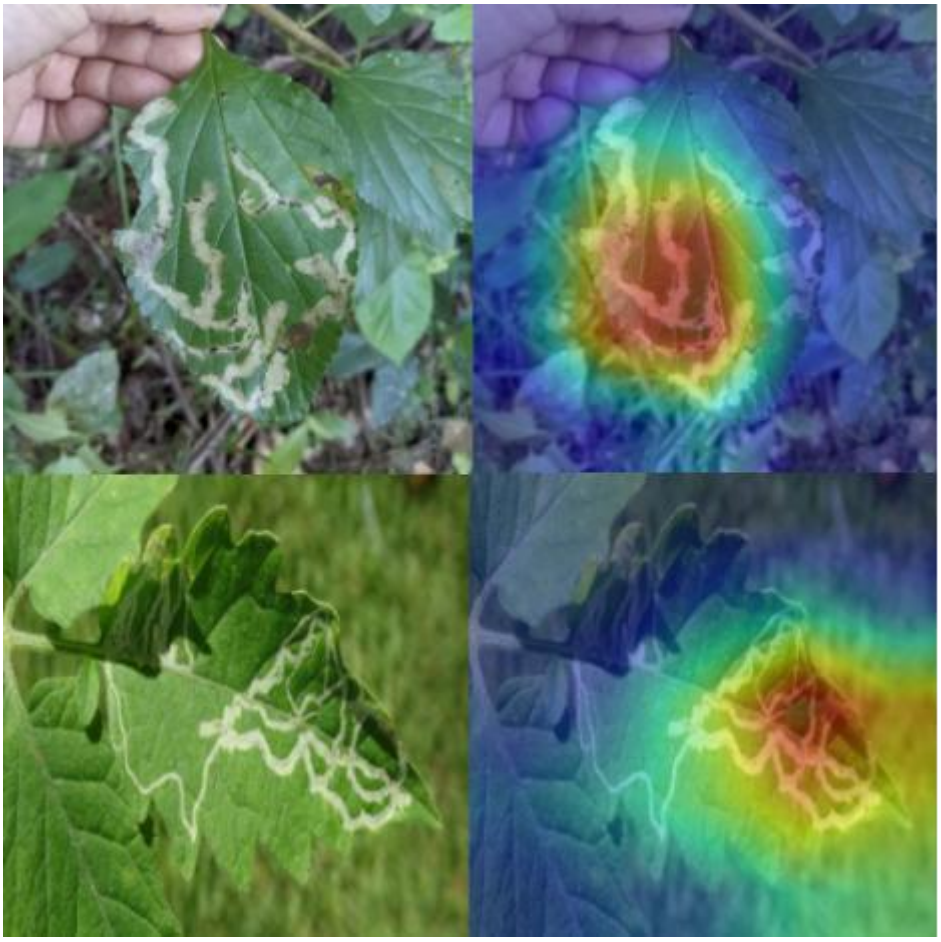
Fonte: Elaborada pelos autores (2024).

Na Figura 8, é possível observar claramente os tons mais quentes (indicando maior ativação do modelo) concentrados nas áreas da folha, enquanto os tons mais frios estão ao redor, no plano de fundo.

Essa distribuição indica que o modelo baseou sua classificação principalmente nas características da folha afetada, priorizando as manchas e bordas visíveis como critérios decisivos.

Figura 9 — Heatmap na folha de tomate com Minador de Folhas (LeafMiner):

Os Minadores de Folhas são insetos que escavam túneis nas folhas, deixando marcas serpenteantes características que prejudicam a capacidade de fotossíntese da planta.



Fonte: Elaborada pelos autores (2024).

Na Figura 9, observa-se que, apesar da complexidade do plano de fundo, onde as colorações podem se confundir com as da folha, o modelo conseguiu priorizar as regiões afetadas pelas escavações dos túneis.

O heatmap destaca os tons mais quentes sobre essas áreas serpenteantes, indicando que o modelo identificou com precisão as marcas específicas associadas à ação dos minadores.

Figura 10 — Heatmap na folha de tomate com Requeima Tardia (Late Blight):

A Requeima Tardia, causada pelo patógeno Phytophthora infestans, é uma doença severa que provoca manchas escuras e úmidas nas folhas, caules e frutos, levando ao rápido apodrecimento

Fonte: Elaborada pelos autores (2024).

Na Figura 10, observa-se que o modelo apresentou maior dificuldade em focar exclusivamente nas regiões mais relevantes para essa classe.

O heatmap destaca áreas de ativação com tons quentes não apenas na folha, mas também em partes do plano de fundo, o que evidencia uma menor precisão na identificação dos padrões específicos da doença.

Figura 11 — Heatmap na folha de tomate com Vírus do Mosaico (Mosaic Virus):

O Vírus do Mosaico causa padrões de manchas amarelas e verdes claras nas folhas, deformações no crescimento e redução na produtividade da planta, sendo transmitido por insetos.

Fonte: Elaborada pelos autores (2024).

Na Figura 11, observa-se que o modelo priorizou o plano de fundo na classificação, como evidenciado pelos tons mais quentes concentrados fora das folhas, enquanto os tons mais frios sobrepõem as folhas.

Essa distribuição sugere que o modelo pode estar captando padrões no plano de fundo que, de alguma forma, contribuem para a classificação desta classe, o que pode indicar um possível viés no conjunto de dados ou características não intencionais associadas ao plano de fundo.

Nota-se que 40% dos heatmaps gerados pelo Grad-CAM concentraram-se em áreas com características coerentes às classes corretas, enquanto outros 40% apresentaram relevância distribuída entre folha e plano de fundo. Em 20% das amostras, as ativações ocorreram em regiões irrelevantes, como o fundo da imagem, indicando possíveis desvios no foco do modelo (Kamakshi; Krishnan, 2023).

Os resultados obtidos com os heatmaps permitiram uma análise detalhada do comportamento do modelo no estudo, evidenciando como ele utilizou as características das imagens para realizar as classificações. Nas classes Requeima Precoce e Minador de Folhas, o modelo mostrou bom desempenho ao focar nas regiões relevantes, como manchas e marcas nas folhas. Contudo, em classes como Requeima Tardia e Vírus do Mosaico, identificou-se uma maior dependência de elementos do plano de fundo, o que sugere que o modelo, em alguns casos, se apoiou em informações não específicas da classe. Esses achados destacam a importância de refinar o conjunto de dados e ajustar os parâmetros do modelo para garantir que as regiões relevantes, como folhas afetadas, sejam consistentemente priorizadas.

4 CONSIDERAÇÕES FINAIS

Com base nos resultados obtidos a partir dos mapas de calor, pode-se concluir que o modelo desenvolvido utilizando o VGG16 se mostrou adequado para a tarefa de classificação, apresentando desempenho satisfatório. A aplicação da técnica Grad-

CAM proporcionou uma camada de explicação visual, permitindo identificar de forma clara as regiões e características que o modelo considerou para suas decisões.

Entretanto, algumas limitações foram observadas, como o uso de áreas irrelevantes, como o plano de fundo na classificação. Isso indica a necessidade de melhorias no pré-processamento dos dados, impactando diretamente o conjunto de dados utilizado para o treinamento do modelo. Além disso, o aprimoramento das tecnologias empregadas pode aumentar a precisão do modelo e oferecer insights mais precisos sobre os fatores utilizados na classificação. Técnicas adicionais, como o LIME (Local Interpretable Model-agnostic Explanations), podem ser exploradas no futuro para complementar a explicabilidade das predições e oferecer uma compreensão mais detalhada das decisões.

REFERÊNCIAS

Aquino, Costa, Filho, 2022: Aquino, Gustavo; Costa, Marly G.F.; Filho, Cicero F.F Costa. Explaining One-Dimensional Convolutional Models in Human Activity Recognition and Biometric Identification Tasks, 2022. Disponível em: https://www.mdpi.com/1424-8220/22/15/5644. Acesso em: 10 out. 2024.

Barbosa, 2023: Barbosa, R. Disponível em: https://w2files.solucaoatrio.net.br/atrio/upe-ppgec_upl/THESIS/272/ppgec_dissertacao_rafaela_20231114150642434.pdf. Acesso em: 08 dez. 2024.

BECKER, Willian Eduardo. Uma abordagem de redes neurais convolucionais para análise de sentimento multi-lingual, 2017. Disponível em: https://repositorio.pucrs.br/dspace/bitstream/10923/13276/1/000490690-Texto%2BCompleto-0.pdf. Acesso em: 15 out. 2024.

BHAGAT, Vikram; KUMAR, Sandeep. An efficient multimedia data retrieval method using deep learning and similarity matching. Multimedia Tools and Applications, v. 82, p. 25029–25052, 2023. Disponível em: https://doi.org/10.1007/s11042-023-17172-1. Acesso em: 4 jul. 2024.

Iman; Arabnia; Rasheed, 2023: Iman, Mohammadreza; Arabnia, Hamid Reza; Rasheed, Khaled. A Review of Deep Transfer Learning and Recent Advancements. Technologies, v. 11, n. 2, p. 145-156, 2023. Disponível em: https://www.mdpi.com/2227-7080/11/2/40. Acesso em: 18 out. 2024.

Kamakshi; Krishnan, 2023: Kamakshi, Vidhya; Krishnan, Narayanan C. Explainable Image Classification: The Journey So Far and the Road Ahead. v. 4, n. 3, p. 123-145, 2023. Disponível em: https://www.mdpi.com/2673-2688/4/3/33. Acesso em: 09 out. 2024.

LYSDAHLGAARD, Mads.Utilizing heat maps as explainable artificial intelligence for detecting abnormalities on wrist and elbow radiographs, v. 234, 2023. Disponível em: https://www.sciencedirect.com/science/article/pii/S1078817423001797. Acesso em: 4 jul. 2024.

Mersha et al., 2024: Mersha, D.; et al.Explainable Artificial Intelligence: A Survey of Needs, Techniques, Applications, and Future Direction. Disponível em: https://arxiv.org/abs/2409.00265v1. Acesso em: 08 dez. 2024.

Mody; Mathew; Jagannathan, 2019: Mody, Mihir; Mathew, Manu; Jagannathan, Shyam. Efficient Pre-Processor for CNN, v. 29, n. 19, p. 1-7, 2019. Disponível em: https://library.imaging.org/ei/articles/29/19/art00009. Acesso em: 19 out. 2024.

Natarajan; Nambiar, 2024: Natarajan, M.; Nambiar, M.Underwater SONAR Image Classification and Analysis using LIME-based Explainable Artificial Intelligence. Disponível em: https://arxiv.org/abs/2408.12837v2. Acesso em: 08 dez. 2024.

NUMPY. Numpy Documentation. NumPy, 2024. Disponível em: https://numpy.org/pt/ Acesso em: 22 out. 2024.

OPENCV. OpenCV Documentation. OpenCV, 2024. Disponível em: https://opencv.org/. Acesso em: 22 out. 2024.

Pereira Junior et al., 2024: Malaquias Pereira Junior, Wanderlei et al. Cracks detection in images of concrete structures using deep neural networks. Revista Matéria, 2024. Disponível em: https://www.scielo.br/j/rmat/a/NxHkMFgFY9WvjnCXCGgYrzj/?lang=pt#. Acesso em: 16 out. 2024.

PIL. PIL Documentation. PIL, 2024. Disponível em: https://pillow.readthedocs.io/en/stable/. Acesso em: 22 out. 2024.

PYTHON. Python Documentation. Python, 2024. Disponível em: https://www.python.org/. Acesso em: 22 out. 2024.

ROBOFLOW. Roboflow, 2024. Disponivel em: https://roboflow.com/
Acesso em: 22 out. 2024.

TAMMINA, S. *Transfer learning using VGG-16 with Deep Convolutional Neural Network for classifying images*. Disponível em: https://www.ijsrp.org/research-paper-1019/ijsrp-p9420.pdf. Acesso em: 4 jul. 2024.

Viana, 2023: Viana, A. Disponível em: https://monografias.ufop.br/bitstream/35400000/6074/8/MONOGRAFIA_Diagn%C3%B3sticoDoen%C3%A7asTomateiros.pdf. Acesso em: 08 dez. 2024.