Interpretable Deep Learning for Banana Leaf Disease Detection Using Dense Convolutional Models

Received 06/30/2025
Review began 07/21/2025
Review ended 08/06/2025
Published 08/26/2025

© Copyright 2025
AR et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI:
https://doi.org/10.7759/s44389-025-08446-w

Bharath AR [1] , Hemalatha N [1] , Sreekumar KM [2]

1. School of Information Science and Technology, St Aloysius Deemed to be University, Mangalore, IND 2. Department of Entomology, College of Agriculture, Padanakkad, IND

Corresponding author: Hemalatha N, hemasree71@gmail.com

Abstract

Banana cultivation often suffers from losses due to leaf diseases with similar visual symptoms, making timely and accurate classification difficult. This study applied a deep learning-based solution using the DenseNet201 architecture to classify four banana leaf conditions: Cordana, Sigatoka, Pestalotiopsis, and healthy. A total of 2,538 images were used, divided into augmented sets for training and validation, and a separate set of unseen original images for testing. Two model variants were developed: one using Global Average Pooling (GAP) and another using feature flattening. The GAP-based model achieved a test accuracy of 96.05%, while the version without GAP reached 97.23%. Despite the slight difference in accuracy, the GAP model demonstrated improved interpretability. Gradient-weighted Class Activation Mapping visualizations were used to observe class activation regions and understand the model's focus during prediction. These explanations revealed that the GAP model attended to more relevant lesion areas, especially in cases where visual similarity between disease types caused confusion. The analysis of misclassifications and confusion matrices supported this observation. The results showed that integrating GAP helps the model generalize better and produce more transparent decisions. This approach strengthens the trustworthiness of deep learning models for agricultural disease diagnosis, particularly when applied to real-world, visually complex data.

Categories: AI applications, Computer Vision, Deep Learning
Keywords: deep learning, banana leaf disease, grad-cam, explainable ai, agricultural applications, dense convolutional networks, computer vision

Introduction

Banana (*Musa* spp.) is a staple food and commercial crop in tropical and subtropical regions, providing both nutritional value and economic support for millions of people. However, banana production is frequently threatened by leaf diseases such as Sigatoka, Cordana, and Pestalotiopsis. These diseases impair photosynthesis and lead to premature leaf death, which ultimately affects fruit yield and quality. Manual disease identification remains the common practice, but it is time-consuming and often inconsistent, especially when symptoms overlap visually across disease types [1,2].

In recent years, convolutional neural networks (CNNs) have demonstrated strong performance in plant disease classification using image data. Models such as AlexNet, ResNet, and DenseNet have achieved high accuracy across various crops, including tomatoes, apples, and citrus [3-5]. DenseNet201, in particular, is well suited for complex image classification tasks due to its dense connectivity and efficient gradient flow [6].

A recent study applied DenseNet201 to the BananaLSD dataset to classify banana leaf diseases and reported a classification accuracy of 98.12 percent [7]. While the result is promising, that work did not assess the model's ability to generalize to completely unseen data. It also lacked a mechanism to interpret model predictions, which is essential for trust and transparency in automated systems.

Explainable AI techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) have become valuable tools in plant disease diagnostics. These methods visually highlight the regions in an image that contribute most to the model's decision, providing insight into the learned features [8,9]. Grad-CAM has also been widely used in medical diagnostics, supporting its relevance in fields where interpretability is essential [10].

Another concern in previous research is the limited focus on generalization performance. Models are often evaluated using augmented or validation data derived from the same distribution as training samples. In contrast, testing on an original, non-augmented dataset helps verify real-world reliability [11].

This study presents a comparative evaluation of two DenseNet201-based architectures for banana leaf disease classification. One model includes Global Average Pooling (GAP) before the final classification layer, while the other uses a flattening operation. Both models are trained on an augmented dataset and tested on a separate set of original images. Grad-CAM is applied to visualize the model's attention during prediction and evaluate interpretability.

This work makes three main contributions. First, it compares the classification performance of DenseNet201 using GAP and feature flattening, allowing for an architectural analysis of both accuracy and interpretability. Second, it integrates Grad-CAM visualizations to highlight regions in the leaf images that influenced the model's decisions, offering transparency in the classification process. Finally, it evaluates the models using a distinct, non-augmented test set to assess generalization performance under real-world conditions.

Materials And Methods

Dataset description

This study uses the publicly available BananaLSD dataset [12], which consists of labeled images of banana leaves categorized into four classes: Cordana, Sigatoka, Pestalotiopsis, and Healthy. The dataset includes a total of 2,538 images, captured under varied lighting conditions and leaf orientations. To enhance model robustness, data augmentation techniques such as horizontal flipping, zooming, brightness adjustment, and rotation were applied. The augmented dataset was divided into training and validation sets using an 80:20 ratio. Additionally, a separate, non-augmented set of original images was reserved exclusively for testing to evaluate the model's generalization ability under real-world conditions.

To enhance transparency and reproducibility, Table *1* provides a detailed breakdown of image counts across classes and dataset splits. While the training and validation sets are nearly balanced, the test set shows a natural skew toward the Sigatoka class, reflecting its prevalence in the real-world dataset.

Class	Training (Augmented)	Validation (Augmented)	Test (Original)
Cordana (0)	321	80	162
Healthy (1)	320	80	129
Pestalotiopsis (2)	320	80	173
Sigatoka (3)	320	80	473
Total	1,281	320	937

TABLE 1: Class-wise image distribution

Model architecture and training setup

The DenseNet201 architecture was selected as the base model due to its proven performance in visual recognition tasks, supported by densely connected convolutional blocks that enable efficient feature reuse and gradient flow [6]. Initially, the model was implemented without GAP, using a flattening operation to convert the final feature map into a vector before passing it to a dense softmax classification layer. This configuration served as the baseline for evaluating classification performance.

To address the limitations observed in interpretability and generalization, a second variant was developed by replacing the flattening layer with a GAP operation. The use of GAP allowed spatial information to be preserved in a condensed form, enabling the model to retain focus on relevant image regions. This architectural change also proved more compatible with the Grad-CAM method, producing clearer and more localized activation maps. Based on qualitative and quantitative analysis, the GAP-based model was finalized for further evaluation due to its better alignment with real-world deployment goals, including transparency and consistency on unseen data.

Both model variants were initialized with pre-trained ImageNet weights to benefit from general-purpose visual features. Initially, the dense classifier was trainable, while deeper layers remained frozen. After early convergence on validation loss, the model was fine-tuned end-to-end.

Training was conducted using TensorFlow. The Adam optimizer was used with a default learning rate of 0.001. Each model was trained for up to 15 epochs, with early stopping enabled to prevent overfitting. A batch size of 32 was used for training and validation, while test set evaluation was performed using a batch size of 1 to enable detailed per-image Grad-CAM visualizations. All input images were resized to

224 × 224 pixels and normalized to a (0, 1) range, matching DenseNet201's expected input format.

Explainability via Grad-CAM

To assess interpretability, Grad-CAM was applied to both model variants [8]. Grad-CAM creates heatmaps that highlight the image regions most influential in a model's decision. These maps were generated for correctly and incorrectly classified samples from all four classes. Comparing the Grad-CAM outputs across both models allowed for an evaluation of how pooling strategies influence the attention mechanism and interpretability of predictions.

Results And Discussion

A comparative evaluation of two DenseNet201 architectures is presented here, analyzing classification performance, learning behavior, and interpretability. One model uses a GAP layer before classification, while the other applies feature flattening. Both were trained on an augmented dataset and tested on a separate, original dataset to reflect real-world conditions.

Classification performance

The classification performance is summarized in Table *2*. The DenseNet201 variant without GAP achieved a test accuracy of 97.23%, slightly higher than the GAP-based model, which reached 96.05%. However, the GAP-based model is more desirable in real-world applications, as it tends to regularize the network, reduce overfitting, and provide more focused feature attention.

Model Variant	Test Accuracy (%)	Peak Validation Accuracy (%)	Training Accuracy (%)
DenseNet201 (GAP)	96.05	98.44	99.84
DenseNet201 (No GAP)	97.23	97.81	99.84

TABLE 2: Model performance comparison

GAP, Global Average Pooling

These results can be compared with the findings of Ünal [7], where DenseNet201 was applied to the same BananaLSD dataset and an accuracy of 98.12% was reported. However, that study evaluated only on augmented validation data and did not assess generalization to an original, non-augmented test set. In addition, interpretability or architectural variations such as Global Average Pooling were not explored. In contrast, the current work integrates GAP and Grad-CAM to support explainability and generalization, offering a more deployment-ready framework despite a slightly lower accuracy. Prior studies have emphasized the importance of testing on independent distributions to avoid overfitting to augmented patterns [13].

Accuracy trends

The training behavior of both models is shown in Figure *1*, where the non-GAP variant displays slightly more stable validation performance during later epochs. Both models converged quickly, but the version without GAP maintained slightly higher validation accuracy during the middle and later training stages. However, this model also showed indications of overfitting, such as overly confident predictions on misclassified samples.

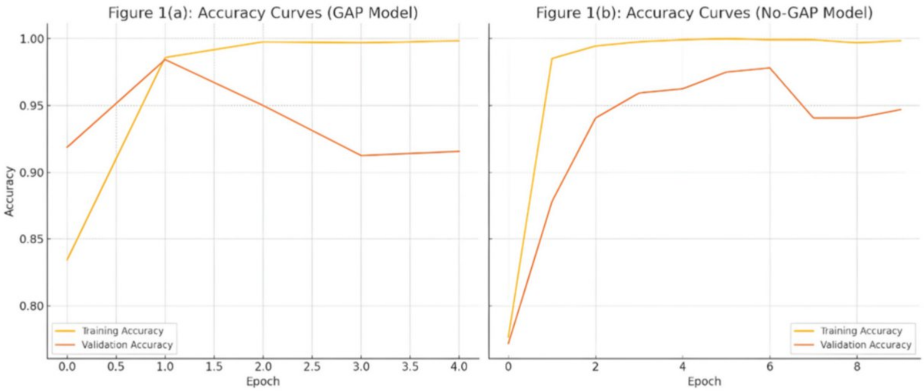

FIGURE 1: Training and validation accuracy curves across epochs for the DenseNet201 models: (a) with GAP and (b) without GAP

GAP, Global Average Pooling

The GAP model exhibited minor fluctuations in validation accuracy, which is often a sign of better generalization. This behavior is desirable when deploying models in variable conditions, such as real farms or diverse lighting environments.

Prediction breakdown by class

The confusion matrices in Figure 2 and Figure 3 show class-wise prediction performance.

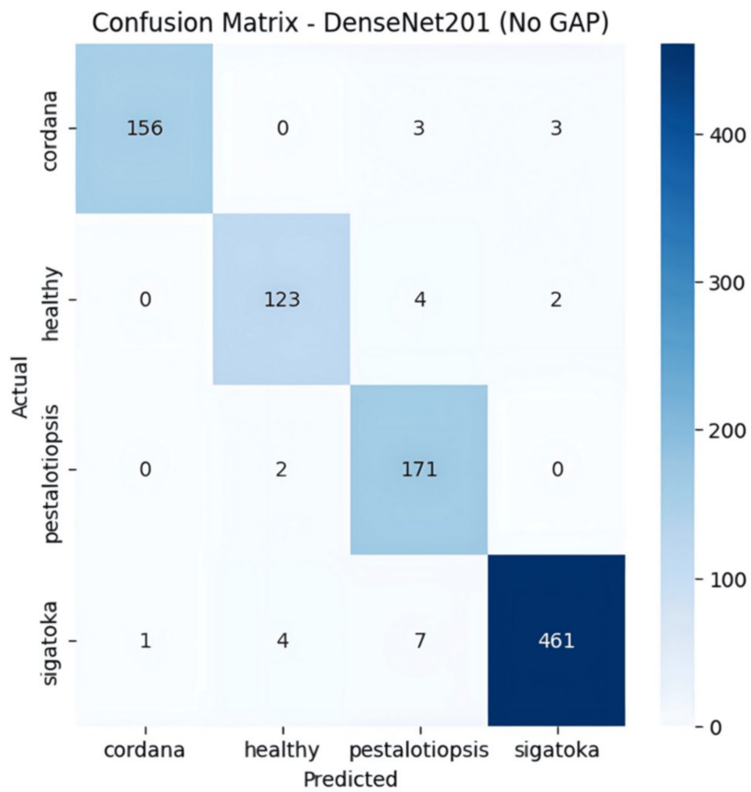

FIGURE 2: Confusion matrix of DenseNet201 model with GAP on the test dataset

GAP, Global Average Pooling

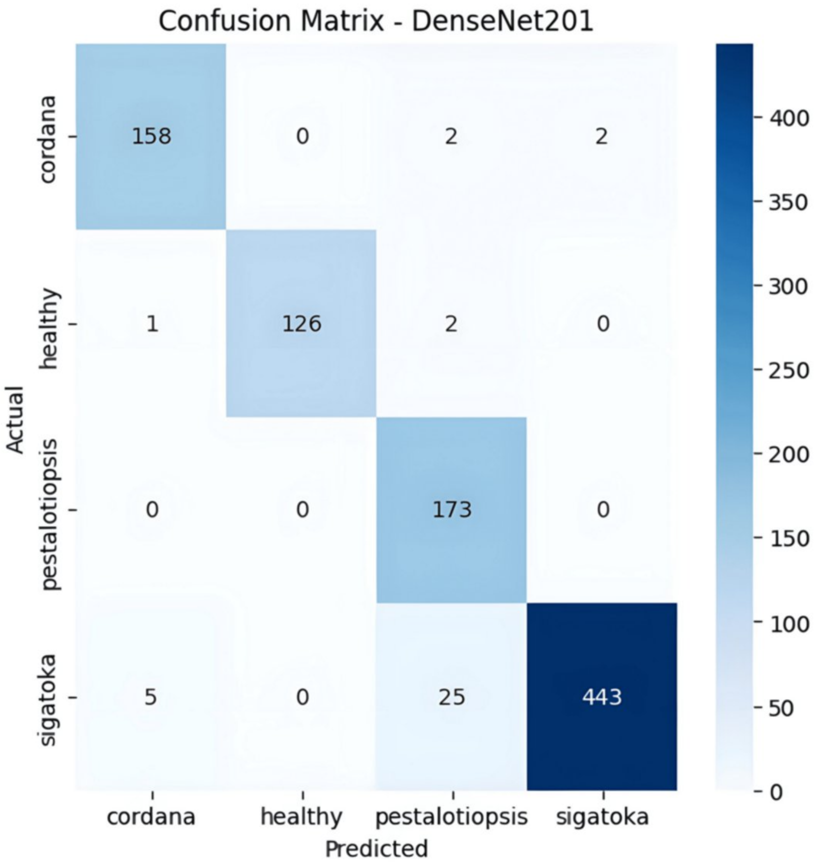

FIGURE 3: Confusion matrix of DenseNet201 model without GAP on the test dataset

GAP, Global Average Pooling

The GAP model achieved strong results across all categories but sometimes confused Cordana with Pestalotiopsis. The no-GAP model performed better in separating these classes but made more errors in distinguishing Healthy and Sigatoka leaves. To provide a detailed breakdown, Table *3* presents class-wise precision, recall, and F1-scores for both models.

Class	GAP Precision (%)	GAP Recall (%)	GAP F1-Score (%)	No-GAP Precision (%)	No-GAP Recall (%)	No-GAP F1-Score (%)
Cordana	95.92	94.00	94.95	97.96	96.00	96.97
Sigatoka	100.00	100.00	100.00	100.00	100.00	100.00
Pestalotiopsis	93.75	93.75	93.75	97.92	97.92	97.92
Healthy	96.08	98.00	97.03	98.04	100.00	99.01

TABLE 3: Class-wise metrics comparison

GAP, Global Average Pooling

These values demonstrate the class-wise robustness of each model. While both performed well, the GAP model displayed better balance in challenging scenarios, particularly where class similarities make distinctions subtle - a challenge previously observed in deep learning-based plant disease classification [14].

The evaluation metrics, viz. accuracy, precision, recall, and F1-score, provide a solid basis for comparison; this study does not include statistical significance testing such as confidence intervals or p-values. Given the relatively small difference in performance between the two models, future work could include hypothesis testing methods (e.g., McNemar's test or bootstrap confidence intervals) to assess whether observed differences are statistically significant. These analysis would further strengthen the reliability and robustness of model comparisons.

Visual attention and interpretability

Grad-CAM visualizations were employed to assess the interpretability of the DenseNet201 models. In correctly classified cases, the GAP-based model demonstrated precise focus on lesion-affected regions, indicating disease-specific learning. In contrast, the model without GAP often responded to broader or irrelevant regions of the leaf surface. Even in misclassified samples, the GAP model maintained attention on meaningful areas, whereas the no-GAP variant tended to misattribute focus. Attention-based analysis methods like Grad-CAM have also been validated in fields such as medical imaging and plant phenotyping, where trust and explainability are critical [15,16].

Representative Grad-CAM results are presented in Figure *4* and Figure *5*. These highlight the GAP model's ability to consistently localize lesion regions, including in incorrect predictions.

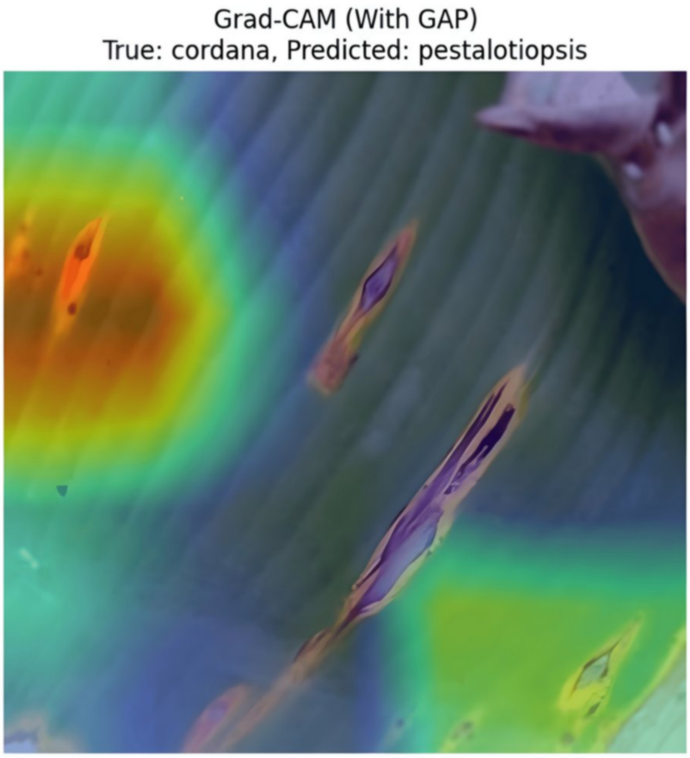

FIGURE 4: Grad-CAM visualization of GAP model for a misclassified Cordana sample predicted as Pestalotiopsis

GAP, Global Average Pooling; Grad-CAM, Gradient-weighted Class Activation Mapping

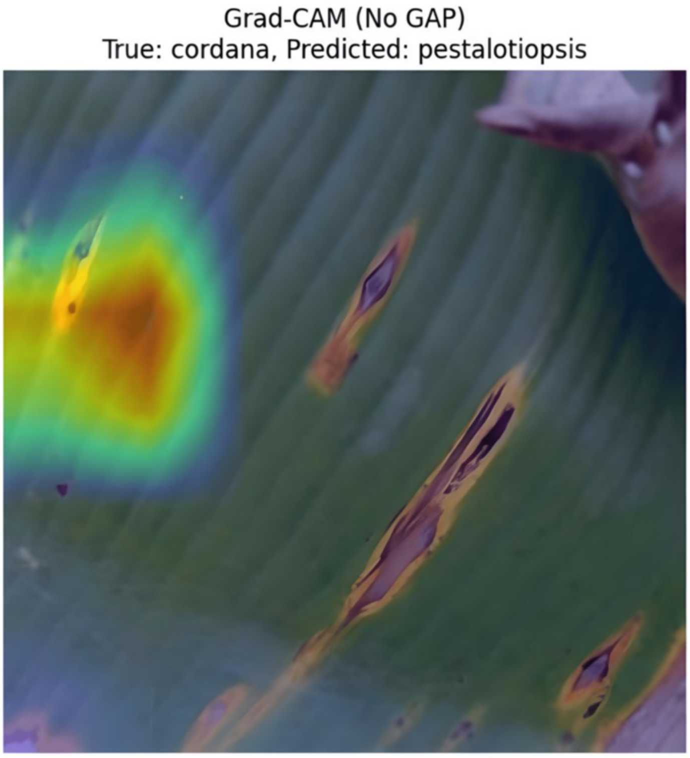

FIGURE 5: Grad-CAM visualization of no-GAP model for the misclassified Cordana sample predicted as Pestalotiopsis

GAP, Global Average Pooling; Grad-CAM, Gradient-weighted Class Activation Mapping

To visualize attention alignment in correctly predicted samples, Figure *6* displays an original image and its Grad-CAM heatmap, illustrating the model’s attention alignment with visible disease symptoms. To further validate attention consistency, additional Grad-CAM visualizations for correctly classified Cordana samples are shown in Figure *7*. These support the conclusion that the GAP model highlights semantically meaningful regions more reliably than its no-GAP counterpart [17].

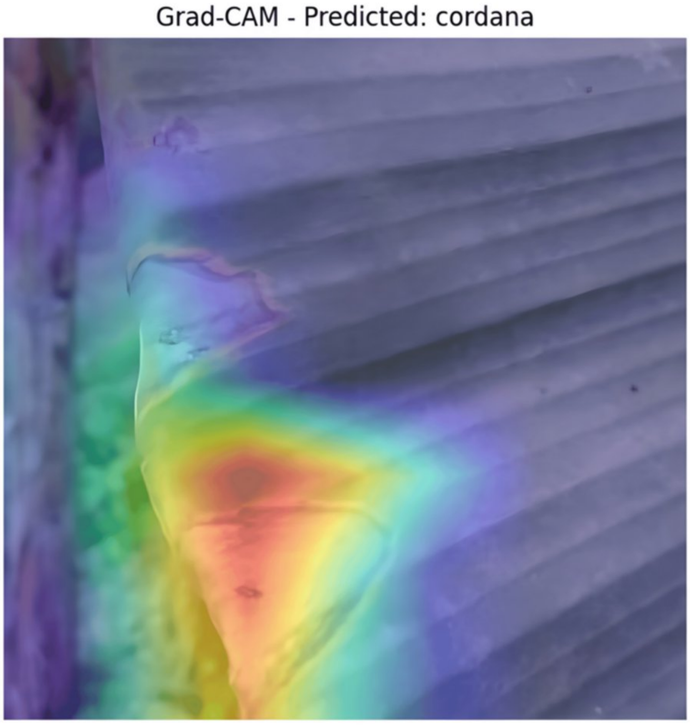

FIGURE 6: Grad-CAM visualization of a correctly classified Cordana sample

Grad-CAM, Gradient-weighted Class Activation Mapping

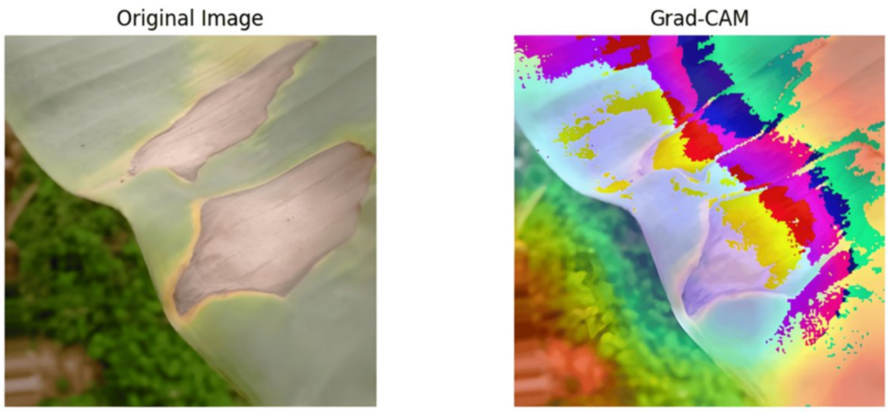

FIGURE 7: Original image (left) and its Grad-CAM output (right), showing lesion-localized attention

Grad-CAM, Gradient-weighted Class Activation Mapping

GAP interpretability advantage

While the no-GAP model achieved a slightly higher test accuracy, it did so at the cost of generalization and interpretability. GAP improved the model's ability to learn stable, spatially aware representations. This is especially important for real-world agricultural diagnostics, where unseen visual variations are common. The risk of overfitting in the no-GAP variant may reduce reliability when applied outside controlled environments.

Practical implications and future work

This comparison highlights that the architectural decision to include GAP not only supports interpretability but also improves generalization. These outcomes are essential for agricultural tools where trust and reliability influence adoption.

Future research should explore how these models perform across a wider variety of crops, disease types, and real-world conditions. Incorporating additional imaging modalities (such as thermal or hyperspectral data) [18], testing under variable lighting, and optimizing for edge-device deployment will further improve usability. Integrating uncertainty quantification alongside attention maps could enhance user confidence and mitigate incorrect actions in response to rare errors.

A limitation of this study is the lack of geographical and environmental metadata associated with the BananaLSD dataset, which was sourced from a publicly available online repository. As such, the dataset may not fully capture visual variations in leaf diseases caused by differing climates, cultivars, or field conditions. This limits the ability to generalize the model across regions. Future work should include datasets collected from multiple geographic zones and diverse growing environments to further validate the model's robustness and practical applicability in global agricultural settings.

Conclusions

This study presented a comparative analysis of two DenseNet201-based deep learning models for banana leaf disease classification, with a specific focus on model interpretability and generalization. One model utilized GAP, while the other relied on a traditional flattening approach. Both were trained on an augmented dataset and evaluated using a separate, original image set to reflect real-world variability. While the model without GAP achieved slightly higher test accuracy, the GAP-based architecture demonstrated more stable attention patterns and greater resistance to overfitting. Grad-CAM visualizations further revealed that the GAP model consistently focused on disease-relevant regions, offering clearer insight into the model’s decision-making process. This highlights the practical advantage of GAP, particularly in agricultural settings where transparency and trust are crucial. The results suggest that incorporating pooling strategies that promote interpretability does not significantly compromise classification performance and may, in fact, enhance the model's usability in field conditions. This balance between accuracy and explainability is essential for developing AI tools that are both effective and reliable in crop disease management.

Future work can extend this framework to other plant species and disease types, explore additional interpretability methods, and test the models under diverse environmental conditions and deployment constraints such as mobile or edge devices. The integration of interpretable deep learning into agricultural diagnostics holds strong potential to support timely and informed decision-making in precision farming.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Bharath AR, Hemalatha N, Sreekumar KM

Acquisition, analysis, or interpretation of data: Bharath AR

Drafting of the manuscript: Bharath AR

Critical review of the manuscript for important intellectual content: Bharath AR, Hemalatha N, Sreekumar KM

Supervision: Hemalatha N, Sreekumar KM

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

The authors acknowledge the availability of the BananaLSD dataset on Kaggle, which was used as the primary data source in this study: https://www.kaggle.com/datasets/armaanoajay/bananalsd

References

1. Bock CH, Poole GH, Parker PE, Gottwald TR: Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. Critical Reviews in Plant Sciences. 2010, 29:59-107. 10.1080/07352681003617285
2. Mahlein AK: Plant disease detection by imaging sensors—parallels and specific demands for precision agriculture and plant phenotyping. Plant Disease. 2016, 100:241-251. 10.1094/pdis-03-15-0340-fe
3. Mohanty SP, Hughes DP, Salathé M: Using deep learning for image-based plant disease detection. Frontiers in Plant Science. 2016, 7:1419. 10.3389/fpls.2016.01419
4. Too EC, Yujian L, Njuki S, Yingchun L: A comparative study of fine-tuning deep learning models for plant disease identification. Computers and Electronics in Agriculture. 2019, 161:272-279. 10.1016/j.compag.2018.03.032
5. Fuentes A, Yoon S, Kim SC, Park DS: A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. Sensors. 2017, 17:2022. 10.3390/s17092022
6. Huang G, Liu Z, van der Maaten L, Weinberger KQ: Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA. 2017, 2261-2269. 10.1109/CVPR.2017.243
7. Ünal C: Incorporating deep learning into the diagnosis of banana leaf spot diseases for the protection of banana crops. Food Science & Nutrition. 2024, 10:1-9. 10.24966/fsn-1076/100204
8. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D: Grad-CAM: Visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy. 2017, 618-626. 10.1109/ICCV.2017.74
9. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN: Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. IEEE Winter Conference on Applications of Computer Vision (WACV). 2018, 839-847. 10.1109/WACV.2018.00097
10. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S: Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017, 542:115-118. 10.1038/nature21056
11. Arman SE, Bhuiyan MAB, Abdullah HM, Islam S, Chowdhury TT, Hossain MA: Banana leaf spot diseases (BananaLSD) dataset for classification of banana leaf diseases using machine learning. Mendeley Data. 2023, 10.17632/9tb7k297ff.1
12. Kamilaris A, Prenafeta-Boldú FX: Deep learning in agriculture: A survey. Computers and Electronics in Agriculture. 2018, 147:70-90. 10.1016/j.compag.2018.02.016
13. Shin HC, Roth HR, Gao M, et al.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Transactions on Medical Imaging. 2016, 35:1285-1298. 10.1109/tmi.2016.2528162
14. Ferentinos KP: Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture. 2018, 145:311-318. 10.1016/j.compag.2018.01.009
15. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A: Learning deep features for discriminative localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016, 2921-2929. 10.1109/CVPR.2016.319
16. Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D: Machine learning in agriculture: A review. Sensors. 2018, 18:2674. 10.3390/s18082674
17. Barbedo JGA: Plant disease identification from individual lesions and spots using deep learning. Biosystems Engineering. 2019, 180:96-107. 10.1016/j.biosystemseng.2019.02.002
18. Sabzi S, Pourdarbani R, Rohban MH, García-Mateos G, Paliwal J, Molina-Martínez JM: Early detection of excess nitrogen consumption in cucumber plants using hyperspectral imaging based on hybrid neural networks and the imperialist competitive algorithm. Agronomy. 2021, 11:575. 10.3390/agronomy11030575