Enhancing Banana Leaf Disease Diagnosis Using Explainable AI on a Simple Convolutional Neural Network

Bharath A R[1], Hemalatha N[2], Sreekumar K M[3]
[1,2]Information Science and Technology St Aloysius (Deemed to be University) Mangaluru, India
[3]Agriculture, College of Agriculture College of Agriculture, Nileshwar Nileshwar, India

Abstract- While high-performance deep learning models have been applied to banana leaf disease detection [1][3], their interpretability remains underexplored. In this study, we deliberately use a baseline Convolutional Neural Network (CNN) with moderate accuracy to demonstrate how Explainable AI (XAI) techniques—such as Grad-CAM and SoftMax confidence analysis—can validate and interpret model predictions. We train a Global Average Pooling (GAP)-based CNN on the Banana LSD dataset [4] and observe a test accuracy of 74.7%. While more advanced models have reported higher performance [5], [6], our focus remains on interpretability and practical relevance. By integrating explainability techniques, we demonstrate that even a basic model can provide reliable support for disease diagnosis, especially in agricultural environments where transparency and resource efficiency are essential.

1. INTRODUCTION

Banana is an important crop in many tropical regions, but it’s vulnerable to leaf diseases that can seriously affect production [1].Convolutional neural networks (CNNs) and other deep learning methods have been used in plant disease detection because they’re good at picking up patterns in image data [2], [3].Still, one of the main problems with these models is that they don’t easily show how or why they make certain decisions, which can be a problem in areas where trust and understanding matter. In one study, Ashoka et al. [5] worked with a deeper model called EfficientNetB0 and got high accuracy, using Grad-CAM to show where the model was paying attention in the images. Even so, many studies tend to focus mainly on model performance, without paying much attention to how or why the model makes certain predictions [6], [7].To address this, our study explores how a simple CNN, when combined with explainable AI (XAI) techniques, can still provide meaningful interpretability—even if it does not match the performance of state-of-the-art networks.

2. RELATED WORK

Past research has explored the use of CNNs on the BananaLSD dataset [4]. Mohanty et al. [2] and Ferentinos [1], for example, demonstrated that these models can identify plant leaf diseases with notable accuracy. In a more recent study, Ashoka et al. [5] combined EfficientNetB0 with Grad-CAM and reached an accuracy of 99.22%, producing visual heatmaps that revealed the model’s attention regions.

Even with such advancements, the use of explainability tools is still uncommon in much of the literature. Aghav Palwe et al. [8] applied Grad-CAM to plant images, but their work didn’t focus on banana leaves specifically. On top of that, high-complexity models like EfficientNet often require hardware that's not readily available in many farming setups.

Our approach follows a different path rather than relying on complex architectures, we focus on a lightweight CNN model and use explainability techniques to better understand how it makes decisions. Our approach takes a different angle by focusing on a lightweight CNN and using explainability tools to understand its behavior. While many existing studies concentrate on high-performance models, simpler networks like ours are rarely explored in terms of their interpretability.

3. METHODOLOGY

3.1 Dataset the BananaLSD dataset [4] includes four categories: Healthy, Cordana, Pestalotiopsis, and Sigatoka. Each image was resized to 128×128 pixels and normalized before training. To deal with class imbalance and help the model learn more effectively, we used a few basic augmentations—like flipping, rotating, and zooming [9].

3.2 Baseline CNN Model The model we used had three convolutional layers. After each one, we added ReLU activation and a max-pooling step to reduce the feature map size. Toward the end, we included a Global Average Pooling (GAP) layer, which helped summarize the features without losing the main spatial cues. We took this approach partly from ideas mentioned by LeCun et al. [3], who highlighted the benefits of keeping models compact, especially for use on limited hardware.

3.3 Hybrid CNN + SVM To evaluate the viability of hybrid modeling, we first extracted high-dimensional features before the Global Average Pooling (GAP) layer and used them as input to an SVM classifier. As shown in Table 1, this configuration produced limited results, with a test accuracy of only 19.53% and a macro F1-score of 0.18. These metrics suggest that without end-to-end training, feature representations are not sufficiently discriminative for classical classifiers.

Table 1. Classification report using high-dimensional features extracted before the GAP layer.

Class	Precision	Recall	F1-score	Support
Cordana	0.11	0.19	0.14	162
Healthy	0.11	0.18	0.14	129
Pestalotiopsis	0.14	0.20	0.17	173
Sigatoka	0.47	0.20	0.28	473
Accuracy			0.1953	937
Macro Avg	0.21	0.19	0.18	937
Weighted Avg	0.30	0.20	0.21	937

We tried using the SVM with features taken after the GAP layer, but the outcome wasn't encouraging. Accuracy dropped to 5.34%, as shown in Table 2. It's possible that the feature set at that point didn't have enough detail for the classifier to make useful distinctions.

Table 2. Classification report using low-dimensional features extracted after the GAP layer.

Class	Precision	Recall	F1-score	Support
Cordana	0.04	0.02	0.03	162
Healthy	0.04	0.09	0.06	129
Pestalotiopsis	0.02	0.06	0.03	173
Sigatoka	0.16	0.05	0.08	473
Accuracy			0.0534	937
Macro Avg	0.07	0.06	0.05	937
Weighted Avg	0.10	0.05	0.06	937

3.4 Explainability Techniques We applied Grad-CAM [10] to visualize class-specific attention regions and used SoftMax confidence scores to quantify prediction certainty. These techniques helped reveal whether the model focuses on disease-relevant leaf regions or extraneous noise.

4. RESULTS

4.1 Performance The baseline model achieved 74.71% test accuracy on clean, non-augmented data. While lower than models like EfficientNet or ResNet [5], this result is consistent and suitable for applications requiring interpretability over peak performance.

4.2 Grad-CAM Visualization Looking at the Grad-CAM outputs, we noticed the model was often paying attention to parts of the leaf that had visible symptoms—like dark areas, spots, or damaged tissue. The model sometimes highlighted the correct regions, despite making an incorrect prediction. Figure 1 shows examples where the CNN's attention was concentrated in regions that are biologically relevant for identifying symptoms.

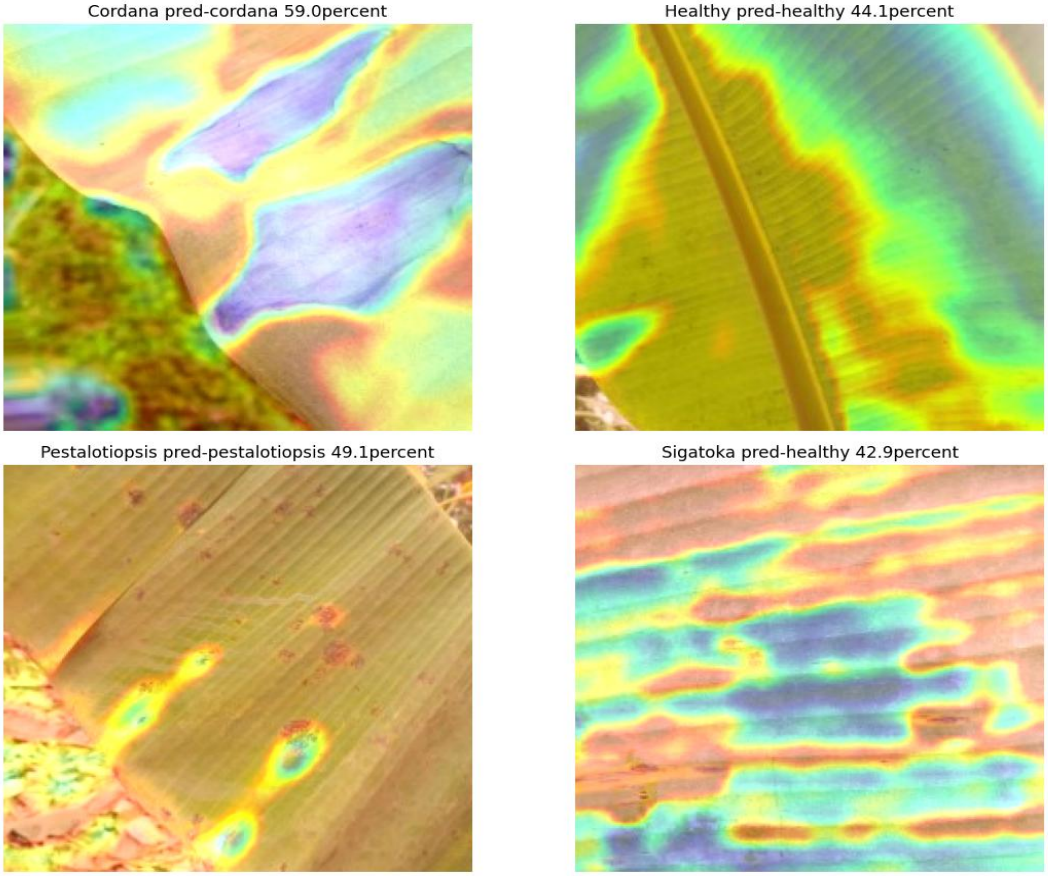
Cordana pred-cordana 59.0percent
Healthy pred-healthy 44.1percent
Pestalotiopsis pred-pestalotiopsis 49.1percent
Sigatoka pred-healthy 42.9percent

Figure 1. Grad-CAM heatmaps highlighting the CNN's attention on disease-relevant regions of banana leaves. This observation is consistent with the work by Aghav Palwe et al. [8], who also used XAI methods to study stress indicators in plants.

4.3 Confidence Analysis When we looked at predictions with low certainty, the SoftMax scores showed confusion between classes—particularly between Healthy and Pestalotiopsis. Figure 2, which presents the raw confusion matrix, reflects this pattern: the model often mixed up these two visually similar categories.

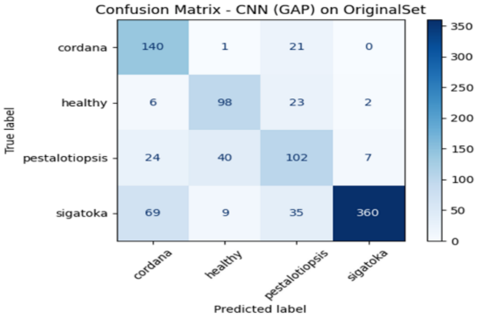

Figure 2. Raw confusion matrix showing prediction distribution across the four banana leaf disease classes

We used the normalized confusion matrix shown in Figure 3 to check how the model performed across the different classes. It showed that predictions for Cordana and Sigatoka were mostly correct. However, the model still made several mistakes when trying to separate Healthy from Pestalotiopsis. These results emphasize the value of incorporating visual interpretability and confidence metrics to better understand model behavior.

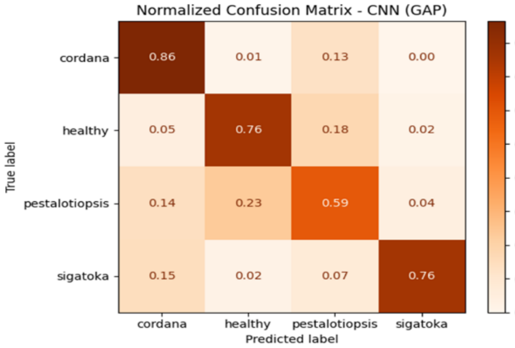

Figure 3. Normalized confusion matrix highlighting per-class prediction accuracy despite class imbalance. One of the samples was misclassified. The model gave 56% confidence for Healthy and 39% for Pestalotiopsis. Figure 4 shows the SoftMax output, where the uncertainty between these two classes is clear. A case like this could be useful to flag during deployment so it can be checked by an expert.

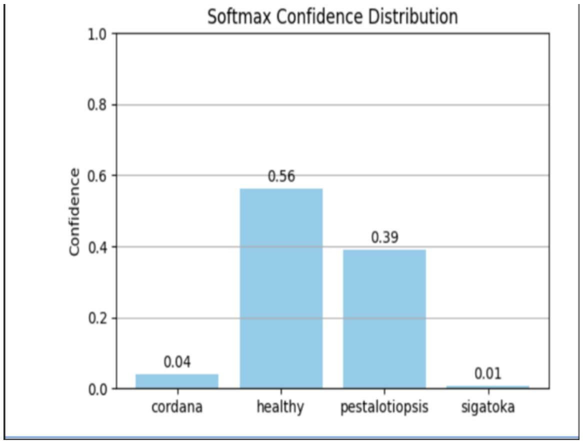

Figure 4. Example softmax output for a misclassified leaf image, showing uncertainty between Healthy and Pestalotiopsis.

5. DISCUSSION

5.1 Model Simplicity and XAI While most research focuses on improving accuracy through architectural complexity [2], [5], we take the opposite route demonstrating that interpretability can validate even modest models, making them more acceptable in real-world deployments.

5.2 Comparison with Prior Work Out of the available studies, it seems that only Ashoka et al. [5] investigated using XAI with the Banana LSD dataset. Their method used a more complex and powerful model, which also needed a lot of computing resources to run. In contrast, our work focuses on applying explainability tools to a simpler CNN, which hasn't been examined much in earlier research.

5.3 Practical Relevance In rural farming setups, models need to be efficient, explainable, and trustworthy. Our work highlights that XAI is a powerful equalizer—enabling lower-accuracy models to be trusted and understood, even when hardware limits prevent deployment of heavier networks [3], [7].

6. CONCLUSION

In this work, we used a simple CNN to detect diseases in banana leaves and looked at how explainable AI methods could make its decisions easier to follow. The model reached 74.71% accuracy, which isn't state-of-the-art, but it was enough to explore how well it understood disease features. The model was simple, but Grad-CAM and confidence scores still gave us some idea of where it was looking and how confident it seemed. That kind of output might be useful when working in places where small models are needed. Later, we want to test other types of models, like transformers, to check if they give clearer results without being too complex.

REFERENCES

[1] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Comput. Electron. Agric.*, vol. 145, pp. 311–318, Jan. 2018, doi: 10.1016/j.compag.2018.01.009.

[2] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Front. Plant Sci.*, vol. 7, p. 1419, Sep. 2016, doi: 10.3389/fpls.2016.01419.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[4] S. E. Arman, M. A. B. Bhuiyan, H. M. Abdullah, S. Islam, T. T. Chowdhury, and M. A. Hossain, "BananaLSD: A banana leaf images dataset for classification of banana leaf diseases using machine learning," *Data in Brief*, vol. 50, Sep. 2023, Art. no. 109608, doi: 10.1016/j.dib.2023.109608.

[5] K. Ashoka, R. B. Raju, and R. R. Rupanagudi, "XAI-integrated EfficientNet model for banana disease detection using Grad-CAM," in *Proc. 5th Int. Conf. Innovative Trends in Information Technology (ICITIIT)*, Kottayam, India, Mar. 2024, pp. 1–6, doi: 10.1109/ICITIIT61487.2024.10580364.

[6] M. J. Karim, M. O. F. Goni, M. Nahiduzzaman, M. Ahsan, J. Haider, and M. Kowalski, "Enhancing agriculture through real-time grape leaf disease classification via an edge device with a lightweight CNN architecture and Grad-CAM,"

Scientific Reports, vol. 14, Art. no. 16022, Jul. 2024, doi: 10.1038/s41598-024-66989-9.

[7] D. Kamilaris and F. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Comput. Electron. Agric.*, vol. 147, pp. 70–90, Apr. 2018, doi: 10.1016/j.compag.2018.02.016.

[8] S. A. Palwe, N. Shukla, M. Rajani, and A. Suri, "Plant disease detection and localization using Grad-CAM," *Int. J. Recent Technol. Eng. (IJRTE)*, vol. 8, no. 6, pp. 3069–3073, Mar. 2020, doi: 10.35940/ijrte.E6935.038620.

[9] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, Art. no. 60, Apr. 2019, doi: 10.1186/s40537-019-0197-0.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations (ICLR)*, May 2021. [Online]. Available: https://arxiv.org/abs/2010.11929