Potato Leaf Disease Classification Using Custom CNNs and Transfer Learning

Bishwas Chaudhary
Birmingham city university

1. Abstract

Plant diseases significantly affect agricultural productivity and global food security, making early and accurate detection essential to reduce crop loss and improve treatment decisions. With advances in deep learning, image-based plant disease classification has become increasingly effective. This study compares two deep learning approaches for potato leaf disease detection: a custom convolutional neural network trained from scratch and a transfer learning model based on MobileNetV2, using a PlantVillage-derived dataset containing Early Blight, Late Blight, and Healthy leaf images.

Both models were trained and evaluated under identical conditions to ensure a fair comparison. Results show that the MobileNetV2-based transfer learning model outperforms the custom CNN in accuracy, generalization, and training efficiency. Additionally, LIME-based explainable AI techniques were applied to interpret model predictions, improving transparency and trust. Overall, the study demonstrates that combining transfer learning with explainability offers an effective and reliable solution for plant disease classification.

Keywords: *Potato leaf disease, Deep learning, Convolutional neural network, Transfer learning, MobileNetV2, Image classification, Explainable AI, LIME, PlantVillage dataset*

2. List of Figures

3. List of Abbreviations

Abbreviation	Full Form and Technical Context
AI	Artificial intelligence
CNN	Convolutional Neural Network
DL	Deep Learning
EDA	Exploratory Data Analysis
FAO	Food and Agriculture Organization of the United Nations
GAP	Global Average Pooling
LIME	Local Interpretable Model-agnostic Explanations
ML	Machine Learning
ReLU	Rectified Linear Unit
SGD	Stochastic Gradient Descent
XAI	Explainable Artificial Intelligence

4. Introduction

1. Background:

Agriculture is essential to global food production, yet plant diseases continue to pose a significant threat to crop yield and quality. Potato, one of the most widely cultivated crops, is highly vulnerable to diseases such as Early Blight and Late Blight. Conventional disease diagnosis depends on expert knowledge and manual inspection, which can be time-consuming, subjective, and often inaccessible to small-scale farmers.

Advances in computer vision and deep learning have enabled automated plant disease detection using leaf images. Convolutional Neural Networks (CNNs) have shown strong performance in visual recognition tasks, including plant disease classification (Mohanty et al., 2016). Transfer learning further enhances accuracy by leveraging pre-trained models on large-scale datasets like ImageNet (Ferentinos, 2018).

2. Research gaps:

Despite promising results, several challenges remain:

- **Black-box behavior:** Deep learning models often lack transparency, limiting trust and real-world adoption (Ribeiro et al., 2016).
- **Generalization issues:** Models trained on controlled datasets may perform poorly in real-field conditions (Ferentinos, 2018).
- **Efficiency constraints:** High-performing models can be unsuitable for deployment on low-resource devices.
- **Limited understandability:** Distinguishing visually similar diseases such as Early Blight and Late Blight remains difficult.

  This study addresses these challenges by comparing a baseline custom CNN with a lightweight transfer learning model and incorporating explainability using LIME.

3. Objectives:

The main objectives of this research are:

- To develop a deep learning-based system for potato leaf disease classification.
- To compare the performance of a custom CNN with a transfer learning model.
- To evaluate both models using robust performance metrics.
- To enhance model transparency through explainable AI techniques.

5. Dataset Description

1. Dataset Introduction:

The dataset used in this study is obtained from the PlantVillage repository, which provides labeled images of healthy and diseased plant leaves (Mohanty et al., 2016). The potato subset consists of three classes: Early Blight, Late Blight, and Healthy. All images are captured under controlled conditions with uniform backgrounds.

Figure 1: Sample images from the PlantVillage potato leaf dataset showing Early Blight, Late Blight, and Healthy classes.

2. Short limitations:

Despite the strong performance achieved by both models, the PlantVillage dataset is collected under controlled laboratory conditions, with uniform backgrounds, consistent lighting, and centrally positioned leaves. Such conditions differ

substantially from real-world agricultural environments, where variations in illumination, occlusion, background clutter, and leaf orientation are common. This discrepancy introduces a potential domain shift between the training data and real-field imagery, which may impact model performance outside controlled settings. Consequently, further validation using in-field datasets or domain adaptation techniques is required to ensure robust deployment in practical agricultural applications.

3. Machine Learning Type Identification

This study focuses on a supervised multi-class image classification task, where labeled leaf images are used to train models to predict disease categories.

6. Exploratory Data Analysis (EDA)

1. Operations:

Exploratory Data Analysis was conducted to understand the dataset and includes:

- Visualization of sample images from each class
- Analysis of class distribution
- Verification of image dimensions and color channels
- Analysis of pixel value distributions and class-wise visual comparison

2. Findings:

Exploratory Data Analysis (EDA) indicated that the dataset is generally well balanced across classes, although the Healthy category contains slightly fewer samples compared to the diseased classes. Despite this minor imbalance, the overall class distribution is suitable for training deep learning models.

The images in the dataset are high quality and consistent in resolution, which helps ensure stable and effective learning. Clear disease-specific visual patterns, such as lesions, spots, and leaf discoloration, are easily observable. These distinct features make the dataset well suited for deep learning-based disease classification and support reliable feature extraction by the models.

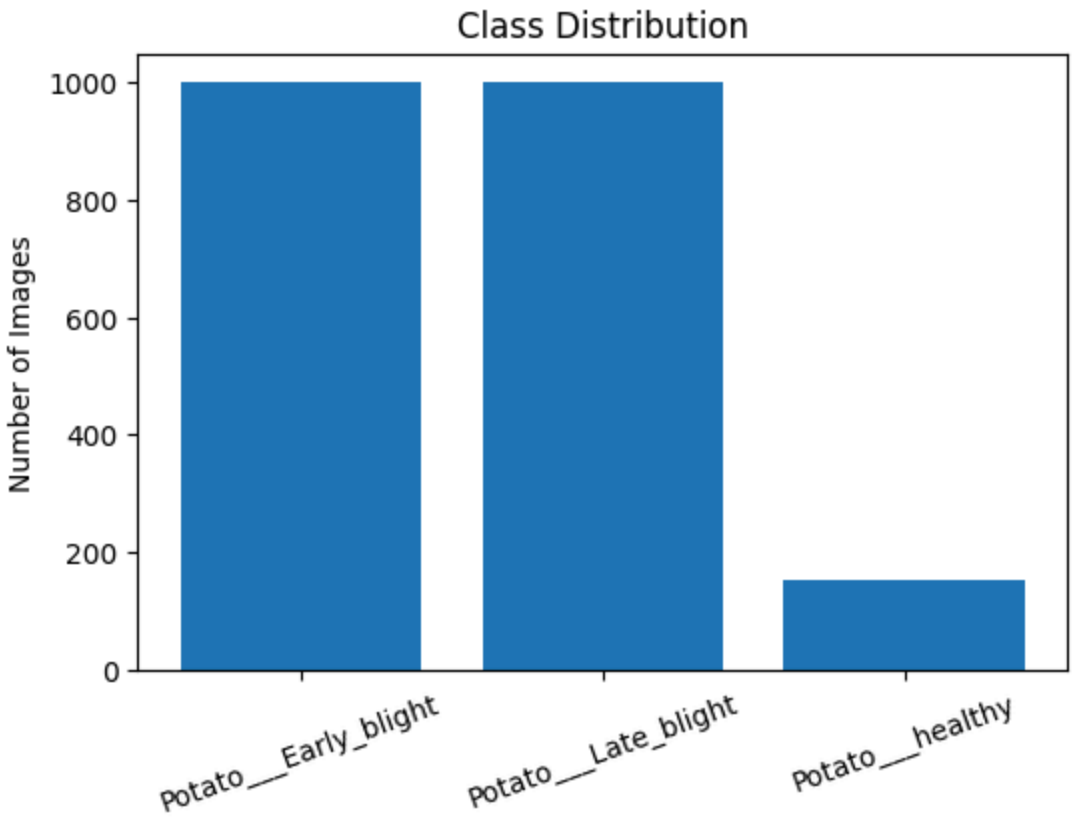


Figure 2: Distribution of images across the three potato leaf classes.

7. Experimental Design

1. Justification of Model 1: Custom CNN:

A custom CNN was implemented as a baseline model to analyze the learning behavior of a network trained from scratch, providing insight into feature extraction capabilities and limitations without relying on external knowledge.

2. Justification of Model 2: Transfer Learning:

Transfer learning with MobileNetV2 was chosen for its efficiency and strong performance in image classification tasks (Too et al., 2019). Its lightweight architecture and low computational cost make it well-suited for this project.

8. Data Cleaning & Preprocessing

Preprocessing steps included:

- Resizing images to 256×256 pixels
- Normalizing pixel values
- Applying data augmentation techniques such as rotation and flipping to improve generalization (Shorten and Khoshgoftaar, 2019)
- Splitting the dataset into training, validation, and test sets using a fixed random seed to prevent data leakage
- Pipeline optimization (caching and prefetching)

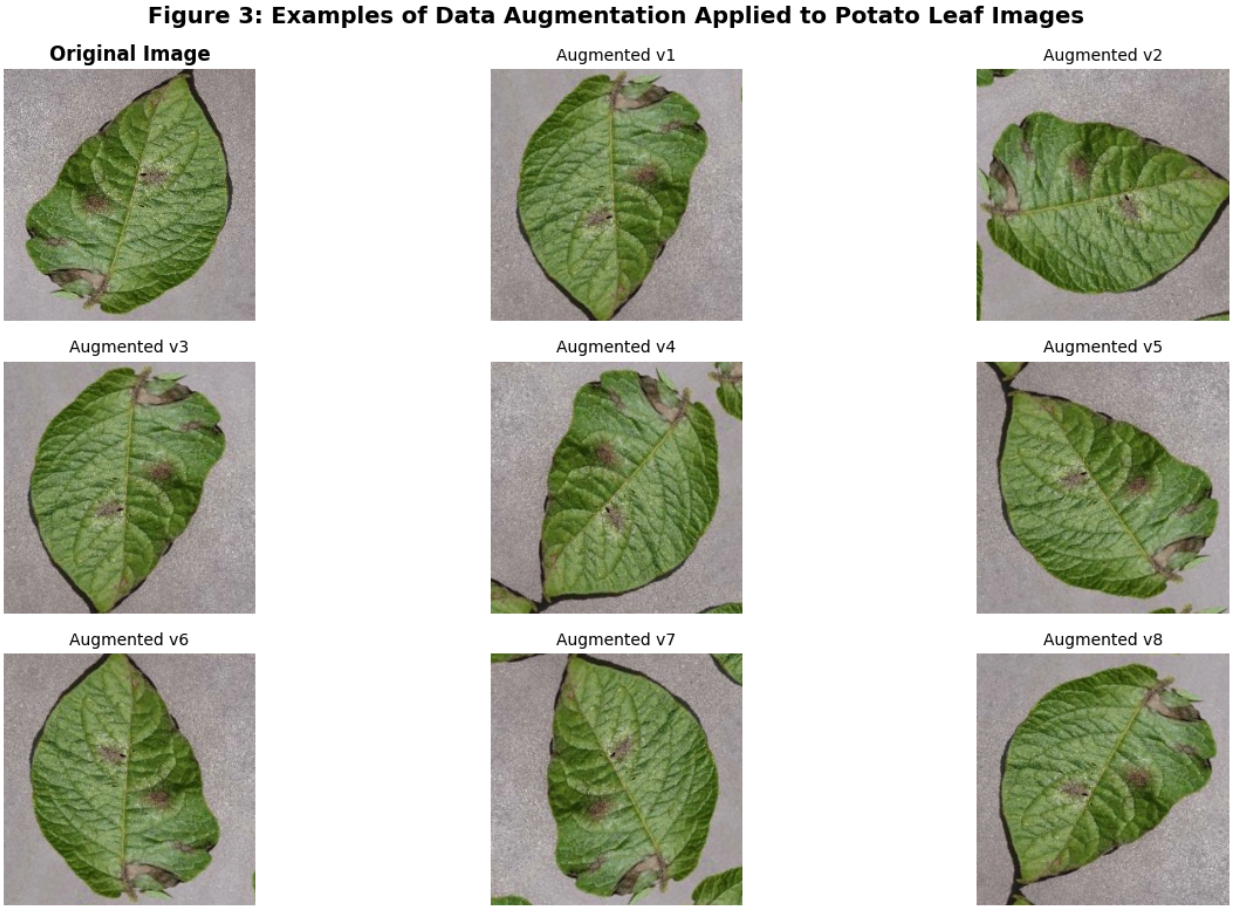


Figure 3: Examples of data augmentation applied to potato leaf images.

9. Model Development

1. Train-Test Split:

The dataset was divided into training, validation, and test sets using deterministic shuffling with a fixed random seed, followed by sequential partitioning with *take()* and *skip()* operations to ensure reproducibility and prevent data leakage.

2. Model Architecture:

A. Model 1: Custom CNN:
- Multiple convolutional layers with ReLU activation
- Max pooling layers for spatial reduction
- Fully connected layers
- Softmax output layer

B. Model 2: MobileNetV2:
- Pretrained backbone with frozen weights
- Global average pooling
- Dropout for regularization
- Fine-tuning with a low learning rate

3. Model Type:

The first model is a custom convolutional neural network (CNN) designed for multi-class image classification. CNNs are effective for visual tasks as they automatically learn hierarchical spatial features from raw pixel data (LeCun et al., 2015).

The second model applies transfer learning using a pretrained MobileNetV2 architecture trained on ImageNet. By reusing learned visual features, transfer learning enables faster convergence and better generalization, especially with limited training data (Howard et al., 2017).

4. Layers:

The custom CNN is composed of multiple convolutional layers with max-pooling to progressively extract spatial features and reduce dimensionality. As the network deepens, the number of filters increases to capture more complex patterns.

After feature extraction, a flattening layer and fully connected dense layers are used for classification. The final layer applies a softmax activation function to output class probabilities for the three disease categories.

In the transfer learning model, the MobileNetV2 backbone is used as a fixed feature extractor during initial training. A global average pooling layer reduces feature dimensionality, followed by a dropout layer to limit overfitting and a dense output layer for classification.

5. Activation Functions:

ReLU activation functions were used in all convolutional and hidden dense layers for their computational efficiency and ability to reduce vanishing gradient issues (Nair and Hinton, 2010). The output layer uses a softmax activation function to convert logits into normalized class probabilities for multi-class classification.

6. Optimizer:

Both models were trained using the Adam optimizer, which combines adaptive learning rates and momentum to achieve faster convergence and stable training (Kingma and Ba, 2015). During MobileNetV2 fine-tuning, a lower learning rate was used to avoid large updates that could disrupt pretrained features.

7. Loss function:

Sparse categorical cross-entropy was used as the loss function since the task involves multi-class classification with integer-encoded labels. It is well-suited for softmax outputs and penalizes incorrect predictions based on their confidence (Goodfellow et al., 2016).

8. Epochs and Early Stopping:

Both models were trained for a maximum of 50 epochs. To prevent overfitting and reduce unnecessary computation, early stopping was applied based on validation accuracy. A patience value of five epochs was used, meaning that training was automatically terminated if validation performance failed to improve for five consecutive epochs. This strategy helped ensure efficient training while retaining the best-performing model parameters.

9. Batch Size:

A batch size of 32 was selected as it provides a balance between gradient stability and computational efficiency, while also ensuring sufficient stochasticity during optimization to improve generalization on unseen data.

10. Evaluation Metrics & Results

Models were evaluated using the following metrics:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion matrix

The transfer learning model outperformed the custom CNN, achieving higher accuracy and better generalization across all classes.

The superior generalization of Model 2 is primarily due to pretrained convolutional features from the large-scale ImageNet dataset, which capture robust visual patterns transferable to plant disease images. Initially freezing the MobileNetV2 backbone acts as implicit regularization by reducing trainable parameters and limiting overfitting. This encourages the model to learn task-specific features only in the classification head, resulting in more stable performance and better generalization on unseen test data compared to the custom CNN trained from scratch.

1. Model Comparison Summary:

Overall, the transfer learning model showed better performance than the custom CNN. It achieved a higher test accuracy of 99.61% compared to 99.22% and also obtained a higher macro-averaged F1-score, indicating more balanced performance across all classes. This improvement is especially important for handling differences between disease categories.

In addition, the MobileNetV2-based model used fewer trainable parameters because most of the pretrained backbone was kept frozen. This reduced the risk of overfitting and allowed the model to train faster. By reusing strong pretrained feature representations, the transfer learning approach generalized better to unseen test data, highlighting its advantage over training a model from scratch for this task.

Model	Test Accuracy	Macro F1	Weighted F1	Wrong Predictions	Total Test Samples
Custom CNN	0.9453	0.9594	0.9452	14	256
MobileNetV2	0.9961	0.9910	0.9961	1	256

Table 1: Comparison Summary between two models

2. Overfitting & Parameters:

The transfer learning approach was implemented in two stages: an initial training phase with a fully frozen MobileNetV2 backbone, followed by a controlled fine-tuning phase where only the final convolutional layers were unfrozen. The frozen training phase acted as a strong regularizer by limiting the number of trainable parameters, while the fine-tuning stage allowed minor adaptation of high-level features using a low learning rate and early stopping. This strategy preserves the generalization benefits of pretrained representations while avoiding the overfitting risks associated with training all parameters from scratch.

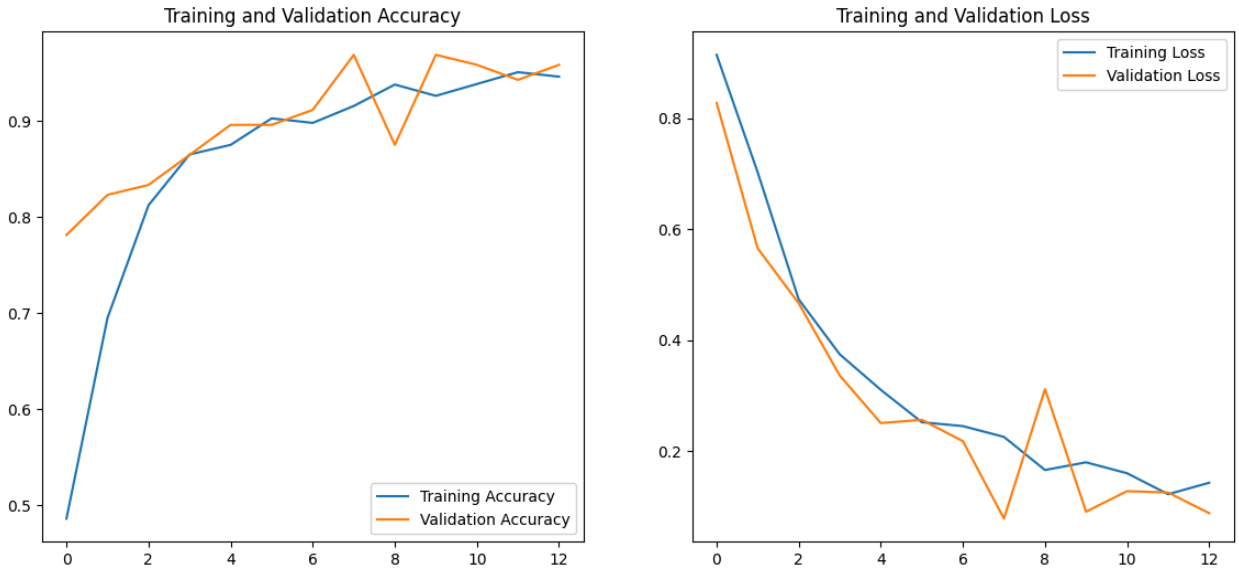


Figure 4: Training and validation accuracy and loss curves for the custom CNN model.

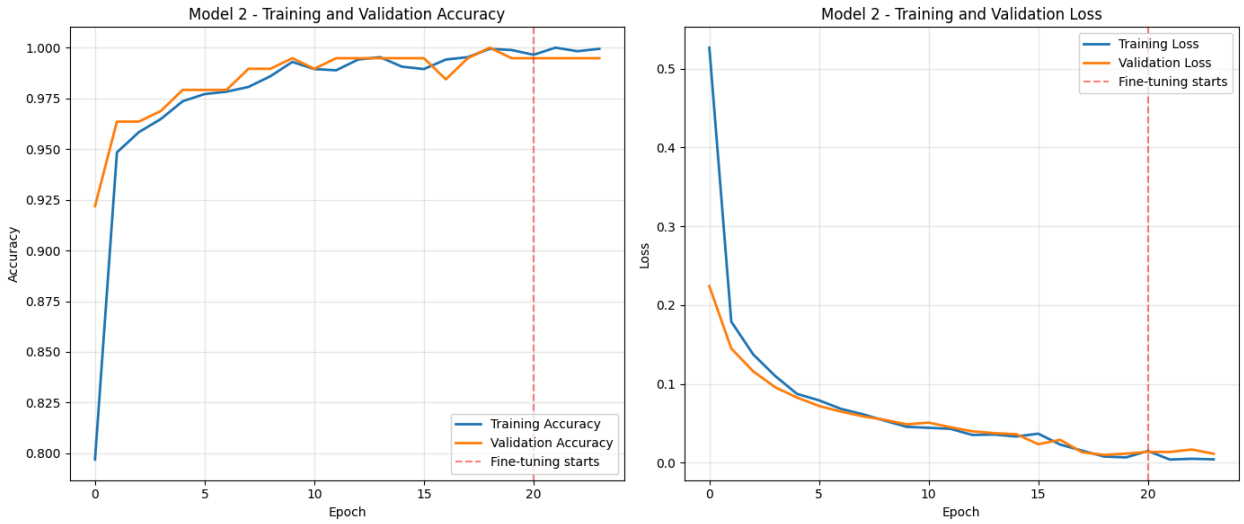


Figure 5: Training and validation accuracy and loss curves for the MobileNetV2 transfer learning model.

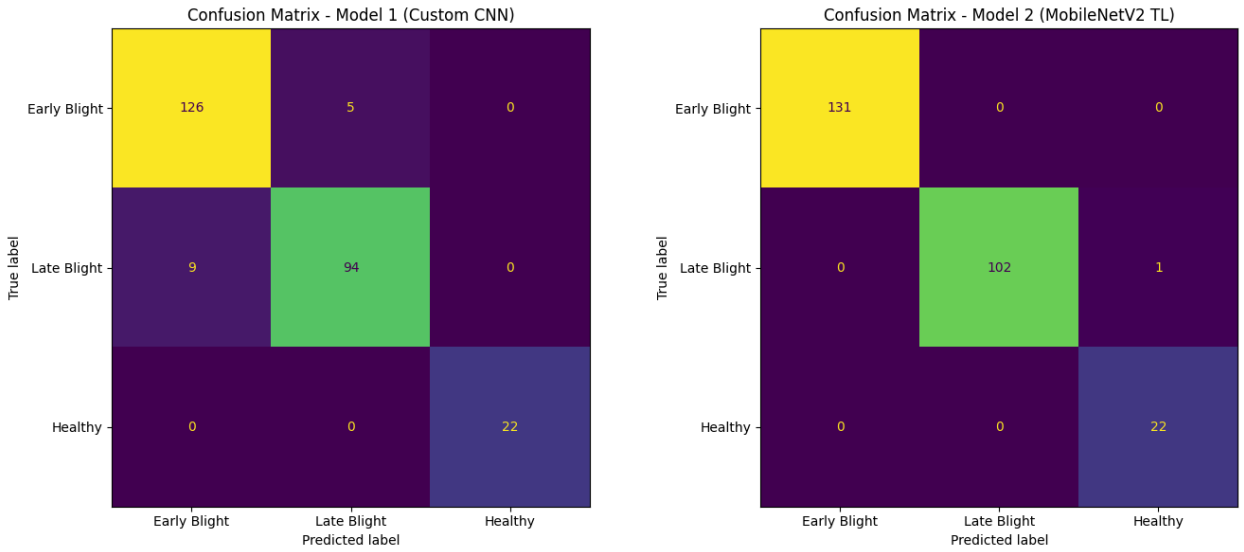


Figure 6: Confusion matrix of the custom CNN and MobileNetV2 model on the test dataset.

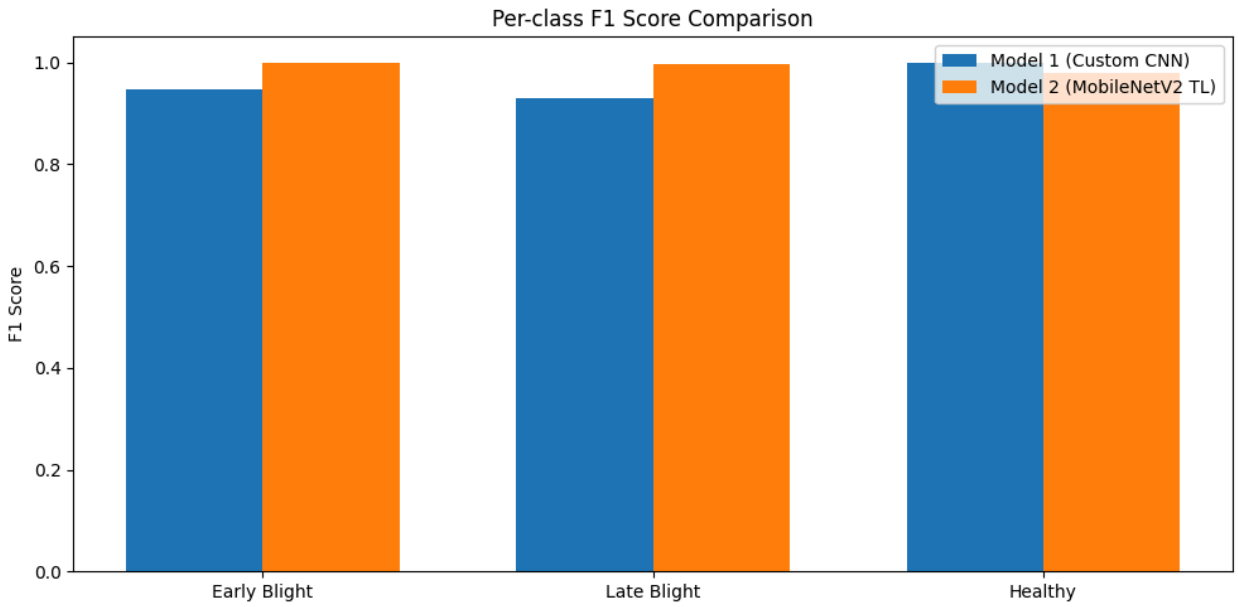
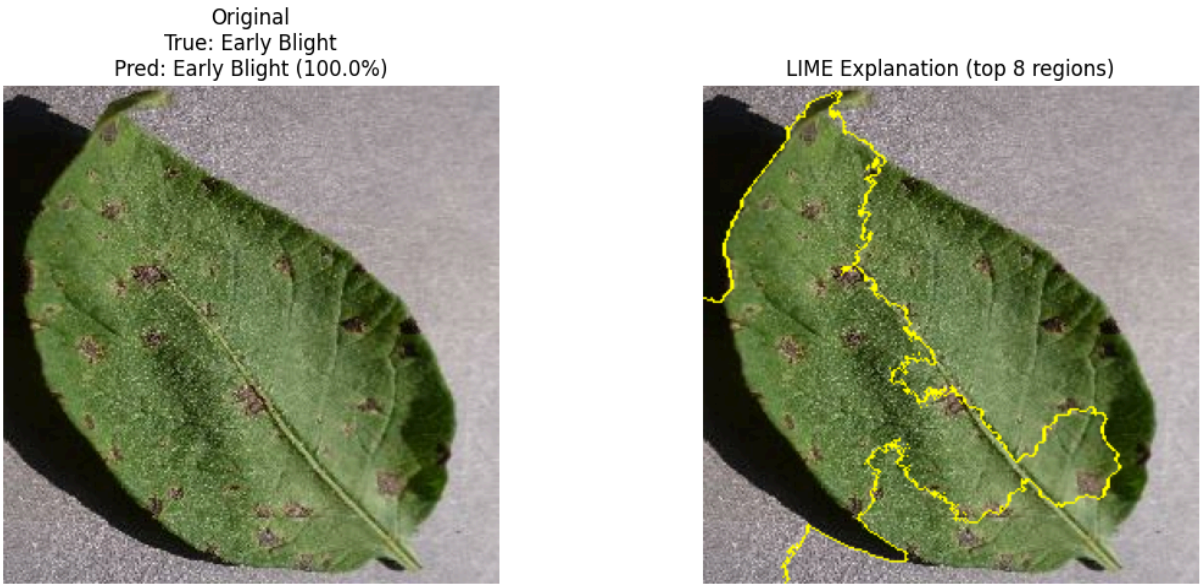


Figure 7: F1 Score of the custom CNN and MobileNetV2 model.

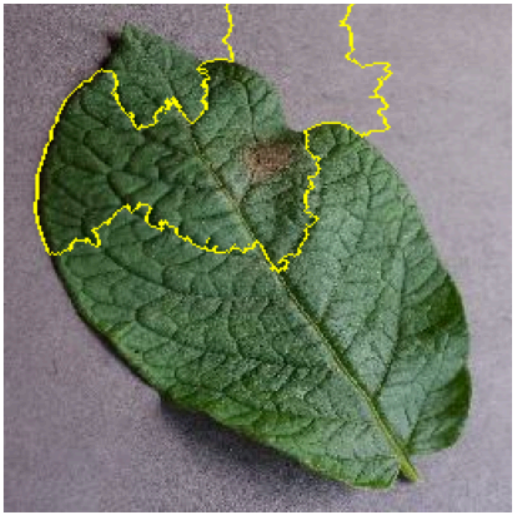
11. Explainable Artificial Intelligence

To enhance transparency, LIME was used to explain individual model predictions (Ribeiro et al., 2016). The explanations emphasized important leaf regions such as lesions and discoloration, aligning with domain knowledge and increasing trust in the model's decisions.



Original
True: Late Blight
Pred: Late Blight (99.8%)

LIME Explanation (top 8 regions)



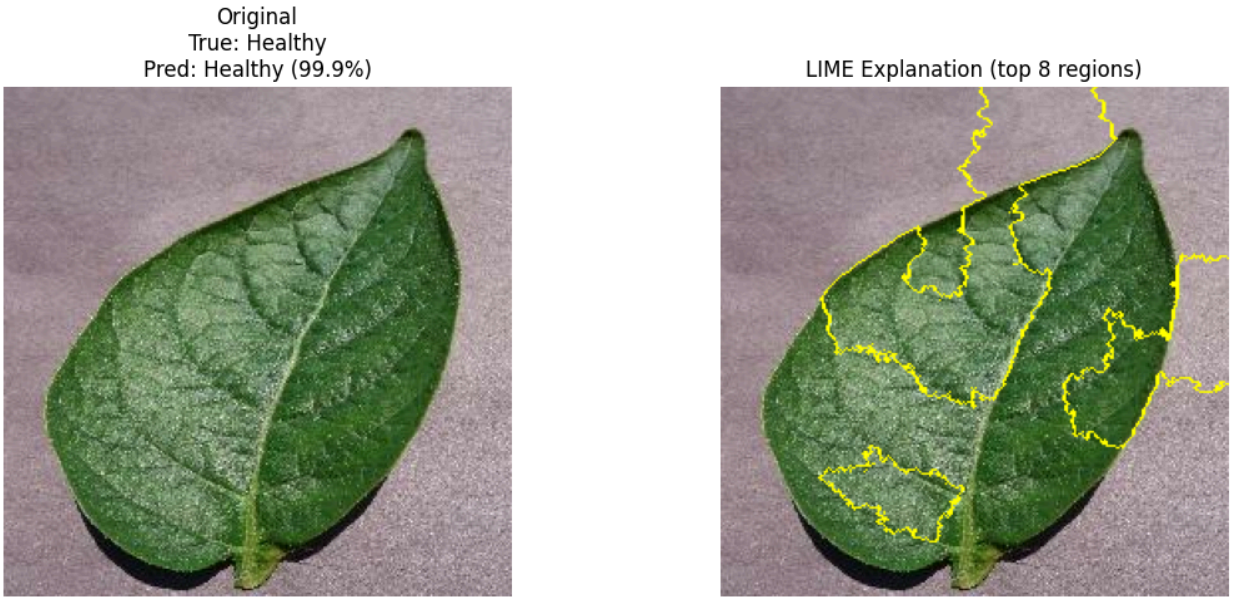


Figure 8: LIME explanations for correctly classified Early Blight, Late Blight, and Healthy samples.

12. Conclusion

1. Summary:

This study demonstrates the application of deep learning techniques for the classification of potato leaf diseases, specifically Early Blight, Late Blight, and Healthy leaves. Two approaches were evaluated: a custom convolutional neural network trained from scratch and a transfer learning model based on MobileNetV2. Experimental results showed that both models achieved high classification performance; however, the transfer learning model consistently outperformed the custom CNN in terms of accuracy, F1-score, and convergence speed. The reuse of pretrained features allowed the MobileNetV2 model to generalize more effectively to unseen data while requiring significantly fewer trainable parameters.

2. Reflection on Learning:

This project provided valuable hands-on experience in the complete machine learning workflow, from dataset exploration and preprocessing to model development, evaluation, and interpretation. Implementing both a custom CNN and a transfer learning approach deepened the understanding of convolutional architectures, optimization strategies, and the impact of model complexity on generalization. Additionally, incorporating explainable AI techniques such as LIME enhanced awareness of model transparency and interpretability, which are increasingly important in real-world AI applications, particularly in sensitive domains such as agriculture.

3. Future Recommendations:

While the results are promising, several directions can be explored to further improve the system. Future work may include training and validating the models on real-field images to address domain shift and improve robustness under natural environmental conditions. Expanding the dataset to include additional potato diseases or other crop types would enhance the applicability of the model. Further explainability techniques, such as Grad-CAM, could be applied to gain deeper insights into the spatial regions influencing model predictions. Finally, deploying the optimized model on mobile or edge devices would enable practical, real-time disease detection for farmers, contributing to more accessible and efficient agricultural decision support systems.

13. References

Boulent, A., Foucher, S., Théau, J. and St-Charles, P.-L. (2019) Convolutional neural networks for the automatic identification of plant diseases. *Frontiers in Plant Science*, 10, pp. 1–18.

Ferentinos, K.P. (2018) Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145, pp. 311–318.

Mohanty, S.P., Hughes, D.P. and Salathé, M. (2016) Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7, pp. 1–10.

Rangarajan, A.K. and Purushothaman, R. (2020) Disease classification in eggplant using transfer learning. *Scientific Reports*, 10(1), pp. 1–11.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD Conference*, pp. 1135–1144.

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

Too, E.C., Yujian, L., Njuki, S. and Yingchun, L. (2019) A comparative study of fine-tuning deep learning models for plant disease identification. *Computers and Electronics in Agriculture*, 161, pp. 272–279.