Potato Leaf Disease Classification Using Custom CNNs and Transfer Learning

Bishwas Chaudhary
Birmingham city university

1. Abstract

Plant diseases significantly threaten agricultural productivity and global food security, making early and accurate detection essential (Ferentinos, 2018; Kamilaris and Prenafeta-Boldú, 2018). Advances in deep learning, particularly convolutional neural networks, have enabled effective image-based plant disease classification through automated feature extraction (LeCun, Bengio and Hinton, 2015; Mohanty, Hughes and Salathé, 2016). This study compares a custom CNN with a MobileNetV2-based transfer learning model for classifying potato leaf diseases using a PlantVillage-derived dataset.

Both models were evaluated under identical conditions, with results showing that the MobileNetV2 model achieves higher accuracy, better generalization, and improved training efficiency, consistent with prior studies (Too et al., 2019; Zhang, Yang and Zhang, 2021). Model predictions were further explained using LIME, enhancing transparency and trust (Ribeiro, Singh and Guestrin, 2016).

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 vital for global food production and security, but plant diseases continue to threaten crop yield and sustainability (Strange and Scott, 2005; FAO, 2019). Potato (Solanum tuberosum), a major staple crop, is particularly susceptible to Early Blight and Late Blight, which can cause severe yield losses if not detected early (Fry, 2008; Savary et al., 2019). Traditional diagnosis relies on expert knowledge and manual inspection, making it time-consuming, subjective, and often inaccessible to small-scale farmers (Polder et al., 2019).

Recent advances in computer vision and deep learning have enabled automated plant disease detection using leaf images. Convolutional Neural Networks (CNNs) have achieved high accuracy by learning hierarchical visual features directly from image data (LeCun, Bengio and Hinton, 2015; Mohanty, Hughes and Salathé, 2016; Too et al., 2019). Transfer learning further improves performance by reusing pretrained features from large datasets such as ImageNet, allowing effective learning even with limited agricultural data (Ferentinos, 2018; Tan and Le, 2019).

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 is widely used in agricultural deep learning research due to its well-annotated and diverse classes (Mohanty, Hughes and Salathé, 2016). The potato subset includes three classes: Early Blight, Late Blight, and Healthy. The images are captured under controlled lighting and uniform backgrounds, which may introduce domain shift when applied to real-field conditions (Barbedo, 2019).

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

2. Short limitations:

Despite strong performance, the PlantVillage dataset is captured under controlled conditions with uniform backgrounds and lighting, which differ from real-world agricultural environments. Variations in illumination, occlusion, background clutter, and leaf orientation in field settings can introduce domain shift and affect

model performance. Therefore, further validation using in-field data or domain adaptation techniques is necessary for reliable real-world deployment.

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 and visualization are essential for understanding dataset characteristics and guiding preprocessing decisions in deep learning pipelines (Goodfellow, Bengio and Courville, 2016). In this study, EDA was conducted to analyze the structure and quality of the dataset through the following operations:

- 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) showed that the dataset is largely balanced, although the Healthy class contains slightly fewer samples. Despite this minor imbalance, the data remains suitable for training deep learning models. The images are high quality and consistent in resolution, with clear disease-specific features such as lesions, spots, and discoloration, enabling effective feature extraction and reliable classification.

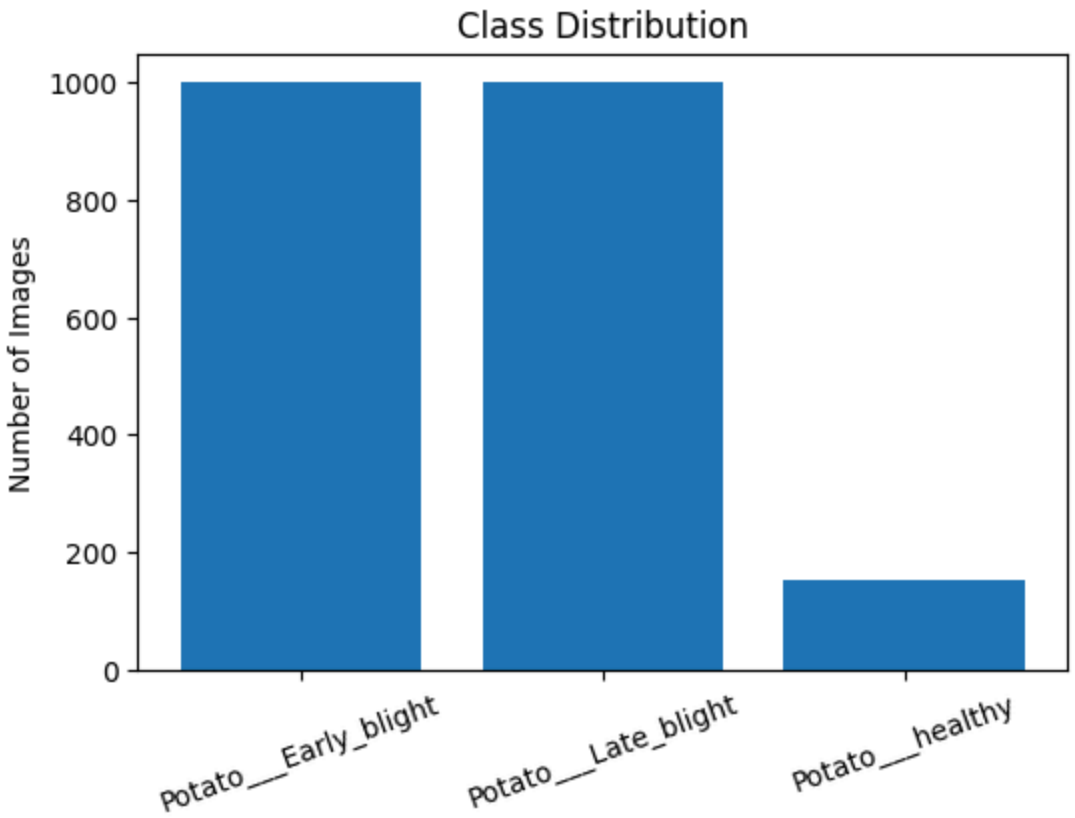


Figure 2: *Distribution of images across the three potato leaf classes.*

7. Experimental Design

1. Justification of Model 1: Custom CNN:

A custom CNN was used as a baseline to study learning behavior and feature extraction when training from scratch. CNNs are effective for image-based tasks because convolutional layers capture local spatial patterns such as edges, textures, and shapes (LeCun, Bengio and Hinton, 2015).

2. Justification of Model 2: Transfer Learning:

MobileNetV2 was chosen for transfer learning due to its lightweight architecture and strong performance in image classification (Too et al., 2019). Such efficient models are well suited for agricultural applications (Sandler et al., 2018; Howard et al., 2017), while pretrained features from ImageNet enable faster convergence and better generalization (Simonyan and Zisserman, 2015; He et al., 2016).

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)

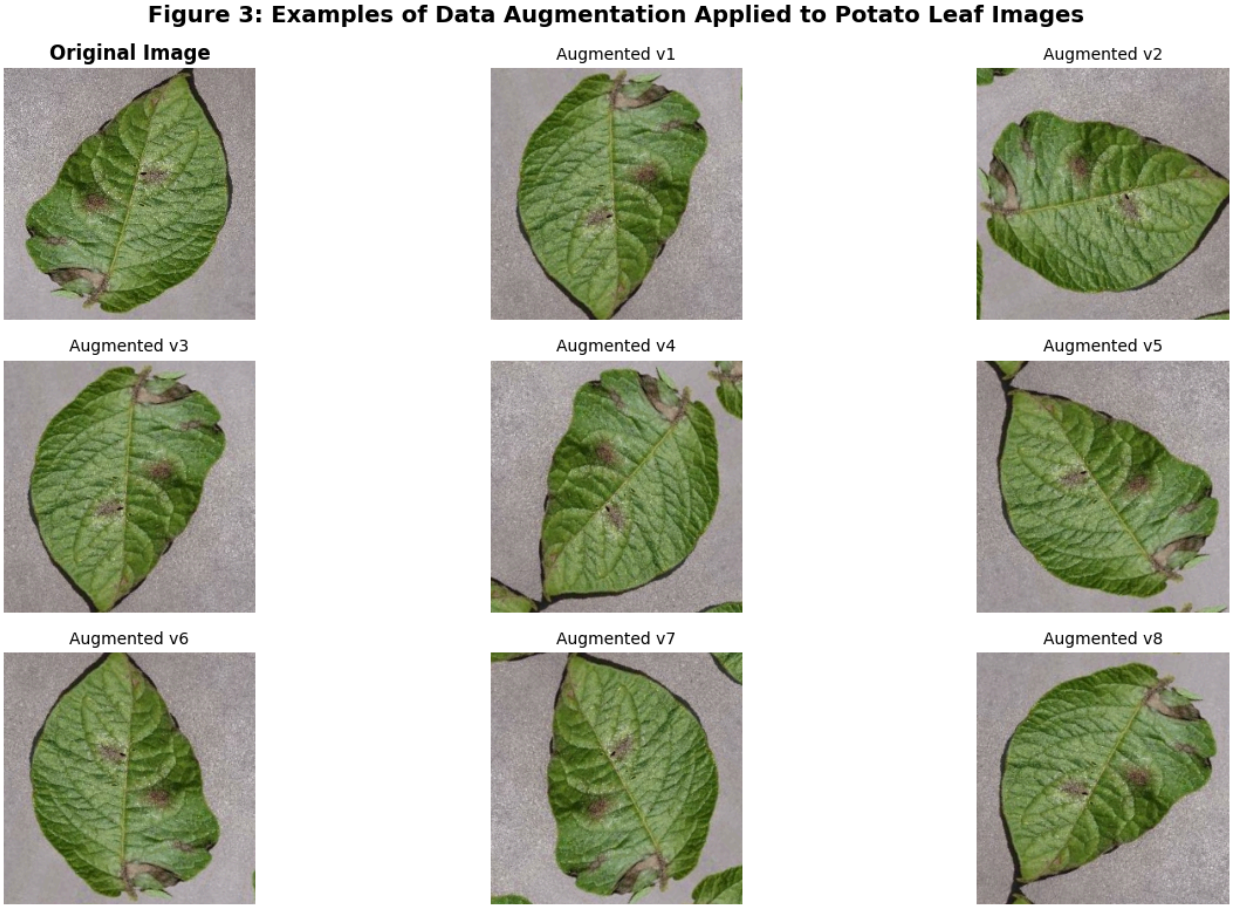


Figure 3: *Examples of data augmentation applied to potato leaf images.*

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 provide fast and stable convergence (Kingma and Ba, 2015). Adaptive optimization methods such as Adam are widely used in deep learning for their training stability (Goodfellow, Bengio and Courville, 2016). During MobileNetV2 fine-tuning, a lower learning rate was applied to preserve 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. Early stopping was applied based on validation accuracy to reduce overfitting and unnecessary computation, with a patience of five epochs. Early stopping acts as an effective regularization strategy by halting training once validation performance no longer

improves, ensuring efficient training and optimal model selection (Buda, Maki and Mazurowski, 2018).

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 phase with the MobileNetV2 backbone fully frozen, followed by a controlled fine-tuning phase where only the final layers were unfrozen. The frozen phase acted as a strong regularizer by limiting trainable parameters, while fine-tuning with a low learning rate allowed limited adaptation without overfitting. This behavior aligns with prior studies showing that transfer learning reduces overfitting risk by restricting the number of trainable parameters (Too et al., 2019; Zhang, Yang and Zhang, 2021).

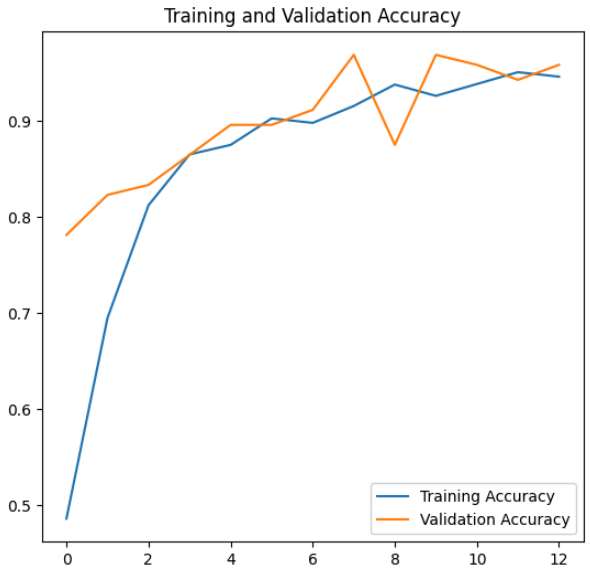


Figure 4: *Training and validation accuracy and loss curves for the custom CNN model.*

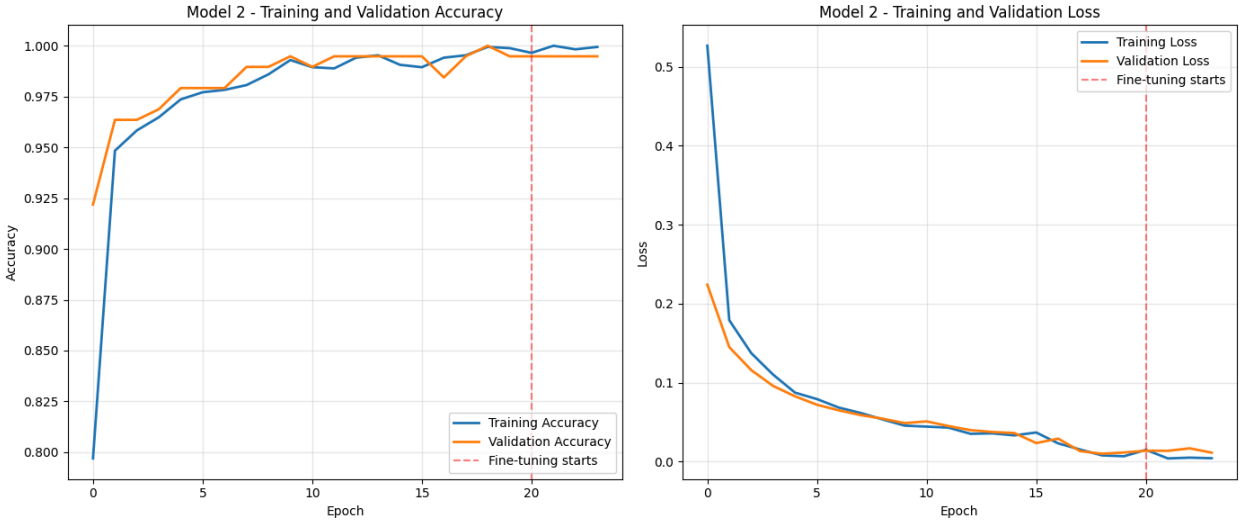


Figure 5: *Training and validation accuracy and loss curves for the MobileNetV2 transfer learning model.*

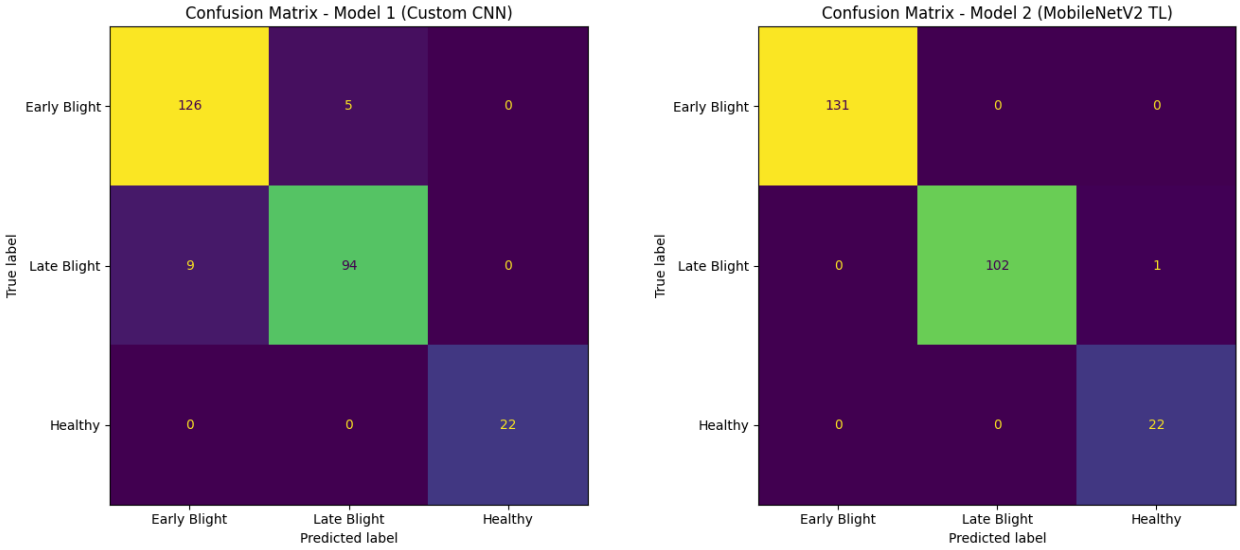


Figure 6: *Confusion matrix of the custom CNN and MobileNetV2 model on the test dataset.*

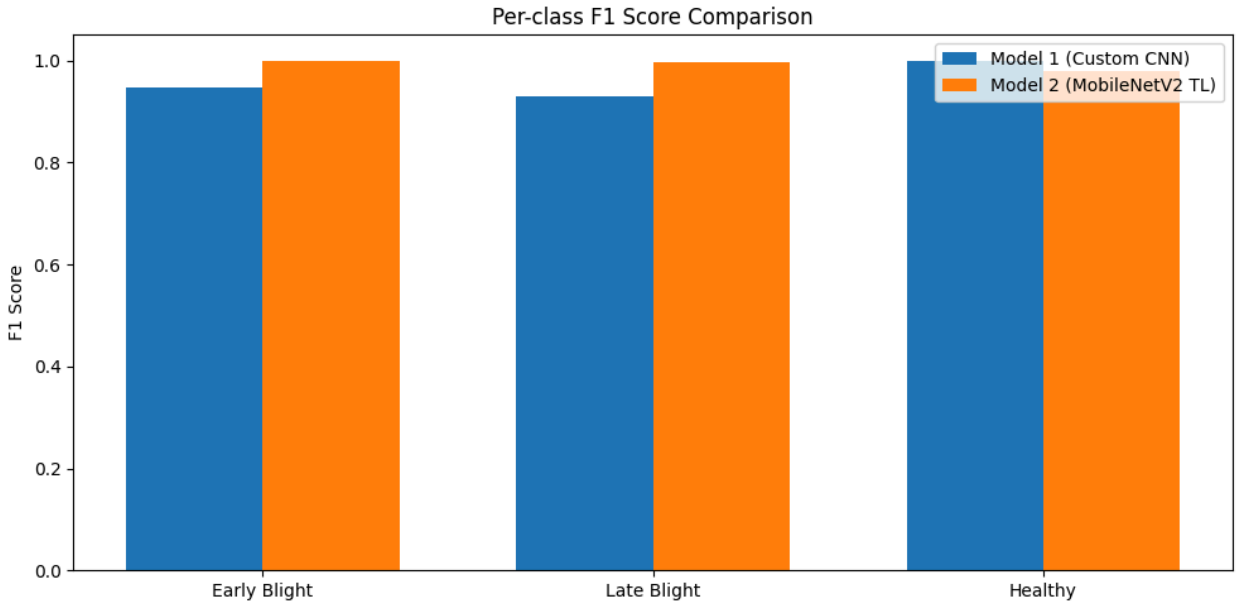


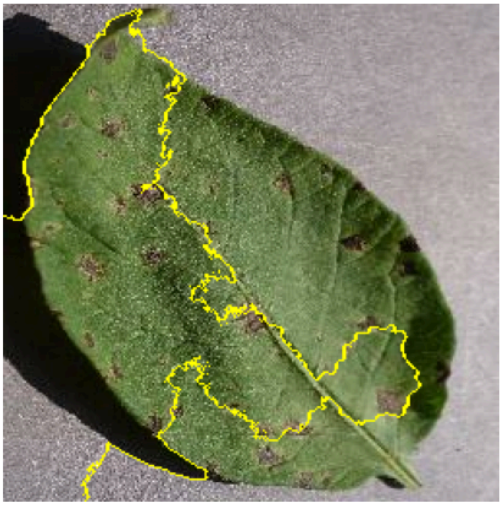
Figure 7: *F1 Score of the custom CNN and MobileNetV2 model.*

11. Explainable Artificial Intelligence

To enhance transparency, LIME was used to explain individual model predictions. Explainable AI techniques are essential for interpreting black-box deep learning models and building trust in automated decision systems (Ribeiro, Singh and Guestrin, 2016). The explanations highlighted important leaf regions such as lesions and discoloration, aligning with domain knowledge and improving trust in model decisions.

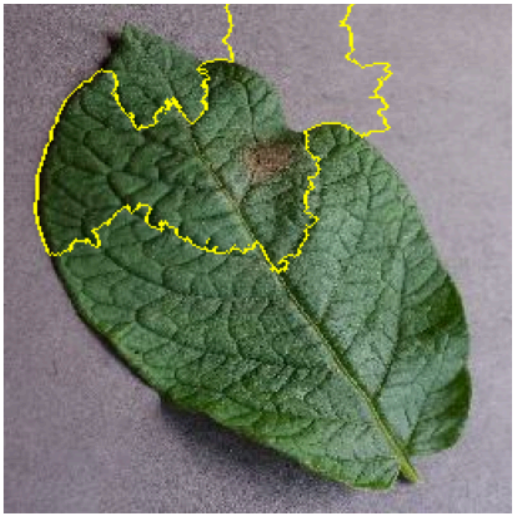
Original
True: Early Blight
Pred: Early Blight (100.0%)

LIME Explanation (top 8 regions)



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

LIME Explanation (top 8 regions)



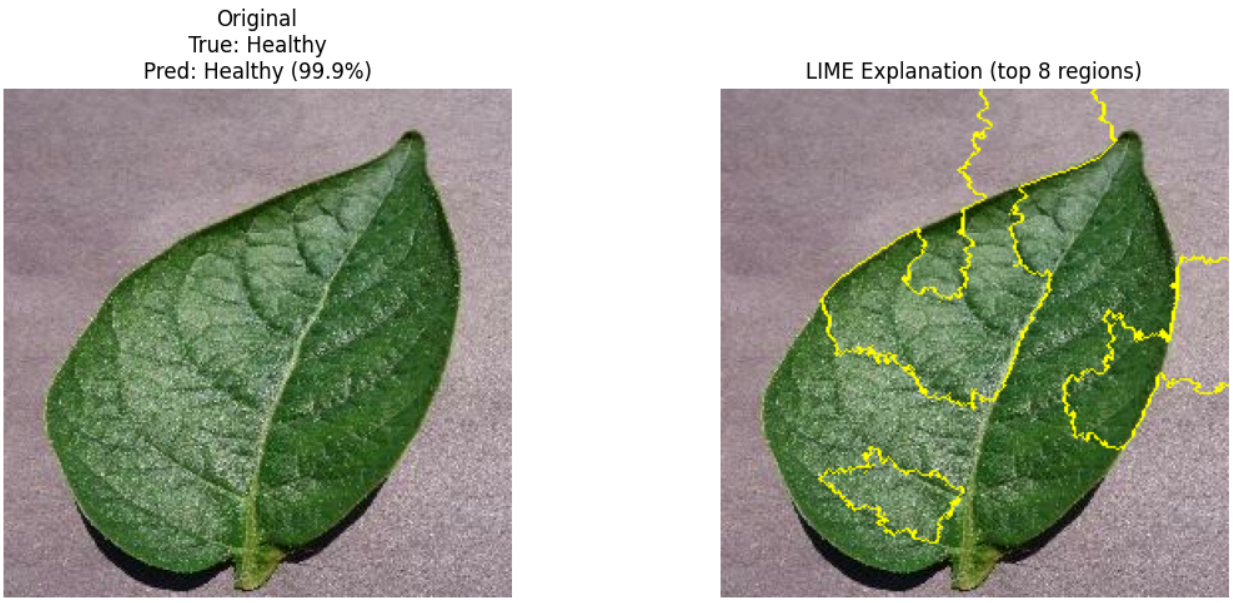


Figure 8: *LIME explanations for correctly classified Early Blight, Late Blight, and Healthy samples.*

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 further improve the system. Future work may include training and validating models on real-field images to reduce domain shift and improve robustness under natural conditions. Expanding the dataset to include additional potato diseases or other crops would enhance applicability. Advanced explainability methods such as Grad-CAM could provide deeper insights into model decisions, while emerging architectures like vision transformers show promise for further improving classification performance (Dosovitskiy et al., 2021). Finally, deploying optimized models on mobile or edge devices would enable practical, real-time disease detection for farmers.

13. References

LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*, 521(7553), pp. 436–444.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) 'ImageNet classification with deep convolutional neural networks', *Advances in Neural Information Processing Systems*, 25, pp. 1097–1105.

Simonyan, K. and Zisserman, A. (2015) 'Very deep convolutional networks for large-scale image recognition', *International Conference on Learning Representations (ICLR)*.

He, K., Zhang, X., Ren, S. and Sun, J. (2016) 'Deep residual learning for image recognition', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

Howard, A.G. et al. (2017) 'MobileNets: Efficient convolutional neural networks for mobile vision applications', *arXiv preprint arXiv:1704.04861*.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.C. (2018) 'MobileNetV2: Inverted residuals and linear bottlenecks', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520.

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.

Mohanty, S.P., Hughes, D.P. and Salathé, M. (2016) ‘Using deep learning for image-based plant disease detection’, *Frontiers in Plant Science*, 7, Article 1419.

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

Barbedo, J.G.A. (2019) ‘Plant disease identification from individual lesions and spots using deep learning’, *Biosystems Engineering*, 180, pp. 96–107.

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.

Buda, M., Maki, A. and Mazurowski, M.A. (2018) ‘A systematic study of the class imbalance problem in convolutional neural networks’, *Neural Networks*, 106, pp. 249–259.

Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning*. Cambridge, MA: MIT Press.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) ‘“Why should I trust you?” Explaining the predictions of any classifier’, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Selvaraju, R.R. et al. (2017) ‘Grad-CAM: Visual explanations from deep networks via gradient-based localization’, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626.

Dosovitskiy, A. et al. (2021) ‘An image is worth 16x16 words: Transformers for image recognition at scale’, *International Conference on Learning Representations (ICLR)*.

Zech, J.R. et al. (2018) ‘Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs’, *PLOS Medicine*, 15(11), e1002683.

Kamilaris, A. and Prenafeta-Boldú, F.X. (2018) ‘Deep learning in agriculture: A survey’, *Computers and Electronics in Agriculture*, 147, pp. 70–90.

Chollet, F. (2017) ‘Xception: Deep learning with depthwise separable convolutions’, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1251–1258.

Zhang, Y., Yang, Q. and Zhang, S. (2021) ‘Transfer learning for agricultural image classification: A review’, *Artificial Intelligence in Agriculture*, 5, pp. 1–19.