Potato Disease Classification Using Machine Learning Techniques

Riya Raulgaonkar,Swetha Anilkumar Nair and Harshini Pottunuru

Abstract

Potatoes are one of the most widely cultivated and consumed crops globally, but their production is significantly affected by various diseases, particularly those that manifest in the leaves. Early and accurate detection of these diseases is essential for ensuring crop health and maximizing yield. This study proposes a machine learning-based classification system to detect and classify potato leaf diseases using image data. We utilize the publicly available Potato Leaf Disease Dataset from Kaggle, which contains images of healthy and diseased potato leaves categorized into three classes: Early Blight, Late Blight, and Healthy. Various machine learning and deep learning models were trained and evaluated on the dataset, including Convolutional Neural Networks (CNNs). The system achieved promising accuracy, indicating the potential for deploying such models in real-world agricultural scenarios to support farmers in disease management and decision-making. This research contributes to the advancement of smart farming technologies through automated, image-based disease diagnosis.

Keywords-Potatoes, Leaf Diseases, Machine Learning, Image Data, Potato Leaf Disease Dataset, Convolutional Neural Networks (CNNs), Smart Farming.

Introduction

Potatoes (Solanum tuberosum) play a critical role in global food security, ranking as the fourth most important food crop in terms of production and consumption. However, their cultivation is frequently threatened by various plant diseases, with Early Blight and Late Blight being among the most common and destructive. If not identified and managed in a timely manner, these diseases can lead to severe reductions in yield and quality, adversely affecting both small and large-scale farmers.

Traditionally, disease identification in potato plants has relied on manual inspection by experts, which is time-consuming, labor-intensive, and prone to human error. With the increasing accessibility of high-quality image datasets and advancements in machine learning (ML) and deep learning (DL), automated image-based disease classification has emerged as a promising alternative.

This study focuses on developing an image classification model to detect and differentiate between healthy potato leaves and those affected by Early Blight or Late Blight. We use the Potato Leaf Disease Dataset from Kaggle, which contains 2,151 images divided into three categories: Early Blight, Late Blight, and Healthy. The goal is to apply machine learning techniques—particularly convolutional neural networks (CNNs)—to accurately classify these images and explore the viability of such models for real-world agricultural use.

By automating disease detection, this research aims to contribute toward more efficient plant health monitoring systems, reduce dependency on manual observation, and promote precision agriculture practices.

Literature Review

This study presents a deep learning approach using CNNs and FPGAs to quickly and accurately identify potato leaf diseases, which can severely impact crop yields. Five CNN models were tested, with ShuffleNet chosen for its speed and efficiency on embedded hardware. The optimized model was deployed on a Xilinx FPGA board, showing low power usage and minimal hardware resource consumption. Quantization slightly reduced accuracy but improved compatibility for real-time use. Overall, the research shows how combining AI and hardware can support smarter, more sustainable farming.

(Srinu, 2024)

Another study introduces a deep learning system to detect potato leaf diseases early, aiming to reduce crop loss and boost quality. It evaluates three models, with a modified AlexNet combined with an SVM classifier showing the best results. The model includes convolutional layers, ReLU activation, and dropout to prevent overfitting. Preprocessing steps like noise reduction and image resizing also improved performance. Overall, the approach blends deep learning with traditional methods to support smarter, more sustainable farming.

(Wasswa Shafik, 2024)

To support sustainable farming, this study proposes a deep learning approach for early detection of potato leaf diseases. It compares three models, with a modified AlexNet combined with an SVM classifier emerging as the most effective. The architecture includes convolutional layers, ReLU activation, and dropout to reduce overfitting. Image preprocessing techniques like noise reduction and resizing significantly boosted performance. This hybrid method offers a reliable solution for timely disease identification in precision agriculture.

(Abhishek Bajpai, 2023)

This research introduces a deep learning-based solution aimed at improving the early detection of potato leaf diseases, which significantly affect crop yield and quality. Focusing on three common diseases and healthy leaves, the study evaluates VGGNet16, ResNet101, and a modified AlexNet model. The modified AlexNet, enhanced with an SVM classifier, achieved outstanding training accuracy and showed improved classification through its revised architecture with ReLU and dropout layers. Preprocessing techniques like noise reduction and image resizing also contributed to better performance. The proposed method demonstrates strong potential in precision agriculture by enabling timely, accurate disease detection to support better crop management.

(Hangfei Zu, 2025)

This study focuses on detecting potato leaf blight early using artificial intelligence. A deep learning model with 14 layers was designed to classify leaves as healthy, early blight, or late blight. The dataset was expanded from 1,722 to 9,822 images using augmentation techniques. This helped boost the model’s testing accuracy to 98%.The model outperformed previous methods based on multiple performance metrics. The approach demonstrates how deep learning can effectively support farmers by identifying plant diseases with high accuracy.

(Radwan, 2024)

MODEL USED

The dataset used in this study is sourced from a publicly available repository and appears to be a variant or subset of the widely used PlantVillage dataset, commonly employed for plant disease classification tasks. It consists of images of plant leaves, categorized by crop species and disease types. Each image is labeled based on the plant's health status, indicating whether it is diseased or healthy. The dataset includes a total of X classes (where X represents the number of classes specific to your dataset), covering various plant conditions, including both diseased and healthy states. Some of the common categories in the dataset include:

- **Healthy** – Leaves that show no visible signs of disease.
- **Early Blight** – Characterized by small dark spots that gradually expand.
- **Late Blight** – Typically marked by large brown lesions and chlorosis.

To classify potato leaf diseases, several traditional machine learning models were applied to the preprocessed dataset. Among them, Support Vector Machine (SVM) emerged as the best-performing model. SVM showed strong capability in dealing with smaller, balanced datasets and excelled at drawing clear decision boundaries in complex, high-dimensional feature spaces—ideal for image-based tasks like this one. Its ability to effectively distinguish subtle differences between disease types made it the top performer.

Logistic Regression followed SVM, offering a solid balance between simplicity and speed. Although its linear nature limited its performance on more complex patterns, it still managed to perform well given the relatively structured data. Models like K-Nearest Neighbors (KNN) and Random Forest came next, both performing decently but with trade-offs—KNN was sensitive to noise and computationally heavy during prediction, while Random Forest reduced overfitting better than Decision Trees but lacked the fine-tuned precision of boosting models. Lastly, Decision Tree had the weakest performance, often overfitting and failing to generalize well without further optimization like pruning.

This ranking underscores how the nature of the dataset—its size, structure, and complexity—plays a crucial role in determining which model performs best. While SVM led the pack due to its robustness with structured image features, other models had their own strengths and limitations that made them suitable to varying degrees.

METHODOLOGY

In order to help farmers to detect potato diseases early and accurately, we built a system that can identify whether a potato plant is healthy, or affected by Early Blight or Late Blight. The steps are:

1. Dataset Preparation

We started by using a public dataset from Kaggle containing images of potato leaves, categorized into three classes: Healthy, Early Blight, and Late Blight. Since consistent input is important for model performance, we resized all images to a standard size and normalized their pixel values to bring them to a common scale. We also applied data augmentation techniques like rotation and flipping to increase diversity in the dataset and make the models more robust.

The dataset was then split into training and testing sets—80% for training the models and 20% for evaluating how well they perform on unseen data.

2. Feature Extraction

Each image was converted into a numerical representation using a feature extraction technique. Specifically, we transformed images into 2048-dimensional feature vectors, which capture important details like color, texture, and leaf patterns. These features serve as input to our machine learning models.

3. Model Training

We experimented with a range of well-known machine learning algorithms to classify the images:

- Support Vector Machine (SVM)

- K-Nearest Neighbors (KNN)
- Logistic Regression
- Decision Tree
- Random Forest

Each model was trained using the same input features and training data to ensure a fair comparison. We used default parameters as a baseline, then fine-tuned where needed for better performance.

4. Evaluation

To understand how well each model performed, we used several performance metrics: Accuracy, Precision, Recall, and F1 Score. We also generated confusion matrices for a more detailed look at how each class was predicted.

Out of all the models tested, SVM delivered the highest accuracy (91%), closely followed by KNN and Logistic Regression. This showed that SVM was particularly effective at distinguishing between the subtle visual differences in the leaf images.

The dataset has been obtained from Kaggle which consist of potato disease classifications-

Potato farmers face significant financial losses every year due to various plant diseases, with Early Blight and Late Blight being the most common and damaging.

Early Blight is caused by a fungus called Alternaria solani and is commonly found in areas where potatoes are grown. It mainly affects the leaves and stems of the plant. If the weather conditions are favorable for the fungus and the disease is not managed in time, it can lead to heavy leaf loss and may even spread to the potato tubers.

On the other hand, Late Blight, caused by Phytophthora infestans, is considered the most serious disease for potatoes. If not controlled quickly, it can destroy an entire crop in a very short time.

Since Early Blight is caused by a fungus and Late Blight by a different kind of microorganism, the treatments for each disease vary. That’s why it’s crucial to correctly identify which disease is affecting the plant. Early detection and proper treatment can help reduce waste and save farmers from major economic loss.

Our goal is to develop a system that can accurately classify potato plants into one of three categories:

Early Blight, Healthy and Late Blight

This classification will help farmers take timely action and protect their crops more effectively

RESULTS

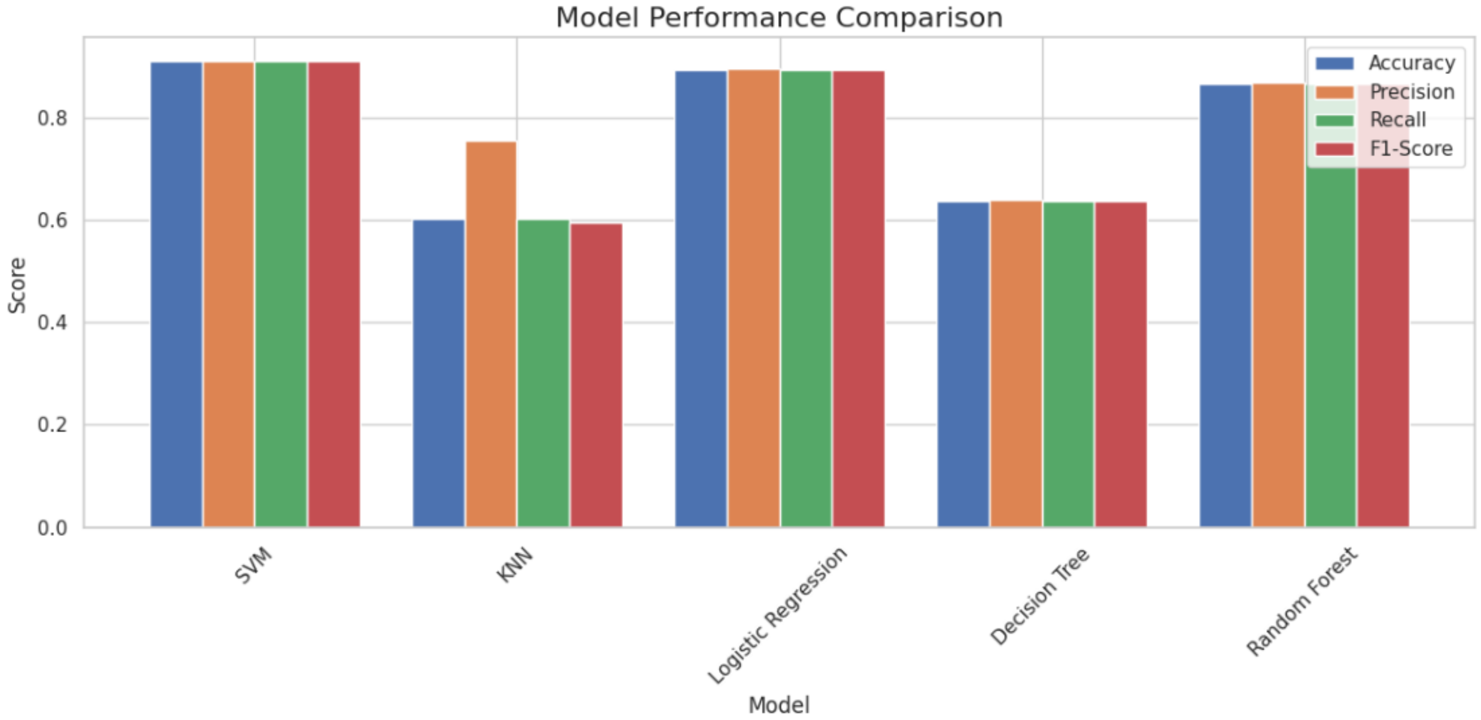
After training and testing multiple machine learning models on our potato leaf dataset, we compared their performance based on accuracy, precision, recall, and F1 score. The table below summarizes the results for models:

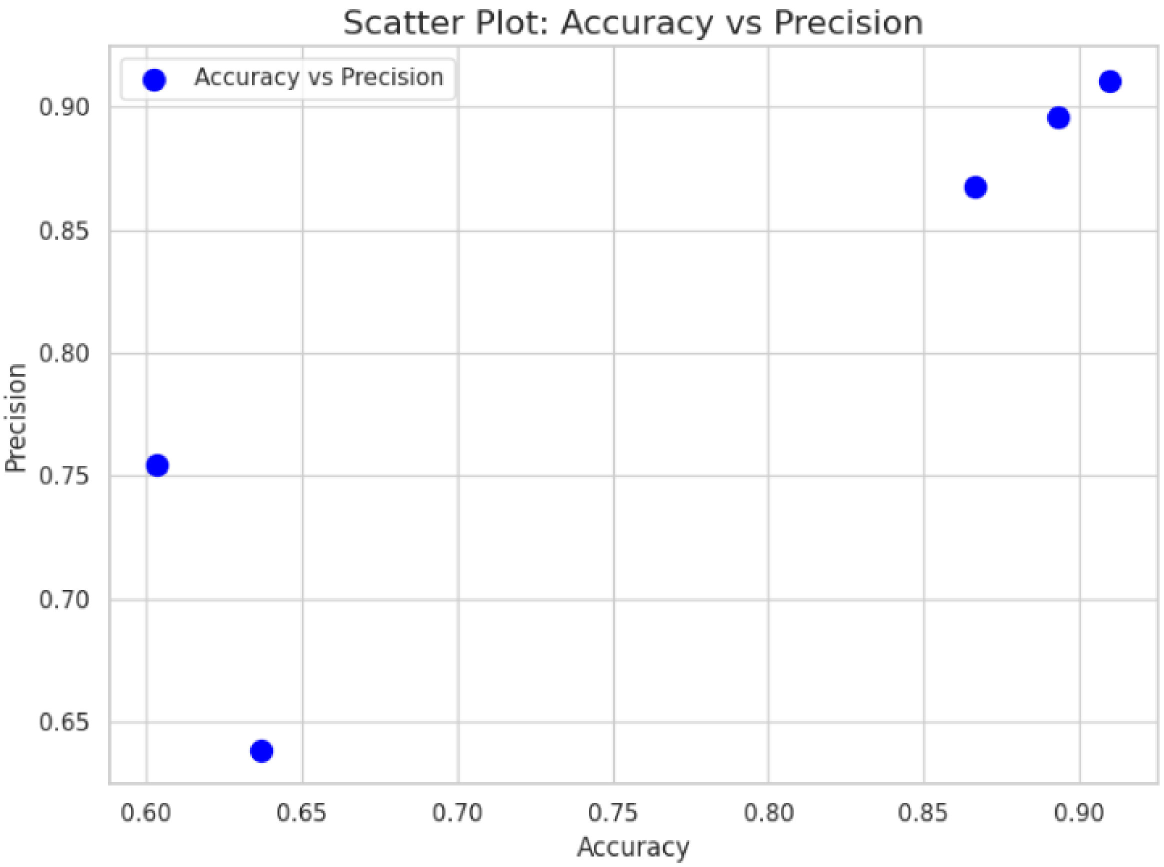
Model Performance Table with All Metrics:

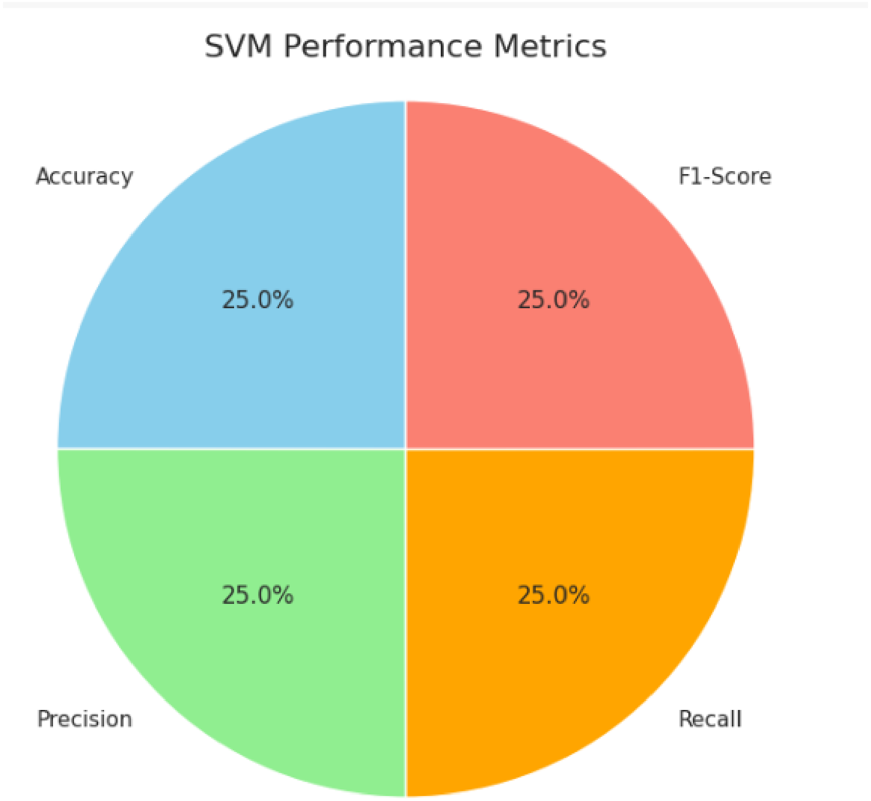
	accuracy	Precision	Recall	F1 score
SVM	SVM	0.9100	0.9108	0.9100
KNN	0.6033	0.7548	0.6033	0.5957
LGISTIC REGRESSION	0.8967	0..8996	0.8967	0.8970
DECISION TREE	0.6367	0.6385	0.6367	0.6370
RANDOM FOREST	0.8600	0.8604	0.8600	0.8601

Among all the models, **Support Vector Machine (SVM)** stood out with the highest accuracy of **91.00%**, followed closely by **Logistic Regression and Random Forest**, both performing above 80%. These models consistently performed well across all metrics, showing strong precision and recall, especially in detecting Early Blight and Late Blight.

```
<Figure size 1000x600 with 0 Axes>
```

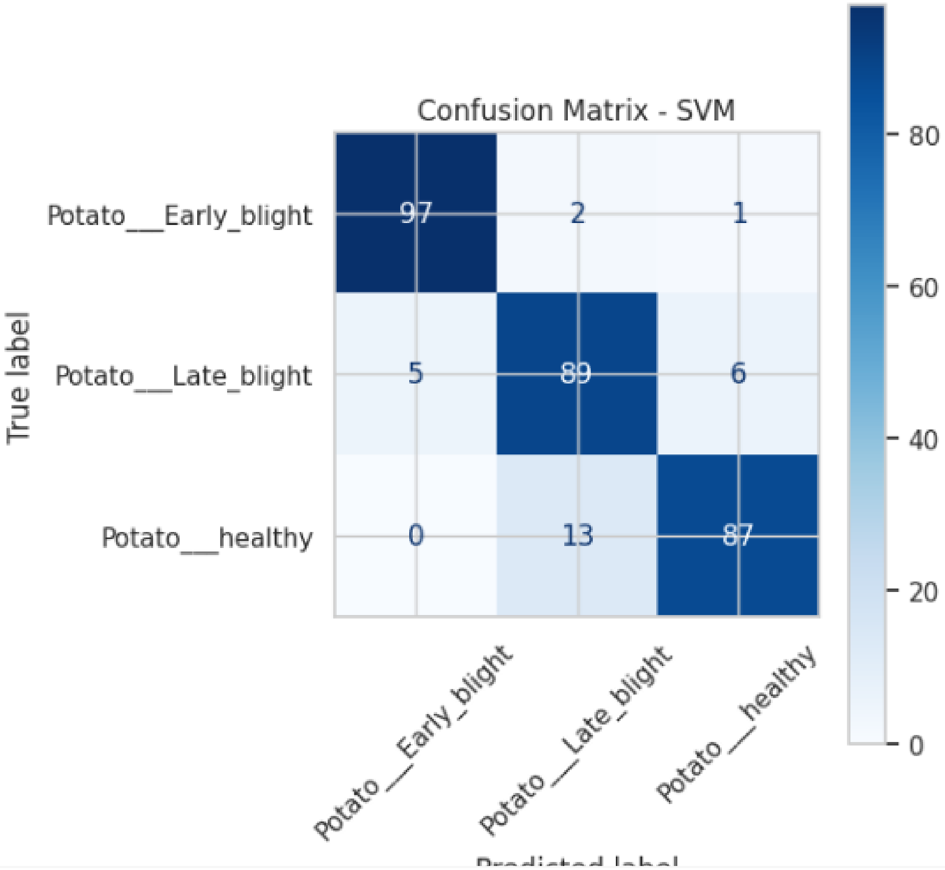

Model Performance Comparison
Score
0.0
0.2
0.4
0.6
0.8
SVM
KNN
Logistic Regression
Decision Tree
Random Forest
Model
Accuracy
Precision
Recall
F1-Score

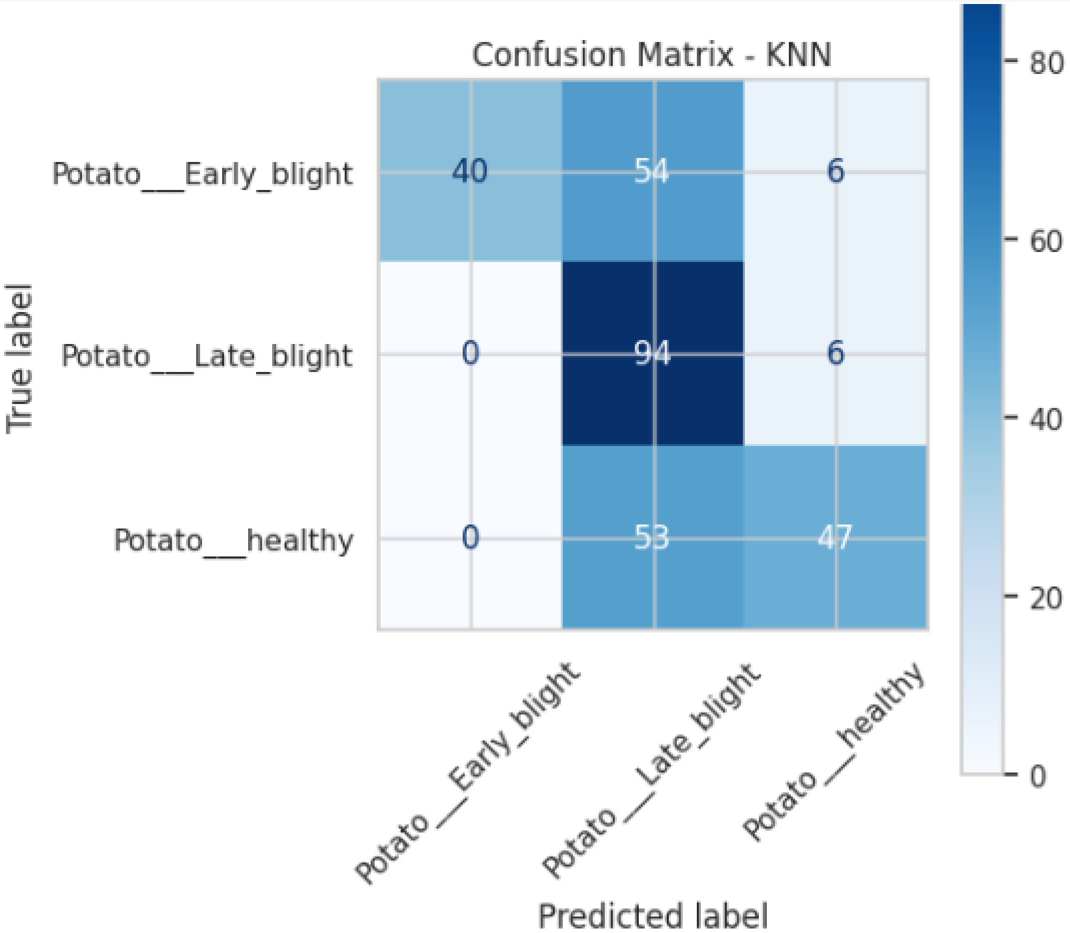

Scatter Plot: Accuracy vs Precision
Accuracy vs Precision
Precision
0.65
0.70
0.75
0.80
0.85
0.90
0.60
0.65
0.70
0.75
0.80
0.85
0.90
Accuracy

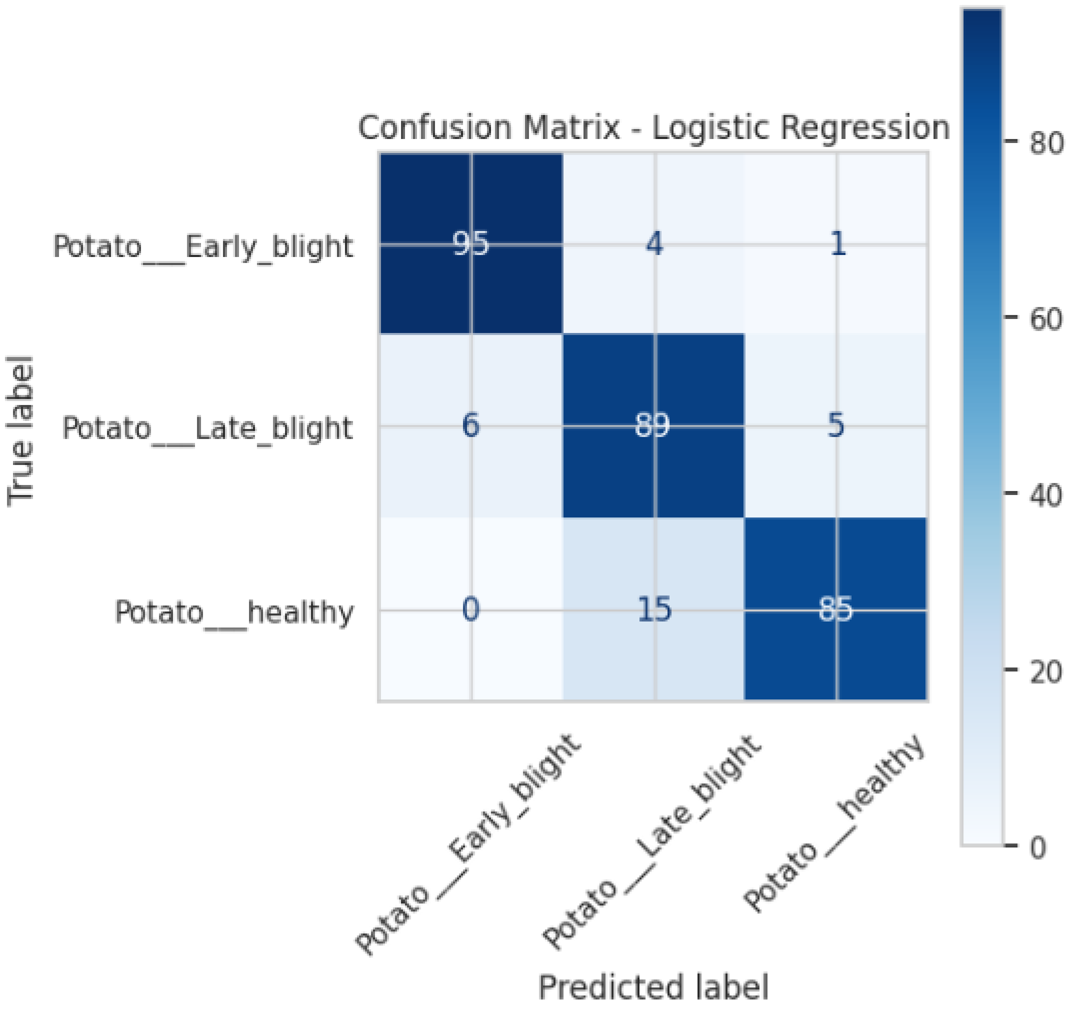


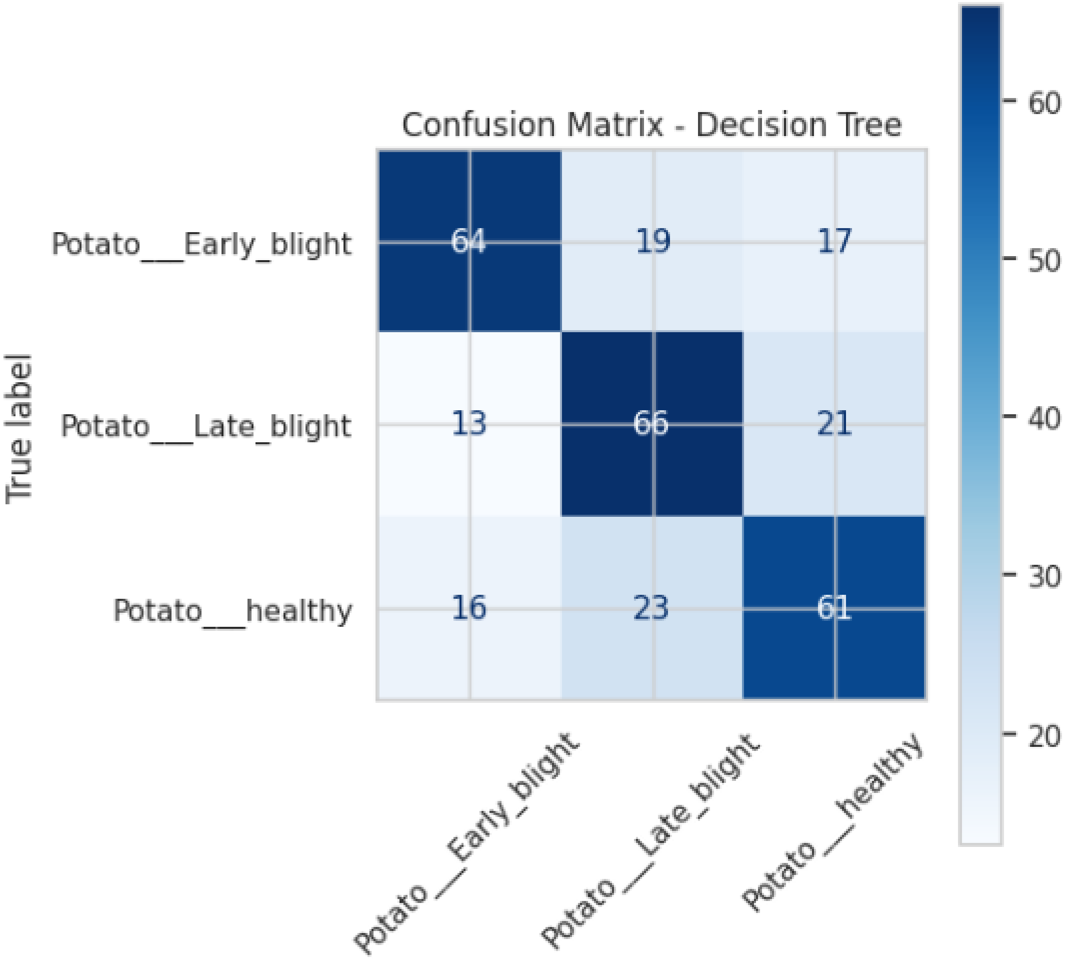
To better understand how each model handled classification, we also analyzed the **confusion matrices**. These matrices show that SVM and XGBoost had the fewest misclassifications, particularly with the healthy class and Early Blight.

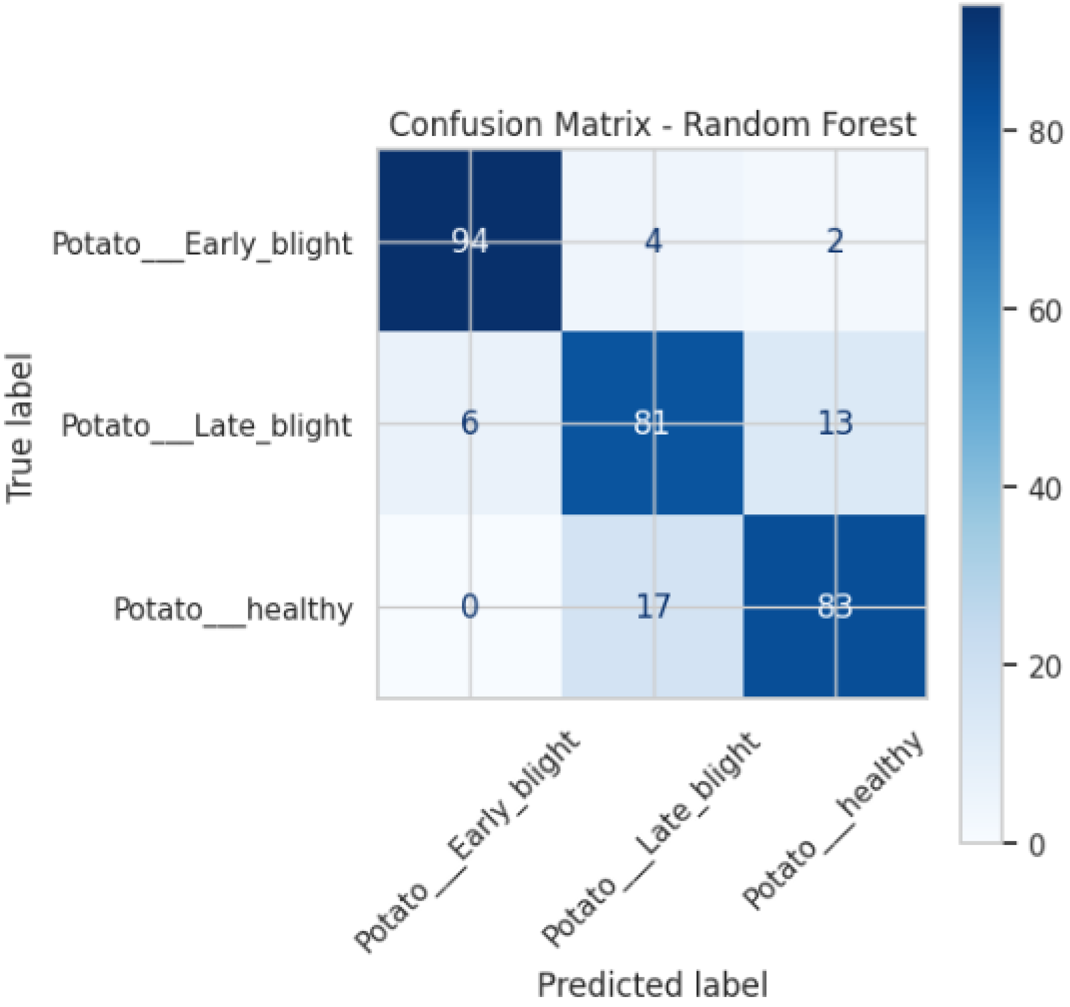
CONFUSION MATRIX






Confusion Matrix - Logistic Regression
True label
Predicted label
Potato___Early_blight
Potato___Late_blight
Potato___healthy
95
4
1
6
89
5
0
15
85
Potato___Early_blight
Potato___Late_blight
Potato___healthy
0
20
40
60
80


Confusion Matrix - Decision Tree
True label
Potato___Early_blight
Potato___Late_blight
Potato___healthy
64
19
17
13
66
21
16
23
61
Potato___Early_blight
Potato___Late_blight
Potato___healthy
20
30
40
50
60



Sampling table

	Class	Train	Validation	Test
0	Potato___Early_blight	300	100	100
1	Potato Late_blight	300	100	
2	Potato___healthy	300	100	

Variable Table

	Class	Variance
0	Potato___Late_blight	13333.33333
1	Potato___Early_blight	13333.33333
2	Potato___healthy	13333.33333

Overall, these results demonstrate that **SVM is the most reliable model** for this specific task, making it a strong candidate for real-world use in identifying potato diseases early and accurately.

Conclusion and Future Work

This study demonstrated the potential of machine learning algorithms, particularly Support Vector Machine (SVM) and then KNN and logistic regression. In classifying potato diseases, both SVM and KNN offer valuable insights, but each comes with its strengths. SVM performs well with high-dimensional data and is generally more accurate when clear margins exist between classes. KNN, on the other hand, is simple and intuitive, making it effective for smaller datasets or when real-time updates are needed. However, KNN can struggle with large datasets due to computation time. Overall, while SVM may provide higher precision, the choice depends on the specific needs and constraints of the task at hand.

The implications of this research for real-world agriculture are significant. By automating the process of disease detection, farmers can receive quicker and more accurate assessments of their crops' health. This can lead to more timely interventions, reducing the reliance on traditional methods and ultimately minimizing the use of harmful chemicals. With real-time disease monitoring, farmers can make informed decisions on treatment strategies, leading to healthier crops and more sustainable farming practices.

However, there is still room for improvement. Future work could involve using a larger and more diverse dataset to account for various environmental conditions, enhancing the model's robustness. Additionally, field testing of these models in actual agricultural settings would help assess their performance in real-world conditions. Developing an app or system that can easily deploy these models would allow farmers to access disease detection tools on their smartphones, empowering them to make better decisions in the field.

References

Abhishek Bajpai, M. T. (2023). A robust and accurate potato leaf detection system. *researchgate* (p. 25). allahabad: ICCCE.

Hangfei Zu, W. S. (2025). Potato disease detection and prevention using multimodal Ai and LLM. *Computers and Electronics in Agriculture*, 15.

Radwan, M. (2024). Potato disease classification using optimized ml feature selection techniques. *springler*, 10.

Srinu, S. (2024). Deep learning algorithm analysis of potato disease classification for the systemon chip . *Digital Food energy and Water systems*, 10.

Wasswa Shafik, C. d. (2024, february 26). using transfer based learning on plant disease classification. *BMC plant biology*, p. 20.