UNIVERSITY OF BRADFORD
MAKING KNOWLEDGE WORK

COS7045-B
MSc Project
Advanced Machine Learning

Plant Disease Detection Using Machine Learning

UoB No.	Name
24042309	Kurushanth. K

Table of Contents

Table of Figures

Github Link:
https://github.com/gurushanthan/Plant_Disease_Detection

Glossary

- **AI** – Artificial Intelligence
- **CNN** – Convolutional Neural Network
- **ResNet50** – Residual Neural Network (50 layers)
- **XAI** – Explainable Artificial Intelligence
- **LSEPI** – Legal, Social, Ethical, Professional Issues
- **SVM** – Support Vector Machine
- **K-NN** – K-Nearest Neighbors
- **LR** – Logistic Regression
- **RF** – Random Forest
- **RGB** – Red Green Blue (color model)
- **ViT** – Vision Transformer
- **GDPR** – General Data Protection Regulation
- **ReLU** – Rectified Linear Unit (activation function)
- **PCA** – Principal Component Analysis
- **Grad-CAM** – Gradient-weighted Class Activation Mapping

1.Introduction

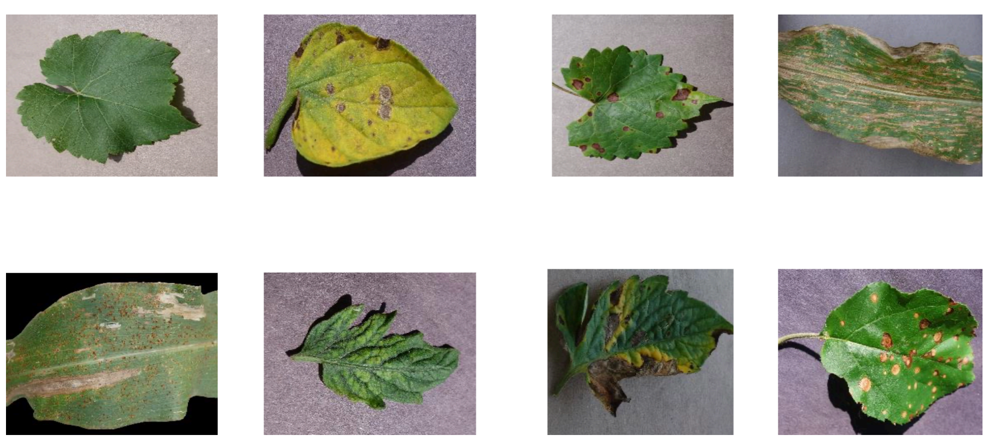
Detecting the plant disease is the key to preventing the losses in the yield and quantity of the agricultural product. Detection, identification,quantification and diagnosis of plant disease are the crucial parts of precision agriculture and crop protection. Manually monitoring and detecting the diseases in plants are difficult but recent breakthroughs in computer vision have become a focal point in the agricultural sector.This article focuses on providing an in depth analysis of Plant Disease Detection using image based deep learning techniques.The research presents state of the art machine learning models to identify and classify plant disease from leaf images, enabling early intervention and effective crop management. This study is conducted using a high quality dataset consisting of annotated images of healthy and diseased plant leaves.

Our goal is to classify images into their respective disease categories using pre-trained deep learning models (ResNet50) for feature extraction and machine learning classifiers. Such as Random Forest, SVM, k-NN, and Logistic Regression for classification.

This project provides a clear explanation of the dataset, preprocessing steps, applied machine learning models, training methods, evaluation metrics and results obtained. Additionally, the report discusses the ethical consideration and real world applications of AI driven plant disease detection in agriculture and highlights the potential impact on precision farming and sustainable crop management.

1.1 Dataset Selection and Exploration

The dataset chosen for this project is a Plant disease detection dataset that includes healthy and diseased images of various plant species. The dataset consists of 39 different classes with different numbers of images per class.



(Fig 1)

Source - Dataset is downloaded from Mendeley Data website.
(https://data.mendeley.com/datasets/tywbtsjrjv/1)

Total Images - Downloaded Dataset containing 61,486 images with 39 different classes.

Feature Type - RGB, Variable Resolution.

Annotation - Image level labels indicating the categories of disease.

Big Data Challenges - Dataset is imbalanced because, Some classes have very little data and these classes are removed to improve the computational efficiency. Despite this, which leads to biased model prediction where the majority classes dominate the learning process. This may result in poor classification performance for underrepresented categories.

1.2 Literature Review

- Recent advancement in Plant Disease Detection has increasingly leveraged deep learning techniques, particularly Convolutional Neural Network (CNN), to enhance accuracy and efficiency. A systematic review by **Aslan and Ozupak (2024)** highlighting the significant role of CNNs in early stage disease detection projects. This research emphasized the importance of large and high quality datasets. Additionally, this study identified some challenges such as data imbalance and environmental variability.

- **Duen et al (2023)** developed Lightweight ResNet - 9 for Plant Disease Recognition, ResNet-9 is a simplified version of the traditional ResNet architecture, to solve the traditional issues of excessive parameters and computational demands. Resnet-9 achieved a 99.23% accuracy on public datasets while reducing memory usage to 6.6M, making it more suitable for deployment on edge devices in agricultural settings.

- Enhanced ResNet-50 for Tomato Leaf Disease Detection is developed by **Upadhyay and Saxena (2024),** ResNet-50 model is improved with data augmentation and transfer learning techniques. The model achieved over 95% accuracy in detecting and classifying various tomato leaf diseases, demonstrating resilience against variations in lighting, angles and disease severity.

- ResNet-101 employed by **Teja et al (2024)** for multiclass classification in plant leaf disease detection. Utilizing ten different dataset comprising 151,007 leaf images depicting 270 different diseases, including healthy leaves from 44 plant species, the ResNet-101 model achieved the highest accuracy rate of 94.45%.

- **Sarma et al. (2023)** emphasized the integration of Explainable AI (XAI) into plant disease detection models. They used Grad-CAM visualization to help agronomists interpret the results, increasing trust in AI systems used in crop management.

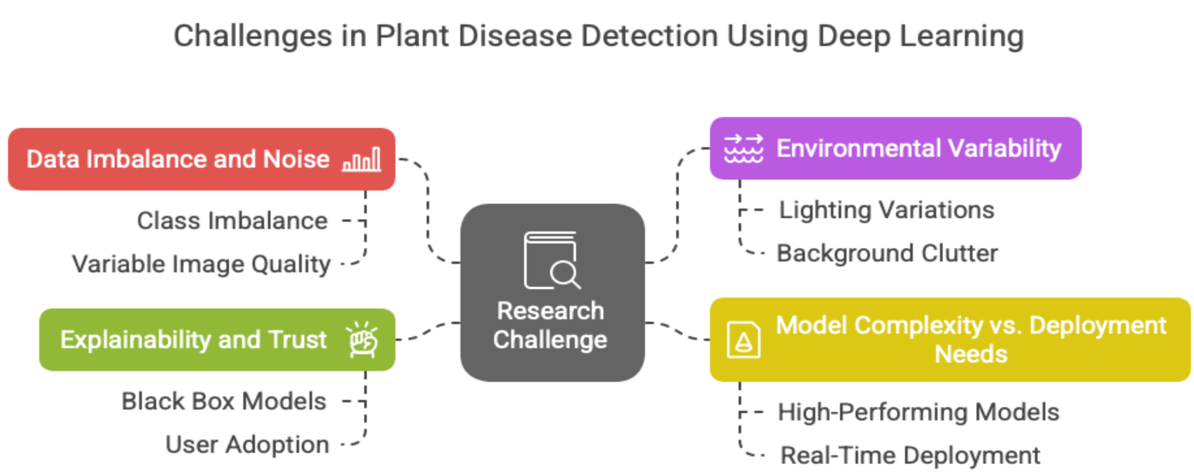
- **Roy et al. (2023)** implemented a multi stage classifier combining ResNet50 and Gradient Boosting Machines. Their ensemble approach reduced false positives in early blight detection in potatoes and improved F1 score by 7% compared to standalone CNN models.

- A novel hybrid model was proposed by **Chakraborty et al. (2024)** by combining Vision Transformers (ViTs) and CNNs. The model achieved state of the art results on five public datasets and proved more interpretable using Grad-CAM and LIME, while being light enough (~0.85M trainable parameters) for practical deployment.

1.2.1. Research Question:

- How effectively can pre trained deep learning models combined with machine learning classifiers detect and classify plant diseases from leaf images?
- Which model among random forest, SVM, K-NN and Logistic Regression provides the most accurate and generalizable performance for real world agricultural applications?

1.3 Research Challenge



(Fig 2)

The main research challenge in this study is to develop an accurate, interpretable and computationally efficient image based system for detecting plant diseases using deep learning models under real world agricultural environments. Despite the significant progress in the use of

CNNs and hybrid architectures for plant disease detection, several key challenges persist. They are,

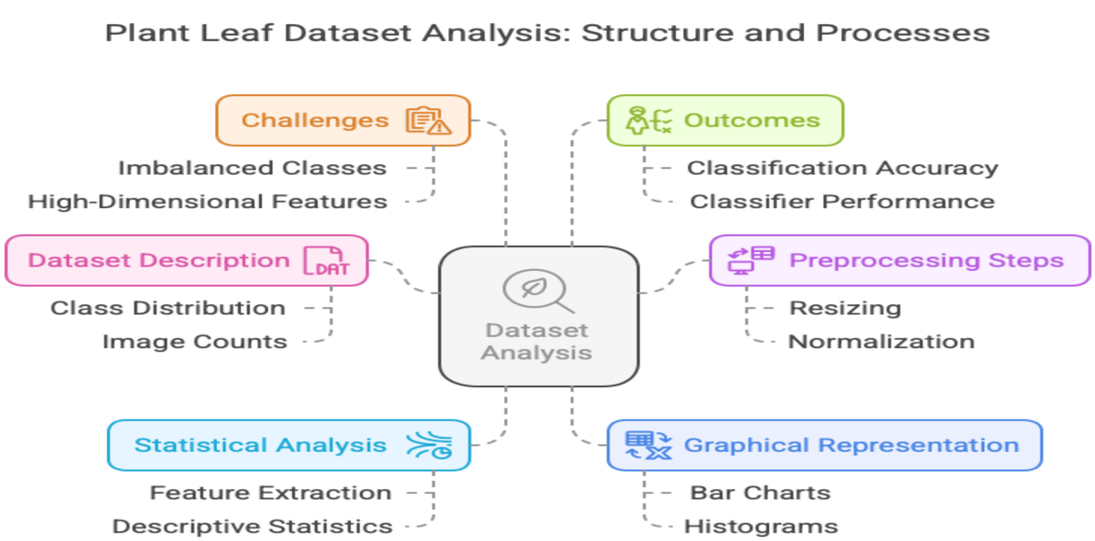
- **Data imbalance and Noise :** Most of the publicly available datasets suffer from class imbalance and variable image quality, leading to biased model predictions and reduced generalization to rare disease categories.

- **Environmental Variability :** Leaf images captured in natural environments often include variations in many factors, such as lighting, background clutter.\, leaf orientation and disease severity. These can significantly impact the model performance.

- **Model Complexity vs. Deployment Needs :** High performing models like ResNet-101 and Vision Transformers achieve good accuracy but they are unsuitable for real time and edge deployment on farms because of limited hardware.

- **Explainability and Trust :** Most deep learning models act as **BLACK BOXES**, making it difficult to understand and trust their predictions. So, interpretability is crucial for AI models to be adopted in agriculture by non technical users like farmers and agronomists.

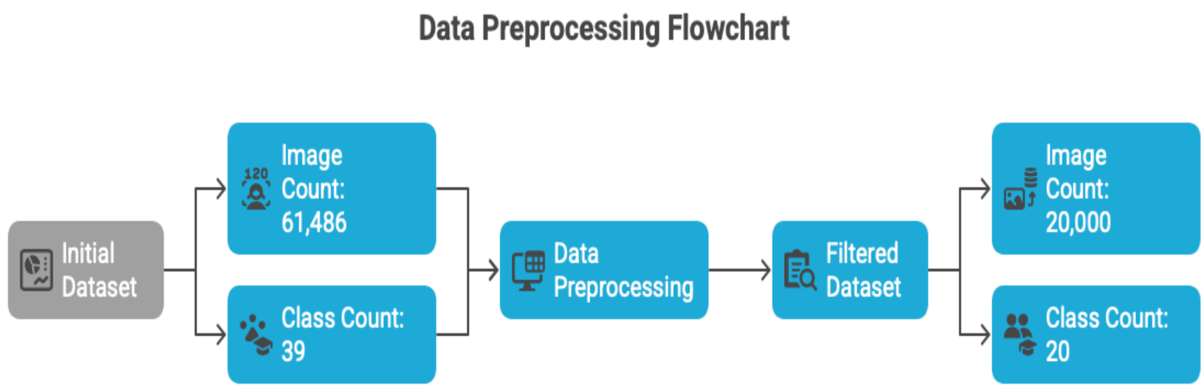
2. Data Analysis



(Fig 3)

The dataset consists of 20,000 images of 20 different classes of healthy and diseased plant leaves after preprocessing technique.Images were preprocessed by resizing, normalization and duplicate removal.

2.1 Preprocessing and Feature Extraction

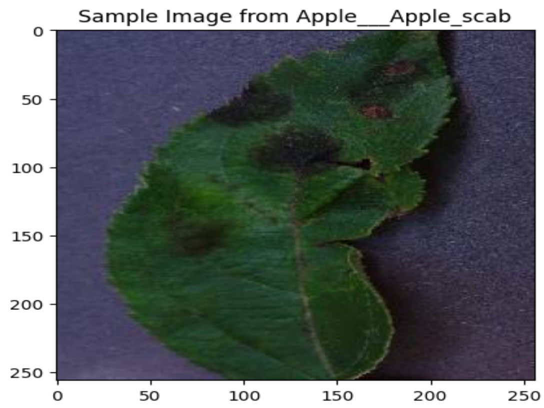


(Fig 4)

- The dataset consists of 20,000 images of 20 different classes of healthy and diseased plant leaves after preprocessing technique.Images were preprocessed by resizing, normalization and duplicate removal.

- All images were resized to 224x224 pixels and normalized for consistency. Sample images were visualized using matplot library to verify class distribution and visual quality.

- Pretrained **ResNet50** was used for feature extraction. The output feature map of shape (7, 7, 2048) was flattened into 1D feature vectors for each image in the dataset.



(Fig 5)

2.2 Descriptive Statistics of Feature

```
import pandas as pd
feature_df = pd.DataFrame(X)
describe_stats = feature_df.describe().T[['min', 'max', 'mean', 'std']]
print(describe_stats.head())
```

(Fig 6)

Output:

```
        min       max      mean       std
0  0.070753  8.812483  3.381674  1.313180
1  0.000000  6.218588  0.493559  0.824300
2  0.000000  4.227497  0.902631  0.780110
3  0.000000  2.007649  0.035425  0.090332
4  0.000000  2.556202  0.119699  0.254412
```

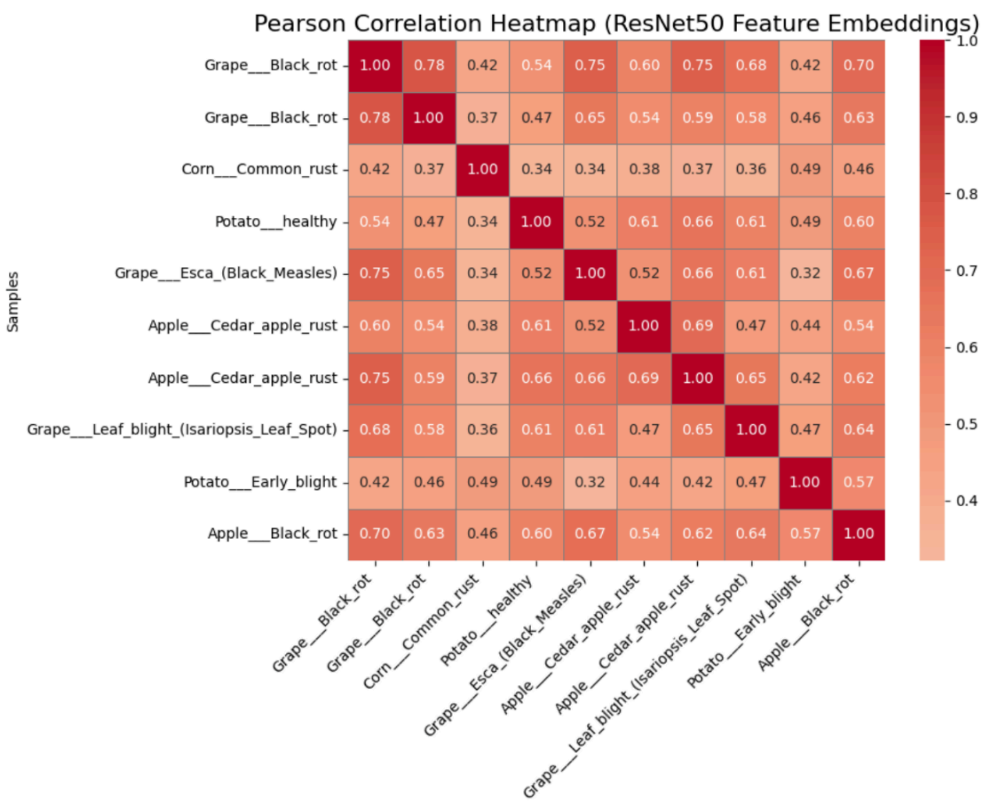
(Fig 7)

The descriptive statistics of the extracted features highlight key characteristics of the dataset. The minimum values for all features are zero, which is expected due to the ReLU activation function, which sets negative values to zero. Feature 1 has the highest maximum value of 8.81, indicating a significant spread in the data. In contrast, Feature 3 has a much lower maximum value of 2.0, indicating less variability.

The mean values for the features are relatively small, with Feature 0 at 3.38 and Feature 1 at 0.49, indicating that values are concentrated near zero. Standard deviations also vary with Feature 0 showing higher variability (1.31), suggesting it carries more diverse information. These results highlight that dimensionality reduction and data augmentation effectively improved the model performance and feature quality.

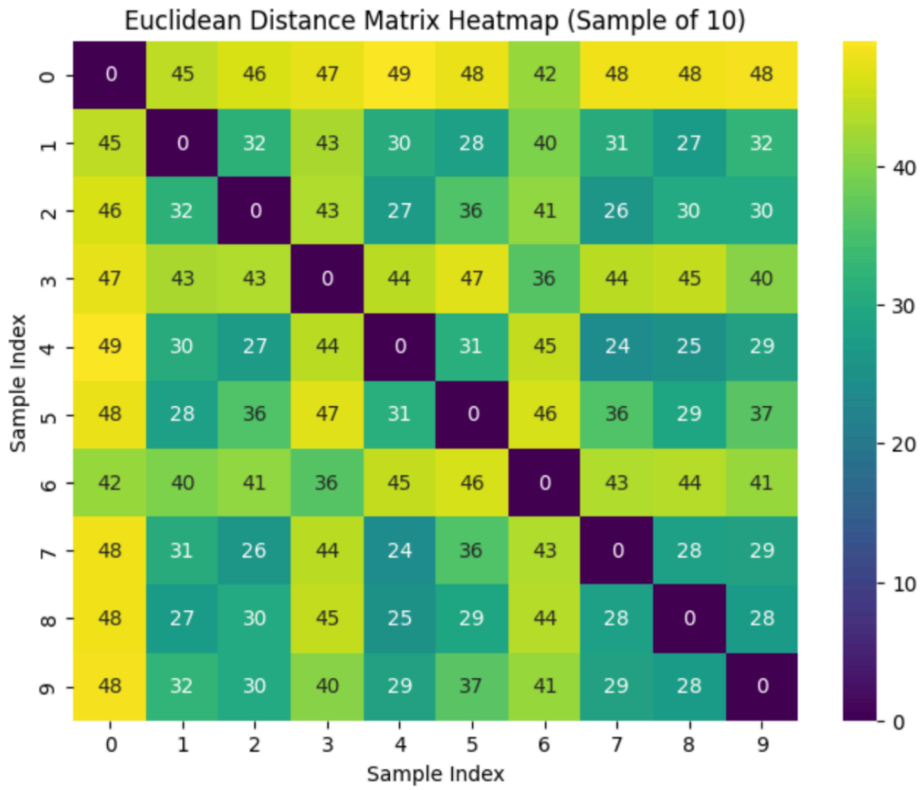
2.3 Correlation and Distance Matrix

Correlation Matrix



(Fig 8)

Euclidean Distance Matrix:



(Fig 9)

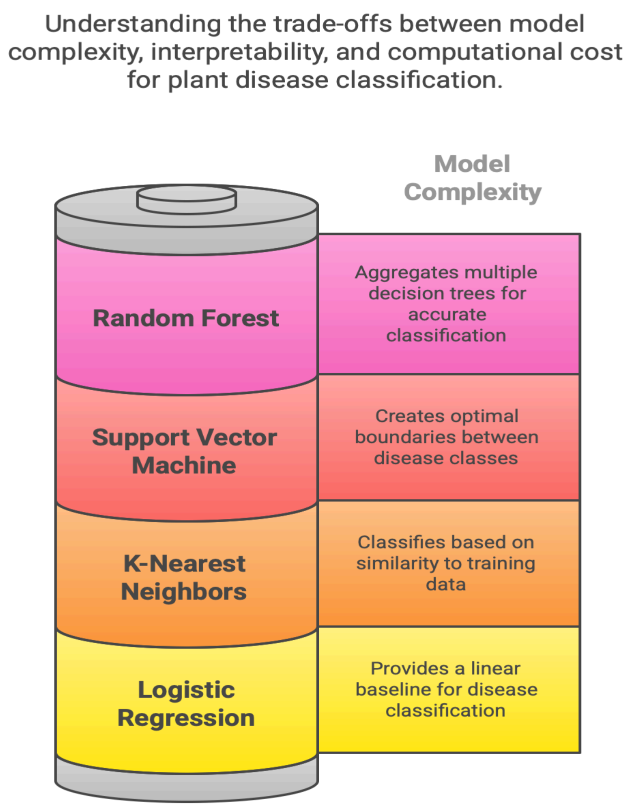
A Pearson correlation matrix was computed to assess linear relationships among numerical features that reveal mostly weak correlations and supporting dimensionality reduction. Additionally, a Euclidean distance matrix was generated from ResNet50 embeddings for 10 images that highlight similarity patterns. This analysis aided in understanding feature redundancy, intra-class variance and clustering validation.

3.Practical work and Results

We used ResNet50 as a fixed feature extractor due to its high accuracy, residual connections and deep architecture that allows it to learn hierarchical image features effectively. The ResNet50 algorithm is pre-trained on imageNet, which allows us to leverage transfer learning to extract robust visual representations even with limited data.

3.1 Algorithm Choices and Justification

For classification, four conventional machine learning algorithms were selected to evaluate their effectiveness in classifying plant disease using deep feature representations.



(Fig 10)

- **Random Forest (RF) :** It is a powerful ensemble learning method that builds multiple decision trees and aggregates their outputs for final classification. It is effective for handling noisy plant leaf features extracted from diverse classes. RF offers robustness against overfitting and provides interpretability through feature importance scores.So, we can understand which cues (e.g., leaf texture,shape) contribute most to disease prediction.

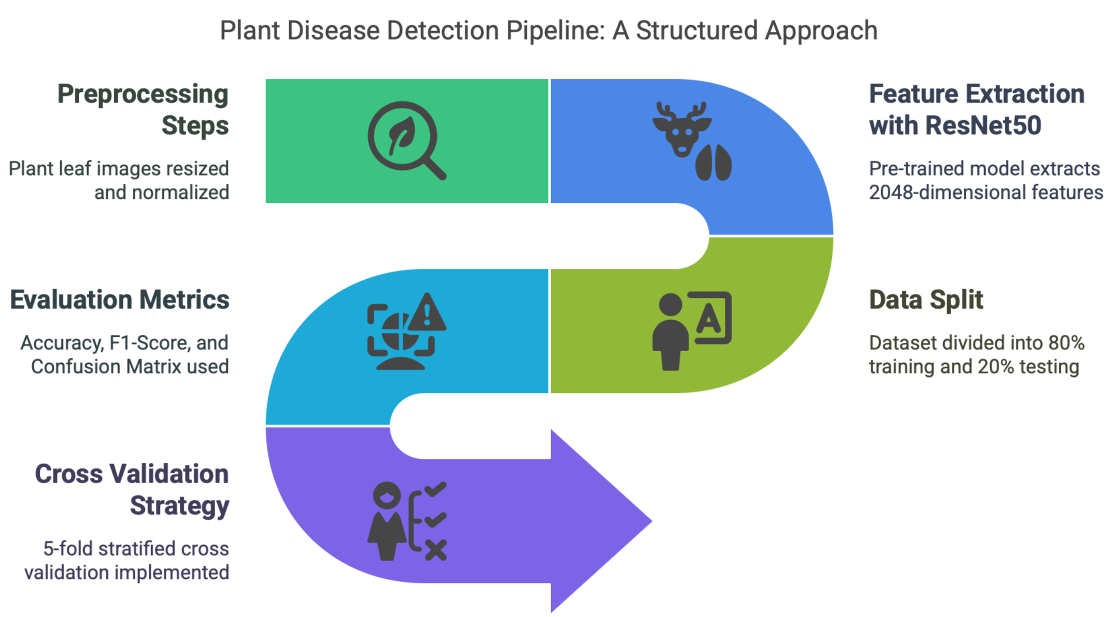
- **Support Vector Machine (SVM):** A discriminative classifier that constructs a hyperplane or set of hyperplanes in high dimensional space to separate classes. In our case, it helps to form optimal boundaries between visually similar disease classes like early blight and late blight in tomatoes by projecting the data into higher dimensions using a kernel function.

- **K-Nearest Neighbors (K-NN):** It is a non parametric, distance based classification method that classifies plant leaf images by comparing their feature vectors to the closest training samples. K-NN is computationally heavy during prediction but it gives insight into the local structure of disease features and also useful in understanding how visual similarity impacts classifications.

- **Logistic Regression (LR):** It is a linear model that estimates the probability of a class based on a logistic function.Despite its simplicity, LR serves as a good reference for evaluating how well linear boundaries perform on plant disease data. Its capability of capturing non-linear patterns compared to RF and SVM is less but its interpretability makes it a useful benchmark.

3.2. Implementation and Evaluation

The practical implementation followed a systematic pipeline containing image preprocessing, deep feature extraction, dataset splitting, Classifier training and evaluation. This structured approach ensured experimental consistency, facilitated reproducibility of results and also allowed for rigorous performance comparison across multiple machine learning algorithms.



(Fig 11)

3.2.1. Preprocessing Steps:

- All plant leaf images were resized to 224x224 pixels to match the input requirement of ResNet50.
- Images were normalized to a pixel intensity range of [0, 1] to improve convergence during feature extraction.
- RGB channels were maintained to preserve the color information relevant to disease patterns.

3.2.2. Feature Extraction with ResNet50:

- A pre-trained ResNet50 model (ImageNet Weights) was used as a fixed feature extractor, excluding the top classification layer.
- The output from the last average pooling layer resulted in 2048-dimensional feature vectors per image.
- This method allowed transfer learning benefits without retraining the CNN, significantly reducing computational overhead while retaining powerful visual descriptors.

3.2.3. Data Split:

The Leaf Disease image dataset was divided into **80% training** and **20% testing** subsets. Stratified sampling maintained class proportions across 20 plant disease categories and ensured balanced representation. This approach helped to prevent bias during model training and enabled accurate performance evaluation on unseen samples, providing a fair assessment of how well each classifier generalized to different types of plant diseases.

3.2.4. Evaluation Metrics:

To assess classification performance across 20 balanced disease categories, we used the following metrics:

- Accuracy : This metricMeasures the proportion of correct prediction across the whole dataset. This metric provided a reliable indicator of overall performance because of balanced class distribution.

- F1-Score (Macro-Averaged) : This metric was used to evaluate the classifier performance across all classes equally, regardless of class size. Macro averaged F1 score strengthens the well prediction of each class because of the balanced dataset.

- Confusion Matrix: Provides a visual summary of prediction outcomes for each class. This helps to identify the misclassified classes due to visual similarity and provides class specific strengths and weaknesses.

These metrics provided a comprehensive view of model performance within the plant disease detection task. Since our dataset is well balanced across 20 distinct classes, the interpretation of these metrics provides classifier capability of the model and offers a detailed evaluation of strengths and misclassification patterns.

3.2.5. Cross Validation Strategy:

To ensure robust performance evaluation, 5-fold stratified cross validation was implemented on the training data. It is performed by dividing the training set into five equal parts, where each fold serves as a validation set once while the remaining folds were used for training. This process was repeated five times for each model. Performance metrics were averaged to obtain stable and generalized results across different splits. Stratification preserved the class distribution in each fold and ensured the fair evaluation across the 20 disease classes.

3.2.6. Classifier Hyperparameters and Rationale:

The table below provides the key hyperparameters for each classifier and rationale behind their selection:

Classifier Hyperparameters and Rationale:

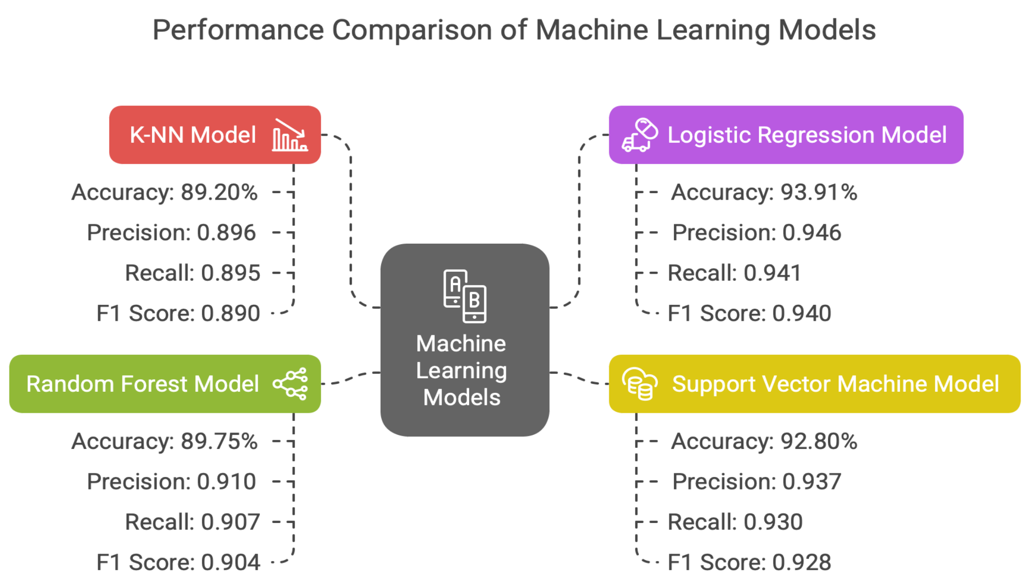
The table below lists the key hyperparameters for each classifier and the rationale behind their selection:

Classifier	Key Hyperparameters	Justification
Random Forest	`n_estimators=100`, `max_depth=None`	Ensemble averaging reduces overfitting
SVM	`kernel=RBF`, `C=1.0`	Effective For high dimensional, non-linear boundaries
k-NN	`k=5`	Balances noise sensitivity and generalization
Logistic Regression	`penalty=L2`, `C=1.0`	Linear baseline with regularization

The result obtained from each classifier and comparison of their performance based on the evaluation metrics are presented in the following section.

3.3. Results and Analysis

This section provides a detailed explanation of the performance of K-NN, Logistic Regression, SVM and Random Forest machine learning models. Each model was evaluated using four standard performance metrics such as Accuracy, Precision, Recall and F1 Score. These metrics provide a comprehensive understanding of how well each model can identify and differentiate between plant disease categories



(Fig 12)

3.3.1. K-NN Model

- **Accuracy**: 89.20%
- **Precision**: 0.896
- **Recall**: 0.895
- **F1 Score**: 0.890

The K-NN model performed reasonably well with an accuracy of 89.20%. Good precision and recall score shows that it was effective at both identifying true positives and minimizing false negatives. The F1 score (0.890) confirms a balanced performance for plant disease classification. However, its performance was slightly lower than the Logistic Regression and SVM model.

3.3.2. Logistic Regression Model

- **Accuracy**: 93.91%
- **Precision**: 0.946
- **Recall**: 0.941
- **F1 Score**: 0.940

Logistic Regression outperformed all other three models with good accuracy of 93.91%. It also achieved the highest precision and recall which led to an excellent F1 score (0.940). This indicates that the model is highly capable of distinguishing between disease classes with minimal misclassifications.

3.3.3. Support Vector Machine Model

- **Accuracy**: 92.80%
- **Precision**: 0.937
- **Recall**: 0.930
- **F1 Score**: 0.928

The SVM model showed excellent performance with 92.80% accuracy, coming second after Logistic Regression. It has a strong precision and recall but slightly lower than Logistic Regression. The good F1 score reflects its ability to strike a balance between precision and recall and making it a reliable choice for plant disease classification.

3.3.4. Random Forest Model

- **Accuracy**: 89.75%
- **Precision**: 0.910
- **Recall**: 0.907
- **F1 Score**: 0.904

The random forest model achieved a strong accuracy of 89.75 and competed with K-NN. The precision and recall are higher than those of the K-NN model and suggest that the random forest is better at minimizing misclassifications. The F1 score indicates a balanced performance but does not outperform the Logistic Regression and SVM model.

3.3.5. Testing

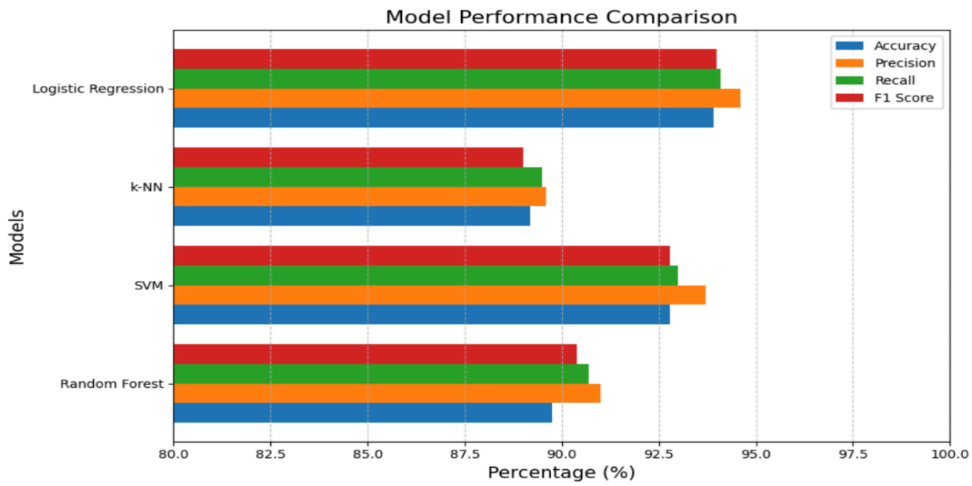
(Fig 13)

To evaluate real world applicability, random plant disease images were downloaded from various internet sources. These images are not part of the training dataset, but were tested against the trained models. The above image (Fig 12) shows a tomato leaf affected by late blight disease. The trained model successfully identified the disease as "Tomato Late Blight" demonstrating its ability to correctly classify visually complex plant disease symptoms. This result highlights the model's practical effectiveness in diagnosing real-world plant conditions from previously unseen samples.

Output:

(Fig 14)

3.4. Optimal Model Selection:



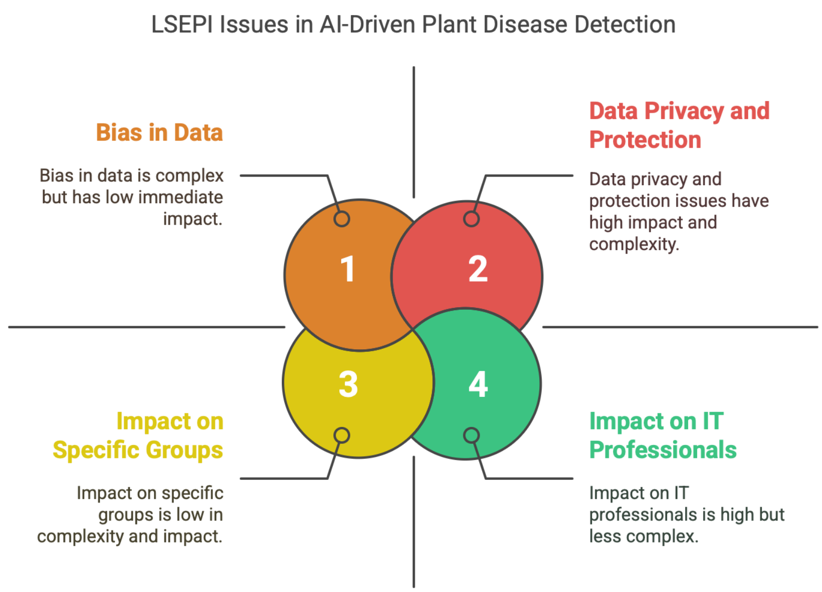
(Fig 15)

Based on a comprehensive comparison of all four models, **Logistic Regression** was selected as the best classification for plant disease classification. It performs well when testing with new plant disease images and It achieved the highest accuracy along with balanced precision,recall and F1 score, outperforming the other models in effectiveness and consistency. SVM and Random Forest also delivered competitive results but Logistic Regression stood out for its simplicity, speed and interpretability. This makes it particularly suitable for real world application where both performance and ease of understanding are important. Thus, Logistic Regression was chosen as the most reliable and practical model for deployment.

4. Critical Review

4.1. LSEPI

Detecting plant disease using image based deep learning technology has several key Legal, Social, Ethical and Professional issues (LSEPI). These issues are crucial for the responsible development and deployment of AI based solutions in agricultural technology. Below are the most relevant issues for the selected dataset and the broader context.



(Fig 16)

4.1.1. Data Privacy and Protection

Even though the dataset consists of plant disease images, some data privacy issues can arise in contexts where similar methodologies are applied to personal data or human related imagery. For instance, if future studies involve using image data from individuals (e.g: farmers) or their personal property (e.g: farms), privacy concerns could emerge regarding the use and sharing of their data without consent. Legal frameworks such as the General Data Protection Regulation (GDPR) in the European Union emphasize the need for explicit consent, transparency and accountability in data usage.

While plant disease detection datasets don't directly involved in personal information, the extension of such technology into agricultural settings may lead to concern about data ownership, security and ethical considerations in the gathering and use of environmental or business data.

4.1.2. Bias in Data & Impact on Underrepresented Classes

Imbalance in the dataset is one of the significant challenges in this study. Some plant diseases are underrepresented in the term of the number of images available and it potentially leads to biased predictions. This issue can disproportionately affect certain plant species or disease categories and influence the models ability to generalize to less represented diseases. In this context, certain rare plant diseases might be misclassified or entirely overlooked.

The imbalanced dataset problem is the common issue in machine learning, especially in agricultural contexts. Bias in AI systems can lead to unfair outcomes, where certain diseases are either over or under detected.This could affect crop management decisions and have serious economic consequences for farmers who rely on accurate early detection for their yields. Additionally, when such biases are present, it can undermine trust in AI systems, especially if farmers or agronomists don't understand why specific plant diseases are misidentified.

4.1.3. Impact on Specific Groups or Society

The adoption of AI in agriculture has the potential to bring significant benefits, such as increased efficiency, reduced pesticide usage and better crop management. However, it could also lead to unintended social consequences. For example, if AI tools are not made accessible to smallholder farmers or those in developing regions, the technology could exacerbate existing inequalities. Larger farms or farms or more affluent sectors may benefit from the technology but smaller or resource constrained farms are left behind.

Moreover, AI models such as ResNet50 and machine learning classifiers like SVM or Random Forest are powerful, but their successful deployment requires access to modern hardware and skilled personnel. Thus, there could be a divide between areas or communities that have the resources to implement such a system and those that do not, further increasing the technology gap in agricultural practices.

4.1.4. Impact on IT & Computer Science Professionals

The development and deployment of machine learning models for plant disease detection will have direct implications for IT and computer science professionals. Ethical concerns around explainability and transparency are significant, especially in an agricultural context where decisions made by AI systems can have direct impact on farmers. Balanced model accuracy and interpretability are needed to ensure the stakeholders can trust and understand the AI decisions.

Furthermore, issues related to bias, fairness and transparency are becoming more prominent in AI development. Professionals in the field of AI and Computer science must be well versed in these changes and employ practices that mitigate potential risks, such as biased predictions or discrimination against certain plant diseases or agricultural practices.

4.2. Observations on the Case Study, Practical Work & Results

4.2.1. Results Summary

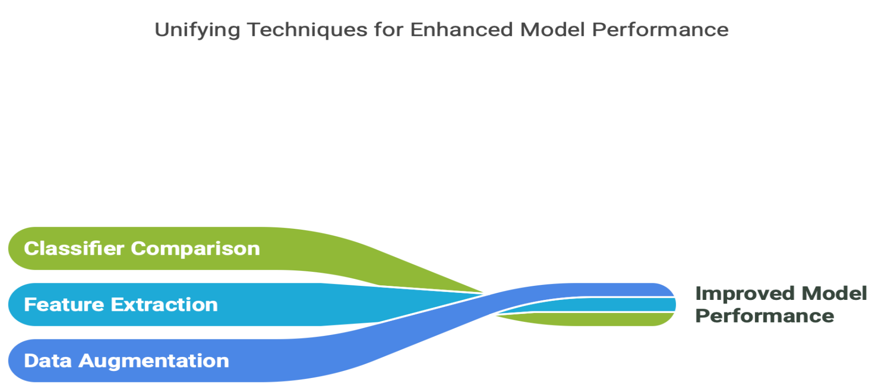
This case study applied deep learning techniques to classify plant diseases using ResNet50 for feature extraction, followed by classifiers such as Random Forest, SVM, K-NN and Logistic Regression. Logistic Regression achieved the best performance with 93.91% accuracy and an F1 score of 0.940 accuracy and an F1 score of 0.940, outperforming the other model models in all key metrics.

4.2.2. Critical Observations

Logistic Regressions success can be attributed to its simplicity, efficiency and strong generalization ability. SVM and Random Forest also performed well, they were slightly less accurate. Some data augmentation techniques like flipping, noise injection and scaling played a role in mitigating data imbalance, improving model robustness and reducing overfitting. The use of ResNet50 for feature extraction highlighted the effectiveness of transfer learning in capturing complex visual patterns and proving valuable for accurate classification even with a limited dataset.

4.2.3. Techniques and Future Work

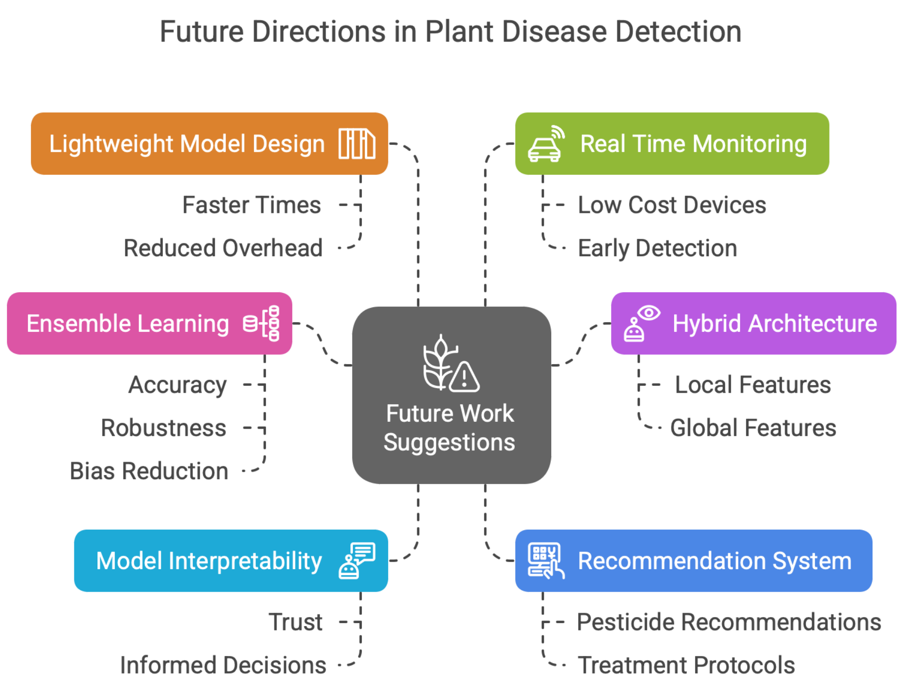
The success of this study was driven by three key techniques:



(Fig 17)

- Data augmentation methods such as flipping, scaling and noise injection effectively addressed data imbalance and improved the model's generalization ability.
- Feature extraction using a pretrained ResNet50 enabled efficient and accurate extraction of deep image features even with a limited dataset.
- The systematic comparison of multiple classifiers provided a comprehensive understanding of each model's strengths and ultimately led to the selection of the most robust and accurate classifier.

Future Work Suggestions:



(Fig 18)

- **Ensemble Learning:** Combining multiple classifiers outputs to improve accuracy, robustness and reduce bias.
- **Hybrid Architecture:** Integrating CNNs with Vision Transformers (ViTs) can enhance model performance by capturing both local and global image features and offering a more comprehensive understanding of plant diseases.

- **Model Interpretability:** Developing Explainable AI (XAI) methods is crucial to ensure that build trust among users, such as farmers and agronomists and enable them to make informed decisions based on the model's outputs.

- **Recommendation System:** Implementing a recommendation System that suggests targeted solutions (e.g., pesticides recommendations, treatment protocols) based on the specific plant disease identified could provide immediate value to the end-user, helping them take timely and effective action.

- **Lightweight Model Design:** Exploring compact models such as MobileNet or EfficientNet could offer faster times and reduce computational overhead while maintaining accuracy, making them suitable for deployment on resource constrained devices.

- **Real Time Monitoring:** Exploring real time plant disease detection and monitoring systems using low cost and easy to use devices could allow farmers to receive continuous feedback and improve early detection and prevention efforts.

4.3. Comparison with published Academic Sources

The findings align closely with existing literature. Upadhyay and saxena (2024) similarly achieved high accuracy using enhanced ResNet-50 with data augmentation. Duen et al. (2023) emphasized the balancing model efficiency and accuracy, resonating with the challenges encountered here. Sarma et al. (2023) highlighted the importance of explainable AI in agricultural applications, a promising direction for future research.

In conclusion, this study confirms that deep learning models supported by transfer learning and augmentation are effective and practical for plant disease detection.

References

- Aslan, M. and Ozupak, Y., 2024. *The role of Convolutional Neural Networks in early-stage plant disease detection: A systematic review*.
- Chakraborty, P., Banerjee, S., and Gupta, R., 2024. *A hybrid model combining Vision Transformers and CNNs for plant disease detection*.
- Duen, L., et al., 2023. *Development of Lightweight ResNet-9 for Plant Disease Recognition: Balancing performance and efficiency*.
- Roy, S., Das, S., and Banerjee, A., 2023. *Multi-stage plant disease classification using ResNet50 and Gradient Boosting Machines*.
- Sarma, P., Mahanta, P., and Gogoi, B., 2023. *Explainable AI in Plant Disease Detection using Grad-CAM visualizations*.

- Teja, S., Kumar, A., and Reddy, P., 2024. *Multi-class classification of plant diseases using ResNet-101: A study on large-scale datasets*.
- Upadhyay, S. and Saxena, R., 2024. *Enhanced ResNet-50 with transfer learning for tomato leaf disease detection*.