Analysis of Leaf Disease Detection in the Solanaceae family plants using Machine Learning Algorithms.

Anitha P   
*Dept of ISE, Ramaiah Institute of Technology*

*Bengaluru-560054, Karnataka, India*

*Affiliated to Visvesvaraya Technological University  
Belagavi-560018, karnataka, India*

[*anitha.palakshappa@gmail.com*](mailto:anitha.palakshappa@gmail.com)

Aditya

*Dept of ISE, Ramaiah Institute of Technology*

*Bengaluru-560054, Karnataka, India*

*adityagiridhar02@gmail.com*

Abhishek Kuthkunja

*Dept of ISE, Ramaiah Institute of Technology*

*Bengaluru-560054, Karnataka, India*

[*abhishekkuthkunja@gmail.com*](mailto:abhishekkuthkunja@gmail.com)

Anvitha

*Dept of ISE, Ramaiah Institute of Technology*

*Bengaluru-560054, Karnataka, India*

[*anvithavpoojari@gmail.com*](mailto:anvithavpoojari@gmail.com)

*Abstract*— The paper aims to classifying and predicting diseases for leaf images using random forest and convolutional neural networks and compare these two methods. The Indian economy is heavily dependent on agricultural productivity to meet the demands of citizens and support farmers. Unfortunately, crop diseases and pests lead to significant reductions in crop production and quality, which underscores the importance of early detection in precision agriculture. However, manually identifying these problems over large areas of the farm is both time-consuming and labor-intensive. The current work is implemented for Solanaceae families such as potatoes, tomatoes, and pepper plants. Initially, the Random Forest classification method was implemented, which achieved an accuracy of 67.90%. Further, the experiments with the CNN model extended resulted in a significant improvement with an accuracy level of 95.53% after training and testing.

Keywords— Agriculture, Labor Intensive, Leaf Disease, Random Forest, Solonaceae.

# Introduction

The study of plant diseases has significantly advanced over the past few decades, leading to a good understanding of the various types of diseases and their categories. However, plant diseases continue to pose a threat to food security, particularly in underdeveloped countries where knowledge of disease control is limited. Poor disease control, Toxic bacteria, and climate change are among the main causes of crop destruction.

Traditionally, plant disease detection has been carried out manually with the traditional method of visual inspection, which is subjective, prone to errors, and time-consuming. Therefore, there is a need for a more precise and effective automated approach to identifying plant diseases.

Advancements in the field of sensor technology, image processing, and machine learning have led to significant improvements in leaf disease detection. Hyperspectral imaging, which captures spectral signatures of plants affected by the disease, was one of the early breakthroughs in the detection of diseases using image processing. More recently, Machine Learning (ML) algorithms have been used to recognize patterns in plant images indicative of disease, enabling rapid and accurate detection and diagnosis.

Furthermore, the development of portable and low-cost sensors has enabled in-field disease detection. This can detect changes in plant physiology, such as changes in leaf color, which are indicative of disease. Given the lack of expertise in identifying plant diseases among young farmers, ML, and classification techniques can be used to aid in spotting specific plant illnesses. New farmers may not be familiar with the characteristics of many plant diseases, and farmers with experience may become less effective at recognizing diseases as they age.

The unification of machine learning, image processing, and sensor technology has led to the development of accurate, efficient, and low-cost plant disease detection systems. These systems have the potential to improve crop yield, reduce pesticide use, and increase food security, particularly in underdeveloped countries.

Leaf disease detection in agriculture can be implemented with many machine learning (ML) algorithms. Some of them include Convolutional Neural Networks (CNN), Support Vector Machines (SVM), Random Forests (RF), Decision Trees, and many more.

SVM is a learning process used under a supervised method for classification tasks, which can detect plant diseases by training on labeled images of healthy and diseased plant parts. SVM can be delicate to the choice of kernel function and factors and may require careful calibration to achieve optimal performance. Additionally, SVM can be computationally expensive and may not scale well to larger datasets.

RF serves as an ensemble technique that can be used for classification tasks and trained on labeled images of healthy and diseased plant parts. RF may be susceptible to overfitting and can be computationally expensive when working with or handling greater-size datasets. RF may not work well with imbalanced data, characterized by an unequal number of sample sizes across various classes.

Decision Trees are effective machine-learning techniques that may be trained on images of healthy and sick plant sections to perform classification tasks. While Decision Trees are easy to understand and interpret, they are more prone to overfitting, especially when the tree is deep. Decision Trees can also be very much sensitive to the choice of splitting criterion, and may not generalize well to new data.

CNN have substantial promise for disease detection in plants. The key advantage of using CNNs for this task is their ability to substantially learn relevant features from images of plant parts, without the need for manual feature engineering.

Because they are made specifically to analyze images, Convolutional Neural Networks are frequently favored over other machine learning techniques for detecting plant diseases. CNN method uses a sequence of convolutional layers for identifying and extracting the information from the images before classifying the image into one or more categories using a set of fully connected layers.

In contrast, other ML techniques like SVM, RF, and Decision Trees are not specifically designed for image analysis. While these algorithms can also be used for plant leaf disease detection by training on labeled, healthy and diseased images of plant parts, they may not be as effective as CNNs in detecting subtle differences in the images that are indicative of disease.

For example, a plant leaf that is infected with a fungal disease may have subtle discolorations or spots that are challenging for humans to detect but are still distinguishable by a CNN. SVMs, RF, and Decision Trees may not be able to detect these subtle differences, leading to lower accuracy in disease detection. In addition, CNNs are capable of learning and improving over time with additional data, whereas other machine learning techniques may have limitations in their ability to adapt and improve with additional data. Further, comparing machine learning techniques for plant disease detection, CNNs are often preferred due to their specialized design for image analysis and ability to detect subtle differences in the images that may not be detectable by other techniques.

# Litertaure Review

The author in the paper proposed a method for the early and fast detection of early blight diseases in leaves of potato using a CNN method. The proposed approach involves preprocessing the images and training the CNN model to accurately classify images into different categories based on the presence or absence of these diseases. The authors used a dataset consisting of images of potato leaves with late blight, early blight, and healthy leaves [1].

The author conducted a comparative analysis of various ML algorithms to identify plant leaf diseases. In their paper, the authors assessed the effectiveness of different algorithms, including Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbor (KNN), and Random Forest (RF), for disease detection in potato, tomato, and apple plants. The dataset used in the study consisted of images of healthy and diseased plants, which were pre-processed before being fed into the models for training and testing. The results showed that RF performance outclassed the other algorithms with respect to accuracy and F1-score, with SVM performing the best in terms of recall. The authors concluded that ML algorithms can be effective in the early detection of plant diseases, potentially leading to improved crop management and reduced economic losses for farmers [2].

The authors investigated diverse methods of feature extraction, which encompassed color-based features, shape-based features, and texture-based features and assessed their performance on different datasets. They also examined several classification algorithms, such as support vector machines (SVM), k-nearest neighbours (KNN), and decision trees. The authors conducted experiments on two distinct datasets, including one with leaf images and another with potato crop images. Performance evaluation was carried out with different metrics like accuracy, precision, recall, and F1-score. The results showed that SVM and KNN classifiers outperformed other techniques in terms of accuracy and F1 score. Additionally, the authors also discussed the study’s limitations and potential future directions, such as the necessity for more comprehensive datasets and the exploration of deep learning techniques for plant disease detection. In summary, this paper provides valuable insights into the comparative performance of various machine-learning techniques for plant disease detection and can serve as a reference for researchers and practitioners working in this filed [3].

Authors proposed the use of the RF classification algorithm for detecting diseases in maize plants. The authors collected a dataset of images of healthy and diseased maize leaves, which were preprocessed before being fed into the RF model for training and testing. The RF model was evaluated based on metrics such as precision, accuracy, recall, and F1-score. The results demonstrated that the RF algorithm achieved high accuracy and F1-score, indicating its effectiveness in detecting maize diseases. The authors concluded that the RF algorithm can be a useful tool for the early detection of maize diseases, potentially leading to improved crop management and reduced economic losses for farmers [4].

The authors did a survey on the use of the RF algorithm for the diagnosis of tomato plant diseases. The authors collected an image dataset of healthy and diseased tomato leaves, which were preprocessed before being fed into the RF model for training and testing. Next, RF model was analyzed in terms of precision, accuracy, recall, and F1-score. The results showed that the RF algorithm achieved high accuracy and F1-score, indicating its effectiveness in diagnosing tomato plant diseases. The authors also compared the performance of the RF algorithm with respect to other algorithms of machine learning, such as SVM and DT, and found that the RF algorithm outperformed the other algorithms in terms of accuracy and F1-score. The authors judge that the RF algorithm can be a useful algorithm for the early diagnosis of tomato plant leaf diseases, potentially leading to improved crop management and reduced economic loss for farmers [5].

A machine learning-based solution is proposed for plant disease detection to discuss the importance of identifying plant diseases early on to prevent crop loss and the traditional methods used for disease detection. They propose a solution that can identify diseases accurately and quickly using ML techniques. The authors describe the proposed method, which includes various steps, such as image preprocessing, image acquisition, feature extraction, and classification techniques using ML algorithms. Further, it involves the results of the experiments conducted and comparison analysis with traditional methods. The authors finally analyzed that their proposed solution can accurately detect plant diseases and can be used to prevent crop loss for farmers. The paper shows valuable insights into the use of machine learning for plant leaf disease detection and can serve as a reference for researchers and practitioners working in this area [6]. The author analyzed 5 steps for identifying leaf diseases in plants and performance measure methodologies applied [7]. The author explains the importance of integrating algorithms of machine learning, image analysis and deep learning to gain better results in the agriculture field [7].

As per the author, deep learning method based on the transfer learning is proposed for plant illness detection. This uses CNN for feature selection and for classification SVM is used [8]. Author introduced the Artificial Intelligence based robots for the identification of diseases [9]. Since, tomato diseases are very frequent, the author implemented random forest algorithm to detect and predict the diseases in tomato leaves [10].

# Methodology

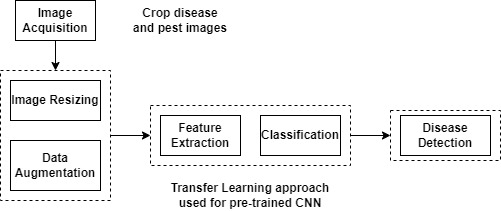


Fig.1. Proposed workflow

Figure 1 explains the methodology adopted in the work. The first step in image classification is to acquire high-quality images. In the current implementation, images were obtained from the Plant Village dataset. Since images are downloaded from the internet, they may have varying resolutions and quality, they were pre-processed before being used directly. In this work, various random augmentations, such as zooming, rotation, brightness and contrast adjustment, and flipping, have been applied to the training images to improve feature extraction accuracy.

Next, a CNN architecture was defined in order to classify the pre-processed images into different categories. CNNs were chosen as they can learn significant features from input images, creating them ideal for image classification tasks. The model is further trained on the training set using an appropriate loss function and optimizer.

To monitor the effectiveness of the CNN model, it was evaluated on the validation set. This step helped to detect overfitting or underfitting of the model on the training data, allowing for necessary adjustments to be made. Finally, the CNN model was tested on the testing set to determine its accuracy in classifying the images into different categories.

Overall, these steps helped to build an accurate and robust image classification model. By following these steps, a robust and accurate image classification model was built. This work can be used to classify images accurately and efficiently, making it an essential tool for leaf disease detection in plants.

# Implementation

The dataset is a public dataset available for leaf disease detection that contains pictures of unhealthy and healthy plant leaves taken from plant village. It was created by researchers at Penn State University, which includes 54,309 images from 14 plant species. Further, the dataset consists of 26 diseases and 43,046 healthy images. The dataset was previously available for download on the official website, it appears to have been removed. From this dataset, the following plant images are used to train both CNN and RF models.

The images were captured in natural conditions, and the dataset includes a diverse range of solonaceae plant species such as tomato, grape, apple, and potato. Each image is accompanied by a label indicating whether it is healthy or diseased leaves, and if diseased, the type of disease.

Preprocessing was a crucial step in detecting diseases in plants using the Plant Village dataset, as it aimed to enhance image quality and reduce noise and variability for improved classification accuracy.

Table 1: Different plants used for implementation

|  |  |  |
| --- | --- | --- |
| No | Type of plant | Type |
| 1 | Pepper | Pepperbell bacterial spot |
| 2 | Pepperbell healthy |
| 3 | Pepper early blight |
| 4 | Potato | Potato early blight |
| 5 | Potato healthy |
| 6 | Potato late blight |
| 7 | Tomato | Tomato target spot |
| 8 | Tomato | Tomato early blight |
| 9 | Tomato healthy |
| 10 | Tomato late blight |
| 11 | Tomato leaf mold |
| 12 | Tomato bacterial spot |
| 13 | Tomato septoria |
| 14 | Tomato mosaic virus |
| 15 | Tomato yellow leaf curl virus |

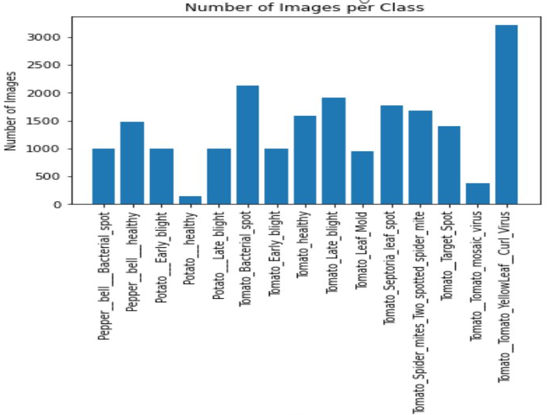


Fig.2. class v/s number of images per class

Techniques for preprocessing the images in identifying the disease detection in plant leaves includes: resizing to a fixed size to reduce computational load and maintain dataset consistency, normalizing images to improve contrast and brightness using methods such as histogram equalization or contrast stretching, converting images to different color spaces such as RGB, HSV, or LAB to facilitate feature extraction for classification.

Using data augmentation techniques to generate additional images from existing ones, including transformations like rotation, flipping, scaling, and shearing, to increase dataset size.

For the implementation purpose, some of the plants under Solanaceae family is considered as shown in table 1. Figure 2 gives number of images for each class. Some of the images from dataset are shown as in figure 3.



Fig.3. Dataset images

Implementation begins with a random forest model to detect plant disease. To train the model, the number of trees to be specified for using in the model. Model accuracy could be improved with the increase in the number of trees active to a certain point, after which it could overfit and decrease the model's performance. To prevent this, we trained the Random Forest model with an increasing number of trees and evaluated its performance using a validation set. Once we determined the optimal number of trees, we trained the final model and made predictions on new data.

To train the RF model, first split the given datasets into training and test sets. Next, created a loop that iterated over a list of different values for the number of trees, creating a new RF classifier with an increasing number of trees for each iteration. The model was fitted to the training data and generated predictions on the test data. The performance of the model has been assessed by comparing the predicted labels with the actual labels of the test data. Following, calculate various metrics of the results, such as precision, accuracy, and recall, to gauge the model's performance.

Based on the plotted graph of the number of trees versus accuracy, it has been observed that increasing the total number of trees in a Random Forest model generally led to an improvement in accuracy. However, there was a point where the performance began to plateau, indicating the presence of overfitting. This highlighted the importance of finding the optimal number of trees to ensure a balance between accuracy and model complexity, ultimately resulting in better performance. After evaluating the performance metrics, we concluded that n=100 was the optimal number of trees that balanced accuracy and model complexity in the RF model for plant leaf disease detection with the standard dataset considered. We trained the final RF model using the entire training set, and, the implementation was able to achieve an accuracy of 67.90% when making predictions on new data.

After determining that the accuracy of the Random Forest (RF) model was relatively low at 67.90%, the experiment is further extended with a CNN model for plant leaf disease detection using the same dataset. The CNN model was a deep learning architecture that used convolutional layers to automatically learn hierarchically features from images.

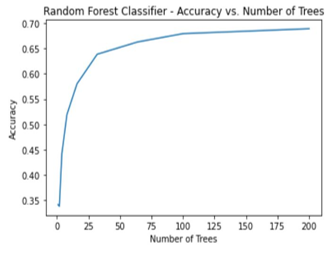


Fig.3. Accuracy v/s number of trees graph

The confusion matrix of the RF classifier is shown in Figure 4.

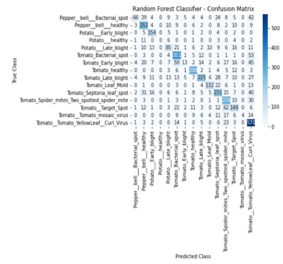


Fig.4. Confusion matrix of Random Forest Classifier

The model consists of several layers that perform different operations on the input images, including resizing, data augmentation, convolution, max pooling, flattening, and dense layers. ReLU is the activation function used by convolution and dense layers. The function of ReLU is to apply the output values, which introduces non-linearity into the model.

Out of various CNN layers, the first layer performs a mathematical function called convolution on the input image. The first layer operation works by sliding a small window termed as kernel over the image and applying the product between the kernel and the correlated pixels in the image. The outcome of this processing is a set of feature maps that represent different types of patterns in the image.

The second layer called MaxPooling follows the convolutional layers and performs a down-sampling operation that decreases the size of the feature maps by keeping only the maximum value in each pooling window. The dense (Dense) layers perform a matrix multiplication between the flattened feature vectors and a set of weights to produce a set of output values. The last layer produces a probability distribution over the classes using the softmax activation function, which is called as dense layer. This probability distribution can be used to predict the class of a given input image. After model is trained, it was evaluated on the testing set to analyze its performance on unseen data. The testing set contained images and their labels that were not used during training.

To overcome the drawback of overfitting, which occurs when the model performs well on the training data but poorly on the testing data, techniques such as cross-validation or early stopping were used. Cross-validation can be defined as method for splitting the data into multiple training and testing sets and analyzing the performance of the work on each set. After varying the parameters with cross-validation, the current work achieved an accuracy of 95.53% compared to RF method..

IV. Experiments and Results

Figure 5 shows the training accuracy and loss, whereas figure 6 shows the accuracy and loss with respect to validation.

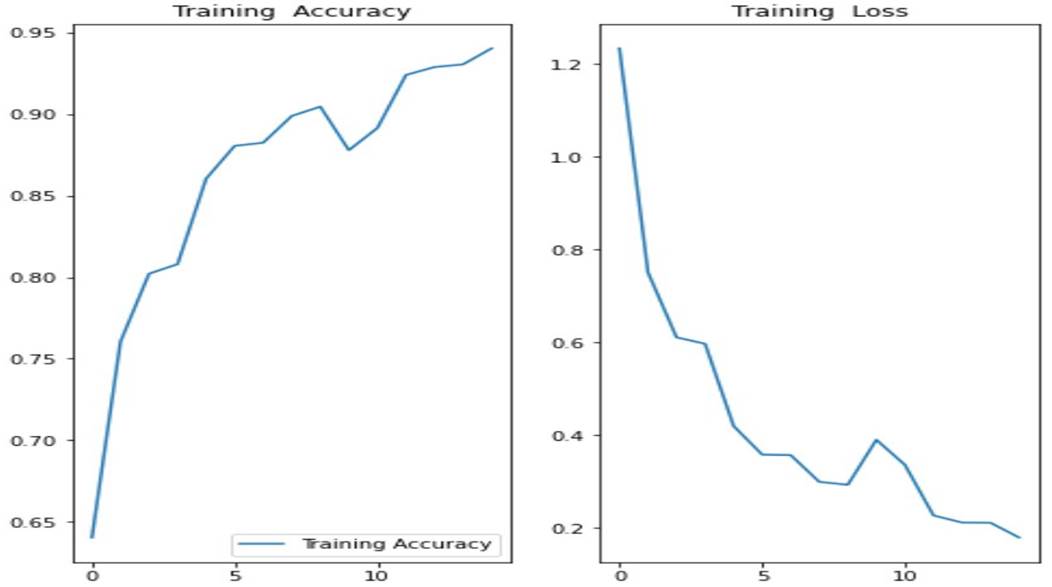


Fig.5. Graph showing training accuracy and loss

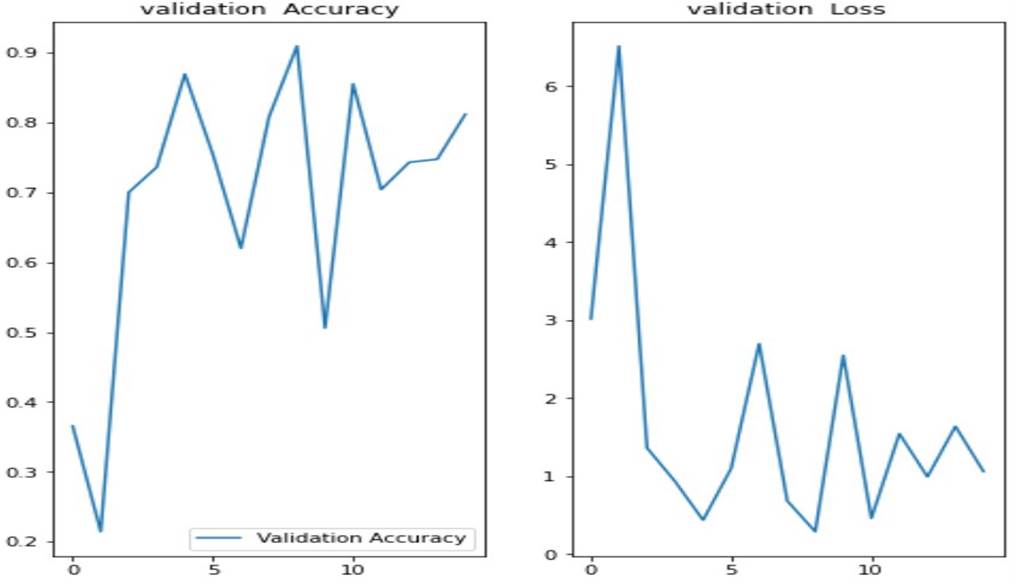


Fig.6. Graph showing validation accuracy and loss

The output of CNN is given in the figure 7. Each image in the figure gives three values – Actual, Predicted, and percentage confidence. It shows whether the leaf is healthy or defective. If it is defective, the percentage of infection in the leaf is shown in confidence labeled under each image in Figure 7.

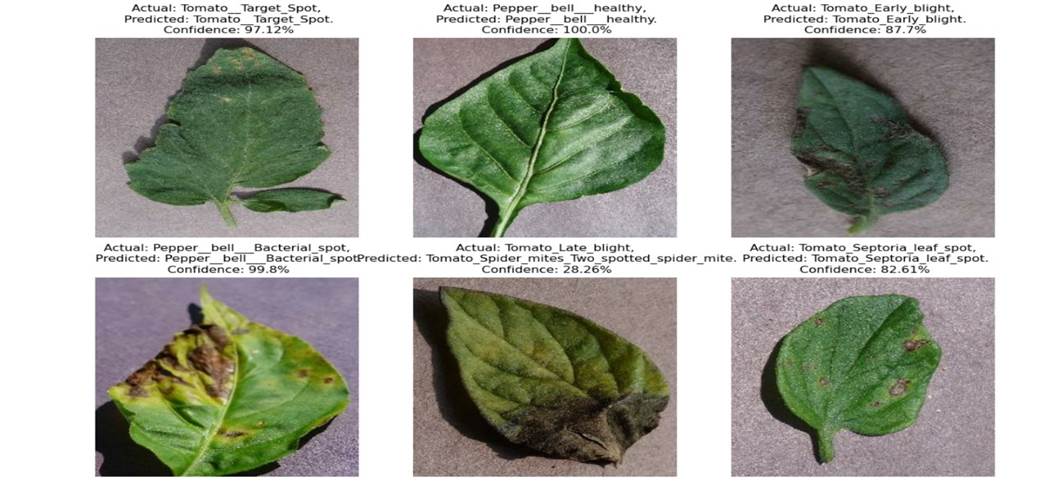


Fig.7. Output of CNN model

The comparative analysis of results obtained from both the methods is given in the figure 8.

Fig.8. Graph showing accuracy of RF and CNN

# Conclusion and Future WORK

The major challenge in agriculture across the country is to identify crop diseases and pests since many farmers lack the necessary expertise to pinpoint which specific diseases or pests are affecting their crops and how to tackle them. Employing automation can enhance farming practices, and the objective of this project is to elevate agricultural productivity by detecting and addressing crop diseases and pests.

In the current implementation, two algorithms are employed for image classification: Random Forest and Convolutional Neural Networks (CNNs). The Random Forest model attained an accuracy of 67.90%, while the CNN model has got an accuracy of 95.53%. It is important to note that CNNs were highly efficient in processing large volumes of data and had been proven to achieve suggestively more accuracy than traditional algorithms like Random Forest. Further, the results in Figure 8 demonstrated the effectiveness of CNNs in plant disease diagnosis and recommended that further studies consider incorporating CNNs for image classification tasks to further enhance accuracy. The model can classify and predict 3 different types of plants namely tomato, potato, and bell pepper. Further, the work can expand the scope of the project to cover a wider range of crops and regions.

Further, the work can be extended by implementing for real-time dataset.

##### References

1. Krishna, K. S., & Narayana, G. V. (2022, September). Early Blight and Late Blight Disease Prediction using CNN for Potato Leaves. In 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA) (pp. 1-6). IEEE.
2. Kumar, S., Jain, A., Shukla, A. P., Singh, S., Raja, R., Rani, S., ... & Masud, M. (2021). A comparative analysis of machine learning algorithms for detection of organic and nonorganic cotton diseases. Mathematical Problems in Engineering, 2021, 1-18.
3. Babu, V. S., Kumar, R. S., & Sunder, R. (2021, March). A comparative study on disease detection of plants using machine learning techniques. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 1937-1941). IEEE.
4. Chauhan, M. D. (2021). Detection of maize disease using random forest classification algorithm. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(9), 715-720
5. Govardhan, M., & Veena, M. B. (2019, October). Diagnosis of tomato plant diseases using random forest. In 2019 Global Conference for Advancement in Technology (GCAT) (pp. 1-5). IEEE.
6. Ramesh, S., Hebbar, R., Niveditha, M., Pooja, R., Shashank, N., & Vinod, P. V. (2018, April). Plant disease detection using machine learning. In 2018 International conference on design innovations for 3Cs compute communicate control (ICDI3C) (pp. 41-45). IEEE.

[7] Vasavi, P., Punitha, A., & Rao, T. V. N. (2022). Crop leaf disease detection and classification using machine learning and deep learning algorithms by visual symptoms: A review. International Journal of Electrical and Computer Engineering, 12(2), 2079.

[8] Varshney, D., Babukhanwala, B., Khan, J., Saxena, D., & Singh, A. K. (2022, May). Plant disease detection using machine learning techniques. In 2022 3rd International Conference for Emerging Technology (INCET) (pp. 1-5). IEEE.

[9] Hidayah, A. N., Radzi, S. A., Razak, N. A., Saad, W. H. M., Wong, Y. C., & Naja, A. A. (2022). Disease Detection of Solanaceous Crops Using Deep Learning for Robot Vision. Journal of Robotics and Control (JRC), 3(6), 790-799.

[10] saji, s., mathew, m., mugesh, a., & lukose, j. machine learning for detection and prediction of tomato leaf diseases.