Plant Disease Prediction Using Machine Learning

EE769 - Course Project, IIT Bombay - April, 2023

Srinivas Bapaiah Naidu Gazula
EE. IIT Bombay
Powai, India
gsbnaidu@ee.iiitb.ac.in

Madhav Prabhakar Nadkarni
EE. IIT Bombay
Powai, India
213070065@iitb.ac.in

Pranav Patel
EE. IIT Bombay
Powai, India
18d070057@iitb.ac.in

***Abstract*—This project explores the use of machine learning and deep learning models for plant disease detection. Four classification techniques, namely L2 logistic regression, random forest, Adaboost decision tree, and voting classifier, were trained and validated using the PlantVillage dataset. In addition, a deep learning model with 12 layers was created to detect plants, and transfer learning was used to generate a disease model. The performance of these models was evaluated using confusion matrices and classification metrics. An user application is also developed which makes use of the models generated. The results showed that the disease model generated using transfer learning outperformed the other traditional machine learning models, with an accuracy of 97% and a F1-score of 0.98.**

***Index Terms*—Resnet-18, Adaboost, Deep Learning, Transfer Learning, F1 Score**

I. Introduction

Plant diseases are a major threat to crop yield and food security, with significant economic and environmental impacts. Traditional methods of disease detection often rely on visual inspection, which can be time-consuming and inaccurate. Machine learning and deep learning models offer a promising alternative, with the potential to accurately and efficiently detect plant diseases based on digital images.

II. Motivation

The motivation behind this project is to develop a reliable and efficient method for plant disease detection using machine learning and deep learning models. By accurately identifying plant diseases at an early stage, farmers can take timely and effective measures to prevent crop loss and reduce the use of harmful pesticides.

III. Background

A. Resnet-18

ResNet-18 [5]is a convolutional neural network architecture that is widely used in deep learning for image classification tasks. The architecture was introduced by Microsoft Research in 2015 and achieved state-of-the-art results on the ImageNet dataset at the time.

ResNet-18 consists of 18 layers, with skip connections that allow for the training of very deep neural networks. The network takes an input image and applies a series of convolutional and pooling layers to extract features from the image. The output of these layers is then fed into a fully connected layer, which produces a probability distribution over the different classes.

One way to use the ResNet-18 architecture for image classification is to extract the features from the network for each image, and then use these features as input to a separate classification model, such as logistic regression, random forest, or a deep neural network. This approach can be useful when the dataset is not large enough to train a deep neural network from scratch, or when transfer learning from a pre-trained model is desired.

B. L2 Logistic Regression

L2 logistic regression is a linear classification technique that uses a logistic function to model the probability of a binary outcome. It involves minimizing the L2 norm of the model coefficients to reduce overfitting.

C. Random Forest

Random forest is an ensemble learning technique that combines multiple decision trees to improve classification accuracy. It randomly selects a subset of features and samples to create each decision tree, and then aggregates the results to make a final prediction.

D. Adaboost Decision Tree

Adaboost decision tree is another ensemble learning technique that combines multiple decision trees, but instead of creating each tree independently, it iteratively adjusts the weights of misclassified samples to improve classification accuracy.

E. Voting Classifier

Voting classifier is an ensemble learning technique that combines multiple classifiers to improve classification accuracy. It aggregates the predictions of each classifier and selects the class with the highest number of votes as the final prediction.

F. Deep Learning Model

Deep learning is a subset of machine learning that uses artificial neural networks to learn complex patterns from data. Deep learning models typically have multiple layers, with each layer learning increasingly abstract features from the input data.

G. Deep Learning vs Traditional Machine Learning

Deep learning and traditional machine learning are two different approaches to solving problems in artificial intelligence. Traditional machine learning typically involves the use of algorithms and statistical models to analyze and make predictions based on data. Deep learning, on the other hand, relies on artificial neural networks that are designed to simulate the human brain's ability to process information.

Deep learning models have shown superior performance in certain applications, such as image and speech recognition, natural language processing, and robotics. However, they require a large amount of data and computing power to train, and can be difficult to interpret and debug.

Traditional machine learning techniques, while less powerful than deep learning in some domains, are often simpler to implement and understand. They can be effective in applications where the data is less complex and the goal is to make predictions based on known patterns in the data.

H. Transfer Learning

Transfer learning is a technique that involves using a pre-trained deep learning model as a starting point for a new task. By fine-tuning the pre-trained model on the new task, transfer learning can significantly reduce the amount of data and training time required for the new model.

IV. Prior Work

Plant Disease prediction was done using machine learning and computer vision where features are extracted using GLCM(Gray Level Co-occurrence Matrix) [3]. A semi-automatic system was designed to detect two diseases of soybean (glycine max) named mosaic virus and leaf spot where the k-means clustering approach was used to extract the combined colour and texture features from the diseased area of soybean leaves and classification was done using the KNN algorithm [4].

V. Experiment Setup

The dataset used in this project is the PlantVillage dataset [2], which contains a total of 88,059 images across 14 plant species and 21 unique disease types. The distribution of training datset is shown in figure-1, figure-2. The distribution of validation datset is shown in figure-3, figure-4. The dataset was preprocessed by resizing the images to 224x224 pixels and extracting the ResNet-18 features from each image using the pre-trained ResNet-18 model available in the torchvision module of PyTorch.

Four different classification techniques were used in this project: L2 logistic regression, random forest, Adaboost decision tree, and voting classifier. For each classification technique, two models were trained: one to detect the type of plant, and another to detect the type of disease. All of the plant and disease models were independent, and there was no transfer of knowledge between them.

In addition to the traditional classification techniques, a deep learning model was also created for plant classification.

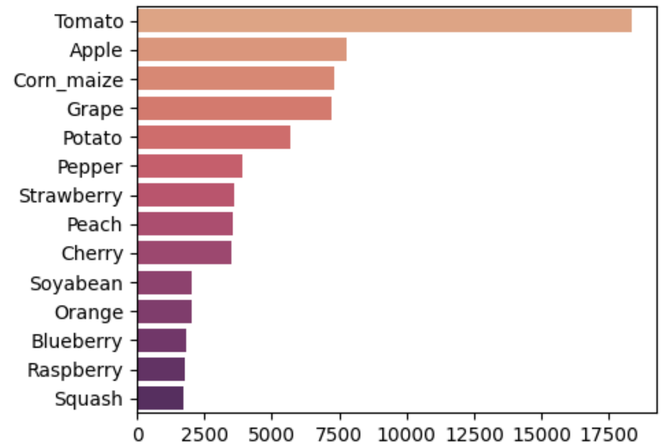

Fig. 1. Distribution of Plants training dataset

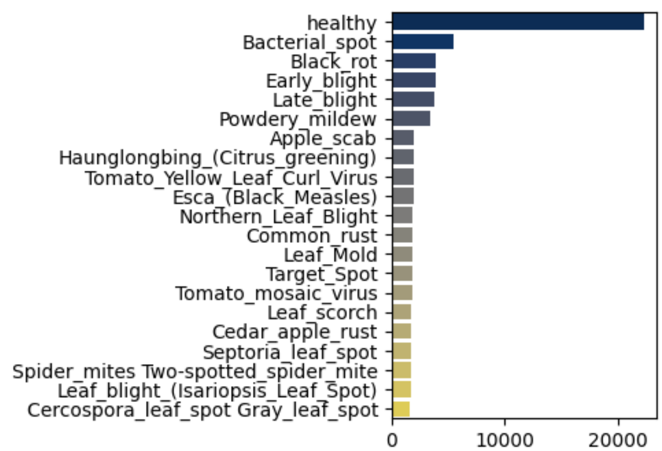

Fig. 2. Distribution of Diseases training dataset

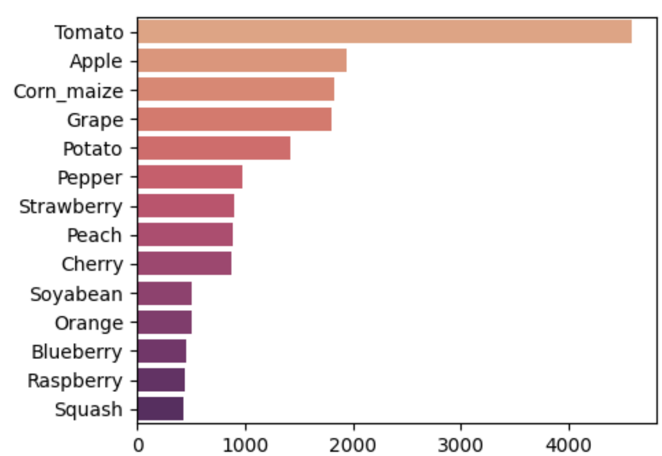

Fig. 3. Distribution of Plants validation dataset

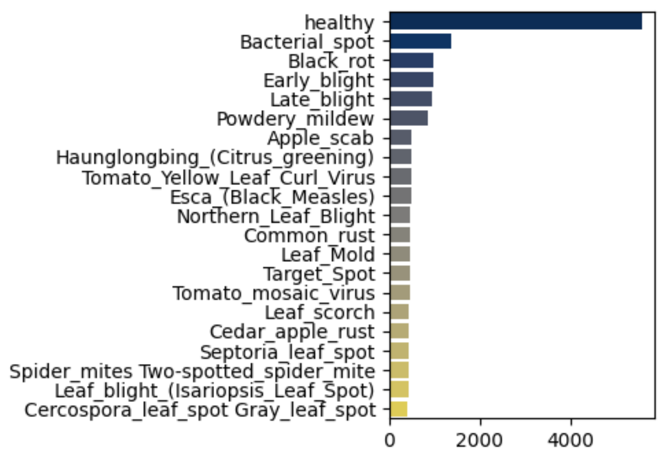

Fig. 4. Distribution of Diseases validation dataset

Furthermore, a separate deep learning model was created for each of 14 plants' diseases using transfer learning, where the penultimate layer of the pre-trained plant model was used as input to the each disease model. The generated 14 disease models were then merged to form a single disease model to detect all diseases. These deep learning models were trained on a GPU for faster training time.

The models were evaluated using several performance metrics, including accuracy, precision, F1-score. GridSearchCV and RandomizedSearchCV were used to optimize hyperparameters for some of the models.

VI. Results

The performance of different models on the PlantVillage validation dataset, testing dataset is summarized in the tables I,II.Based on the tables, we can make the following comparison analysis between the results:

- Classification Accuracy: In traditional ML classifiers, the classification accuracy of the plants model is higher than that of the diseases model for all the traditional classifiers. This indicates that the plant model is better at distinguishing between different types of plants compared to the diseases model in distinguishing between different types of diseases. But in deep learning models, accuracy of disease model is greater than that of plant model. This is because of transfer learning from plants model to diseases model.
- Precision: The precision values of both models vary across different traditional classifiers, but the plant model generally has higher precision values than the disease model. This means that the plant model is better at identifying true positives for each class, compared to the disease model. But it is reverse in case of deep learning models.
- Recall: The recall values of both models also vary across different traditional classifiers, but overall, the plant model has higher recall values than the disease model. This indicates that the plant model is better at correctly identifying positive instances for each class, compared to the disease model. But it is reverse in case of deep learning models.
- F1 Score: The F1 score is a harmonic mean of precision and recall, and gives an overall measure of the classifier's performance. Based on the F1 scores, we can observe that the plant model outperforms the disease model in most cases. The F1 score of the plant model is higher than that of the disease model for all the traditional classifiers. But it is reverse in case of deep learning models.

TABLE I
Results-Validation Dataset

model	*type*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
L2 Logistic Regression	plants	98.54%	99%	99%	99%
	diseases	96.39%	96%	96%	96%
Random Forest Tree	plants	94.91%	95%	95%	95%
	diseases	88.08%	89%	88%	88%
Decision Tree	plants	75.75%	76%	76%	76%
	diseases	68.07%	68%	68%	68%
Adaboost-Decision Tree	plants	90.24%	92%	90%	90%
	diseases	82%	84%	82%	81%
Voting Classifier	plants	95.97%	96%	96%	96%
	diseases	90.39%	91%	91%	90%
Deep Learning	plants	79%	81%	79%	77%
	diseases	97%	97%	97%	97%

All the precision, recall, f1-score values are weighted averages.

TABLE II
Results-Testing Dataset

model	*type*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
L2 Logistic Regression	plants	93.94%	100%	94%	96%
	diseases	91%	100%	91%	95%
Random Forest Tree	plants	90.91%	95%	91%	91%
	diseases	66.67%	84%	67%	70%
Decision Tree	plants	72.73%	91%	73%	77%
	diseases	54.55%	79%	55%	63%
Adaboost-Decision Tree	plants	96.97%	97%	97%	97%
	diseases	75.76%	93%	76%	79%
Voting Classifier	plants	96.97%	97%	97%	97%
	diseases	78.79%	94%	79%	83%
Deep Learning	plants	82%	100%	82%	84%
	diseases	97%	100%	97%	98%

All the precision, recall, f1-score values are weighted averages.

The Accuracy of Decison tree diseases model is very less. Hence Adaboost Decision tree models are introduced to adjust the weights of the misclassified samples accordingly. Hence we can observe a significant improvement in accuracy. But when Random forest tree disease model is added inside Adaboost classifier, there is no much improvement in accuracy.

The results show that deep learning performed well on the diseases classification with high accuracy, precision, recall, and F1-score values, while the traditional machine learning models performed better on the plants classification.

VII. Deployment

The binaries which are generated during the training of models are saved. These binaries are used to perform predictions on new data. The uploaded image will be converted into a data array of Resnet18 features and then this data array is given to the selected model to get the predictions. The deployed models are at [1]. The figure-5 shows how the result appears in the user application.

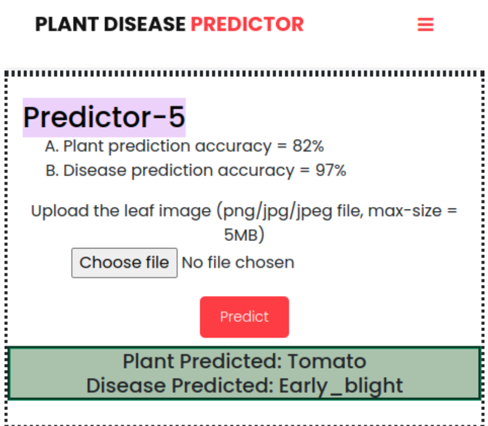

Fig. 5. Output by deployed model in user application [1]

VIII. Conclusion

In this project, we developed different models for plant disease classification using machine learning and deep learning techniques. The deep learning diseases model achieved the highest accuracy, precision, recall, F1-score compared to other disease models. Transfer learning is an effective approach for disease classification, as demonstrated by the high accuracy of the transfer learning model. The deep learning disease model achieved high accuracy and can be used for plant disease diagnosis.

Acknowledgement

We thank Prof. Amit Sethi, the instructor of the course EE769-Introduction to Machine Learning in Spring 2023, for his continuous guidance and support to our project.

References

[1] Plant Disease Predictor. URL: https://www.ee.iitb.ac.in/course/~gsbnaidu/index.php
[2] spMohanty. (2018). PlantVillage-Dataset.
[3] Pranesh Kulkarni, Atharva Karwande, Tejas Kolhe, Soham Kamble, Akshay Joshi, & Medha Wyawahare. (2021). Plant Disease Detection Using Image Processing and Machine Learning.
[4] Venus Patel, Noopur Srivastava, Manish Khare, (2021). Plant Disease Detection Using Image Processing and Machine Learning. In Futuristic Trends in Networks and Computing Technologies (pp. 549–561). Springer Nature Singapore.
[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (pp. 770-778).