Cotton Leaf Disease Detection using transfer Learning and Traditional Classifiers

Diya Sapra
Dept. of Computer Science
Vellore Institute of Technology, Chennai, India
Dia.sapra@gmail.com

ABSTRACT

The agriculture industry's production and food quality are both impacted by plant diseases.The detection of disease at an earlier stage is crucial for raising the quality of agricultural output and stopping the overall plant extinction. In this paper, a Transfer Learning-based automated crop disease recognition system is proposed. The framework being proposed includes two key stages: First, the deep features are retrieved using a a pre-trained Visual Geometry Group (VGG16) convolutional neural networks (CNN) model; Second, The extracted features are plugged into the traditional classifiers as (random forest and SVM) for final Classification of the diseases. For the algorithm's evaluation, a four-class dataset is chosen, with three classes containing diseased cotton leaf images and one class for healthy cotton leaf images. The proposed approach has an average accuracy of approximately 98% (Random Forest) and 99%. (SVM).

Keywords—leaf diseases, CNN, Transfer Learning, Image Classification, Random Forest, SVM, Deep Learning

I. INTRODUCTION

Plant diseases reduce the number and quality of crops, jeopardising food security and causing significant financial losses. Traditionally, identification of diseases has been done manually, which is unreliable, cost effective, and time-consuming. A vast number of crops are affected with diseases each year, which causes enormous economic losses and a scarcity of or a rise in the demand for food in the market. Therefore, it's critical to detect plant diseases early because they have a significant impact on how crops are produced. Computer vision (CV) and machine learning (ML) techniques for plant diseases can have a significant impact on crop output and quality.

India is among the world's top three cotton producers. Therefore to avoid any quality degradation or overall loss in cotton produce a CNN-based model is proposed in this research for the early detection and categorisation of diseases in cotton plants. In this study, a deep learning-based system is used to classify a few well-known cotton leaf diseases, such as bacterial blight, curl virus, and fusarium wilt.The primary Deep Learning tool in this work is CNN. The term "deep learning" refers to the use of artificial neural network topologies with several processing units. Convolutional neural networks have been the most effective kind of image analysis models till date. CNNs have numerous layers that, to a limited extent, use convolution filters to alter their input. In order to develop a reliable and accurate model, a Transfer-Learning based model is proposed in this research. The diseases are classified using conventional classifiers like random forest and SVM and a feature extractor built on the CNN-based VGG-16 model.

SVM, Random Forest, and VGG16 are used in this model. Visual Geometric Group, or VGG, is a deep convolutional neural network model that has already been trained. It has learned how to recognise common features in images after being trained on more than 10 million images. 13 of the 16 layers in the VGG16 are

convolutional layers, and the remaining 3 levels are fully connected layers. The design is relatively straightforward; it consists of two contiguous blocks of two convolution layers, followed by max pooling, three contiguous blocks of three convolution layers, max pooling, and three dense layers as the final layer. However, in this model, the last three dense layers are discarded, and is replaced with a conventional (Random forest/SVM)classifier. The Random forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction. Another classifier used in this work is support vector classifier (SVCor SVM). SVM selects the decision boundary that maximises the distance between all classes from the nearest data points. The maximum margin classifier or maximum margin hyper plane is the terminology for this decision boundary produced by SVMs. In other words, the data points on one side of the line will all be assigned to one category, while the data points on the other side of the line will be assigned to a different category. Therefore VGG16 is simply utilised as a feature extractor and the classification of the diseases is done by Random Forest classifier and/or Support Vector Classifier.

A confusion matrix is plotted to evaluate the proposed algorithm, and features including precision, recall, and F1 score are computed to assess the model's performance. The parameters true positives, true negatives, false positives, and false negatives, where true positives and negatives are accurately predicted, while false positives and false negatives are incorrectly predicted, can be used to calculate these variables. The precision measures how many accurately anticipated positive observations there were compared to all the predicted positive observations. The ratio of accurately anticipated positive observations to all of the class's observations is known as recall. As the weighted average of Precision and Recall, the F1 score accounts for both false positives and false negatives.

II. RELATED WORK

Based on its well-known applications in areas like object categorization, medical imaging, and agriculture, feature extraction plays a crucial role in the age of ML and pattern recognition. For identifying diseases using the leaf database, there are many methods documented in the literature. Plant Village dataset, a notable collection with thousands of photos of crops leaves, is available to the public.

Five main processes make up the comprehensive automated system, but only three levels—data normalisation, usable feature extraction and reduction, and classification—are involved in the recognition process. For the detection of plant diseases, colour feature extraction using several colour spaces is extremely important. J.Anitha R et al. [1] proposed a model that used CNN as feature extractor and ODNN for disease detection. In ODNN two level of weight optimisation is employed to boost the performance by using Improved Butterfly Optimisation Algorithm and achieves an accuracy of 99 percent. Yushan Z et al. [2] proposed a novel approach named Multi-Context Fusion Network (MCFN) which uses a standard CNN backbone as feature extractor from 50k images Next, contextual data is gathered from image collection sensors as knowledge to classify crop diseases and minimise false positives in their ContextNet. In order to combine visual characteristics with contextual features to predict disease, a connected network is used and achieves a good identification accuracy of 97.5 percent. While another paper used Naive Bayes, Random forest, Decision tree, Gradient Boosting to classify the data set of rice leaf diseases and measured by the accuracy, precision and recall of each algorithms. Where Random forest has the highest accuracy of 69.44 [3]. Another paper [4] proposed a model that works in two stages, first stage is a tailor-made light weight classification model(Xception model(CNN)) which classifies the input images into diseased, healthy or damaged categories and the second stage (detection stage) processing starts only if any disease is detected in first stage. Detection stage performs the actual detection and localisation of each symptom from diseased leaf images. YoloV4, Faster-CNN and EffcieDet are used for detection and achieved mAP of 42.01%,41.1%,32.1% respectively. In [5] authors present a five-layered CNN model for automatic detection of pepper plant disease as healthy or bacterial with 99.99% accuracy on 25 epochs. In [6] CNN(VGG19) is used for feature extraction and partial least squares (PLS) for combing features.The

most discriminant features are finally plugged into the ensemble baggage tree classifier for final recognition. Experimental results were an Average accuracy of 90.1%. In [7] a CNN based model combined with Generative Adversarial Networks (GAN) is proposed while in [8] a Mask R-CNN model is suggested for the autonomous segmentation and detection of tomato plant leaf disease with an outcome of mAP, F1-score, and accuracy as 0.88, 0.912, and 0.98, respectively. In [9] MFaster R-CNN is proposed by improving the Faster R-CNN algorithm used in [4]; the random gradient boosting algorithm is used to optimize the training model. In [10] the authors proposed a Non-dominated Sorting Genetic Algorithm (NSGA-II) based image clustering model detecting the disease area, PCA and multi-class SVM are used for feature reduction and identifying the disease in the tea leaves, respectively with an average accuracy of 83%. In [11], U-Net arch is used to identify the bean leaves and classification is done using 5 deep learning models i.e. Densenet121, ResNet34, ResNet50, VGG-16, and VGG-19; this gave precision of 98.45% for a 3 class dataset. In [12] the authors proposed Deep ensemble neural network with transfer learning for fine tuning and augmentation. In [13] ,images are preprocessed using a histogram pixel localization technique with a median filter and the segmentation is done through a region-based edge normalization. Here an integrated system is formulated for feature extraction using Gabor-based binary patterns with convolution recurrent neural network. Finally, a region-based convolution neural network is used to identify the disease area by extracting and classifying features in order to increase disease diagnostic accuracy. The proposed CRNN–RCNN model achieves an accuracy of 98% for banana dataset. Paper [14] proposed a Fuzzy Based Function Network (FBFN) for identification and classification ,Scale-invariant feature transform method and Firefly algorithm for optimisation. [15] provides a thorough survey compiling methods for identifying leaf stresses that can be found in a range of plants. In this paper, 14 renowned convolutional neural network architectures are used to examine 45 deep learning-based strategies presented for 33 different crops. On the basis of the type of stress, the size of the dataset, the training/test size, the strategies were split into vegetables, fruits, and other crops.

2.1 Motivation

Cotton is one of the most significant cash crops in India, where it is grown extensively by many farmers. Additionally, given that India is one of the top three cotton-producing countries, it is crucial to identify cotton leaf infections as soon as possible. The aforementioned methods mainly rely on handcrafted elements like colour and texture. Complex structure plant data is not a good fit for these traits. In this research, a cutting-edge deep learning system for the detection of cotton leaf diseases is proposed. Additionally, the suggested approach for feature extraction does not include any preprocessing or segmentation procedures as it uses a pre-trained CNN model (VGG16).

III. PROPOSED METHODOLOGY

The four key stages of CV and image processing-based crop infection recognition include preprocessing of the original data, segmenting the infected areas, feature extraction, and classification of disease. The extraction and merging of multi-layer characteristics is the main goal of this work. An automated recognition system is still faced with a number of limitations, including the following: i) Changes in symptom shapes for illness regions; ii) Variation in illumination and environment; iii) Changes in symptom features of crops symptoms, such as their colour and texture, make confusion for correct recognition into their appropriate class; and iv) During the process of picture capturing, the visual quality of an image becomes degraded because of various notable factors such as noise and blurriness.These include effects on the initial pixel values of the image that later led to the production of irrelevant features, v) similarity between colour information from healthy and symptomatic regions, vi) training the system with a pretty sizeable dataset, and (vii) the occurrence of unrelated and reiterated information that lowers recognition accuracy. In this paper, a novel method for identifying crop diseases using deep learning is proposed. Transfer learning-based feature extraction from a pre-trained CNN model and using the retrieved features as inputs

to conventional classifiers (RF/SVM) to create an image classification solution are the two key contributions.

In the areas of image identification, image segmentation, emotion recognition, etc., deep learning techniques have produced remarkably accurate and precise results. As the layer count rises, the deep learning technique learns features at various levels of abstraction. In transfer learning, CNN weights are initialised using data from a trained network. Studies have shown that transfer learning outperforms training the model from scratch in terms of performance.

There are two main phases to the hybrid CNN framework that is proposed. The first stage uses the pre-trained CNN model (VGG16) for deep learning feature extraction. Visual Geometric Group, or VGG, is a deep convolutional neural network model that has already been trained. It has learned how to recognise common features in images after being trained on more than 10 million images. 13 of the 16 layers in the VGG16 are convolutional layers, and the remaining 3 levels are fully connected layers. The design is relatively straightforward; it consists of two contiguous blocks of two convolution layers, followed by max pooling, three contiguous blocks of three convolution layers, max pooling, and three dense layers as the final layer. The architecture of VGG16 is illustrated in Fig[1] below.

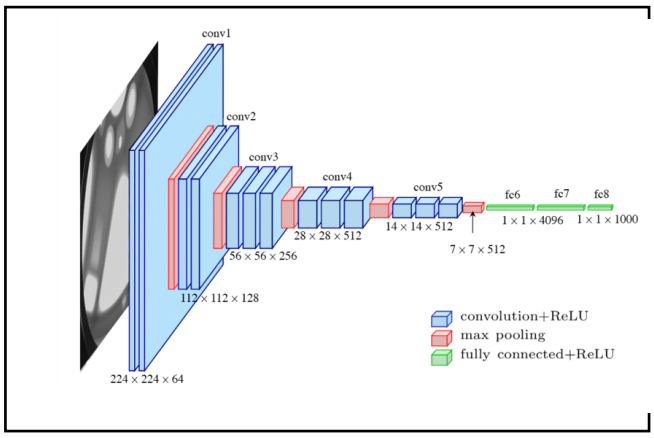

Fig [1]. Architecture of VGG16

But in this model, the final three dense layers are removed and swapped with a conventional (Random forest/SVM) classifier. A randomly chosen portion of the training data is used by the Random forest classifier to generate a collection of decision trees. It simply consists of a collection of decision trees (DT) from a randomly chosen subset of the training set, which are subsequently used to decide the final prediction. Support vector classifier is another another classifier utilised in this study (SVCor SVM). The decision boundary chosen by SVM maximises the separation of all classes from the closest data points. The name for this decision boundary created by SVMs is the maximum margin classifier or maximum margin hyper plane. To put it another way, the data points on different sides of the line will all be categorised into different groups. Therefore, VGG16 is only used as a feature extractor, while Support Vector Classifier or Random Forest classifier are used to classify the diseases. Fig [2] shows the proposed system architecture.

3.1. Convolutional neural network (CNN)

Convolutional neural networks (CNNs) is a more sophisticated machine learning approach than neural networks and are significantly more useful because they require less training parameters. As shown in Fig. 2 [36], CNN is made up of several layers, including convolutions, pooling, and fully connected. It requires a X x H x three-dimensional Matrix input as its input. Convolutional layer X x h x 3 and kernel size K are related, while a convolutional layer's threshold is t. Eqs. (1)–(3) define the primary formulation of a convolutional layer.

$$X_n = X \times K = \sum_{i=1}^{3} (xi+ki)+t \quad (1)$$

$$(2)\ G_n = \frac{G - g + 2 \times z}{D} + 1$$

$$(3)\ K_n = \frac{K - k + 2 \times z}{D} + 1$$

where K stands for the kernel size, D for the amount of convolutions, and t for the threshold. The ReLu activation layer is then carried out as follows:

$$y = \max(0, X)$$

This equation illustrates how, when this function is used, all negative numbers are transformed to zero. For minimizing the dimensionality of the acquired

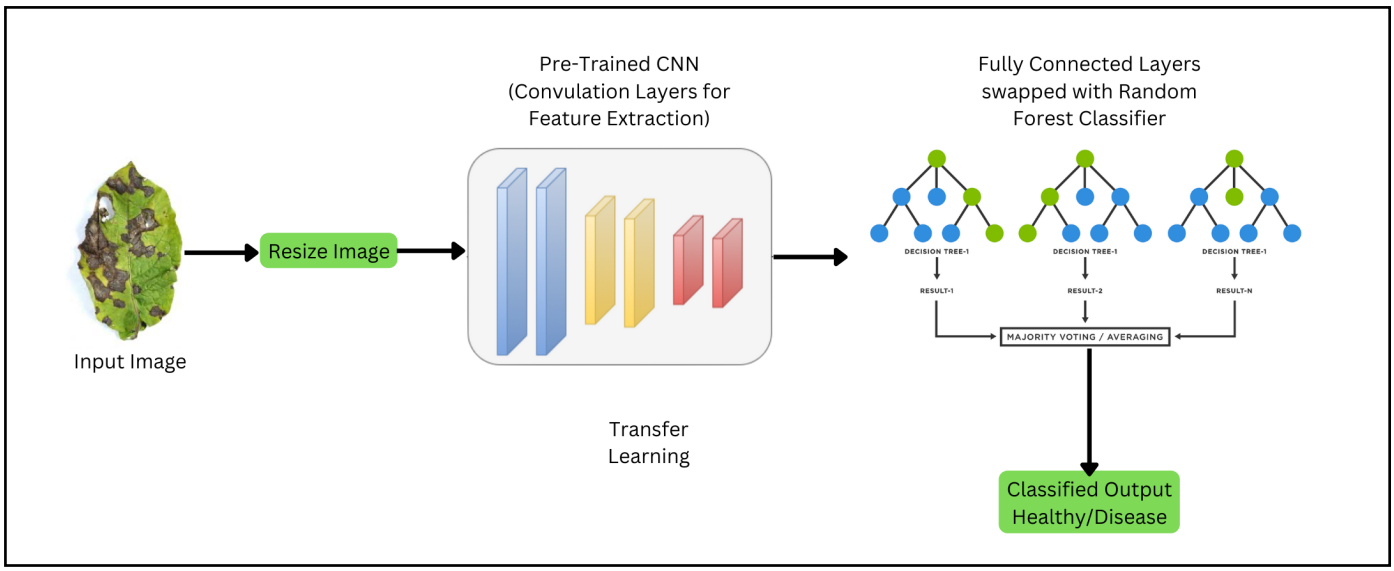

Fig [2]. Proposed system architecture for Random Forest classifier

features from the preceding layers, a further layer known as the pooling layer is used. The most frequently used pooling layers are the minimum, maximum, and average. The main significance of the pooling process is we can get the most similar features. Fig. [4] shows a sample of the pooling procedure.

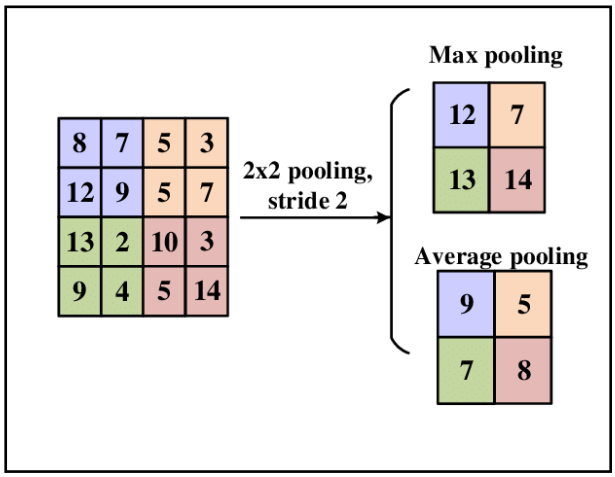

Fig [4]. Max and average pooling example

The following eqs. give the mathematical definition of a pooling layer:

$$P_{max} = Maxf(.,.,.) \quad (5)$$

$$P_{min} = Minf(.,.,.) \quad (6)$$

$$P_{avg} = \frac{1}{n}\sum f(.,.,.) \quad (7)$$

The CNN model's last layer is made up of fully connected layers. Through this layer, the high-level features are being extracted. Formulation of the FC Layer is Defined Mathematically as:

$$L^{in} = X \times W + D \quad (8)$$

$$L^{out} = ReLU(L^{in}) \quad (9)$$

The FC layer is defined using these generalised equations as follows:

$$L_0^{out} = X \quad (10)$$

$$Liin = Li\text{-}1out \times W + Di \quad (11)$$

$$L_i^{out} = F_i(L_i^{in}) \quad (12)$$

where, i is the no. of layers and F is the activation function.

3.2. Feature Extraction

For crop leaf diseases, a number of handcrafted features, including shape, colour, texture, and are retrieved. However, these features do not perform well for a sizeable amount of training and testing data. Recent study has also shown that these traits are inaccurate for complicated crop conditions.

More recently, CNN has been used to extract deep learning features, which are then used to more effectively describe crop disease symptoms than by hand-drawn patterns. Recently, a number of deep

learning models have been introduced, and some well-known pre-trained CNN models used for the recognition process are VGG, AlexNet , YOLO etc.

The VGG16 pre-trained CNN model is utilised in this study to automatically recognise excellent learned features. For feature extraction, three successive layers, referred to as FC layers six, seven, and eight are employed. The Transfer Learning algorithm is used to carry out this procedure. The final three layers of the original model are eliminated, and a new model is trained on a subset of cotton leaf samples.

IV. RESULT AND DISCUSSION

In this section, tabular findings and graphical plots are used to display the general tentative results and their discussion:

4.1. Experimental setup and evaluation measures:

The proposed approach is executed in the spyder Python environment utilising the feature extractor VGG16 and the classifier random forest and Support Vector Classifier. There are 1592 cotton leaf photos in all. Four classes—Curl virus, Fussarium wilt , Bacterial Blight, and Healthy leaves—are used to categorise the dataset. The process of acquiring photos involves resizing and RGB conversion. After that, an array is created from the list and normalised pixel values between 0 and 1. Then, without altering the weights, VGG16 is used to extract the features, and random forest/SVM is used to classify the images. The system's performance is assessed using its accuracy, precision, recall,

and f1 score. All results are calculated, and performance is assessed using four metrics: accuracy, F1 score, precision, and confusion matrix. The 80% - 20% training - testing dataset has a 98.29% accuracy (random forest) and 99.31% accuracy (SVM). In Fig [5] and Fig [6] Confusion matrix is shown for both the classifiers.

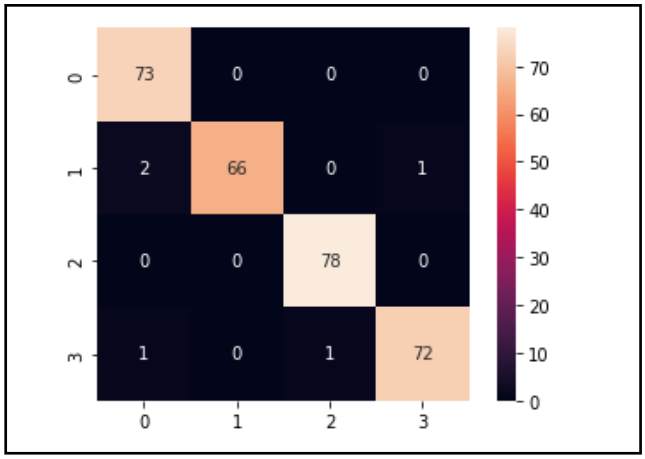

Fig [5]. Confusion Matrix for random forest classifier

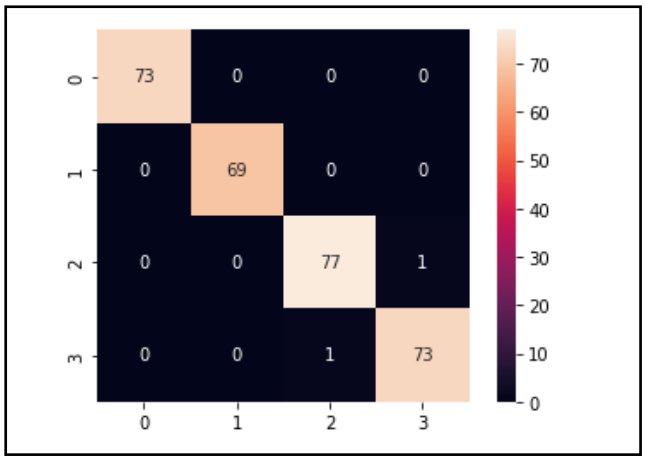

Fig [6]. Confusion Matrix for support vector classifier

4.2. Dataset Description:

1592 total images from the following classes are used: (i) Curl virus; (ii) fussarium wilt; (iii) Bacterial Blight; and (iv) Healthy leaves. The used data's description and sample photographs for each class are shown in Table 1.

1. Bacterial Blight

Xanthomonas citri pv. malvacearum is the root behind cotton's bacterial blight, also known as seedling blight, angular leaf spot, and boll rot (formerly referred to as Xanthomonas campestris pv. malvacearum and Xanthomonas malvacearum). Up until the 1955 introduction of resistant cotton cultivars, bacterial blight devastated cotton-growing regions throughout the 1940s and 1950s. Since this disease thrives in highly warm, humid situations, which some places are

experiencing this growing season, environmental factors have a significant impact on the severity of the disease.

2. Fusarium Wilt

In susceptible cotton cultivars, fusarium might result in severe symptoms. These include a general wilt, which is more prominent on warm days, as well as lower leaf border yellowing and necrosis. Infected plant tissue exhibits a brownish discoloration of the vascular system.All stages of the crop are impacted by the illness. The seedlings' cotyledons, which first become yellow and then brown, show the first signs of the disease. Following wilting and drying of the seedlings, a brown ring can be seen at the base of the petiole. The initial sign of the disease in both young and mature plants is a yellowing of the leaf margins and the area surrounding the veins, or a discoloration that begins at the margin and moves toward the midrib. The leaves eventually lose their turgidity, begin to droop, turn brown, and fall off.

2. Curl Virus

The Cotton Leaf Curl Geminivirus is the main cause of the primary cotton disease in Asia and Africa (CLCuV). Infected cotton leaves curl upward and have thickened veins and enations that resemble leaves on the underside. Early season infections result in stunted plants and much lower yields. The whitefly, Bemisia tabaci, which transmits the virus and is difficult to manage due to the presence of numerous virulent viral strains or related species, is the source of the disease. Because the complex of viruses that causes CLCuD has a higher rate of recombination, the issue is made even more difficult. The presence of alternative host crops like tomato, okra, etc., and the use of mixed farming methods have made the problem worse by promoting the emergence of new virus strains and vectors.

Cotton leaf disease dataset			
Diseases		***Number of images***	***Image***
Bacterial Blight	Training	333	
	Testing	74	
Fussarium Wilt	Training	308	
	Testing	80	
Curl Virus	Training	310	
	Testing	76	
Healthy	Training	329	
	Testing	82	

Table 1. Details of four classes with sample images of each class

Performance measures	***Random Forest Classifier***	***Support vector classifier***
Accuracy	98.299	99.31
Precision	98.33	99.31
F1 score	98.295	99.31

Table 2. shows precision, recall, f1 score in 80% - 20% training – testing dataset

Strength of Proposed Method:

Deep learning is far superior to traditional machine learning with loads of training data. But, for limited training data traditional machine learning (e.g. Random Forest or SVM) may outperform deep learning. Features must be retrieved or engineered for image processing applications to improve accuracy. For image processing applications features need to be extracted / engineered for improved accuracy. Alternatively, features can be extracted from convolutional filters that are part of convolutional neural networks.

In order to provide an image classification solution, this study goes through the process of extracting features using convolutional filters and applying them as inputs to a conventional Random Forest classifier. This method was created specifically for the image classification of custom datasets. With standard machine learning and feature engineering, the method adopted achieves greater accuracy (e.g., Random Forest, svm).

V. CONCLUSION

Only a limited fraction of research has been done on the rapid and automatic diagnosis and classification of plant diseases. A model based on transfer learning has been suggested in this paper to categorise pathogens in cotton leaves. The two basic steps of the suggested methodology are deep feature extraction and final recognition. The VGG16 model uses three successive layers (FC6, FC7, and FC8) for feature extraction, which are then utilised by traditional classifiers (random forest and SVC) for final recognition.

The cotton leaf disease dataset was used to calculate all of the results, which had an average accuracy of 98.8% in the 80% - 20% training - testing dataset. We are confident in saying that the suggested feature selection technique has played a crucial role in enhancing the overall classification accuracy based on the attained classification accuracy.A confusion matrix is also created to assess the model. Using huge datasets, the model's performance can be further enhanced.More plant diseases from the Plant Village dataset will be taken into consideration in the future, and the feature selection approaches will be further investigated to reduce the error rate of the current approach.

REFERENCES

1. J.Anitha R, R.Uma, A.Meenakshi ,P.Ramkumar “Meta-Heuristic Based Deep Learning Model for Leaf Diseases Detection” , Neural Processing Letters, Volume 54, June 2022, pp 5693–5709
2. Yushan Z, Liu L, Chengjun X, Rujing W, Fangyuan W, Yingqiao B, Shunxiang Z “An effective automatic system deployed in agricultural Internet of Things using Multi-Context Fusion Network towards crop disease recognition in the wild.” Applied Soft Computing, Volume 89, Apr 2020
3. Panuwat M, Nutnicha T “Image Classification of Rice Leaf Diseases Using Random Forest Algorithm”, March 2021 (conference paper- DAMT and NCON)
4. Asif Iqbal Khan, S.M.K. Quadri , Saba Banday, Junaid Latief Shah , “Deep diagnosis: A real- time apple leaf disease detection system based on deep learning” , Computers and Electronics in Agriculture, June 2022.
5. Hassan Mustafa, Muhammad Umer, Umair Hafeez, Ahmad Hameed · Ahmed Sohaib, Saleem Ullah, Hamza Ahmad Madni “Pepper bell leaf disease detection and classification using optimized convolutional neural network”, Multimedia Tools and Applications, September 2022
6. Farah S, Md. Attique Khan, Md. Sharif, Mamta M., Lalit M, Sudipta Roy “Deep neural network features fusion and selection based on PLS regression with an application for crops diseases classification”, Applied Soft Computing , February 2021
7. Bhavana N., Sanjay T. “Cross-dataset learning for performance improvement of leaf disease detection using reinforced generative adversarial networks”, International Journal of Information Technology, Aug 2021.
8. Prabhjot K, Shilpi H, Vinay G, Mukund P, Santar P, “An approach for characterization of infected area in tomato leaf disease based on deep learning and object detection technique”, Engineering Applications of Artificial Intelligence,Vol 115, Oct 2022

9. Jie He, Tao Liu, Liujun Li, Yahui Hu, Guoxiong Z , “MFaster R-CNN for Maize Leaf Diseases Detection Based on Machine Vision”, Springer, Apr 2022
10. Somnath Mukhopadhyay, Munti Paul, Ramen Pal, Debashis De, “Tea leaf disease detection using multi-objective image segmentation”, Multimedia Tools and Applications, Jan 2021
11. Sudad H. Abed,Alaa S. Al-Waisy, Hussam J. Mohammed,Shumoos Al-Fahdawi “A modern deep learning framework in robot vision for automated bean leaves diseases detection”, International Journal of Intelligent Robotics and Applications, June 2021.
12. Sasikala V.,V. Sistla, Venkata K.,“Transfer learning-based deep ensemble neural network for plant leaf disease detection”, Journal of Plant Diseases and Protection, May 2021
13. K. Seetharaman & T. Mahendran, “Leaf Disease Detection in Banana Plant using Gabor Extraction and RCNN”, Journal of The Institution of Engineers , June 2022
14. Siddharth Singh, Uday Singh, Sanjeev J, “Automated Plant Leaf Disease Detection and Classification Using Fuzzy Based Function Network”, Wireless Personal Communications, Dec 2021
15. Serosh K, Muhammad Ajmad, Muhammad Ali, Abdul Mannan , “Use of deep learning techniques for identification of plant leaf stresses: A review”, Sustainable Computing: Informatics and Systems 28, September 2020