



# An Enhanced Random Forest Classifier to detect Crop Disease with Texture and Shape Features OF Corn Images (ERFCTS)

V.Praba<sup>1\*</sup>, Dr. K. Krishnaveni<sup>2</sup>

<sup>1\*</sup>Research Scholar, Department of Computer Science, Sri. S. Ramasamy Naidu Memorial College (Affiliated to Madurai Kamaraj University, Madurai), Sattur, Virudhunagar District, Tamil Nadu 626203, India. E-mail:praba@srmcollege.ac.in

<sup>2</sup>Head and Associate Professor, Department of Computer Science, Sri. S. Ramasamy Naidu Memorial College (Affiliated to Madurai Kamaraj University, Madurai), Sattur, Virudhunagar District, Tamil Nadu 626203, India. E-mail:kkrishnaveni@srmcollege.ac.in

**Citation:** V.Praba, Dr. K. Krishnaveni, (2024) An Enhanced Random Forest Classifier to detect Crop Disease with Texture and Shape Features OF Corn Images (ERFCTS), *Educational Administration: Theory and Practice*, 30(5), 3720-3727

Doi: 10.53555/kuey.v30i5.3524

## ARTICLE INFO

## ABSTRACT

Crop diseases pose a significant threat to global food security, affecting crop yield and quality. Early detection and accurate diagnosis of these diseases are crucial for effective disease management. A novel random forest classification for crop disease detection using texture and shape features is proposed in this research. The input dataset contains corn images with three different types of diseases namely Corn Blight, Rust and Gray Leaf Spot as well as Healthy images. These images are pre-processed using image processing techniques, binarized and the feature vector for each image is created and stored as training feature vector. The Random Forest Classifier is trained with this feature data set. The feature set with two images of each class is taken as test data set, test feature vector is calculated, mapped against training feature vector and finally classified by Random Forest Classifier. The performance of the proposed system is evaluated using the classification metrics such as accuracy, precision, recall, and F1-score. Experimental results demonstrate that the effectiveness of the proposed approach in detecting and classifying the crop diseases is 97% accuracy, thereby aiding farmers and agricultural stakeholders to take disease management strategies timely.

**Key words:** Crop, Corn Blight, Rust, Gray Leaf Spot, Healthy, Random Forest Classifier.

## I. INTRODUCTION

Differentiating common and more serious infections in crops is a difficult task for all farmers. It takes a long period to go to agribusiness office and discover what the infection may be. Close continuous monitoring of crops is required to identify the diseases. Numerous recent technologies have emerged to minimize post-harvest processing, reinforce agricultural sustainability and exploit the productivity. Though the laboratory-based approaches such as polymerase chain reaction, gas chromatography, mass spectrometry, thermography and hyper spectral techniques have been employed for disease identification, they are highly time consuming not cost effective. In recent times, effective server based or mobile based approaches are employed to identify and recognize the crop diseases automatically. Several factors of these technologies being high resolution camera, high performance processing and extensive built in accessories.

Machine learning and Deep learning algorithms play a vital role in recognizing the crop diseases with high rate of accuracy. Various researches have taken place under the field of machine learning for plant disease detection and diagnosis, such as traditional random forest, artificial neural network, Support Vector Machine (SVM), fuzzy logic, K-means method, etc. Random Forest overcomes the disadvantage of over fitting of their training data set and it handles both numeric and categorical data well. Hence, crop disease detection using Random Forest technique with texture and shape features is proposed here.

This paper is organized into the following sections. Section II reviews the existing crop disease detection techniques. The proposed methodology is discussed in Section III. Experimental results are analysed in Section IV. Finally, the conclusion and future work is presented in Section V.

## II. LITERATURE SURVEY

Various crop disease detection methods using image processing and machine learning algorithms existing in the literature are reviewed and described in this section.

Gui et al. [1] divided the early Soybean Mosaic Virus disease (SMV) into 0, 1, and 2 degrees according to its severity. In the case of a small number of experimental soybean samples, they proposed a novel SVM early detection method which combined convolutional neural network and Support Vector Machine (CNN-SVM) and achieved an accuracy rate of 96.67% on the training set and 94.17% on the testing set.

Huang et al. [2] introduced a system titled "Identification of multiple plant leaf diseases using neural architecture search," which presents an innovative approach utilizing image analysis and deep learning algorithms for real-time disease detection. The system employs Neural Architecture Search (NAS) guided by Bayesian Optimization (BO) and is validated using a dataset of 54,306 plant disease images. This method achieves high accuracy even with unbalanced data, showcasing its efficacy in simplifying network architecture design and precisely identifying plant diseases. These findings hold significant promise for enhancing disease control strategies in agriculture.

S. Ramesh et al. [3] proposed a plant disease detection system using machine learning approaches to identify the healthy and diseased leaves. The system utilizes datasets that are generated and trained using Random Forest for the classification of healthy and diseased leaves. This process includes dataset creation, feature extraction using Histogram of an Oriented Gradient, training the classifier and displaying the result. The final result gives 70% accuracy level [3].

Liang et al. [4] presents a new approach for rice blast disease recognition using deep CNNs, leveraging a dataset of 2906 positive and 2902 negative samples. The experiments show that CNN outperforms compared to other traditional methods like Local Binary Pattern Histogram (LBPH) and Wavelet Transform (Haar-WT) in terms of accuracy, AUC and ROC curves. Both CNN with Softmax and CNN with SVM show comparable performance, suggesting the efficacy of CNN in rice blast disease recognition, promising practical applications.

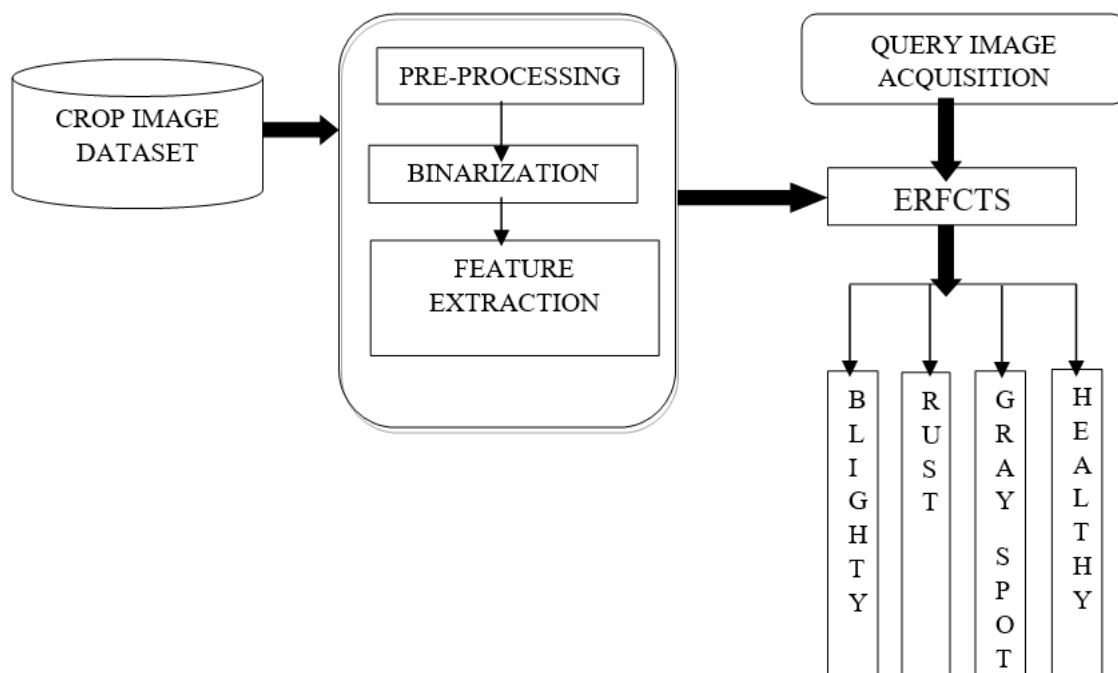
P. R. Rothe et al. [5] introduces "Cotton Leaf Disease Identification using Pattern Recognition Techniques". The input data set is created from the fields at the Central Institute of Cotton Research in Nagpur, as well as cotton fields located in the Buldana and Wardha districts. Snake segmentation and Hu's moments are used as distinctive attributes yields 85% of classification accuracy.

Sannakki et al. [6] proposed Diagnosis and classification of grape leaf diseases approach using neural networks that utilized image processing and artificial intelligence techniques to diagnose grape leaf diseases. Techniques like thresholding and K-means clustering are employed for image segmentation and noise reduction. The system achieved optimal results with Feed-forward Back Propagation Neural Network classification after identifying the diseased portions from the segmented images.

Wang et al. [7] discusses digital image recognition's role in reducing reliance on agricultural experts for plant disease identification. Four neural network models (BP, RBF, GRNN, PNN) were tested to differentiate wheat stripe rust and leaf rust, and grape downy mildew and powdery mildew. Results show high accuracy in disease identification, particularly with BP networks, GRNNs, and PNNs, achieving 100% accuracy for wheat diseases and between 94.29% and 100% for grape diseases.

## III. METHODOLOGY

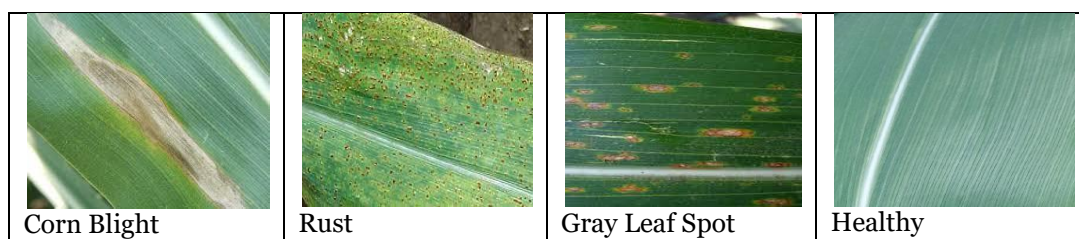
An enhanced random forest classifier to detect and classify the crop diseases based on shape and texture features is proposed in this research work. The input data set contains the healthy and different diseased corn images. In the training phase, the dataset the healthy and diseased images are labelled as 0, 1, 2 and 3 where 0 - Corn Blight, 1 - Rust, 2-Gray leaf spot and 3- Healthy. The texture and shape features of these images are extracted and stored as a feature dataset. Two images from each crop class are selected and stored as a testing data set. The testing feature dataset is then computed, mapped with the training feature set, and subjected to classification using Random Forest Classifier. The schematic diagram of **ERFCTS** is shown in Figure1.



**Figure 1. Schematic Diagram of ERFCTS**

**3.1 Dataset**

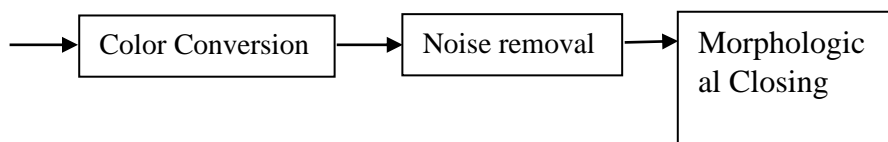
The Kaggle dataset containing 342 corn images of four different categories (Healthy, Corn Blight, Rust and Gray Leaf Spot) is taken as an input data set. In these 80% of the dataset is used to train the pre-trained models and 20% is used for testing. Out of 342 crop images, 274 samples were used for training and 68 for testing. The crop image samples are shown in Figure 2.



**Figure 2. Sample Corn Leaf images**

**3.2 Preprocessing**

During image acquisition, the presence of shadows on the images can potentially influence the process of computing the feature values. To mitigate the impact of shadows, the following preprocessing procedure is applied.



**Figure 3. Flow Diagram for Pre-processing**

**1) Color Conversion:**

The input leaf image in RGB format is converted into a grayscale image using (1).

$$Y = 0.2989 R + 0.5870 G + 0.1140 B \tag{1}$$

**2) Noise removal process:**

*Linear interpolation* is a fundamental technique in image processing that plays a crucial role in maintaining the quality, accuracy, and visual appeal of images during various operations, particularly in resizing and geometric transformations. Linear interpolation strikes a balance between simplicity and effectiveness, making it efficient for real-time applications while maintaining edge details and features in the processed images. It has two methods *Up sampling* and *Down sampling*. *Up sampling is applied to resize the grayscale image into 10 times larger than its original size. Then Gaussian Blur operation is applied to*

smooth the image that is to reduce the noise particles. Finally down sampling is applied to resize the pre-processed image back to its original dimensions and the result is shown in Figure 3.

### 3) Morphological Operation

Perform morphological closing on the thresholded image, which helps to close gaps in the white regions (foreground).

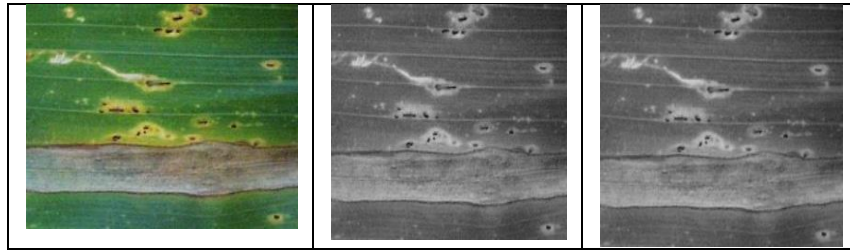


Figure 4. (a) Original Image (b) Gray scale image (c) After Noise Removal

### 3.3 Binarization

Otsu's thresholding is applied to convert the image into a binary format, to easily identify the disease affected area of the image.

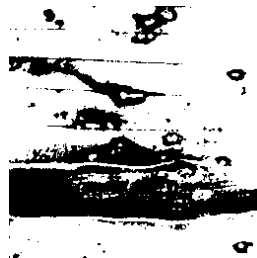


Figure 5 (a) Binarization

### 3.4 Feature Extraction

Each crop exhibits distinct characteristics namely shape, color and texture that differentiate it from others. This research work focuses only the shape and texture features of crop.

#### 1) Shape Features

The shape features are extracted from the contours of the images. The procedure to extract the contours is as follows.

In the initial phase, the outline points of the binary image are found. Using these points, the shape features  $area(A)$  and  $perimeter(P)$  are computed as below.

$$A = \frac{1}{2} \left| \sum_{i=1}^{N-1} (x_i y_{i+1} - x_{i+1} y_i) + x_N y_1 - x_1 y_N \right| \text{---(2)}$$

Where  $A$  is the area,  $(x_i, y_i)$  are the coordinates of the contour points, and  $N$  is the total number of points in the contour.

The  $perimeter (P)$  can be calculated by summing the Euclidean distances between consecutive contour points:

$$P = \sum_{i=1}^{N-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} + \sqrt{(x_N - x_1)^2 + (y_N - y_1)^2} \text{---(3)}$$

#### 2) Texture Features

GLCM (Gray-Level Co-Occurrence Matrix) is the most common methods for extracting informational features. GLCM is a statistical approach that captures texture features by considering the spatial arrangement of pixels. This method builds a gray-level co-occurrence matrix (GLCM), also known as a gray-level spatial dependence matrix. GLCM functions evaluate the image's texture by measuring how frequently pair of pixels with specific values and spatial relationships appears. Statistical measures are then derived from this matrix. It's essential to note that GLCM functions don't provide information about the shape of objects but focus on the spatial relationships between pixels. The texture features Contrast, Dissimilarity, Homogeneity, Energy and Correlation are extracted as follows.

(i) **Contrast** measures the differences in gray levels between pixels and their neighbouring reference points

$$Contrast = \sum_{i,j=0}^{N-1} P_{ij} (i - j)^2 \text{---(4)}$$

(ii) **Dissimilarity** measures the distance between pairs of pixels within the region of interest.

$$Dissimilarity = \sum_i \sum_j |i - j| p(i, j) \text{---(5)}$$

(iii) **Homogeneity** is typically inversely linked to contrast and reflects the likeness of off-diagonal elements in the GLCM.

$$Homogeneity = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(i-j)^2} \text{----- (6)}$$

(iv) **Energy** Provides the sum of squared elements in the GLCM. Also known as uniformity or the angular second moment.

$$Energy = \sum_{i,j=0}^{N-1} (P_{ij})^2 \text{----- (7)}$$

(v) **Correlation** shows how the gray levels in the GLCM are linearly related to each other.

$$Correlation = \sum_{i,j=0}^{N-1} P_{ij} \frac{(i-\mu)(j-\mu)}{\sigma^2} \text{----- (8)}$$

Where:

- I = row number.
- j = column number.
- P=Pixels of the image
- μ = mean of the GLCM (an estimate of all pixel intensities that contribute to the GLCM).
- σ = variance of the pixel intensities that contribute to the GLCM.
- C = (I,j)<sup>th</sup> element of the normalized GLCM.

**Table 1 Features of Sample Images**

S.No	Area	Perimeter	F1	F2	Contrast	Dissimilarity	Homogeneity	Energy	Correlation	Label
1	41631.5	1029.179	0.27	0.73	855.8818	25.91448	1.796879	1.414109	3.854875	0
2	16.5	18.24264	0.62	0.38	1720.353	45.86581	1.425861	1.211257	3.619154	0
3	37377	1328.333	0.59	0.41	1227.683	35.28662	1.485825	1.202266	3.773151	0
4	36609.5	1414.139	0.02	0.98	314.7235	3.741891	3.815064	3.703258	2.493234	1
5	25750.5	1493.453	0.02	0.98	333.6495	4.105103	3.813295	3.695044	2.424426	1
6	43080.5	1611.002	0.3	0.7	2885.054	50.19133	1.814771	1.488093	3.531735	2
7	37289.5	1535.747	0.32	0.68	664.6673	16.48908	2.209125	1.659207	3.86453	2
8	2804.5	361.7817	0.02	0.98	1750.389	19.94543	2.366632	1.961717	3.844344	3
9	64675	1044.828	0.98	0.02	178.3989	14.64846	1.184402	0.151974	3.920319	3

**3.5 Random Forest Classification**

Random Forest is a simple and flexible machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in Machine Learning. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

**Procedure:**

**Input:** Dataset *Train*// **Training dataset with class labels.**

**Output:** Classification Accuracy ACC

**RFC**

**Step 1:** Read the dataset has number of rows “r” and number of columns “m”

**Step 2:** Select samples from the sample set using Bootstrap sampling.

**Step 3:** Build decision tree for selected subset of features

**Step 2:** Randomly select a subset of the data from Dataset to build decision tree.

**Step 3:** Repeat Step1 and Step2 for times to build decision trees.

**Step 4:** Each decision tree will generate an output.

**Step 5:** Final output is considered based on Majority Voting for Classification.

Once the feature vector for each image is constructed and potentially normalized, they serve as the input data for training the random forest classifier. Each feature vector, along with its corresponding class label, is used to train the random forest classifier.

During the training process, each decision tree in the random forest is trained on a random subset of the feature vectors. This subset may be selected using techniques such as bagging, where random samples of the feature vectors are drawn with replacement. The decision trees are trained recursively, with each split optimizing a chosen criterion Gini impurity to maximize the purity of the resulting child nodes.

Once the random forest classifier is trained, it can be used to make predictions on new, unseen data. For each new data point represented as a feature vector, the random forest aggregates the predictions made by each individual decision tree and outputs the final prediction based on a majority vote.







**IV EXPERIMENTAL ANALYSIS AND RESULTS**



The training data set is created with 342 leaf images of one healthy and three different plant leaf diseases. The Morphological shape features and the texture feature for each plant leaf are computed and stored as a training feature data set. From the training set, two leaf images of each class are taken as query images and



feature vector for each image is generated, validated against the training data set and classified by Random Forest Classifier. The classification results are given in Table 2.

**Table 2: Validated Results with Proposed Model**

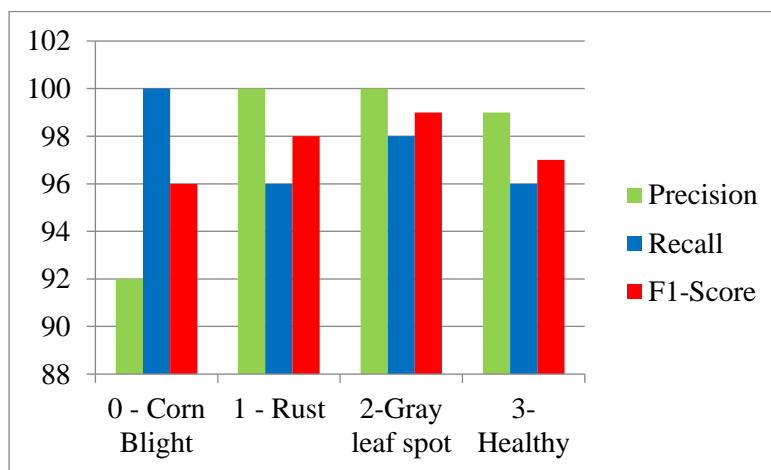
S.No	Images	Disease Classified
1		Correctly identified the Imag1.jpg is: Corn_(maize)___Blight
2		Correctly identified the Imag2.jpg is: Corn_(maize)___Blight
3		Correctly identified the Imag3.jpg is: Corn_(maize)___Common_rust
4		Correctly identified the Imag4.jpg is: Corn_(maize)___Common_rust
5		Incorrectly identified the Imag5.jpg is: Corn_(maize)___Healthy
6		Correctly identified the Imag6.jpg is: Corn_(maize)___Gray_Leaf_Spot

7		Correctly identified the Imag7.jpg is: Corn_(maize)____Healthy
8		Correctly identified the Imag7.jpg is: Corn_(maize)____Healthy

The result of the proposed ERFCTS is evaluated based on the following classification metrics given in Table 3 and figure 6 shows the graphical representation these classification metrics. Table 4 specifies the confusion matrix details.

**Table 3 Random Forest Classification Metrics**

Class Labels	Precision	Recall	F1-Score	Support
0 - Corn Blight	0.92	1.00	0.96	92
1 - Rust	1.00	0.96	0.98	89
2-Gray leaf spot	1.00	0.98	0.99	84
3- Healthy	0.99	0.96	0.97	77



**Figure 6. Random Forest Classifier Confusion Metrics**

**Table 4 Confusion Matrix**

Label	0	1	2	3
0	92	0	0	0
1	4	85	0	0
2	1	0	82	1
3	3	0	0	74

**Accuracy-** Accuracy quantifies the frequency with which a classification model makes correct predictions. It can be defined as the ratio of the number of accurate predictions to the total number of predictions

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$= (92+242)/( 92+242+8+0)=0.97$$

**Precision**–Precision describes the number of correctly predicted cases that are actually positive.

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \\ = 92/(92+8)=0.92$$

**Recall (Sensitivity)** – Recall describes the number of actual positive cases that the model correctly predicted.

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \\ = 92/(92+0)=1$$

**F1 score** – The F1 score is the harmonic mean of Precision and Recall. The F1 score ranges from the ideal value of 1 to the worst value of 0. It can be computed using the following formula:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\ = 2 \times \frac{0.92 \times 1}{0.92 + 1} = 0.96$$

**Specificity**- In contrast to recall, specificity can be defined as the number of negatives returned by the classification model. It can be easily calculated by the confusion matrix using the following formula

$$\text{Specificity} = \frac{TN}{TN + FP} \\ = 242/(242+92)=0.72$$

**Support** - Support can be defined as the number of true response samples in each class of target values.

**Support**=92

## V. CONCLUSION

Corn crop disease detection using shape and texture features with random forest classifier is performed in this research. The experimental analysis shows that the proposed method outperforms and yield random forest classifier performed better by giving 97% of accuracy. The results are very encouraging and the future work will be geared towards using a dataset with more features and high-performance computing facilities focussing the corn crops. This research contributes to the advancement of precision agriculture by leveraging machine learning and image analysis techniques for efficient crop disease detection and monitoring.

## References

1. J. Gui, J. Fei, Z. Wu, X. Fu, and A. Diakite, "Grading method of soybean mosaic disease based on hyperspectral imaging technology," *Inf. Process. Agricult.*, vol. 160, no. 7, pp. 1–6, Nov. 2020.
2. J.-P. Huang, J.-X. Chen, K.-X. Li, J.-Y. Li, and H. Liu, "Identification of multiple plant leaf diseases using neural architecture search," *Trans. Chin. Soc. Agricult. Eng.*, vol. 36, no. 16, pp. 166\_173, Aug. 2020.
3. S. Ramesh et al., "Plant Disease Detection Using Machine Learning," 2018 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C), 2018, pp. 41-45, doi: 10.1109/ICDI3C.2018.00017.
4. W.J. Liang, H. Zhang, G.-F. Zhang, and H.-X. Cao, "Rice blast disease recognition using a deep convolutional neural network," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, Feb. 2019.
5. P. R. Rothe and R. V. Kshirsagar, "Cotton Leaf Disease Identification using Pattern Recognition Techniques", *International Conference on Pervasive Computing (ICPC)*, 2015.
6. S. S. Sannakki, V. S. Rajpurohit, V. Nargund, and P. Kulkarni, "Diagnosis and classification of grape leaf diseases using neural networks," in *Computing, Communications and Networking Technologies (ICCCNT)*, 2013 Fourth International Conference on. IEEE, 2013, pp. 1–5.
7. H. Wang, G. Li, Z. Ma, and X. Li, "Application of neural networks to image recognition of plant diseases," in *Systems and Informatics (ICSAI)*, 2012 International Conference on. IEEE, 2012, pp. 2159–2164.