Multi-Descriptor Enabled Leaf Disease Detection Using Machine Learning Methods

C.Venkata Sudakar[1], G. Umamaheswara Reddy[2]

[1]Assistant Professor, Department of Electronics and Communication Engineering, Sree Vidyanikethan Engineering College, Tirupati-517102, Andhra Pradesh, India

[2]Professor, Department of Electronics and Communication, Engineering, Sri Venkateswara University College of Engineering, Sri Venkateswara University, Tirupati, 517502, India.

ABSTRACT

India is an agricultural country. Two-thirds of the Indian populations work in agriculture, making it the backbone of the country's economic system. Leaf disease is a major challenge in agriculture that affects crop yield and quality. In recent years, advancements methods for accurately and effectively detecting leaf diseases have been developed thanks to advances in machine learning and computer vision. The objective of this work is to create a machine learning-based Random Forest Regression Algorithm for automatically detecting leaf diseases. In the suggested methodology, leaf pictures are acquired, subjected to pre-processing which includes RGB image and HSV image conversion, multi-descriptor feature extraction includes Hu moments, Haralick, color histogram and classification using Random Forest Regression Algorithm. Metrics including accuracy, precision, recall, and f_1-score are used to assess the system's performance by implementing in Python 3.8. The findings of the study indicate that the proposed system achieves precision 98 %, recall 98 %, f_1-score 98%, test accuracy 97.81 %, while validation accuracy is 95.93 %. The developed system has potential applications in precision agriculture, enabling farmers to detect and treat plant diseases early, thereby reducing crop losses and increasing yields.

Keywords: Leaf Disease Detection, Random Forest regression, Python, Google Colab

I. INTRODUCTION

Agriculture is the most significant part of our economy. Many diseases cause plant leaves to deteriorate, which affects agricultural yield. Finding leaf disease is crucial. Crop diseases are a significant issue because they significantly lower the quantity and quality of agricultural goods [4]. As this is the only method that offers the possibility of finding diseases at an early stage, an automatic system for detecting plant diseases has an obvious advantage in monitoring large fields. Farmers' low production is a result of their ignorance about leaf disease. Being able to identify leaf disease is essential since productivity affects profit and loss.

Rural farmers might think it's challenging to discern the illnesses that could harm their crops. They find it difficult to attend the agricultural office to inquire about the potential sickness. Our main goal is to identify the disease that is introduced in a plant by analyzing its structure using image processing and machine learning. The damage of crops and plant components by pests and diseases lowers food output and increases food insecurity. Additionally, a lot of less developed countries have a poorer understanding of disease and pest control. Toxic pathogens, inadequate disease control, and dramatic climatic changes are all major causes of decreased food production.

The accuracy of the results has been increased by employing modern methods such as machine learning (ML)and deep learning(DL) algorithms. Numerous research have been undertaken to detect and diagnose plant illnesses utilizing traditional machine learning approaches such as random forests (RF), artificial neural networks(ANN), support vector machines (SVM), fuzzy logic(FL), the K-means method [13], and convolutional neural networks (CNN) [14].

II. LITERATURE REVIEW

P.R. Rothe and R.V. Khirsagar [16], published a research paper on Cotton leaf disease detection using pattern recognition technologies. In order to teach an adaptive neuro-fuzzy inference system, Hu's moments are extracted from the segmented images using the active contour model. It is discovered that the classification precision is 85%.

Saradhambal, j.et al [15], "Plant disease detection and its solution using image classification" is a proposal made in 2018. The leaves' diseased region was predicted using the k-mean clustering technique. This essay's objective is to identify plant illnesses and offer treatments for them. It displays the proportion of the leaf that is affected. Here, the leaf's diseased area is divided into segments and analysed.

Alina Forster, j.et al [3], Hyperspectral Plant Disease Forecasting Using Generative Adversarial Networks.Cycle-Consistent Generative Adversarial Network was employed.Using hyperspectral pictures, this technique forecasts the growth of powdery mildew on barley leaves.The model can understand how the disease is spreading because it can see it visually and in the corresponding reflectance spectra. Additionally, this model produces acceptable outcomes for a prediction over a seven-day period.

K.Rajesh Babu, j.et al [4], proposed automatic Plant disease detection and classification using image processing methods. Selected characteristics are extracted, and these features are used to train support SVM and ANN classifiers. The results were satisfactory.

Xulang Guan, j.et al [5], developed "A Novel Method of plant leaf disease detection based on deep learning and convolutional neural network" by combining four CNN models. The method's accuracy rate of 87% was much greater than the results obtained by using just one CNN model.

III.METHODOLOGY

To ascertain if the leaf is healthy or diseased, certain procedures must be followed. Preprocessing, feature extraction, categorization, and classifier training fall under this category. The Fig.1 shows the Multi-descriptor enabled leaf disease detection using Random Forest regression classifier.

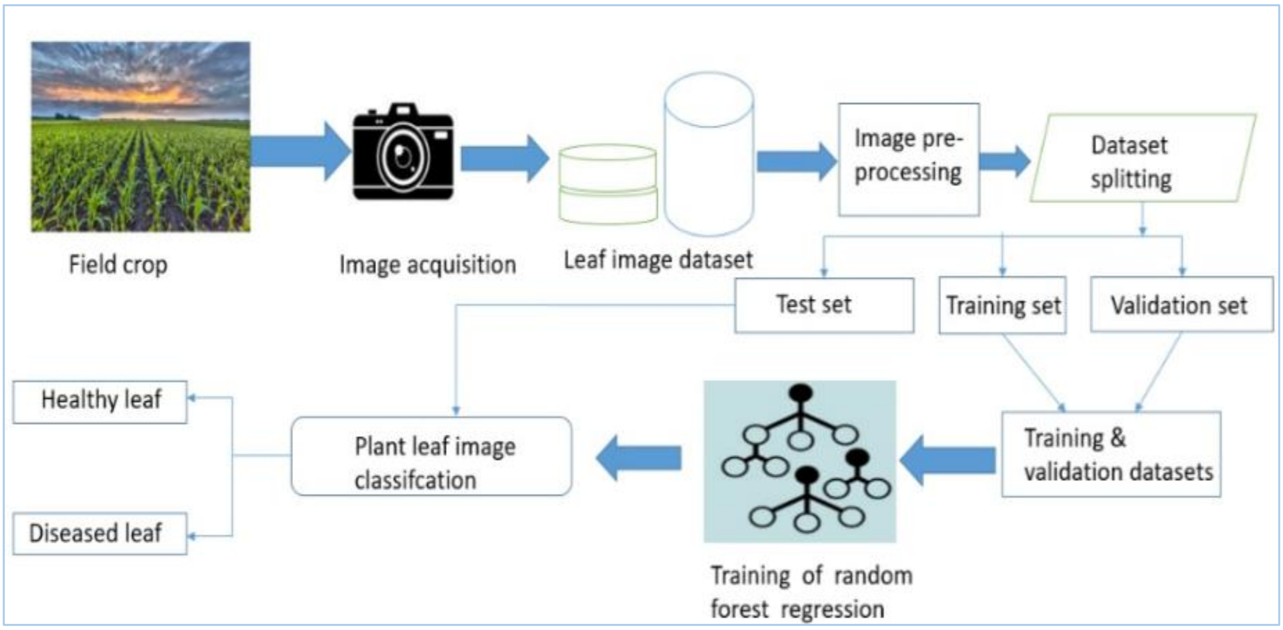


Fig.1 Architecture of RFregressionsystem for leaf disease detection

Data Collection

The dataset is collected from kaggle database in allowed file extensions .jpg, .png format using the web link https://www.kaggle.com/datasets/emmarex/plantdisease.

The collected data is labeled using symptom- adaptive annotation strategies to help machines understand what exactly is in it and what important [7] is. This annotated data is then used for model training [8]. During the training process data augmentation used to artificially increase the training data set

Image Preprocessing

Image preprocessing is a technique used to enhance and transform the raw leaf images before feeding them into the machine learning model. These techniques include image resizing, normalization, filtering, and segmentation by eliminating noise, emphasizing key features, and standardizing the input data, image preprocessing helps to increase the accuracy and performance of the machine learning mode [9-11].

Random Forest Regression expects input data in RGB format as shown in Fig. 7(a) and Fig. 7(b), whereas BGR is commonly used in computer vision applications. Hence, converting from BGR to RGB is necessary to ensure compatibility with the machine learning model as shown in Fig.2 (a) and Fig.2 (b).

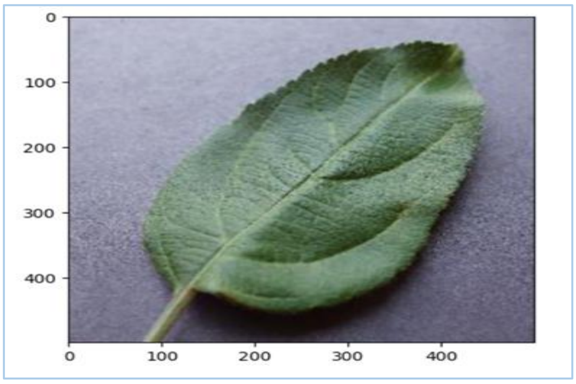


Fig.2 (a) RGB image of Healthy leaf

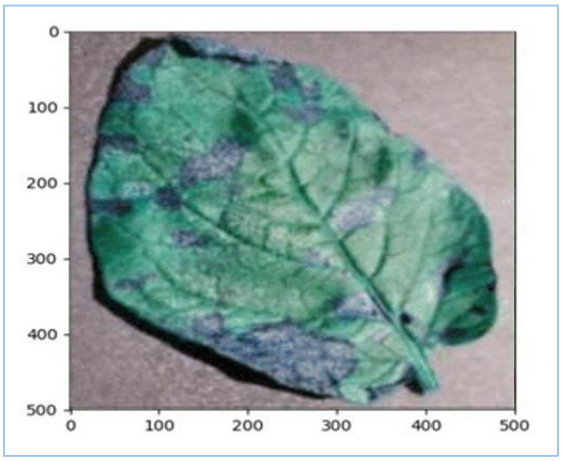


Fig.2 (b) RGB image of diseased leaf

Feature Extraction Techniques

Random forest algorithm is a supervised learning (ensemble) method for issues like classification, regression, and others that develop a forest of decision trees during the course of the training phase. In contrast to decision trees, random forests, which can handle both categorical and numerical data, do not suffer from the disadvantage of overfitting their training data set.

The histogram of oriented gradients (HOG) is a component descriptor used in computer vision and image processing for object recognition. We are employing three component descriptors here: Hu moments, Haralick fabric, and Color histogram

Hu moments: Images that have the vital attributes of an image's pixels help characterize the items. In this instance, Hu moments help to describe the specific leaf's shape [2]. There is only one channel available for computing Hu moments. The RGB data must first be transformed to grayscale before the Hu moments can be calculated.

Haralick Texture: Healthy and sick leaves typically have distinct textures. For the purpose of separating the textures of healthy and sick leaves in this instance, we employ the Haralick texture feature. The frequency at which the pixel I fills the area next to the pixel J is used to compute texture in the adjacency matrix, which holds each (I, J) pixel's location. To compute the Haralick texture, the image must be gray scaled.

Color Histogram: The image's colours are represented by the colour histogram. Prior to computing the histogram, RGB is transformed into HSV colour spaceas depicted in Fig. 3(a) and Fig. 3(b). It is required to convert the RGB image to HSV because the HSV model closely mimics how the human eye recognises the colours in a picture [1]. The histogram plot gives a description of the number of pixels that are usable within the designated colour ranges.

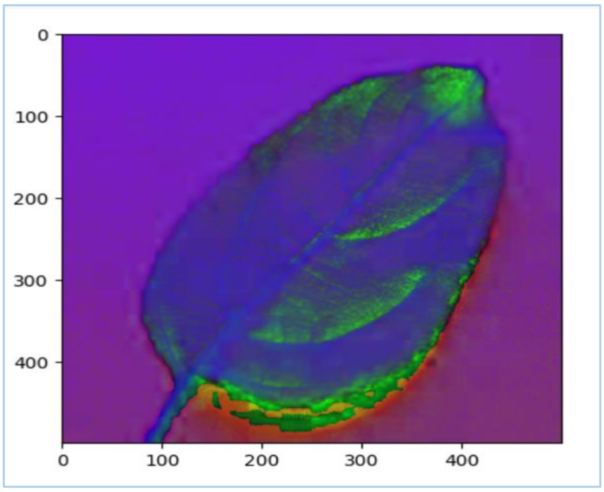


Fig. 3 (a) HSV image of healthy leaf

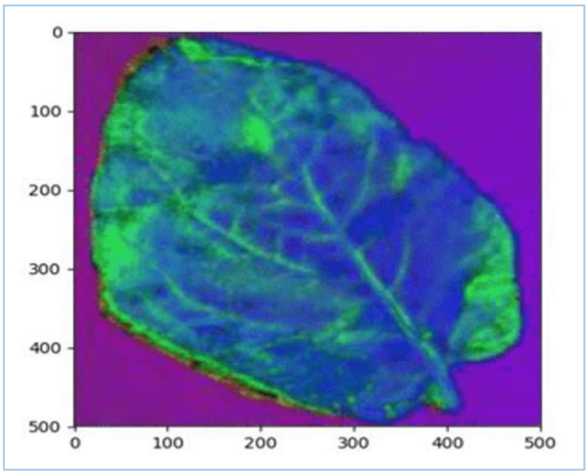


Fig.3 (b) HSV image of Diseased Leaf

As shown in the Fig. 4(a) and Fig. 4(b), Image segmentation of leaf is typically performed as a preprocessing step to extract the leaf region from the background [6].

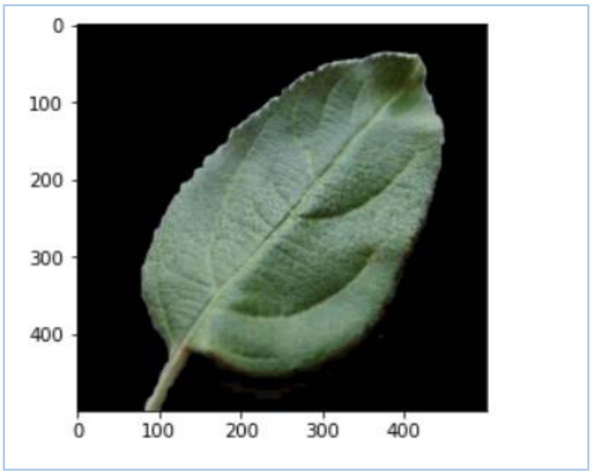


Fig.4 (a) Segmented Image of healthy Leaf

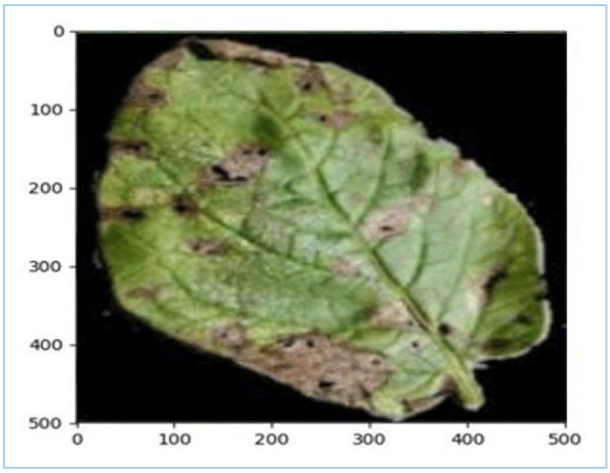


Fig.4 (b) Segmented Image of Diseased Leaf

Data Training: 1300 samples of input data are collected from Google and sent to the model to be trained, known as pre-trained data, in order to train the system. The system analyses the leaf using the pre-trained data. The data set must be split into two parts: data sets for testing are 320 samples and data sets for training are 980 samples.

Random Forest regression classifier

Random forests classifier is used to implement the technique in this case. They are adaptable and may be used for both regression and classification procedures. Compared to other machine learning techniques like SVM, Gaussian Naive Bayes, logistic regression, and linear discriminant analysis (LDA), Random forests algorithm (Fig.5) is the more accurate with fewer sets of image data.

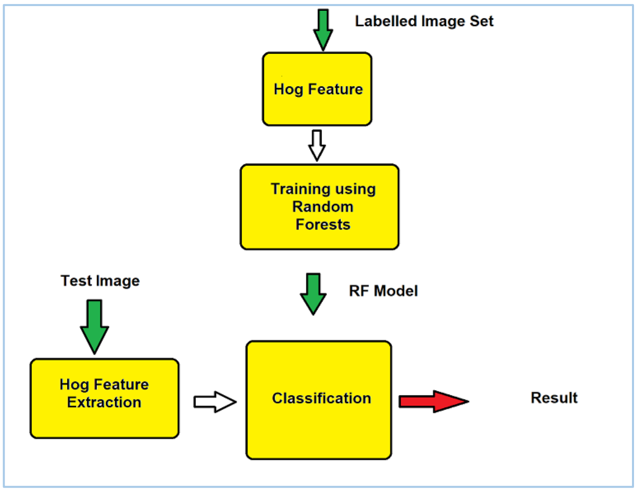


Fig.5 Random Forest regression frame work

Mathematical Expressions:

Over the training process, random forests (RF) create a range of distinctive decision trees. Using the regression, the predictions from all trees are combined to create the final forecast.

Scikit-learn uses the Gini Importance to determine a node's importance using Equation (1), there are only two child nodes in the binary tree:

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \quad (1)$$

Where,
ni_j = Importance of node j,
w_j = Quantity of samples whose weights reach node j
C_j= Impurity value of node j
$w_{left(j)}$ = A left child node on node j divides the node.
$w_{right(j)}$ = A right child node on node j divides the node.

The decision tree's features are given weights according to the Equation (2).

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ features\ i} ni_j}{\sum_{k \in all\ nodes} ni_k} \qquad (2)$$

Where,
fi_i= the importance of trait i
ni_j = Node j's relevance

These can then be normalized to a value between 0 and 1 by dividing by the sum of feature importance as shown in Equation (3)

$$norm\ fi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j} \qquad (3)$$

The final relevance of the feature at the Random Forest level is determined by the average over all trees using Equation (4). When the significance of each characteristic for each tree is calculated, and the sum is divided by the number of trees, the result is:

$$\text{RF}fi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T} \qquad (4)$$

Where,
RFfi_i= computed the feature's relevance from all the trees in the Random Forest model.
$normfi_{ij}$= The relevance of the tree j's normalized feature for i
T= number of trees overall

The division of the nodes in this case is based on a subset of randomly chosen attributes. By randomly selecting a subset of attributes at each node, the approach reduces the correlation between the trees and improves the overall performance of the random forest.

The characteristics used to divide nodes include leaf color, leaf shape, texture, vein patterns, size, symmetry, and the presence of spots.

The labeled datasets are isolated from the training and test data. HoG feature extraction is used to create the feature vector for the training dataset. A RF classifier is then trained using the produced feature vector. Additionally, as shown in Fig. 5, the trained classifier receives the testing data feature vector produced by HoG feature extraction for prediction.

HoG feature extraction transforms labeled training datasets into the corresponding feature vectors. The training datasets are where these extracted feature vectors are kept. Moreover, a Random Forest classifier is used to train the trained feature vectors.

The test image's feature vectors are retrieved using HoG feature extraction, as shown in Fig.5.The stored and trained classifier receives these created feature vectors in order to forecast the outcomes.

IV. RESULTS AND DISCUSSIONS

In this paper leaf images are collected from the website and given to the proposed system. This algorithm predict the input leaf is either diseased or healthy based on and finds which algorithm is the best for identifying leaf disease.

It produces a confusion matrix, as depicted in Fig.6, from which we can derive the values for true positives (TP) are 153, true negatives (TN) are 160, false positives (FP) are 5, and false negatives (FN) are 2. These numbers allow us to compute the precision, recall, and f_1-score. A measure of precision is the percentage of number true positives (TP) out of all positive predictions (TP + FP). The model makes fewer false positive predictions when it has a high precision score.

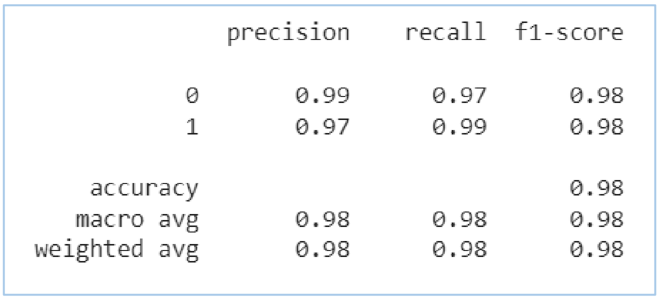
Recall is a measurement of how often true positives (TP + FN) occur. Low false negative predictions are a sign of a model with a high recall score. The f_1-score is a harmonic mean of precision and recall [5] is computed using Equation (5) and illustrated in Fig.6

$$f_1\text{-score} = 2x\frac{\text{Precision x Recall}}{\text{Precision + Recall}} \qquad (5)$$

where,

$$\text{Precision}= \frac{TP}{\text{TP + FP}}$$

$$\text{Recall}= \frac{\text{TP}}{\text{TP + FN}}$$



```
              precision    recall  f1-score

           0       0.99      0.97      0.98
           1       0.97      0.99      0.98

    accuracy                           0.98
   macro avg       0.98      0.98      0.98
weighted avg       0.98      0.98      0.98
```

Fig.6 Precision, recall and f1-score values

The Equation (6) below used for determine the Random Forest (RF) model's accuracy

$$\text{Overall accuracy} = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} \qquad (6)$$

As shown in the Fig. 7 and 9, the accuracy score obtained from the accuracy score function represents the percentage of instances in the test set that were correctly categorized. In this case, accuracy score of 0.9782 indicates that model correctly classified 97.81% of the instances in the test set.

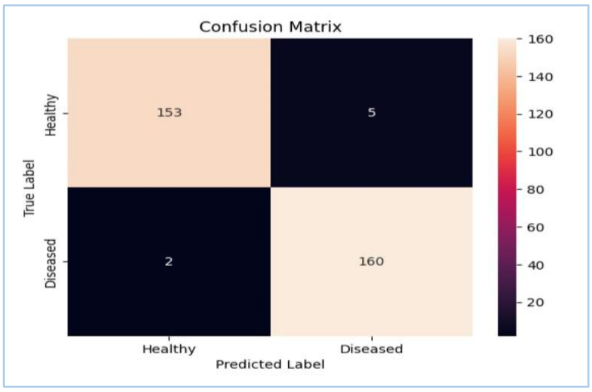


Fig.7 Confusion Matrix of Random Forest

The test accuracy achieved is 97.8% and the validation accuracy is 95.9%.The classification methods accuracies are compared in Table 1.

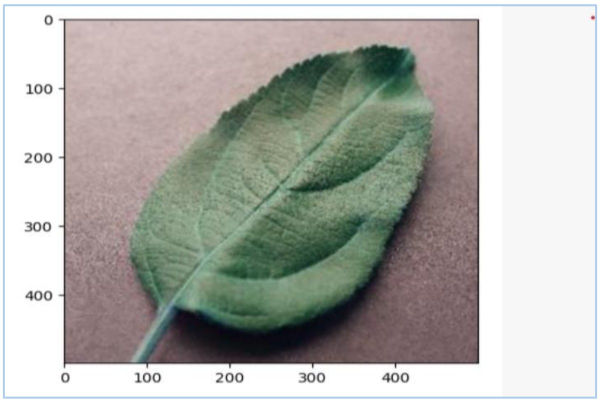


Fig.8 (a) Input leaf Image

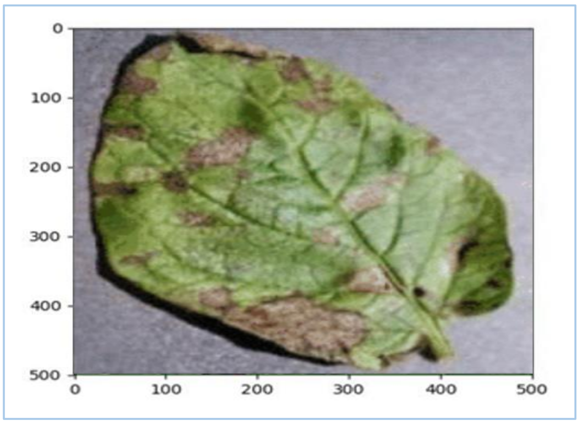


Fig.8 (b) Diseased leaf

After preprocessing Fig.1 (a) and Fig.1 (b), the algorithm found Fig.1 (a) as healthy image and Fig.1(b) as diseased image, those are illustrated in Fig.8(a) and Fig.8(b) respectively.

Table.1 Accuracy comparison

Machine Learning Algorithms	Accuracy
LDA	90.31
KNN	92.26
SVM	91.95
RF Regression	95.93

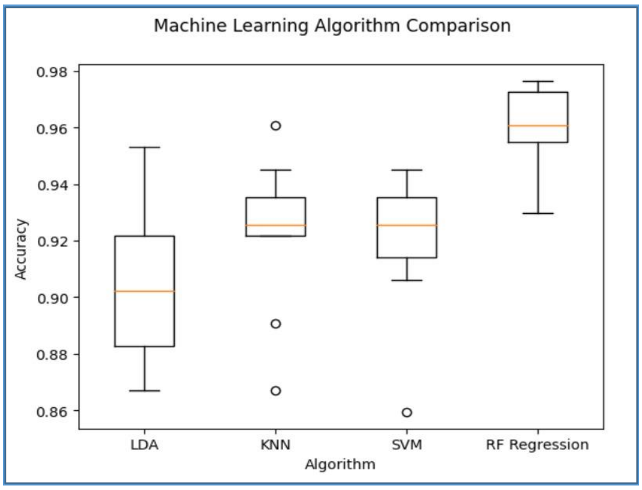


Fig.9 Machine Learning Algorithm Comparison

V. CONCLUSION

This project seeks to identify anomalies in plant leaf. The methodology involved collecting a dataset of leaf images from various sources. Using the collected data set Random forest regression algorithm has been trained and it uses cross-validation to evaluate the leaf is diseased or healthy. Random forest regression algorithm has produced precision as 98 %, recall as 98 %, F1-score as 98%, and test accuracy as 97.81 % and validation accuracy as 95.93 % in leaf image data set classification.

Future scope includes the user-friendly system implementation using deep learning methods to display name of the leaf disease and suggest the remedy for the diseases to increase the agriculture productivity. The accuracy achieved is 95%, and the limitation of this leaf disease detection is that it cannot identify the disease's name. As a result, it may be further developed to do so, as well as to offer treatments for the disease and detect its symptoms, in order to increase accuracy.

Additionally, the model could be applied to real-time monitoring systems to detect and alert farmers of potential disease outbreaks in their fields.

VI. REFERENCES

1. Dahiya, Sachin, Tarun Gulati, and Dushyant Gupta. "Performance analysis of deep learning architectures for plant leaves disease detection." *Measurement: Sensors* 24 (2022): 100581. https://doi.org/10.1016/j.measen.2022.100581

2. Harakannanavar, Sunil S., Jayashri M. Rudagi, Veena I. Puranikmath, Ayesha Siddiqua, and R. Pramodhini. "Plant leaf disease detection using computer vision and machine learning algorithms." *Global Transitions Proceedings* 3, no. 1 (2022): 305-310. https://doi.org/10.1016/j.gltp.2022.03.016
3. A. Förster, J. Behley, J. Behmann and R. Roscher, “Hyperspectral Plant Disease Forecasting Using Generative Adversarial Networks”, IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, pp. 1793-1796, doi: 10.1109/IGARSS.2019.8898749,2019.
4. Sushil R. Kamlapurkar, “Detection Plant Leaf Disease Using Image Processing Approach”, International Journal _ of Scientific and Research Publications, vol. 6, no. 2, pp. 73-76, February 2016.
5. X. Guan, “A Novel Method of Plant Leaf Disease Detection Based on Deep Learning and Convolutional Neural Network”, 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, pp. 816-819, doi: 10.1109/ICSP51882.2021.9408806,2021.
6. E. Vamsidhar, P. Jhansi Rani and K. Rajesh Babu , “Plant Disease Identification and Classification using Image Processing”, International Journal of Engineering and Advanced Technology, vol. 8, no. 3S, pp. 442-446, February 2019.
7. Simranjeet kaur, Geetanjali Babbar and Gagandeep, “Image Processing and Classification A Method for Plant Disease Detection”, International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 9S, pp. 868-871, July 2019.
8. R. Kundu, U. Chauhan and S. P. S. Chauhan, “Plant Leaf Disease Detection using Image Processing” 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), Gautam Buddha Nagar, India, 2022, pp. 393-396, doi: 10.1109/ICIPTM54933.2022.9754170.
9. R. Bora, D. Parasar and S. Charhate, “Plant Leaf Disease Detection using Deep Learning: A Review”, 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-6, doi: 10.1109/I2CT54291.2022.9824925.
10. C. Venkata Sudhakar , G. Umamaheswara Reddy, “Land use Land cover change Assessment at Cement Industrial area using Landsat data-hybrid classification in part of YSR Kadapa District, Andhra Pradesh, India”,International Journal of intelligent systems and applications in engineering IJISAE, 2022, 10(1), 75–86. https://doi.org/10.18201/ijisae.2022.270
11. Venkata Sudhakar C. , Umamaheswara Reddy G., Usha Rani N., “Delineation and evaluation of the captive limestone mining area change and its influence on the environment using multispectral satellite images for industrial long-term sustainability”,Cleaner Engineering and TechnologyVolume 10, October 2022, 100551. . https://doi.org/10.1016/j.clet.2022.100551
12. Gautam Kaushal1, Rajni Bala2,” GLCM and KNN based Algorithm for Plant Disease Detection”, International Journal of Advanced Research in Electrical,Vol.6, Issue7,July2017
13. Sudhakar CV, Reddy GU. “Land use/land cover change assessment of Ysr Kadapa District, Andhra Pradesh, India using IRS resourcesat-1/2 LISS III multi-temporal open source data.” International Journal of Recent Technology and Engineering, Volume-8 Issue-3, 2019. DOI: 10.35940/ijrte.C6067.098319
14. Venkata Sudhakar, C and Reddy, Umamaheshwara and Rani, Usha, Delineation of the Captive Limestone Mine Boundaries Using Multispectral Satellite Images Through the Use of NDVI and Google Earth Image Template Matching (April 23, 2022). Proceedings of the International

Conference on Innovative Computing & Communication (ICICC) 2022, Available at SSRN: http://dx.doi.org/10.2139/ssrn.4091402

15. Sanjay B. Dhaygude and Nitin P. Kumbhar, “Agricultural plant Leaf Disease Detection Using Image Processing”, International Journal of Advanced Research in Electrical Electronics and Instrumentation Engineering, vol. 2, no. 2, pp. 599-602, January 2013.
16. P.R. Rothe and R. V. Kshirsagar, “Cotton Leaf Disease Identification using Pattern Recognition Techniques”, International Conference on Pervasive Computing (ICPC), 2015.