ISSN 2395-8618

Plant Leaf Disease Recognition Using Random Forest, KNN, SVM and CNN

Bijaya Kumar Hatuwal, Aman Shakya, and Basanta Joshi

Abstract—**Huge loss in crop production occurs every year due to late identification of pant diseases in developing countries like Nepal. Timely and correct identification of such diseases with less dependency in related field expert can be more effective solution to the problem. Plants suffer from various diseases and correctly identifying them by observing the leaves is major challenge especially if they have similar texture. Consideration of plant leaf color and various texture features is extremely important to correctly predict the defect in plant. The aim of this work is to classify and predict given disease for plant images using different machine learning models like Support Vector Machine(SVM), k-Nearest Neighbors (KNN), Random forest Classifier (RFC), Convolutional Neural Network and compare the results. Image features like contrast, correlation, entropy, inverse difference moments are extracted using Haralick texture features algorithm which are fed to SVM, KNN and Random Forest Algorithms whereas CNN directly feeds upon images as input. Among the used models CNN produced highest level of accuracy of 97.89% and RFC, SVM and KNN had accuracy of 87.43%, 78.61% and 76.96% respectively for sixteen different image categories used.**

Index Terms—**Plant disease, Haralick texture, support vector machine, k-nearest neighbor, random forest classifier, convolutional neural network.**

I. Introduction

FOR BETTER agricultural productivity health of plant is a primary concern. Diseases hinder the normal state of a plant and potentially modify or interrupt its vital mechanisms such as fertilization, photosynthesis, transpiration, germination etc. Time and again plants get various diseases depending upon the factors like environment, season, soil, bacteria and others. The conventional process of plants disease detection with bare eyes observation method is tiresome and is non-effective. So identifying the plant disease for farmers requires help of related filed specialist most of the time. Hiring the related filed experts may cost farmers heavily and use of pesticides without knowledge will degrade the quality land and harm the living organisms. So utilizing available technology to identify the plant diseases may be a viable solution.

Plant disease recognition by visual method is cumbersome task, less accurate and can be applied only in limited areas. Some general diseases in plants are early and late scorch, yellow and brown spots and some are bacterial and virus diseases. Some research works are done identifying plant disease but they have not covered the broader categories of plant diseases and image features for training purpose to obtain much accurate results for large set of images. Depending upon the texture (shape, size, roughness, intensity, etc.) plant diseases can be predicted in some cases and in others it may require further test. This research is focused on classifying and recognizing images based on the plant leaf textures such as contrast, correlation, entropy, inverse difference moments and leaf colors (red, green and blue). There are sixteen different plant leaf categories considered as healthy or other predefined disease for this work. The models for CNN, KNN, RFC and KNN will be developed and the results will be compared.

This papaer is organized with introduction to emphasize the need of this research work to be done which is followed by reivew section to provide the overview of the works done previously, their limitaions and the further enhancements that will be incorporated by this research work in the plant leaf diesease recgonition. The working meachinism, variuos algorithms used and mathematical represenations used in the work is explanined in the methodology section. The result and discussion section describes the findings of the research work. The overall findings and the possible future enhacements is summarized in the conclusion and future works section.

II. Literature review

Various researchers have used machine learning and image processing techniques for identifying the diseases on different types of plants. The Authors of paper [1] investigated using k-means clustering method for Brinjal leaves with image processing techniques to identify plant leaf disease. The authors performed histogram equalization to increase image quality prior clustering process. Color Co-occurrence Method (CCM method) was used to extract the color and texture features. The features were trained using k-means clustering algorithm with three clusters as infected object, infected leaf and the black background of leaf. However, the features are not sufficient to classify much larger classes of images and clusters with subtle change in colors.

The authors of the paper [2] proposed converting RGB image into HSV and perform color based subtraction of unwanted background by retaining pixel having G value more than R and B values for plant leaf disease classification. The

Manuscript received on June 20, 2020, accepted for publication on July 31, 2020, published on December 30, 2020.

Bijaya Kumar Hatuwal is with the Department of Computer and Electronics Engineering, Himalaya College of Engineering (of Tribhuvan University), Chyasal, Lalitpur, Nepal (e-mail: bjkat28@gmail.com).

Aman Shakya and Basanta Joshi are with the Department of Computer and Electronics Engineering, Institute of Engineering Pulchowk (of Tribhuvan University), Lalitpur, Nepal (e-mail: aman.shakya@ioe.edu.np, basanta@ioe.edu.np).

https://doi.org/10.17562/PB-62-2

ISSN 2395-8618

connected elements in the image are discovered out from the cluster based background subtraction and the immense part of the image is kept and other part is removed. They used SVM which created the hyper planes in high dimensional space for categorizing the data points into different classes.

The authors of the paper [3] proposed KNN as an effective method in identifying leaf diseases for agronomical crop images. They used luminance and linear characteristics image to detect skeleton of leaves to determine whether the leaf is of grape or not. Then, GLCM (Gray-Level Co-Occurrence Matrix) features are extracted and diseases are classified by using the obtained grape leaf images. However the detection and recognition was only for grape specific and could not perform well for other species of plant.

The authors of the paper [4] performed Convolutional Neural Network operation for plant disease detection using python API They resized image to 96x96 resolution for image processing. Data augmentation technique was used to rotate, flip, shit images horizontally and vertically. Adam optimizer was incorporated using categorical cross-entropy. They trained the image with 75 epochs using 32 batch sizes for 35000 images. Similarly, the authors of paper [5], proposed framework ResNet50, ResNet101, DenseNet161, and DenseNet169 as their Deep Neural Network (DNN) framework to detect disease in rice plant. Images were resized as 224 × 224 pixels, the batch size was set to 64 , epoch to 15 and the learning rate was set a constantly of 0.0001. The DenseNet161 produced the best results with an accuracy of 95.74%.

The authors of the paper [6] investigated on using k-means clustering for the image segmentation of grape leaf disease. Shape, color and texture were extracted as main features. Linear Support Vector Machine (LSVM) was used for classification purpose. The images were classified into two classes Downy and Powderly using the extracted nine texture features and nine color features for all three segmented parts of single leaf image.

The authors of the paper [7] proposed using three feature descriptors Hu moments, Haralick Texture and color histogram for plant disease classification using various machine learning algorithms Logistic regression, Support vector machine, k-nearest neighbor, CART, Random Forests and Naive Bayes. The accuracy of machine learning models Support vector machine 40.33%, k-nearest neighbor 66.76% and random forest 70.14 was quite low.

Most of the papers focused on same species of plant for the disease prediction which has almost similar texture. Also comparison among the various algorithms to obtain the better results is done by few authors where the number of image feature considered was fewer and the overall accuracy was low. This study will consider multiple features like contrast, correlation, inverse difference moments, entropy and red, green and blue colors for training and prediction purpose of plant leaf images to get higher accuracy. The sixteen different plant image classification categories will be considered. The machine learning models like SVM, KNN, CNN and RF will be used for training and testing purpose and the accuracy results among the models will be compared.

III. Methodology

There are four major stages involved in the proposed approach as image acquisitions and preprocessing, image features extraction, model training and testing and given input image prediction. Fig. 1 and Fig. 2 represents the detailed flow diagram of the methodology.

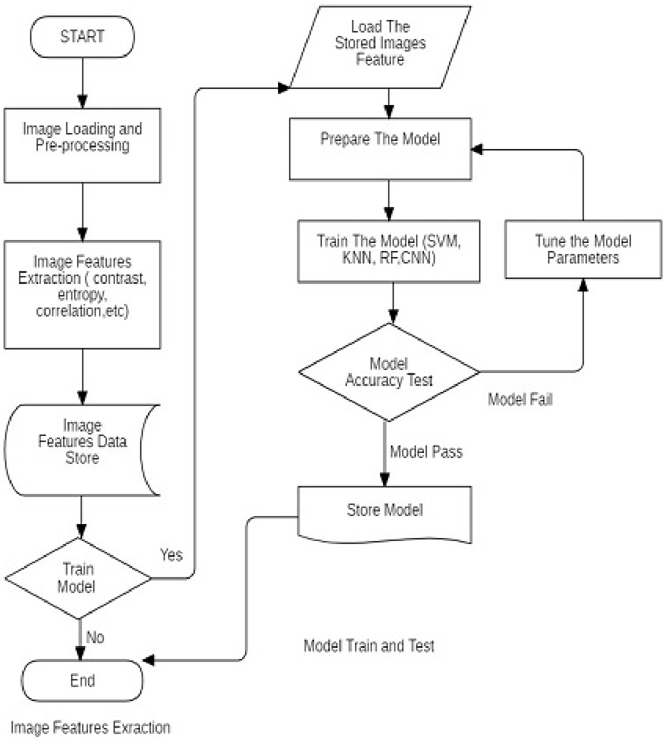

START
Image Loading and Pre-processing
Image Features Extraction (contrast, entropy, correlation,etc)
Image Features Data Store
Train Model
Yes
No
End
Image Features Exraction
Load The Stored Images Feature
Prepare The Model
Train The Model (SVM, KNN, RF,CNN)
Tune the Model Parameters
Model Accuracy Test
Model Fail
Model Pass
Store Model
Model Train and Test

Fig. 1. Image feature extraction and model preparation

A. Image Acquisitions

The images were collected for various plants species and diseases. The images were placed in jpg format. The source images were taken from the Kaggle plant village dataset. The train (folder containing images for training purpose of the models) and valid (folder containing images for validation purpose of the model) folder consists of images in ratio of 80 to 20 for training and testing purpose respectively for sixteen different categories which is shown in Fig. 3' .

B. Image Features Extraction

In total ten properties from color and textures are generated as the features from the images. The mean and standard deviation of each color channel red (R), green (G) and blue (B) are calculated. Then blurring is done after converting

ISSN 2395-8618

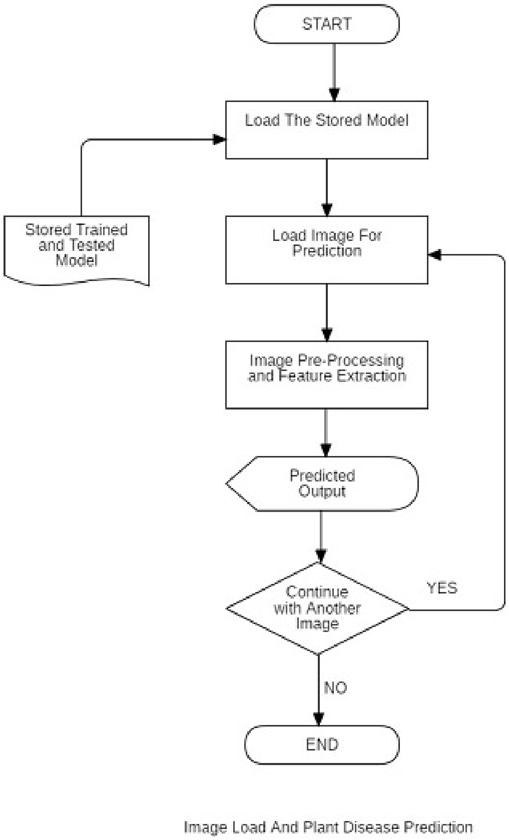
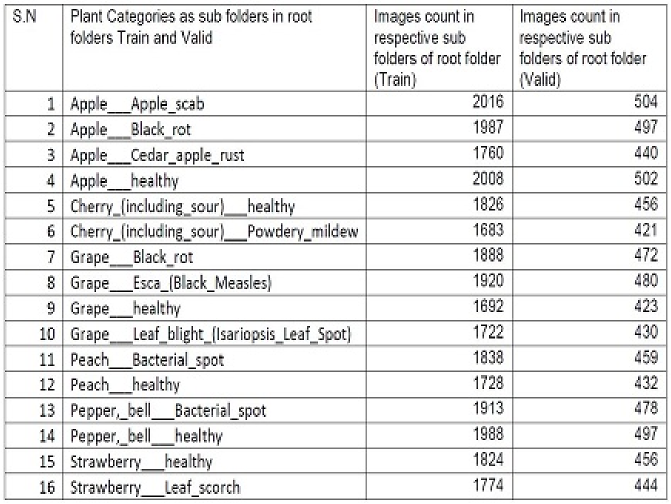

START
Load The Stored Model
Stored Trained and Tested Model
Load Image For Prediction
Image Pre-Processing and Feature Extraction
Predicted Output
Continue with Another Image
YES
NO
END
Image Load And Plant Disease Prediction

Fig. 2. Plant disease prediction with given image



S.N	Plant Categories as sub folders in root folders Train and Valid	Images count in respective sub folders of root folder (Train)	Images count in respective sub folders of root folder (Valid)
1	Apple___Apple_scab	2016	504
2	Apple___Black_rot	1987	497
3	Apple___Cedar_apple_rust	1760	440
4	Apple___healthy	2008	502
5	Cherry_(including_sour)___healthy	1826	456
6	Cherry_(including_sour)___Powdery_mildew	1683	421
7	Grape___Black_rot	1888	472
8	Grape___Esca_(Black_Measles)	1920	480
9	Grape___healthy	1692	423
10	Grape___Leaf_blight_(Isariopsis_Leaf_Spot)	1722	430
11	Peach___Bacterial_spot	1838	459
12	Peach___healthy	1728	432
13	Pepper,_bell___Bacterial_spot	1913	478
14	Pepper,_bell___healthy	1988	497
15	Strawberry___healthy	1824	456
16	Strawberry___Leaf_scorch	1774	444

Fig. 3. Plant disease images counts in different categories for training and testing

the image into gray scale to reduce the noise level in the image. Gaussian noise is very common kind of noise that is likely to arise for case of any image due to poor illumination or high temperature or transmission [8]. The texture based feature extraction is performed using Haralick texture features algorithm which extracts contrast, correlation, inverse difference moments, and entropy from the images converted as grayscale. The Haralick texture algorithm uses gray-level co-occurrence matrix (GLCM) to calculate the features. GLCM is a matrix that represents the relative frequencies of a pair of grey levels present at certain distance d apart and at a particular angle θ [9]. Extraction of textural information from images containing highly directional characteristics is majorly dependent on selection of correct angle θ [9]. The mathematical equatioin of the features used are represented by Equations (2)–(5).

$$G = \begin{bmatrix} P(1,1), & P(1,2), & \cdots & ,P(1,D_g) \\ P(2,1), & P(2,2), & \cdots & ,P(2,D_g) \\ \vdots & \vdots & \ddots & \vdots \\ P(D_g,1) & P(D_g,2) & \cdots & P(D_g,D_g) \end{bmatrix} \tag{1}$$

$$\mathit{Contrast} = \sum_{n=0}^{Dg-1} x^2 \left\{ \sum_{i=1}^{Dg} \sum_{j=1}^{Dg} P(i,j) \right\}, |i-j| = n \tag{2}$$

where Dg is numbers of gray levels that can be represented by a matrix G having dimension Dg as shown in Equation (1) with any pixel point (i,j) and P(i,j) represents the probability of presence of pixel pairs at certain distance d at angle θ in GLCM image.

$$\mathit{Correlation} = \frac{\sum_{i=1}^{Dg} \sum_{j=1}^{Dg} (i,j)\, P(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{3}$$

where $\mu_x \mu_y$ are means and $\sigma_x \sigma_y$ are standard deviations of Px and Py the partial derivative function.

$$\mathit{Inverse\ Difference\ Moments} = \sum_{i=1}^{Dg} \sum_{j=1}^{Dg} \frac{1}{1+(1-j)^2} P(i,j) \tag{4}$$

$$\mathit{Entropy} = -\sum_{i=1}^{Dg} \sum_{j=1}^{Dg} P(i,j)\, log\,[P(i,j)] \tag{5}$$

C. Model Training and Testing

The image features extracted were split in the ratio eighty to twenty for the training and testing purpose for the SVM, KNN, K-means clustering and random forest.

There are many ways to solve Multi-class classification problems for SVM such as Directed Acyclic Graph (DAG), Binary Tree (BT), One Against-One (OAO) and One-Against-All (OAA) classifiers [10]. Constructing an optimal hyper plane regarded as the decision surface using the input samples to make the two sides' margin largest is the main mission of support vector machine [11]. To

ISSN 2395-8618

perform the multi class classification using SVM we have used one-versus-rest method and Gaussian radial basis function. Radial basis function (RBF) kernel (is a positive parameter for controlling the radius) [10] which is given by Equation (6): As SVM cannot perform multiclass classification at once, one-versus-rest method is used by SVM which does binary operation with each dataset to finally make multiclass classification.

$$K\left(xi,xj\right)=exp\left(\frac{-\left\|xi-xj\right\|^{2}}{2\sigma^{2}}\right) \tag{6}$$

where k is the kernel function , x_i = (x_{i1}, x_{i2}, . . . , x_{iN}) corresponds to the attribute set for the ith sample in each sample tuple represented by (x_i, x_j) in N training data of a binary classification.

Random Forest can be used for classification and regression. The proposed methodology uses classification using Random forest. It is an ensemble method as uses the average or voting from multiple decision trees to reach the decision. The problem of over fitting is also reduced drastically by reducing variance in this algorithm.

K Nearest Neighbor (KNN) can be used for both classification and regression operation. In pattern acknowledgement, the (KNN) k nearest neighbors algorithm is a non-parametric method used for classification and regression [12]. KNN performs classification based on the majority voting on similarity to K nearest number of neighbors calculated using distance functions. Some most common distance functions are Euclidean, Manhattan, Minkowski and Hamming distance. For categorical variables Hamming distance is used and for continuous variables Euclidean, Manhattan and Minkowski distance calculation are used.

$$Euclidean\ Distance=\left(\sum_{i=1}^{k}\left(x_i-y_i\right)^{2}\right)^{1/2} \tag{7}$$

$$Hamming\ Distance\ \left(H_D\right)=\sum_{i=1}^{k}\left\|x_i-y_i\right\| \tag{8}$$

where $x = y \Rightarrow D = 0, x \neq y \Rightarrow D = 1$, x =($x_1$,$x_2$,$x_3$,. . .,$x_k$) and y=($y_1$,$y_2$,$y_3$,...,$y_k$) are the points in the space and D is the distance between the two points x and y.

Convolutional Neural Network (CNN) is preferred as a deep learning method in this study. CNN, which can easily identify and classify objects with minimal pre-processing, is successful in analyzing visual images and can easily separate the required features with its multi-layered structure [13]. The major layers in CNN consist of convolutional layer, pooling layer, activation function layer and fully connected layer. Python library scikit-learn was used with initial image width and height equal to 180px. Multiclass classification with sixteen number of classes was done. The number of epoch used was calculated as: (Total number of images) mod (Batch size). Convolutional Kernel of size (3, 3), max pooling matrix of size (2, 2) and ReLu was used as an activation function in each convolutional layer. The convolutional layer output was flattened with 0.5 dropout value. Sigmoid activation function was used in fully connected layer with dense value as 16 was used for the experiment.

D. Given Input Image Prediction

The input image absolute file path is provided as an input. For the given image, feature extraction is performed. The extracted featured is used for the prediction using the previously saved machine learning models. The predicted plant category (healthy or disease) name as shown in Fig. 5 is displayed on the integrated development environment console or in browser.

IV. Results And Discussion

For support vector machine we achieved the accuracy of 78.61% where regularization parameter C is set to 100, gamma to 0.0001 and tolerance in optimization to 0.001. The given hyper parameters values were changed manually to get the best possible accuracy. Fig. 4 and Fig. 5 show the graphical user interface for image feature extraction and plant disease prediction. Table. I. Shows the precision, recall, f1-score and support of testing images for SVM based classification

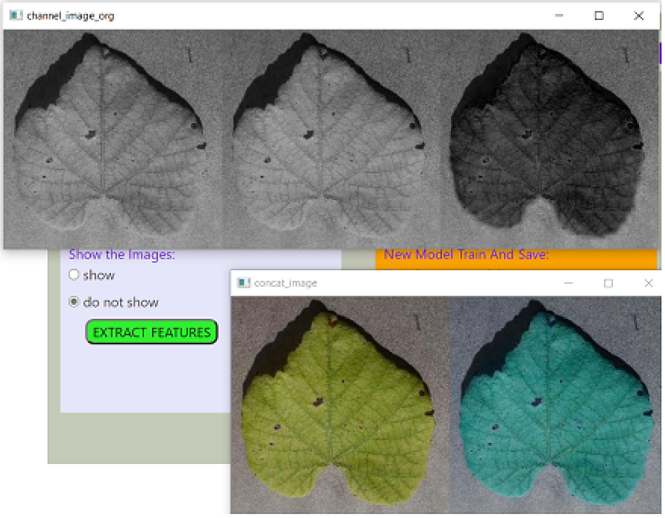


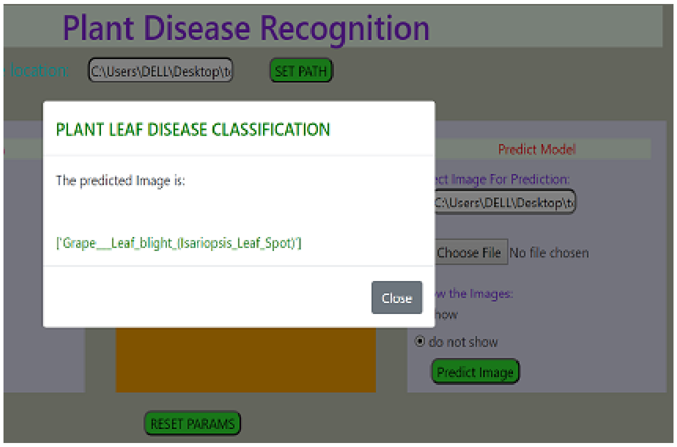
Fig. 4. Image feature extraction

Feature extraction process of image with different color channel (red, green and blue) and original image is shown in Fig. 4.

In KNN the nearest number of neighbors k value is used as 5 and an accuracy of 76.969% is obtained. Though best accuracy can be obtained at k=1 we used k=5 to prevent use of single value voting for prediction. The plot for KNN is show in Fig. 6 and Fig. 7.

The plot Fig. 6 shows the multiclass classification of the images of sixteen different categories in testing. It shows the number of times the image of given category is categorized in which category moving along the truth and predicted axes as

ISSN 2395-8618


Plant Disease Recognition

location: C:\Users\DELL\Desktop\t SET PATH

PLANT LEAF DISEASE CLASSIFICATION

The predicted Image is:

['Grape___Leaf_blight_(Isariopsis_Leaf_Spot)']

Close

Fig. 5. Plant disease prediction using SVM

TABLE I
SVM CLASSIFICATION REPORT FOR TESTING DATA

Labels	*Precision*	*Recall*	*F1-Score*	*Support*
Apple___Apple_scab	0.75	0.68	0.71	418
Apple___Black_rot	0.62	0.73	0.67	380
Apple___Cedar_apple_rust	0.79	0.78	0.79	326
Apple___healthy	0.88	0.85	0.87	411
Chery__Powdery	0.82	0.88	0.85	346
Cherry__healthy	0.91	0.96	0.93	356
Grape___Black_rot	0.63	0.75	0.69	356
Grape__(Black_Measles)	0.75	0.71	0.73	365
Grape__blight_(Isariopsis)	0.81	0.83	0.82	327
Grape___healthy	0.83	0.87	0.85	340
Peach___Bacterial_spot	0.81	0.75	0.78	384
Peach___healthy	0.91	0.92	0.91	341
Pepper_bell_Bacterial_spot	0.68	0.59	0.63	388
Pepper_bell_healthy	0.76	0.68	0.72	410
Strawberry___Leaf_scorch	0.89	0.85	0.87	371
Strawberry___healthy	0.84	0.88	0.86	395

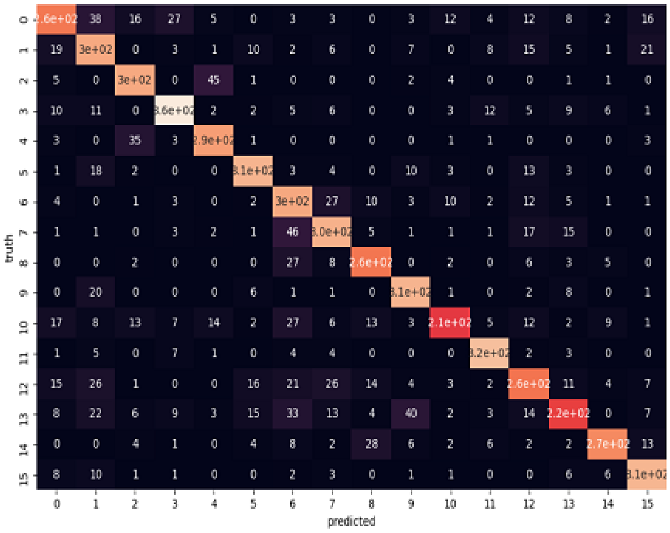


Fig. 6. Heatmap plot for KNN

shown in the figure. The weighted average value for precision, recall, f1-score and support are 0.78, 0.77, 0.77 and 5914 respectively for testing images for KNN.

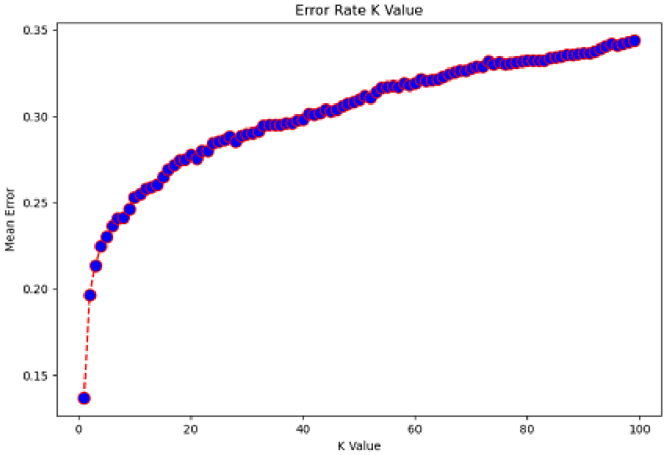


Fig. 7. Elbow criterion plot for KNN

The Elbow criterion plot in Fig. 7 shows the mean error for the given K value in the iteration. The least mean error value as shown is for k 1 but we took K value as 5 to prevent single value voting which might consider the outlier data as true positive value.

Random Forest model with accuracy 87.436% is created with 250 numbers of estimators and the heatmap plot as shown in Fig. 8. The weighted average value for precision, recall and, f1-score is 0.88 and support value is 5914 for testing images for KNN.

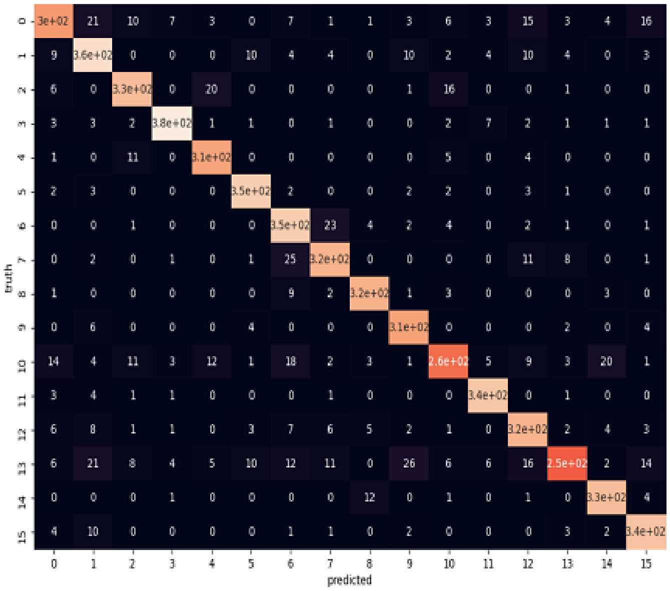


Fig. 8. Heatmap plot for random forest

Convolutional Neural Network model has training accuracy of 97.89% and validation accuracy of 99.01% which is trained

ISSN 2395-8618

for 147 epochs with 29567 and 7391 images for training and validation respectively. The CNN model was trained using the Google Colaboratory GPU named as device:GPU:0. The model accuracy and loss plot for training and testing is shown in Fig. 9.

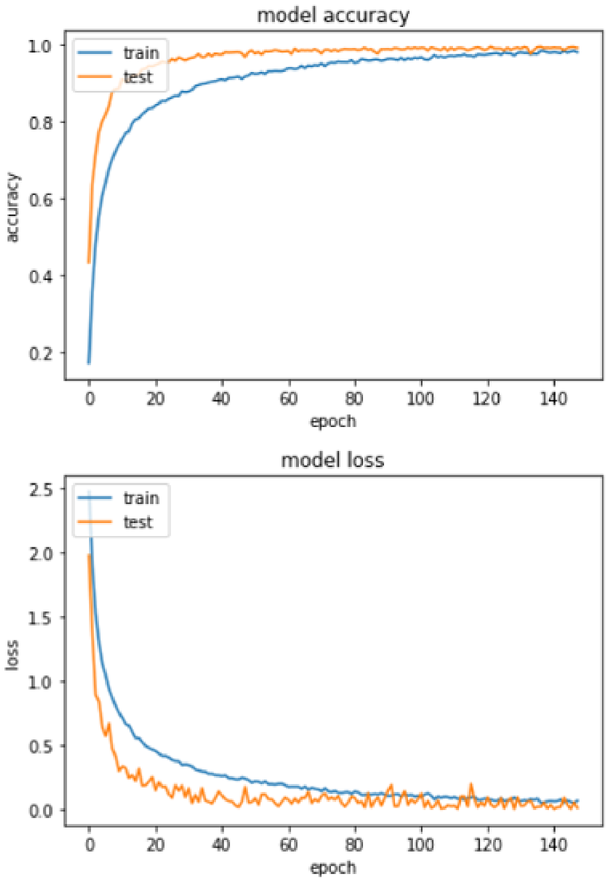


Fig. 9. CNN accuray and losss plot for train and test

TABLE II
TESTING DATA OUTPUT COMPARISON OF DIFFERENT MODELS

Model Name	***Accuracy in %***
Convolutional Neural Network (CNN)	97.89%
Random forest (RF)	87.43%
Support Vector Machine (SVM)	78.61%
K Nearest Neighbor (KNN)	76.96%

The CNN produced the result with much accuracy than other machine learning model but the time and the physical resource like RAM and CPU (Google Colaboratory GPU named as device:GPU:0) used by CNN is high compared to other models. The obtained accuracy of work for CNN is 2.52% higher than the accuracy obtained by author of paper [4] which was 94.74%. The accuracy of SVM, KNN and RF obtained in our work is higher than the authors of paper [7] where SVM, KNN and RF produced the highest accuracy of 40.33%, 66.76% and 70.14% respectively.

V. Conclusion And Future Works

This work presents the various plant leaf diseases recognition using Haralick feature extraction technique and machine learning models like SVM, KNN, Random forest and CNN. Contrast, correlation, inverse difference moments, entropy and images RGB color standard deviation are extracted image features for this work. Among the given models CNN produced the highest level of accuracy 97.89% followed by Random forest 87.436%, SVM 78.61% and KNN 76.969%.

Future works can be done to include more plant species with different diseases and texture characteristics. Further improvements in the given models prediction can be done by extracting much distinct features from the plants leaves. Grid search or other algorithms can be used to find the best optimal value of the hyper parameters used in different models used in efficient way.

References

[1] R. Anand, S. Veni, and J. Aravinth, “An application of image processing techniques for detection of diseases on brinjal leaves using k-means clustering method,” in *2016 International Conference on Recent Trends in Information Technology (ICRTIT)*, Chennai, 2016, pp. 1–6, doi: 10.1109/ICRTIT.2016.7569531.

[2] C. G. Dhaware and K. H. Wanjale, “A modern approach for plant leaf disease classification which depends on leaf image processing,” in *2017 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, 2017, pp. 1–4, doi: 10.1109/ICCCI.2017.8117733.

[3] N. Krithika and A. G. Selvarani, “An individual grape leaf disease identification using leaf skeletons and KNN classification,” in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, 2017, pp. 1–5, doi: 10.1109/ICIIECS.2017.8275951.

[4] S. V. Militante, B. D. Gerardo, and N. V. Dionisio, “Plant leaf detection and disease recognition using deep learning,” in *2019 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)*, Yunlin, Taiwan, 2019, pp. 579–582, doi: 10.1109/ECICE47484.2019.8942686.

[5] S. Mathulaprangsan, K. Lanthong, D. Jetpipattanapong, S. Sateanpattanakul, and S. Patarapuwadol, “Rice diseases recognition using effective deep learning models,” in *2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*, Pattaya, Thailand, 2020, pp. 386–389, doi: 10.1109/ECTIDAMTNCON48261.2020.9090709.

[6] P. B. Padol and A. A. Yadav, “SVM classifier based grape leaf disease detection,” in *2016 Conference on Advances in Signal Processing (CASP)*, Pune, 2016, pp. 175–179, doi: 10.1109/CASP.2016.7746160.

[7] S. Ramesh, R. Hebbar, M. Niveditha, R. Pooja, B. N. Prasad, N. Shashank, and P. Vinod, “Plant disease detection using machine learning,” in *2018 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C)*, Bangalore, 2018, pp. 41–45, doi: 10.1109/ICDI3C.2018.00017.

[8] G. Mukherjee, A. Chatterjee, and B. Tudu, “Study on the potential of combined GLCM features towards medicinal plant classification,” in *2016 2nd International Conference on Control, Instrumentation, Energy & Communication (CIEC)*, Kolkata, 2016, pp. 98–102, doi: 10.1109/CIEC.2016.7513746.

[9] S. Singh, D. Srivastava, and S. Agarwal, “GLCM and its application in pattern recognition,” in *2017 5th International Symposium on Computational and Business Intelligence (ISCBI)*, Dubai, 2017, pp. 20–25, doi: 10.1109/ISCBI.2017.8053537.

[10] F. F. Chamasemani and Y. P. Singh, “Multi-class support vector machine (SVM) classifiers—an application in hypothyroid detection and classification,” in *2011 Sixth International Conference on Bio-Inspired Computing: Theories and Applications*, Penang, 2011, pp. 351–356, doi: 10.1109/BIC-TA.2011.51.

ISSN 2395-8618

[11] S. Xiaowu, L. Lizhen, W. Hanshi, S. Wei, and L. Jingli, “Image classification via support vector machine,” in *2015 4th International Conference on Computer Science and Network Technology (ICCSNT)*, Harbin, 2015, pp. 485–489, doi: 10.1109/ICCSNT.2015.7490795.

[12] M. P. Vaishnnave, K. S. Devi, P. Srinivasan, and G. A. P. Jothi, “Detection and classification of groundnut leaf diseases using KNN classifier,” in *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, Pondicherry, India, 2019, pp. 1–5, doi: 10.1109/ICSCAN.2019.8878733.

[13] M. Sardogan, A. Tuncer, and Y. Ozen, “Plant leaf disease detection and classification based on CNN with LVQ algorithm,” in *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, Sarajevo, 2018, pp. 382–385, doi: 10.1109/UBMK.2018.8566635.