Comparative Study Of Machine learning models For Healthy Plant Detection

Pranita Chaudhary Assistant Professor, Department of CSE(AIML) PCCOE, Pune
pranita.chaudhary@ pccoepune.org

Sujata Swami Student, Department of CSE(AIML) PCCOE, Pune
sujata.swami23@pccoepune.org

Adinna Thaware Student, Department of CSE(AIML) PCCOE, Pune
adinna.thaware23@pccoepune.org

Vipul Phatangare Student, Department of CSE(AIML) PCCOE, Pune
vipul.phatangare23@pccoepune.org

Diksha Bhosale Student, Department of CSE(AIML) PCCOE, Pune
diksha.bhosale23@pccoepune.org

Abstract: ***This paper contains high concern to the healthy plant detection in agriculture to produce crops in healthy conditions with minimal financial loss. This work compares six machine learning algorithms for the detection of healthy plants, including SVM, Random Forest, XGBoost, AdaBoost, Gradient Boosting, and Decision Tree. An image dataset of healthy and unhealthy plants is employed like mango, bell pepper, potato and Pongamia Pinnata. All these images are processed and converted into numerical data. The dimension of image data is too large; hence, PCA reduces features from 49152 to 50 for improved performance, computation and reduces execution time. Performances of all algorithms will be tested based on accuracy, precision, recall, and processing time. Based on the above, this research finding would present the best method of machine learning for the practical detection of healthy plant, thus bringing successful information to both farmers and agricultural experts improving crop management strategies.***

Keywords: SVM , Random forest, XGBoost , AdaBoost, Gradient Boosting, Decision Tree , PCA , Image Processing.

INTRODUCTION

Plant diseases are one of the major dangers in agriculture in terms of yield and quality of crops. Because of the increased demand for food each day, there is a need for efficient methods of early detection of healthy plant. Identification methods in relation to plant diseases have been conventionally taking a lot of time and therefore require expertise in knowledge related to them, which slows up the response process. However, recent developments in machine learning offer new prospects for automation. A variety of machine learning algorithms such as SVM, Random Forest, XGBoost, AdaBoost, Gradient Boosting, and Decision Tree are tested to determine their relative performance regarding plant disease detection in this study. For PCA feature reduction, an attempt is made that makes the analysis easier while retaining all the important information regarding image data complexity. We are currently evaluating the performance of these algorithms in order to find the best and most efficient methods for the detection of healthy plant. This work benefits the field of agricultural technology as well as helps farmers in making wise judgments about saving their crops and increasing productivity to a healthy extent.

LITERATURE REVIEW

Bansilal Bairwa, Vivek Kumar, Zulqurnain Omar, Asif Sayed, Vishal Kumar (2023) [6]. Tomato Plant Disease Detection Using Image Processing for Agriculture Application. The primary aim of the paper is to propose a detection method in tomato plants based on advanced techniques of image processing. The method of acquiring the images of tomato plants and processing them through the extraction of their texture and color features followed by the identification of early blight, late blight, and bacterial spot through diagnostics will be discussed. It is going to help farmers through proper identification of diseases and early detection, thus promoting improved crop management and less usage of pesticides. The system was tested on an immense dataset and showed significant accuracy in the classification of disease.

H. K. Kondaveeti, K. G. Ujini, B. V. V. Pavankumar, B. S. Tarun, and S. C. Gopi. (2023) [3]. Ensemble Learning for Plant Disease Detection. In this work, six base models were proposed for the ensemble learning-based system to detect plant diseases: Inception V3, MobileNet, MobileNetV2, VGG16, GoogleNet, and ResNet50. The base models were trained on a dataset of 87,000 images of healthy and diseased plant leaves from 38 classes and obtained accuracies ranging between 72.7% and 97.2%. Soft and hard voting classifiers increased the system accuracy to as high as 97.8% with soft voting and 98.3% by using hard voting. This ensemble learning approach demonstrates promise for utilization as a means of combining multiple models into efficient and accurate plant disease identification.

Abdul Hafiz Bin Abdul Wahab, Rahimi Zahari, Tiong Hoo Lim. Detection of Chilli Plant Diseases Using K-Means Segmented Support Vector Machine [6]. An Artificial Intelligence-based image processing algorithm that can identify diseases in chili plants by analyzing leaf images is proposed in this paper. Image segmentation is carried out with the k-means clustering algorithm and different SVM algorithms are compared to test the classification in proposed work. These characteristics of the processed images are used to classify them in categories like background, healthy, and unhealthy (Cucumber Mosaic). The results have shown how SVM succeeds in making a differentiation between healthy plants and diseased ones and would be useful in detecting plant diseases in real-time with precision. Research by Vishnu S Babu et al.(2021) conducted the study for comparative analysis of different machine learning techniques for plant disease detection with a concentration on the Indian agriculture sector [4]. Their work shows potential in the accurate classification of plant diseases through machine learning but also identified needs or challenges for large amounts of datasets for training models. These studies emphasize the use of image processing and machine learning based techniques in conditions of early disease detection, therefore avoiding crop losses. The major limiting factor here is the availability of large datasets. Moreover, as reflected from the literature, more work should be done to adapt these techniques for specific crops and regional agricultural practices.

Kumud, Deepa Gupta, Sujeet Kumar, Methily Johri, Aditri Ashish (2021)[1]. Plant Disease Identification and Detection Using Machine Learning Algorithms. The given paper is centered on the rather vital problem of diseases in plants, causing huge agricultural losses and a lack of food security. Researchers used open-source data sets of images of healthy and infected leaves, applying machine learning and computer vision techniques for detector training. The study proves the capability of the model using Random Forest to assure accuracy at a confidence level of 95%. The approach could be useful in helping farmers manage crop production effectively and become proactive in the prevention of disease threats so that food can be assured to meet the populating needs of a growing population.

METHODOLOGY

1.1 Description of the Dataset:

The dataset contains 5660 images of diseased and healthy leaves from four plants - mango, bell pepper, potato, and Pongamia Pinnata. The images captured in this dataset are under different lighting conditions and angles that portray several diseases of the leaves. The images are of varying resolutions and quality thus necessitating preprocessing to standardize them.

Fig 1 . Images of diseased and healthy plant leaves from 'Plant Village' dataset.

3.2 Data Preprocessing

To prepare a dataset for the training of the machine we applied the following preprocessing steps:

Resized images: The images were resized to a standard size, for instance, 128x128 pixels, which means equal dimension and easier to work with images and not losing significant features either.
Normalization: Pixel value is scaled from 0 to 1 such that the probability distribution could be equal and increase the chance of convergence.

$$X' = \frac{X - Xmin}{Xmax - Xmin} \quad \text{(i)}$$

Data augmentation: We use techniques such as rotation, flipping, zooming, and shifting to artificially increase the sample size for our training set. These improve the generalization of the network and reduce overfitting.

3.3 Feature Extraction:
The raw images were converted into a number, suitable for machine learning using the techniques of computer vision. Each resized image contained 49152 features (128x128x3). To reduce the dimensionality, thus reducing the required computation, we applied PCA which decreased the features from 49152 to 50 features while retaining most of the information.

3.4 Supervised Algorithms:
On the extracted features, we trained several traditional supervised learning algorithms on leaf diseases classification. SVM: It is a typical classifier with the ability to create a hyper plane that classifies the different classes of the unhealthy plant. Decision Tree: A simple yet very effective model based on which it splits the data into a tree- based structure for classification using the importance of the feature.

3.5 Ensemble Technique Algorithms:
A few more algorithms are implemented to reduce errors and improve accuracy :
a. Ensemble Method : Random Forest is a decision tree ensemble technique which decreases the variance and improves the performance by averaging the predictions of many trees.
b. XGBoost, Gradient Boosting and AdaBoost: These algorithms build models sequentially in that each new model focuses on correcting the errors of the previous ones. XGBoost and Gradient Boosting have been particularly distinguished by the high accuracy they attain for classification tasks.
c. Custom Boosting Algorithm: We have designed and trained a custom boosting algorithm using 5 different base models of SVM classifiers. The predictions of the same were put together to boost the accuracy. It is also possible through such an ensemble approach to make more robust decisions using multiple models to make a decision.

3.6 Final Model:
After training and evaluating all of the models, we chose an algorithm showing performance through accuracy, precision, recall, and F1-score. Then, we fine-tuned the hyperparameters of the models to optimize their performance. The algorithm will be finalized based on generalization properties toward unseen data as well as effectiveness to find unhealthy plant precisely in different species of plants.

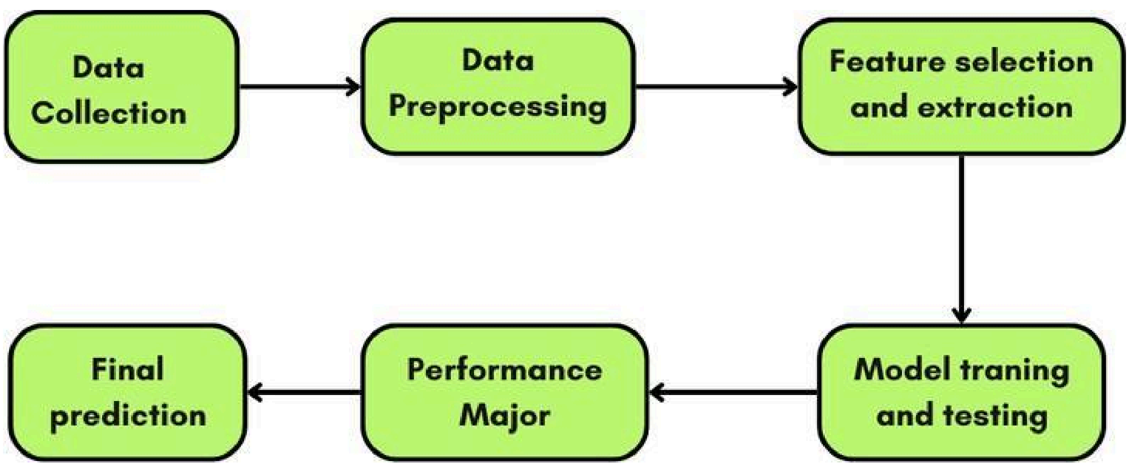

Fig 2. Flowchart of Standard Workflow for plant disease detection

RESULTS

In this research, the performances of several machine learning algorithms for plant disease detection using leaf images have been compared on the basis training , testing, mean squared error, precision, recall, and F1-score.

Performance measures	Formulas	
F1 Score	TP/(TP +½(FP+FN))	(ii)
Precision Call	TP/(TP+FP)	(iii)
Recall	TP/(TP+FN)	(iv)
MSE	$1/n\ \Sigma(y - y')^2$	(v)

Table 1. Evaluation Measures

The best performance is taken by the SVM model with training accuracy and testing accuracy amounting to 97.50% and 93.63%, respectively, and the lowest MSE of 0.0636. For both models, MSEs were significant, though the lower value for Random Forest 0.1773 compared to that of the Decision Tree model, which stood at 0.1926. Training accuracy for this model also stood pretty high at 88.94% against 88.26% for the Decision Tree model. Testing accuracy was at 82.27% for the Random Forest model and 80.74% for the Decision Tree model. AdaBoost achieved a training accuracy of 92.02% and testing accuracy of 89.22% with an MSE of 0.1337. For XGBoost, it performed equally with its training accuracy as 91.31% and testing accuracy at 86.63%, whereas the MSE was at 0.1078. For Gradient Boosting, training accuracy is at 88.56%, whereas testing accuracy is 86.69%. The MSE is equal to 0.1331.Finally, the SVM custom boosting algorithm provided 95.08% training accuracy, 89.16% testing accuracy, and an MSE of 0.1084, which was very effective as well.

Algorithms	Training accuracy	Testing accuracy	MSE
SVM	97.50%	93.63%	0.0636
Decision Tree	88.26%	80.74%	0.1926
Random Forest	88.94%	82.27%	0.1773
AdaBoost	92.02%	89.22%	0.1337
Gradient Boost	88.56%	86.69%	0.1331
Xgboost	91.39%	86.63%	0.1078
Custom SVM boost	95.08%	89.16%	0.1084

Table 2: Summary of Model Performance in Plant Disease Detection I

The table above shows a summary of various machine learning algorithms applied for the identification of plant diseases in terms of training accuracy, testing accuracy, and mean squared error (MSE). It can be observed that the SVM gave better results in performance, at a 93.63% testing accuracy and at the lowest MSE score of 0.0636, indicating a minor prediction error. The Decision Tree model was not as accurate during testing with 80.74% accuracy and a greater MSE of 0.1926 as the Random Forest model had with a testing accuracy of 82.27% and a marginally better MSE of 0.1773. Both the ensemble methods, AdaBoost and XGBoost, had exceptional results with 89.22% and 86.63% testing accuracies, respectively, along with relatively moderate MSE values. Gradient Boosting had a similar outcome with a testing accuracy of 86.69% and an MSE of 0.1331. Finally, the Custom SVM Boost algorithm showed great effectiveness, with a testing accuracy of 89.16% and a mean squared error (MSE) of 0.1084, thus placing it as a competitive method along with the standard SVM.

Algorithms	F1 score	Precision	Recall
SVM	0.930	0.930	0.935
Decision Tree	0.895	0.890	0.895
Random Forest	0.785	0.765	0.875
AdaBoost	0.885	0.880	0.885
Gradient Boost	0.855	0.845	0.870
Xgboost	0.855	0.845	0.865
Custom SVM boost	0.885	0.895	0.875

Table 3: Summary of Model Performance in Plant Disease Detection II

The table represents the performance of various machine learning algorithms used in the detection of plant diseases based on F1-score, precision, and recall. SVM is found to be the best model that has the highest F1-score (0.930), precision (0.930), and recall (0.935) values to illustrate its robust capabilities in balancing accuracy and consistency in terms of prediction. The Decision Tree did well with an F1-score of 0.895, with precision and recall values at 0.890 and 0.895, respectively. The Random Forest, on the other hand, did poorly with an F1-score of 0.785, having a lower precision of 0.765 but a respectable recall of 0.875. The ensemble methods of AdaBoost and Gradient Boost had similar performance with F1-scores of 0.885 and 0.855, respectively, while having balanced precision and recall metrics. Custom SVM Boost presented competitive performance; the algorithm managed to achieve an F1-score equal to 0.885, coupled with precision of 0.895 and recall 0.875, placing it as an efficient substitute for plant disease detection. The output of XGBoost was an F1-score equal to 0.855, with lower precision 0.845 and recall 0.865.

CONCLUSION

In this research, we compared various machine learning algorithms on a task related to plant disease detection from leaf images. The results show that the SVM succeeded the best with the highest accuracy level of 93.63% for testing and the lowest mean square error for this task, which indicates the SVM is highly effective for this kind of classification task. They were effective again, with a high level of accuracy and good generalization. Moreover, the custom SVM boosting algorithm of ensemble techniques and AdaBoost produced good results. Other models such as Random Forest and Decision Tree produced relatively lower performance, but if all things are considered, XGBoost and Gradient Boosting were good results concerning the combination of accuracy and error reduction. The most reliable in the detection of unhealthy plant using image-based features were the SVM and the boosting methods. This could possibly be helpful in early diagnosis of the disease and ultimately contribute to better agricultural decisions.

REFERENCES

[1] Kumud, D. Gupta, S. Kumar, M. Johri, and A. Ashish, “Plant Disease Identification and Detection Using Machine Learning Algorithms,” 2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART), pp. 106–111, Dec. 2023, doi: 10.1109/smart59791.2023.10428282.

[2] S. Sinha, A. R. Sinha, C. Ghose, H. Bhardwaj, and J. Singh, “Automated Detection of Potato Plant Diseases using Machine Learning Algorithms,” 2023 International Conference on Recent Advances in Science and Engineering Technology (ICRASET), pp. 1–6, Nov. 2023, doi: 10.1109/icraset59632.2023.10420350.

[3] H. K. Kondaveeti, K. G. Ujini, B. V. V. Pavankumar, B. S. Tarun, and S. C. Gopi, “Plant Disease Detection Using Ensemble Learning,” 2023 2+nd International Conference on Computational Systems and Communication (ICCSC), pp. 1–6, Mar. 2023, doi: 10.1109/iccsc56913.2023.10142982.

[4] V. S. Babu, R. S. Kumar, and R. Sunder, “A Comparative Study on Disease Detection of Plants using Machine Learning Techniques,” 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 1937–1941, Mar. 2021, doi: 10.1109/icaccs51430.2021.9441844.

[5] M. Jhuria, A. Kumar, and R. Borse, “Image processing for smart farming: Detection of disease and fruit grading,” 2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013), pp. 521–526, Dec. 2013, doi: 10.1109/iciip.2013.6707647.

[6] Abdul Wahab, R. Zahari, and T. H. Lim, “Detecting diseases in Chilli Plants Using K-Means Segmented Support Vector Machine,” 2019 3rd International Conference on Imaging, Signal Processing and Communication (ICISPC), pp. 57–61, Jul. 2019, doi: 10.1109/icispc.2019.8935722.

[7] B. Bairwa, V. Kumar, Z. Omar, A. Sayed, and V. Kumar, “Tomato Plant Disease Detection Using Image Processing for Agriculture Application,” 2024 IEEE International Conference on Big Data & Machine Learning (ICBDML), pp. 270–276, Feb. 2024, doi: 10.1109/icbdml60909.2024.10577337.

[8] A. S. Tulshan and N. Raul, “Plant Leaf Disease Detection using Machine Learning,” 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–6, Jul. 2019, doi: 10.1109/icccnt45670.2019.8944556.

[9] P. Bose, S. DUTTA, V. Goyal, and S. K. Bandyopadhyay, “Leaf Diseases Detection of Medicinal Plants based on Image Processing and Machine Learning Processes ,” Jul. 2021, doi: 10.20944/preprints202107.0638.v1.