A Hybrid Approach to Identify Plant Leaf Disease using Machine Learning and Image Processing

Prabhjot Kaur
Department of Computer Science and Engineering
IET Bhaddal Technical Campus ,Ropar
Punjab, India

Barinderjit Kaur
Department of Computer Science and Engineering
IET Bhaddal Technical Campus ,Ropar
Punjab,India

ABSTRACT
One of the main factors affecting agricultural output decrease globally is plant disease and crop losses must be avoided through early diagnosis of these diseases.Machine learning (ML) and Image processing techniques have shown great potential in automating plant disease identification. The recent developments in the field of image processing and ML for plant leaf disease identification, include pre-processing, image acquisition,classification, andfeature extraction. Additionally, we provide a summary of the various ML algorithms utilized for plant leaf disease classification, including supervised, unsupervised, and deep learning algorithms. Furthermore, we discuss some of the challenges faced in plant leaf disease identification using MLand image processing techniques, such as the need for large-scale datasets having 256x256 size of images and the generalization of models across different plant species and environmental conditions. Machine learning algorithms can be used to learn from multiple sources of data. Fusion methods may be used to merge data from several sources to produce anML model that is more reliable and precise. To get better results, we will apply fusion techniques that can be utilized to improve the robustness and accuracy of plant leaf disease identification by combining multiple sources of information. Fusion techniques involve combining multiple sources of information, such as different types of images or features, to create a more comprehensive representation of the plant leaf and its disease.

General Terms
Machine Learning, Image Processing, Decision Tree, Hybrid Approach, Deep learningapproaches

Keywords
Plant leaf disease, classification, Feature extraction, Plant Leaf Disease, Image Segmentation

1. INTRODUCTION
Agriculture has been one of the key forces for economic growth in India. When a farmer selects a crop, he or she takes into account the kind of soil, the climate in the area, and the product's economic value.Growing populations, shifting weather patterns, as well as political uncertainty prompted the agricultural sectors to start exploring for new techniques to raise food production.This makes it possible for researchers to search for new, highly productive ideas. Using information technology and PA (Precision Agriculture), farmers may acquire knowledge and data to make the best decisions for high farm output.This cutting-edge technology offers efficient systems for increasing agricultural yield. Agriculture may experience economic prosperity by using this cutting-edge technology.PA has a broad range of uses like weed identification, plant pest identification,detection of plant diseases andcrop yield production, etc. Pesticides are utilized by farmers to control pests, avoid disease, and increase crop yield.Because ofpoor yield, crop diseases,industrial agriculture, and economic losses, farmers are having issues. As a result, emphasis is placed on the necessity to characterize an illness' diagnosis and severity as appropriate [1].

To deliver treatments that manage illnesses, a rapid and accurate disease diagnosis is crucial. It mostly depends on the distinctive symptoms shown by a damaged plant.Disease, as previously mentioned, may harm any component of a leaf, including the stem, flower, root, and many more.Leaf examination is seen to be the best method for diagnosing plants, nevertheless. The primary damage to the diseased leaf is deformation in shape, color, size, etc. It is outdated, time-consuming, and inappropriate to diagnose plant leaf disease using the traditional procedure, which includes experts.ICT use in agriculture is seen as a developing field that significantly alters the scope of rural and agricultural development. Artificial intelligence is a rapidly developing area of ICT that has a lot of potential for use in the agricultural industry.Additionally, computer vision develops the capability for computers to detect and comprehend data from digital images. ML has become a powerful computing paradigm in artificial intelligence that supports the resolution of several challenging computer vision tasks. ML enables computers to learn without the need for human involvement [2].

Studies have shown that DL (deep learning) approaches are efficient ways of classifying plant diseases. Using computer image processing and DL technology, this is to acquire rapid and accurate identification. Enhancing the correctness, reliability, and accuracy of image processing for detecting and classifying plant illness has been a top priority.Learners engaged in the growing process as well as skilled experts might benefit greatly from an automated system designed to aid in the diagnosis of plant diseases by the existence and obvious indicators of the plant. The researchers used a hybrid strategy to obtain a representation of plant disease utilising visualization approaches [3].The methodology in the study involves some key stages: reading image data, pre-processing of images, feature extraction, image segmentation,classification of images using the hybrid approach, and disease prediction. The data set of plant images includes images of Normal,Gray-spot, Black-mold, Late-mold, Bacterial spot, and Powdery-mildew as given in Figure 1. The results of the hybrid approach are compared with the CNN detection method.

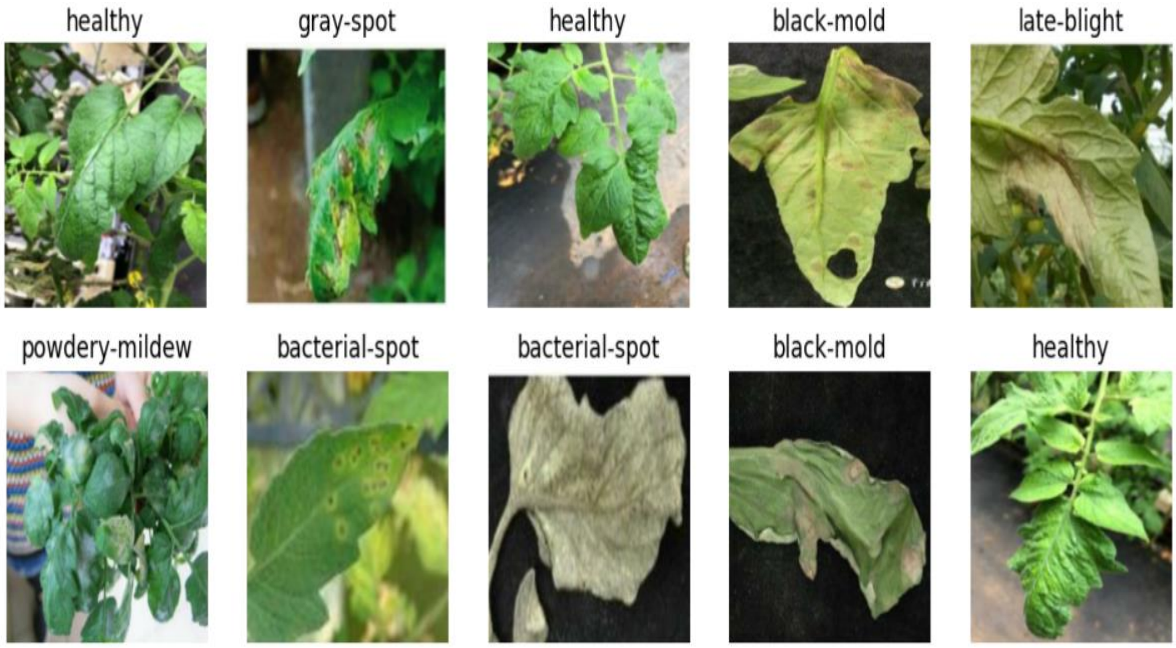

Fig 1: Dataset Images

2. LITERATURE REVIEW

Although ML algorithms are used in many different sectors, feature engineering is still the biggest challenge. With the development of DNN (deep neural network) and image processing, plant pathology may now access promising findings without time-consuming feature engineering. DNNs greatly improve the accuracy of picture categorization. The many DL methods used by researchers to identify plant diseases are presented in this section.

AlexNet was trained to identify previously unknown plant diseases by Mohanty et al. [4]. Model accuracy was drastically reduced since the conditions for testing and training images were different. Disease may sometimes be seen on the lower sides as well as the top sides of the leaves.Progressively Growing GANs is a different architecture designed by Tero K. et al. that utilizes progressive resolution complexity to train the networks. The model is trained using this technique's increasing resolution.For example, the resolution range begins with an input size of 4 x 4 and goes up to 8 x 8 until reaching an output resolution of 1024 x 1024. To enable the plant network to precisely and successfully apply the appropriate ML techniques and best-practice guidelines for various biotic & abiotic stress features, A. Singh et al. [6] provide a complete explanation and simple-to-use scientific classification of ML methodologies.S. Radhakrishnan et al [7]'s visualization and machine learning techniques are used to organize the backwoods land on the terrain dataset provided by the ASTER imaging equipment in order to interpret the accumulated information utilizing Box Plot & Heat Map.

P. Sharma et al. [8] proposed the AI-based automated PLDD and characterization for quick and easy disease diagnosis, characterization, and application of anticipated solutions.T.V. Reddy, and K. Sashirekhak [9] describe many PD categories and advanced ML & image processing algorithms to detect PD. This review also identifies important research gaps that will aid in future research into PA recognition.Malvika Ranjan, et al. [10] provide a technique of determination that is mostly visual but also calls for accurate judgment and scientific procedures. The impure leaf picture is noticed. Color segmentation leads to the creation of HSV characteristics.After being trained, the ANN (Artificial Neural Network) performs 80% more accurately than previous approaches in classifying ill and healthy samples from samples.This work has a little accuracy fault, which accounts for 80% of it. The neural network's accuracy in detecting plant leaf diseases is limited and has some detection mistakes.

An overview of the most notable conventional approaches for plant disease detection methods was provided by Sankaran S et al. in their publication [11]. These approaches include methods for detecting plant diseases based on spectroscopy, imaging, and volatile profiling. The advantages and limitations of different strategies are compared in the study.For the purpose of identifying different pests and illnesses in leaves, Babu [12] suggested and created the feed-forward neural network with a backpropagation model. The gradient descent parameters for ANNs that are required to minimize the error function may be calculated with the use of the backpropagation approach.A model for distinguishing yellow leaf curl disease-infected tomato leaves was put out by Mokhtar U. et al. in [13]. The classification model was built using 200 photos of diseased and healthy tomato leaves and an SVM with various kernels. The model has an average classification accuracy of 90%. With plant leaf photos and various quantities of data, Grinblat GL et al. [14] developed a comparable deep CNN technique for several plant identification tasks. Deep CNNs may also be used to identify pests and diseases that affect plants. This method was used to find pests and diseases in tomato plants.

3. RESEARCH METHODOLOGY

- Image Accumulation: Figure 1 shows that dataset accumulation is a crucial first step to take. During this stage, raw photos pertaining to plant leaf disease are gathered.[15]

- Image Pre-processing: Images must be pre-processed to get rid of the noise. Additionally, it is required for an image's HSV conversion. Prior to applying any computational method to a picture, pre-processing improves its quality by eliminating background noise and balancing the intensity of the image's many components.

- Image Segmentation: Following that, picture segmentation is carried out. Typically, undesired or background items are present around a picture. Techniques for picture segmentation must be used for that. Image segmentation is a method used in digital picture processing that divides the digital image into different relevant or focused parts needed for the issue domain.It often operates on pixels with comparable properties. It is utilised to separate and emphasise the foreground as well as the background of a picture. The proper characteristics may be extracted when the extraction is conducted.

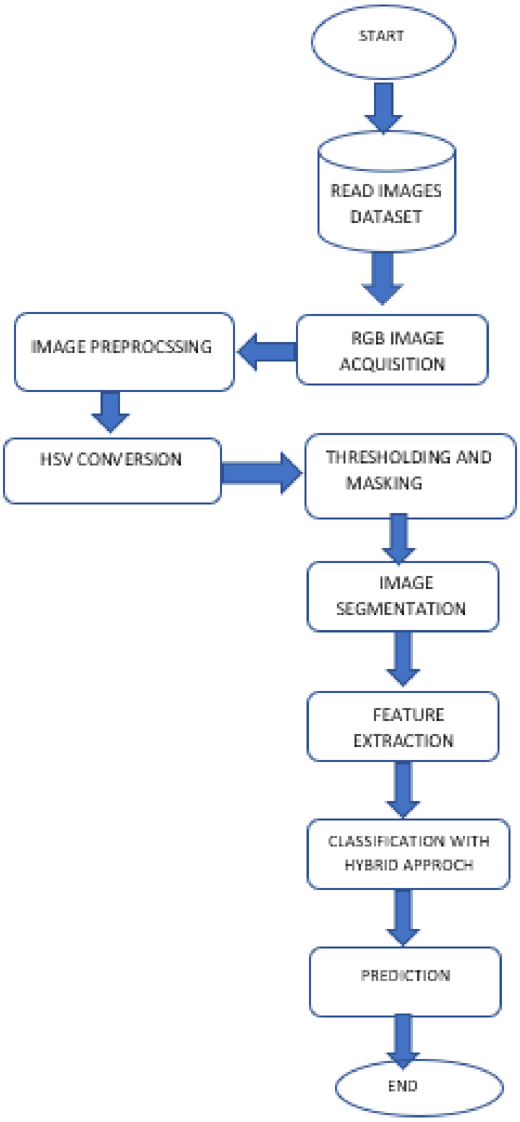

Fig. 2: Flowchart of Proposed approach

- Feature Selection and Extraction: A variety of feature extraction approaches are utilized to extract the relevant characteristics that may be used for classification once the diseased leaf has undergone the proper pre-processing and segmentation. Features may be thought of as measurable qualities that are acquired from the different focused areas of the picture. The first stage in every ML-based application is the identification and extraction of features. The collected characteristics significantly influence the ML algorithms' overall accuracy.

- Image Classification: The techniques of Principal Component Analysis are used for classifying the images. A dimensionality reduction approach called PCA (Principal Component Analysis) is used to convert high-dimensional data into a lower-dimensional space while preserving the most crucial details or patterns in the data. PCA is used in this instance to decrease the features from 150,528 to 300. The hybrid approach makes use of two systems XGBoost and Decision Tree. [16]X-Axis and Y-axis in Figure 3 are:

(a) X-Axis: Number of Components. The X-axis represents the number of components or features after applying PCA. It ranges from 1 to 300, indicating the different choices for the number of components you want to keep.

(b) Y-Axis: Proportion of Variance Explained. The percentage of variation explained by the chosen number of components is shown on the Y-axis. It shows how much information or variability in the original dataset is captured by the reduced feature set.

The graph is created by calculating the proportion of variance explained for each number of components and plotting it against the corresponding number of components. The proportion of variance explained is a measure of how well the selected components capture the original dataset's variability. [17]

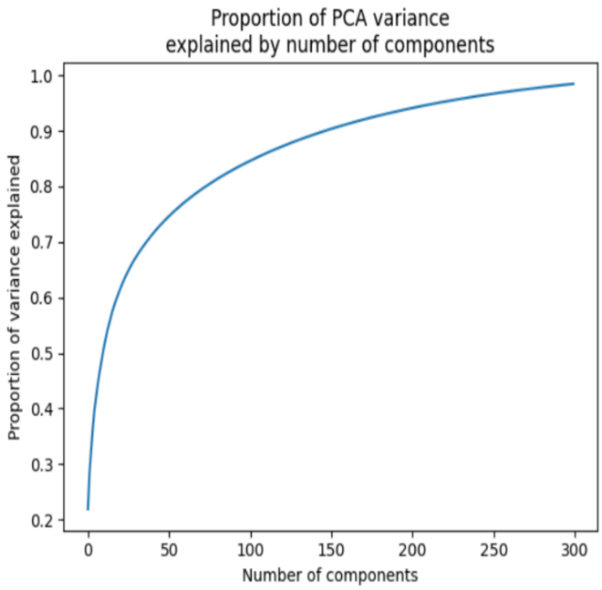

Fig.3: Fusion Technique-PCA Result

LDA is a dimensionality reduction technique commonly used for classification tasks. It aims to find a lower-dimensional representation of the data while maximizing the separation between different classes. In your case, it seems that you applied LDA to fuse the 300 features into two new features (coordinate1 and coordinate2) for visualization purposes. After applying the LDA fusion technique, you plotted the classification results on a graph. The x-axis signifiesthe values of the first new feature (coordinate1), and the y-axis signifies the values of the second new feature (coordinate2). Each point on the graph corresponds to one of the 371 images in your dataset. By plotting the data in this way, you can observe how the different classes are separated or clustered in the two-dimensional space defined by coordinate1 and coordinate2. The goal of this visualization is to understand if the fusion technique applied by LDA has successfully improved the separation between the different classes, making them more distinguishable or discriminative. Depending on the specific dataset and the effectiveness of the fusion technique, you may observe distinct clusters or patterns on the graph, where images belonging to the same class are closer to each other and well-separated from images belonging to other classes. This would indicate that the LDA fusion technique has successfully enhanced the discriminative power of the data.

XGBoost and Decision Tree: The model was trained and validated using the scalable tree-boosting machine learning classifier XGBOOST. It is shown in Eq. (1) [18] as follows:

$$S = \sum_{x=1}^{N} fx\ (P)$$

N specifies the number of estimators, where *fx (P)*=the xth tree in the forest. P stands for the feature vector, and S stands for the classification value. One of the most significant algorithms utilised in the development of expert systems is the decision tree.due to the way a classification or regression model was implemented within the framework. It reduces a dataset to a relatively tiny subset. A decision tree's ultimate conclusion may simply be converted to a rule by sequentially mapping from root to leaf nodes. The study that makes use of a decision tree method is shown below.[19]

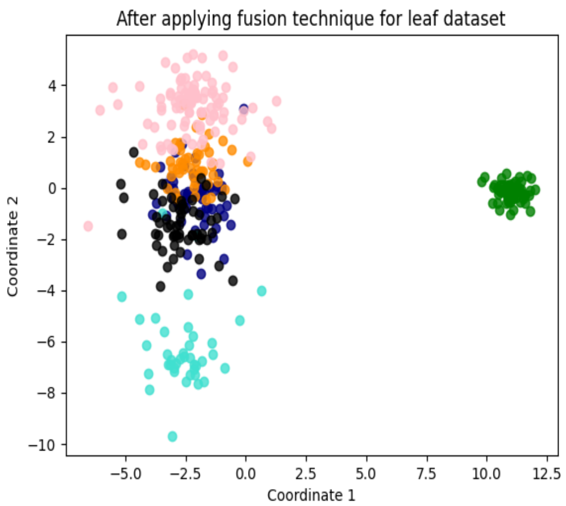

Fig.4: Result after Fusion Techniques

```
Confusion Matrix :-
 [[48  0  0  0  0  0]
 [ 0 36  0  0  0  0]
 [ 0  0 44  0  0  0]
 [ 0  0  0 51  0  0]
 [ 0  0  0  0 43  0]
 [ 0  0  0  0  0 74]]
Classification Report :-
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        48
           1       1.00      1.00      1.00        36
           2       1.00      1.00      1.00        44
           3       1.00      1.00      1.00        51
           4       1.00      1.00      1.00        43
           5       1.00      1.00      1.00        74
```

Fig.5: Confusion Matrix after applying Hybrid Approach (XGBoost and Decision Tree) on resultant data of fusion technique

To evaluate how effectively a classification model is doing, a confusion matrix is utilized. It summarises the results of the model's predictions by comparing the actual values of the target variable with the predicted values.Typically, the confusion matrix has 4 cells: TP, TN, FP, and FN, true positives, true negatives,false positives, and false negatives, respectively[20].

4. RESULTS

In this particular section, we present the comparison of the results of the approaches used in the study. Two approaches were employed: Convolutional Neural Network (CNN), which has been implemented in earlier research works, and a novel hybrid approach proposed in this paper.

The accuracy achieved by the CNN approach was found to be 99.06%. However, through the implementation of the newly proposed hybrid approach, a remarkable accuracy of 100% was attained. This substantial improvement in accuracy showcases the efficacy and potential of the hybrid approach in addressing the research problem at hand.

These results highlight the superiority of the hybrid approach over the CNN method and emphasize its significance in advancing the field. The attainment of perfect accuracy signifies the successful integration of various techniques and features within the hybrid model, leading to enhanced performance and improved classification outcomes.

In this research, the hybrid approach implemented is a combination of XGBoost and Decision Tree algorithms. Figure 6 illustrates the comparison of accuracy between these two approaches.

Figure 7 displays a comparison of the F1-score between the proposed hybrid model and the base paper results. The F1-score is a metric that combines precision and recall, providing a balanced measure of a model's performance.

According to the results presented in Figure 7, the CNN approach in the base paper achieved an F1-score of 98.8. On the other hand, the hybrid model proposed in this study demonstrated a perfect F1-score of 100. This significant improvement in the F1-score highlights the enhanced performance and effectiveness of the hybrid approach over the CNN method.

The comparison depicted in Figure 7 visually illustrates the superiority of the hybrid model in terms of F1-score when compared to the base paper results. The perfect F1-score attained by the hybrid model signifies its ability to achieve a balance between precision and recall, resulting in improved classification accuracy.

These findings further support the assertion that the hybrid approach presented in this research paper outperforms the existing CNN method and provides a more accurate and reliable solution for the problem at hand. The results depicted in Figure 7 emphasize the advantages of the proposed hybrid model and its potential for practical applications in the field.

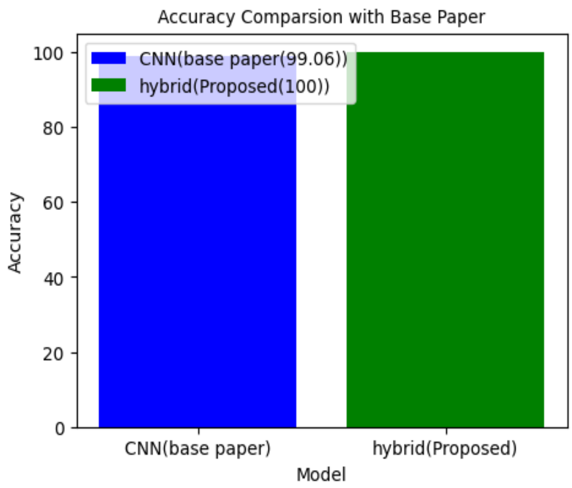

Fig 6: Compare Accuracy of Hybrid Approach(XGBoost and Decision Tree) with Base Paper Result

```
Text(0.5, 1.0, 'Comparsion  f1 with Base Paper')
```

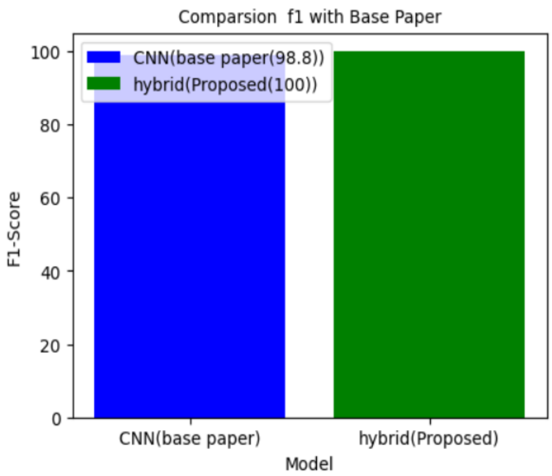

Fig. 7: Comparison F! -Score of proposed models with Base Paper Result.

5. CONCLUSION

In conclusion, this research paper introduced a novel hybrid approach for identifying plant leaf diseases by leveraging ML and image processing techniques. The study compared the performance of the proposed hybrid approach with the conventional CNN method used in earlier research works.

The results clearly demonstrated the superiority of the hybrid approach in terms of accuracy. The hybrid model achieved a remarkable accuracy rate of 100%, surpassing the 99.06% accuracy achieved by the CNN approach. This substantial improvement highlights the effectiveness of combining XGBoost and Decision Tree algorithms within the hybrid model, resulting in enhanced classification outcomes.

Moreover, the comparison of F1-scores between the proposed hybrid model and the base paper results further solidified the superiority of the hybrid approach. The hybrid model attained a perfect F1-score of 100, outperforming the CNN approach which achieved an F1-score of 98.8. These findings emphasize the capability of the hybrid approach to achieve a balance between precision and recall, leading to more accurate and reliable disease identification.

The successful implementation of the hybrid approach signifies its potential for practical applications in the field of plant leaf disease identification. By leveraging the strengths of both ML and image processing techniques, the hybrid model offers an innovative and robust solution to address the challenges associated with accurate disease detection.

Overall, this research paper contributes to the advancement of the field by introducing and validating a hybrid approach that outperforms the conventional CNN method. The results validate the effectiveness and potential of the proposed hybrid model for identifying plant leaf diseases, offering promising prospects for future research and practical implementations.

6. REFERENCES

[1] Sujatha, R., Chatterjee, J. M., Jhanjhi, N. Z., & Brohi, S. N. (2021). Performance of deep learning vs machine learning in plant leaf disease detection. Microprocessors and Microsystems, 80, 103615.

[2] Ganatra, N., & Patel, A. (2020). A multiclass plant leaf disease detection using image processing and machine learning techniques. International Journal on Emerging Technologies, 11(2), 1082-1086.

[3] Jasim, M. A., & Al-Tuwaijari, J. M. (2020, April). Plant leaf disease detection and classification using image processing and deep learning techniques. In 2020 International Conference on Computer Science and Software Engineering (CSASE) (pp. 259-265). IEEE.

[4] Mohanty, R., Wankhede, P., Singh, D., & Vakhare, P. (2022, May). Tomato Plant Leaves Disease Detection using Machine Learning. In 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 544-549). IEEE.

[5] Tero K, Timo A, Samuli L, Jaakko L. Progressive growing of GANs for improved quality, stability, and variation. CoRR, abs/1710.10196, 2017.

[6] Singh, B. Ganapathy Subramanian, A.K. Singh, S. Sarkar, Machine learning for high-throughput stress phenotyping in plants, Trends Plant Sci. 21 (2) (2016) 110–124.

[7] S. Radhakrishnan, A.S. Lakshminarayanan, J.M. Chatterjee, D.J. Hemanth, Forest data visualization and land mapping using support vector machines and decision trees, Earth Sci Inform (2020) 1–19.

[8] P. Sharma, P. Hans, S.C. Gupta, Classification of Plant Leaf Diseases Using Machine Learning and Image Pre-processing Techniques, in: 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, 2020, pp. 480–484.

[9] T.V. Reddy, K. Sashirekhak, Examination on advanced machine learning techniques for plant leaf disease detection from leaf imagery, J Crit Rev 7 (5) (2020) 1208–1221.

[10] M. Ranjan1, M. R. Weginwar, NehaJoshi, Prof. A.B. Ingole, detection and classification of leaf disease using artificial neural network, International Journal of Technical Research and Applications eISSN: 2320-8163, Volume 3, Issue 3 (May-June 2015), PP. 331-333.

[11] Sankaran S, Mishra A, Ehsani R, Davis C. A review of advanced techniques for detecting plant diseases. Comput Electron Agric 2010;72(1):1–13. doi:10. 1016/j.compag.2010.02.007.

[12] Babu BSR MSP. Leaves recognition using back propagation neural network-advice for pest and disease control on crops. IndiaKisan 2007:13. http: //www.indiakisan.net/web/leafrecgnition.pdf (accessed October 12, 2019).

[13] Mokhtar U, Ali MAS, Hassanien AE, Hefny H.

Identifying two of tomatoes leaf viruses using support vector machine. In: Mandal JK, Satapathy SC, Kumar Sanyal M, Sarkar PP, Mukhopadhyay A, editors. Information Systems Design and Intelligent Applications. New Delhi: Springer India; 2015. p. 771–82.

[14] Grinblat GL, Uzal LC, Larese MG, Granitto PM. Deep learning for plant identification using vein morphological patterns. Comput Electron Agric 2016;127:418–24. doi:10.1016/j.compag.2016.07.003.

[15] Falk, K. G., Jubery, T. Z., Mirnezami, S. V., Parmley, K. A., Sarkar, S., Singh, A., ... & Singh, A. K. (2020). Computer vision and machine learning enabled soybean root phenotyping pipeline. Plant methods, 16, 1-19.

[16] Ganatra, N., & Patel, A. (2020). A multiclass plant leaf disease detection using image processing and machine learning techniques. International Journal on Emerging Technologies, 11(2), 1082-1086.

[17] Gupta, S. K. (2019). Identification of plant leaf diseases using a nine-layer deep convolutional neural network.

[18] Kannan, E. (2022). An Efficient Deep Neural Network for Disease Detection in Rice Plant Using XGBOOST Ensemble Learning Framework. International Journal of Intelligent Systems and Applications in Engineering, 10(3), 116-128.

[19] Rajesh, B., Vardhan, M. V. S., & Sujihelen, L. (2020, June). Leaf disease detection and classification by decision tree. In 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184) (pp. 705-708). IEEE.

[20] Hassan, S. M., Jasinski, M., Leonowicz, Z., Jasinska, E., & Maji, A. K. (2021). Plant disease identification using shallow convolutional neural network. Agronomy, 11(12),2388.