International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 / https://doi.org/10.14445/22315381/IJETT-V70I10P229
Volume 70 Issue 10, 298-306, October 2022


*Original Article*

# A Hybrid Multi-class Classification Model for the Detection of Leaf Disease using XGBoost and SVM

Rina Mahakud[1]*, Binod Kumar Pattanayak[1], Bibudhendu Pati[2]

[1]*Department of Computer Science and Engineering, Siksha 'O' Anusandhan Deemed to be University Institute of Technical Education and Research, Bhubaneswar, India*
[2]*Department of Computer Science, Rama Devi Women's University, Bhubaneswar, India*

[1*]*Corresponding Author : rinamahakud@gmail.com*

Received: 26 June 2022 Revised: 19 August 2022 Accepted: 15 October 2022 Published: 20 October 2022

***Abstract*** - *The primary source of human nutrition is generated from plants. Plants get affected by the disease, from crop farming to the production of foods. So, leaf disease identification is a crucial task in the farming industry. Various machine learning models are developed and evaluated by multiple researchers to identify leaf disease with significant results. This article compares the multi-class classification result of different state-of-art machine learning methods (SVM, LR, RF, K-NN, DT, Extra Tree) with hybrid models. The model's performance is measured by precision, accuracy, F1 score, and confusion matrix. The experimentation shows that the hybrid model (MCLXGB) provides an impressive result of 93.22%, whereas the Decision Tree gives the least effective result of 74.57% accuracy.*

***Keywords*** - *Leaf Disease, Data Augmentation, Multi-Class Classification, XGBoost, Machine Learning.*

## 1. Introduction

The Indian economy is directly related to agricultural products. About 70% [1] of the population in India depends on cultivation for their income. Product quality and quantity can be maintained only when diseases are prevented early. It is challenging to detect and identify plant diseases at a very early stage. Plants are getting affected by diseases seasonally due to climate change which is classified as biotic and abiotic. Various factors, such as environmental effects, changes in climate, unsuitable crop selection, weeds, etc., are responsible for the plants' infections. The use of pesticides harms plants and harms nature [2]. With the rapid advancement of technology, farmers still follow traditional farming methods in identifying disease plants by physical visualization. But sometimes, these old techniques cannot provide accurate detection of infection and a stage of infection due to equal symptoms shown by the leaf. This process leads to the development of wrong-controlled strategies for diseased cultivation. Thus, the quality and quantity of the product cannot be maintained [3]. Experts can accurately classify the type of disease based on symptoms, which is very expensive and time-consuming. Therefore, an automatic monitoring system should be applied to enhance the quality of the product. So, automatic monitoring and recognition systems are developed for detecting and locating the disease [4].

The issues mentioned above can be resolved by using computer vision (CV), the internet of things (IoT), artificial intelligence (AI), machine learning technology (ML), and deep learning (DL). The ML approach, a subset of AI, is used in various fields. ML technologies are applied in agriculture stock market prediction, healthcare monitoring, human behavior evaluation, and disease detection [5].

### *1.1. Motivation and Contribution*

The present work is based on the use of ML to identify tomato leaf disease. The production of the crop is dependent on different climate-changing factors. Thus, with climate change, the plant is susceptible to disease caused mainly due to bacteria, fungus, and viruses [6]. In the past decade, a human has mostly performed disease identification. It is very difficult and costly for the farmer to consult experts in a remote area. The production of tomato crops can be prevented if the disease is detected earlier. To overcome this problem, an automated identification system should be developed. Therefore, automated disease identification models need to be designed. Recently with the advancement of technology, ML [7], DL [8], and Transfer Learning (TL) [9] have contributed significantly to the solution of classification problems with high accuracy. Most ML techniques are used on balanced datasets to categorize whether plants are infected or not. However, the proposed hybrid model was more accurate when it was concentrated on unbalanced datasets.

The significant contribution of this work is summarized as follows:

- Classify multi-class tomato leaf disease using traditional as well as hybrid approaches of ML methods.
- Creating a hybrid model that incorporates cutting-edge and learning-based methods (XGBoost plus SVM)
- An automated leaf disease detection technique is proposed to identify tomato leaf illness, which will help the farmers increase the yield quality in less time.

This manuscript is divided into different sections, i.e., a summary of related work, discusses materials and methodology, proposed model, experimental setup and result analysis, and finally, a conclusion and scope for future direction.

## 2. Related Work

This section presents some machine learning-based algorithms related to the plant disease identification domain. From the related study, a few research gaps were also represented.

Karthick Manoj et al. [10] proposed an efficient pixel replacement-based segmented method to enhance the IoT and ML environment classification. Here author implements an SVM classifier to obtain an overall accuracy of 92.325%. But the author can take a more significant number of data both for training and testing purposes which can improve the model's accuracy. Panigrahi et al. [11] proposed a classification model to classify the diseased maize plants at a very early stage. The author collects the dataset from the plant village dataset of 3,823 images and splits the dataset to 90% for training and 10% for testing. Compared with other SVM, K-NN, DT, and NB ML-based classifiers, their experimentation approach achieved an accuracy of 79.23%. But authors can use a high-dimensional dataset for better disease detection accuracy. Jaisakthi et al. [12] proposed a machine learning-based automated disease identification and classification model. Here authors train the model with three classifiers such as SVM, RF, and AdaBoost, among which the SVM classifier performs better with an accuracy of 93% compared to others. Vijayalakshmi et al. [13] proposed an early prediction of plant disease models based on IoT, ML, and image processing methodology. IoT-enabled cameras can capture diseased leaf images. Then, extract the ROI features from captured images to improve the image quality. Finally, the author uses an SVM classifier to classify the disease with 92% accuracy.

Ashourloo et al. [14] focused on disease diagnosis using machine learning techniques such as partial least square regression (PLSR), v-SVR (v-support vector regression), and GPR (Gaussian process regression). The proposed model is trained by a small collected dataset of wheat rust crops. From the comparative analysis, it was found that GPR provides better performance on a small sample dataset. Kumar et al. [15] stated a multi-label classification model for plant disease prediction. The proposed machine learning model provides 98% accuracy in predicting the disease. In a similar vein, Fenu et al. [16] investigated the use of machine learning to forecast the severity of potato late blight disease. With a 96 percent accuracy rate, ANN is used to forecast the disease severity, while SVM is used to classify the disease with a 98 percent accuracy rate. A classifier was created by Bhatia et al. [17] to predict tomato powdery mildew illness at a very early stage.

Compared to another classifier, the suggested classifier, the medium gaussian support vector machine (MGSVM), does statistical analysis and offers greater performance with an accuracy of 94.74 percent. A machine learning-based disease detection algorithm on different classifiers is proposed in reference [18]. The author considers 100 images of 15 types of disease leaves. In their proposed work KNN classifier provides the best performance than other classifiers.

A multi-class classification system for detecting potato leaf disease was proposed by Singh et al. [19]. The author reports a 95.99% accuracy rate for the K-mean image segmentation, GLCM feature extraction, and subsequent disease classification. However, they can offer a comparison of the disease classification with other ones that already exist. Maria et al. [34] describe an automated disease detection methodology. The author of the proposed methodology contrasted transfer learning techniques with more conventional machine learning techniques. Image segmentation, feature extraction, and disease classification are all done using machine learning techniques. Herein, a number of classifiers are used for classification, but random forest provides better accuracy, about 81.68%, which is further improved to 90.08% by using transfer learning.

Considering the above literature analysis on plant disease identification and classification, numerous researchers extensively use machine learning algorithms. Most of the plant disease detection methods use ML methods to be low in accuracy. Hence, enhancing the classification methods by incorporating state-of-art learning methods can provide superior performance. In this experimentation, some of the parameters are tuned, and the model is evaluated through some of the statistical measures like accuracy, precision, f1 score, and confusion matrix, proving the performance of the proposed model.

## 3. Methodology for Disease Classification

The proposed model performs three major steps (a) Data Pre-Processing, (b)Data Augmentation, and finally, (c) Classification. The workflow of the proposed methodology is shown in Figure 1.

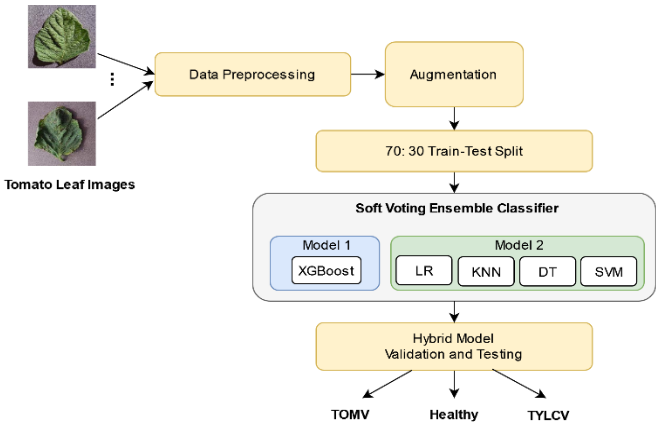


**Fig. 1 Workflow of the proposed model**

### *3.1. Dataset Description*

This experiment's dataset required for analysis contains healthy and diseased images. These images were collected from the "plant village" dataset. This dataset has 6660 images, comprising 82 nos. healthy and 6578 diseased. Out of 6578 diseased images,1110 images are Mosaic virus, and 5468 are yellow leaf curl virus. All images used in the dataset are resized to 227x227x3. Some sample images from the dataset used in the proposed work are represented in 3 classes, as shown in figure 2. The model was trained by randomly splitting the samples into 80% for training and 20% for testing.

(a)Tomato mosaic virus (b)Tomato yellow leaf curl (c)Tomato Healthy

**Fig. 2 Sample of tomato leaf images**

### *3.2. Data Augmentation*

In the deep learning model, a huge amount of data is used to train a pre-trained model, which can avoid overfitting issues in an adequate amount of data. Thus, data augmentation is applied. Data augmentation is an exploiting technique used to enhance the dataset. Data augmentation can be done by simply flipping, inverting, and rotating the original images [34]. Herein, the sample of images is increased to six times through the data augmentation process. Table 1 shows the used tomato leaf disease class in the study.

## 4. Methodology

The multi-class classification shows the classification result in different categories of class. In the dataset, each sample category is mapped into one class label. Depending on the features, classification result is generated. Many algorithms are now implemented in the ML domain to train the model with a training dataset that predicts the classification result for the testing dataset. In this proposed work, we have considered three classes based on images taken to classify the disease.

**Table 1. Details of tomato leaf disease image dataset used in the experiment**

| **Leaf diseases** | Tomato mosaic virus (ToMV) | Tomato yellow leaf curl virus (TYLCV) | Total |
|---|---|---|---|
| **Original images** | 1110 | 5468 | 6578 |
| **No of augmentations images** | 910 | 5268 | 6178 |
| **Training images** | 5460 | 31608 | 37068 |
| **Testing images** | 200 | 200 | 400 |

### *4.1. Linear-support vector machine (Linear-SVM)*

SVM [21–23], the most popular supervised learning technique, is utilized to address classification and regression issues. This algorithm analyzes the input labeled data through an optimal hyperplane. The two or more classes of data maximize the marginal distance on both sides of the hyperplane, and those data close to the marginal line are taken as the winning class.

### *4.2. Random Forest (RF)*

[24–25] propose that RF is a classifier that resolves regression and classification problems using an ensemble approach. This algorithm is conceptualized based on the decision tree (DT) algorithm. This approach uses several DTs during the training process to make a decision. In the RF approach, the whole training dataset is divided into subsets so the different DTs can be formed, and then these DTs are ensembled. The main drawback of the DT algorithm is overfitting which can be solved by taking advantage of RF.

### *4.3. Logistic Regression (LR)*

LR [26] is a predictive ML-based analytic algorithm that solves regression and classification problems. To solve the classification problem, a cost function is established, from which an optimal parameter is obtained through the optimization method, and at the end, the model is validated. So, LR can be represented in Eq. (1).

$$G = \frac{1}{1 + e^{-f(x)}} \tag{1}$$

Where G=probability of occurrence of the class lies the value between 0 to 1, $f(x)$=Function consisting of features x related to corresponding coefficients α and $f(x)$ can be represented in the following Eq. (2).

$$f(x) = x_0 + x_1\alpha_1 + x_2\alpha_2 + \cdots + x_K\alpha_K + \varepsilon \tag{2}$$

Where $\varepsilon$ represents an error

### 4.4. K-Nearest Neighbors (K-NN)

K-NN [35] is a simple, robust, and oldest non-parametric classification algorithm. This algorithm is used for solving statistical estimation of the multi-class classification problem and pattern recognition. K-NN classifier calculates the most relevant related vectors present in the neighbors based on their comparability. In this algorithm mostly, the distance function is calculated by using the Euclidean equation, which is shown in Eq. (3)

$$E(P,Q) = \sqrt{\sum_{P=1}^{n}(P_K - Q_K)^2} \quad (3)$$

Where E (P, Q) defines the distance measured between vectors P & Q, K defines the positive integer, and n = quantity of highlights in the vector.

### 4.5. Decision Tree (DT)

It is stated by Wani et al. [28] that a decision tree is one of the successful algorithms that perform better in solving classification and regression problems. Mostly DT resolves the overfitting problem. Random forest algorithm constructed by summing multiple decision trees. DT can be defined as in Eq. (4)

$$R = Q_2 log_2(Q_2) - Q_1 log_2(Q_1) - Q_0 log_2(Q_0) \quad (4)$$

Where R denotes, the sample taken in the problem, Q2, Q1, and Q0 are proportions of values of three classes.

### 4.6. XGBOOST

XGBOOST [29-31]is a scalable tree-boosting machine learning classifier used to train and validate the model. It can be represented in Eq. (5)

$$S = \sum_{x=1}^{N} f_x(P) \quad (5)$$

Where $f_x(P)$= $x^{th}$ tree in the forest, N defines no. of estimators. S represents the classification value, and P denotes the feature vector.

### 4.7. Extra Tree

According to Xie et al., the Extra tree strategy is based on a decision tree [36]. With the use of this technique, the challenging classification problem may be broken down into a series of choices that resolve the over-fitting issue in a model. This model can be defined through Eq. (6)

$$E(n) = 1 - \sum_{j=1}^{L}(\frac{P_j}{Q})^2 \quad (6)$$

Where E(n) represents Node, L is the classes of a sample taken in test data, Pj is the sample size, and Q is the sample size in node.

## 5. Experimental Set Up and Result Analysis

The proposed research model, an image-based ML approach, is considered for classifying tomato plant diseases. The system has a configuration with processor: 3.70 GHz AMD Ryzen 5 5600X 6-Core, RAM:16.0 GB, HDD: 250 GB SSD &1 TB, GPU: NVIDIA GeForce RTX 3060 has been used for experimental study. The success of the classification models is assessed using performance measures like accuracy, precision, recall, F1 score, and others. The proposed research model uses the steps listed below from Algorithm 1.

**Algorithm 1: Proposed Model (MCLXGB)**

**Input:** Diseased images taken from the plant village dataset

**Output:** Performance Metrics and classification report

1. *Image dataset Acquisition*
2. *Data pre-processing*
3. *Splitting the dataset into Train (70%) and test (30%)*
4. *Modeling for multi-class classification*
5. *Calculating the classification metrics*
6. *Plot the confusion matrix and mis-classification Error*

The plant village dataset is used to assess and test the effectiveness of the suggested approach. The results are also compared with the hybrid model classifiers like K-NN, LR, DT, SVM, Extra Tree, RF, and the hybrid model.

We used the 5-fold cross-validation in this experiment to address the stratified classification problem. The folds are chosen to include nearly equal amounts of the target class in each fold. By adjusting the parameter, the classifier's performance is maximised. In the suggested model, the parameters are tuned to simplify the model, resulting in improved accuracy and model performance. The following parameters were selected for our proposed model: "base score=0.5," "gamma=0," "learning rate=0.1," "max delta step=0," "max depth=3," "n estimators=100," "objective='binary:logistic'," "reg alpha=0," and "reg lambda=1."

A supervised ensemble machine learning classifier that can manage the non-linear behavior of the signal is called Extreme Gradient Boosting (XGBoost) [33]. The reference [29-30] observed that XGBoost is implemented in solving binary classification problems. So, in this research, we have implemented a multi-class classification problem using this algorithm and found that the MCLXGB algorithm successfully classifies the plant disease. Figure 3 displays the XGBoost model's train-test error throughout several epochs. This investigation leads to the conclusion that the model converges after 60 epochs. Table 2 displays the XGBoost model's experimental performance.

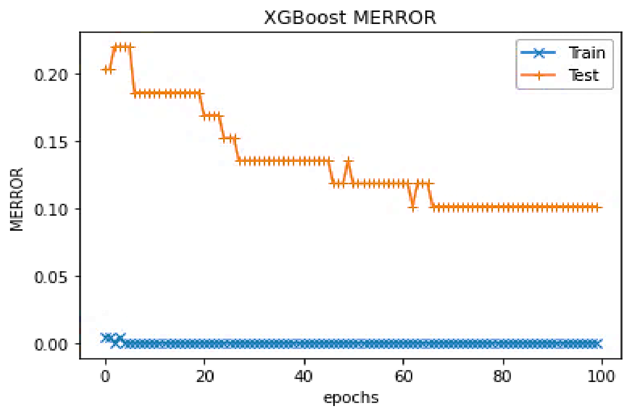


**Fig. 3 Train-Test Error**

Precision, recall, f1-score, and accuracy are some measurement factors that can be used to assess the classifier's performance. Table 2 shows the test accuracy scores for each class using several machine-learning classifiers. Table 2 shows that the suggested model achieves the highest accurate value and the best accuracy score. Compared to hybrid machine learning-based classifiers, traditional machine learning algorithms are straightforward but produce the worst results after training. This process involves a multi-class classification issue. As a result, criteria including precision, recall, and F1 score are used to evaluate each class. Comparing the suggested model to every conventional machine learning method, accuracy is provided at 93.22 percent. Values in the table that are bolded and highlighted show the model's best results.

### *5.1 Confusion Matrix*

To determine the experimental performance of the suggested model, the confusion matrix table of each class based on the classifier is employed, which provides transparent information about mapping correct and erroneous (misclassification) classes on test data. Each cell in the confusion matrix can be mapped using the calculation procedure for True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Figure 4-5 shows the confusion matrix of the hybrid proposed and state-of-the-art machine learning approaches.

**LR + XGBoost**

| | | **Actual** | | | |
|---|---|---|---|---|---|
| | | (TYLCV) | (ToMV) | Healthy | ∑ |
| **Predicted** | (TYLCV) | 17 | 2 | 0 | 19 |
| | (ToMV) | 3 | 18 | 0 | 21 |
| | Healthy | 0 | 0 | 18 | 18 |
| | ∑ | 20 | 20 | 19 | |

**RF + XGBoost**

| | | **Actual** | | | |
|---|---|---|---|---|---|
| | | (TYLCV) | (ToMV) | Healthy | ∑ |
| **Predicted** | (TYLCV) | 18 | 2 | 0 | 20 |
| | (ToMV) | 2 | 17 | 0 | 19 |
| | Healthy | 0 | 1 | 19 | 20 |
| | ∑ | 20 | 20 | 19 | |

**KNN + XGBoost**

| | | **Actual** | | | |
|---|---|---|---|---|---|
| | | (TYLCV) | (ToMV) | Healthy | ∑ |
| **Predicted** | (TYLCV) | 18 | 5 | 0 | 22 |
| | (ToMV) | 2 | 14 | 0 | 16 |
| | Healthy | 0 | 1 | 19 | 20 |
| | ∑ | 20 | 20 | 19 | |

**MCLXGB (Proposed Model)**

| | | **Actual** | | | |
|---|---|---|---|---|---|
| | | (TYLCV) | (ToMV) | Healthy | ∑ |
| **Predicted** | (TYLCV) | 18 | 1 | 0 | 19 |
| | (ToMV) | 2 | 18 | 0 | 20 |
| | Healthy | 0 | 1 | 19 | 20 |
| | ∑ | 20 | 20 | 19 | |

**Fig. 4 Hybrid proposed approach**

**Table 2. Test accuracy score of each class with different machine learning classifier**

| Classifiers | Class | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|---|
| K-NN | 0 | 0.95 | 0.70 | 0.81 | 0.8474 |
| | 1 | 0.65 | 0.93 | 0.76 | |
| | 2 | 0.95 | 1.00 | 0.97 | |
| LR | 0 | 0.85 | 0.85 | 0.85 | 0.8983 |
| | 1 | 0.90 | 0.86 | 0.88 | |
| | 2 | 0.95 | 1.00 | 0.97 | |
| DT | 0 | 0.70 | 0.67 | 0.68 | 0.7457 |
| | 1 | 0.60 | 0.63 | 0.62 | |
| | 2 | 0.95 | 0.95 | 0.95 | |
| SVM | 0 | 0.85 | 0.89 | 0.87 | 0.8983 |
| | 1 | 0.85 | 0.85 | 0.85 | |
| | 2 | 1.00 | 0.95 | 0.97 | |
| XGBoost | 0 | 0.90 | 0.86 | 0.88 | 0.8983 |
| | 1 | 0.80 | 0.89 | 0.84 | |
| | 2 | 1.00 | 0.95 | 0.97 | |
| ExtraTree | 0 | 0.85 | 0.89 | 0.87 | 0.8813 |
| | 1 | 0.85 | 0.81 | 0.83 | |
| | 2 | 0.95 | 0.95 | 0.95 | |
| RF | 0 | 0.85 | 0.85 | 0.85 | 0.8644 |
| | 1 | 0.75 | 0.83 | 0.79 | |
| | 2 | 1.00 | 0.90 | 0.95 | |
| LR + XGBoost | 0 | 0.85 | 0.85 | 0.85 | 0.8983 |
| | 1 | 0.90 | 0.86 | 0.88 | |
| | 2 | 0.95 | 1.00 | 0.97 | |
| RF + XGBoost | 0 | 0.90 | 0.90 | 0.90 | 0.9152 |
| | 1 | 0.85 | 0.89 | 0.87 | |
| | 2 | 1.00 | 0.95 | 0.97 | |
| KNN + XGBoost | 0 | 0.90 | 0.78 | 0.84 | 0.9152 |
| | 1 | 0.70 | 0.88 | 0.78 | |
| | 2 | 1.00 | 0.95 | 0.97 | |
| Proposed Model (MCLXGB) | 0 | **0.90** | **0.95** | **0.92** | **0.9322** |
| | 1 | **0.90** | **0.90** | **0.90** | |
| | 2 | **1.00** | **0.95** | **0.97** | |

In the table-2, Tomato yellow leaf curl virus (TYLCV) represents class 0, Tomato mosaic virus (ToMV) represents class 1, and Healthy represents class 2.

Table 2 shows that the proposed model provides better accuracy than other state-of-art methods. To evaluate the performance of the proposed model, another parameter confusion matrix is also presented in Figure 4 and Figure 5. Figure 4 shows the hybrid approach results, whereas Figure 5 shows the results of all state-of-art classifier methods. In this analysis hybrid proposed model MCLXGB represents a more accurately classified result. The confusion matrix of this proposed model for class 0 18 samples is correctly classified as TYLCV, but only one sample is misclassified as ToMV. In class 1, two samples are misclassified as ToMV; similarly class-2, only one sample is misclassified as ToMV. So, the result shown in figure 5 confirms that the proposed model MCLXGB can perform better in classifying plant disease.

| | | Actual | | | |
|---|---|---|---|---|---|
| | | (TYLCV) | (ToMV) | Healthy | ∑ |
| Predicted | (TYLCV) | 19 | 7 | 1 | 27 |
| | (ToMV) | 1 | 13 | 0 | 14 |
| | Healthy | 0 | 0 | 18 | 18 |
| | ∑ | 20 | 20 | 19 | |

**K-NN**

| | | Actual | | | |
|---|---|---|---|---|---|
| | | (TYLCV) | (ToMV) | Healthy | ∑ |
| Predicted | (TYLCV) | 17 | 3 | 0 | 21 |
| | (ToMV) | 3 | 15 | 0 | 18 |
| | Healthy | 0 | 2 | 19 | 21 |
| | ∑ | 20 | 20 | 19 | |

**DT**

| | | Actual | | | |
|---|---|---|---|---|---|
| | | (TYLCV) | (ToMV) | Healthy | ∑ |
| Predicted | (TYLCV) | 17 | 2 | 1 | 20 |
| | (ToMV) | 3 | 18 | 0 | 21 |
| | Healthy | 0 | 0 | 18 | 18 |
| | ∑ | 20 | 20 | 19 | |

**LR**

| | | Actual | | | |
|---|---|---|---|---|---|
| | | (TYLCV) | (ToMV) | Healthy | ∑ |
| Predicted | (TYLCV) | 14 | 7 | 0 | 21 |
| | (ToMV) | 6 | 12 | 1 | 19 |
| | Healthy | 0 | 1 | 18 | 19 |
| | ∑ | 20 | 20 | 19 | |

**DT**

| | | Actual | | | |
|---|---|---|---|---|---|
| | | (TYLCV) | (ToMV) | Healthy | ∑ |
| Predicted | (TYLCV) | 17 | 2 | 0 | 19 |
| | (ToMV) | 3 | 17 | 0 | 20 |
| | Healthy | 0 | 1 | 19 | 20 |
| | ∑ | 20 | 20 | 19 | |

**SVM**

| | | Actual | | | |
|---|---|---|---|---|---|
| | | (TYLCV) | (ToMV) | Healthy | ∑ |
| Predicted | (TYLCV) | 18 | 3 | 0 | 21 |
| | (ToMV) | 2 | 16 | 0 | 18 |
| | Healthy | 0 | 1 | 19 | 20 |
| | ∑ | 20 | 20 | 19 | |

**XGBoost**

| | | Actual | | | |
|---|---|---|---|---|---|
| | | (TYLCV) | (ToMV) | Healthy | ∑ |
| Predicted | (TYLCV) | 17 | 2 | 0 | 19 |
| | (ToMV) | 3 | 17 | 1 | 21 |
| | Healthy | 0 | 1 | 18 | 19 |
| | ∑ | 20 | 20 | 19 | |

**Extra Tree**

**Fig. 5 State-of-art machine learning methods**

## 6. Conclusion

In the AI domain, ML and DL approaches are major performers. In our experimentation, algorithms of various ML models are taken into consideration. The measuring metrics like precision, recall, F1 score, accuracy, and the result of the confusion matrix are considered evaluation factors for the performance of the proposed model. On comparison of state- of -art, we achieved DT as 74.57%> K-NN as 84.74%>RF as 86.44%>Extra Tree as 88.13%> XGBoost, SVM, LR as 89.83%> and hybrid models LR + XGBoost as 89.83%> KNN + XGBoost as 91.52%> RF + XGBoost as 91.52%> SVM + XGBoost as 93.22%. In future work, real-time images can be captured regularly using IoT modules to analyze the data to benefit the farmers.

## Acknowledgments

We want to thank the reviewers of this manuscript for their valuable and effective efforts toward improving our manuscript. In addition, we are thankful to Dr. Millee Panigrahi and Dr. Dayal Kumar Behera for their support in preparing the manuscript.

## References

[1] JAWani, S. Sharma, M.Muzamil, S. Ahmed, S.Sharma, and S. Singh, "Machine Learning and Deep Learning Based Computational Techniques in Automatic Agricultural Diseases Detection: Methodologies, Applications, and Challenges," *Archives of Computational Methods in Engineering*, pp. 1-37, 2021.

[2] M.Ouhami, A.Hafiane, Y.Es-Saady, El Hajji, M., & R.Canals, "Computer Vision, Iot and Data Fusion for Crop Disease Detection using Machine Learning: A Survey and Ongoing Research," *Remote Sensing*, vol. 13, no. 13, pp. 2486, 2021.

[3] A.Almadhor, H. T. Rauf, M. I. U. Lali, R.Damaševičius, B.Alouffi, & A. Alharbi, "AI-Driven Framework for Recognition of Guava Plant Diseases through Machine Learning from DSLR Camera Sensor based High Resolution Imagery," *Sensors*, vol. 21, no. 11, pp. 3830, 2021.

[4] W.Haider, A. U.Rehman, N. M.Durrani, & S. U. Rehman, "A Generic Approach For Wheat Disease Classification and Verification Using Expert Opinion for Knowledge-Based Decisions," *IEEE Access*, vol. 9, pp. 31104-31129, 2021.

[5] T.VanKlompenburg, A. Kassahun, & C.Catal, "Crop Yield Prediction using Machine Learning: A Systematic Literature Review," *Computers and Electronics in Agriculture*, vol. 177, pp. 105709, 2020.

[6] P.Alagumariappan, N. J.Dewan, G. N.Muthukrishnan, B. K. B.Raju, R. A. A Bilal, & V.Sankaran, "Intelligent Plant Disease Identification System using Machine Learning," *Engineering Proceedings*, vol. 2, no. 1, pp. 49, 2020.

[7] S.Vijayalakshmi, G Balakrishnan, & S. N. Lakshmi, "Early Prediction of Plant Disease Using AI Enabled IOT," In *Data Science and Security*, pp. 303-309, 2021.

[8] K. R. Aravind, P. Raja, R. Aniirudh, K. V. Mukesh, R. Ashiwin, and G. Vikas, "Grape Crop Disease Classification using Transfer Learning Approach," In *International Conference on ISMAC in Computational Vision and Bio-Engineering*, pp. 1623-1633, 2018.

[9] Ferentinos, Konstantinos P, "Deep Learning Models for Plant Disease Detection and Diagnosis," *Computers and Electronics in Agriculture*, vol. 145, pp. 311-318, 2018.

[10] R.Karthickmanoj, J.Padmapriya, & T.A Sasilatha, "Novel Pixel Replacement-Based Segmentation and Double Feature Extraction Techniques for Efficient Classification of Plant Leaf Diseases," *Materials Today: Proceedings*, 2021.

[11] K. P.Panigrahi, H.Das, A. K.Sahoo, & S. C. Moharana, "Maize Leaf Disease Detection and Classification using Machine Learning Algorithms," In *Progress in Computing, Analytics and Networking*, pp. 659-669, 2020.

[12] S. M.Jaisakthi, P. Mirunalini, and D. Thenmozhi, "Grape Leaf Disease Identification using Machine Learning Techniques," *International Conference on Computational Intelligence in Data Science (ICCIDS)*, pp. 1-6, 2019.

[13] S.Vijayalakshmi, G. Balakrishnan, and S. Nithya Lakshmi, "Early Prediction of Plant Disease Using AI Enabled IOT," In *Data Science and Security*, pp.303-309, 2021.

[14] D.Ashourloo, H.Aghighi, Ali Akbar Matkan, Mohammad Reza Mobasheri, and A.M. Rad, "An Investigation into Machine Learning Regression Techniques for the Leaf Rust Disease Detection using Hyperspectral Measurement," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 9, pp. 4344-4351, 2016.

[15] M.Kumar, A.Kumar& V. S. Palaparthy, "Soil Sensors-Based Prediction System for Plant Diseases Using Exploratory Data Analysis and Machine Learning," *IEEE Sensors Journal*, vol. 21, no. 16, pp. 17455-17468, 2020.

[16] Gianni Fenu, and Francesca MaridinaMalloci, "An Application of Machine Learning Technique in Forecasting Crop Disease," In *Proceedings of the 2019 3rd International Conference on Big Data Research*, pp. 76-82, 2019.

[17] AnshulBhatia,Anuradha Chug, and AmitPrakash Singh. "Statistical Analysis of Machine Learning Techniques for Predicting Powdery Mildew Disease In Tomato Plants," *International Journal of Intelligent Engineering Informatics*, vol. 9, no. 1, pp. 24-58, 2021.

[18] P.Prathusha, K. E. Murthy, and K. Srinivas, "Plant Disease Detection Using Machine Learning Algorithms," In *International Conference on Computational and Bio Engineering*, pp. 213-220, 2019.

[19] A. Singh, &H.Kaur, "Potato Plant Leaves Disease Detection and Classification using Machine Learning Methodologies," In *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, pp. 012121, 2021.

[20] MargeshKeskar, Dhananjay D Maktedar, "Evolutionary Computing Driven ROI-Specific Spatio-Temporal Statistical Feature Learning Model for Medicinal Plant Disease Detection and Classification," *International Journal of Engineering Trends and Technology*, vol. 70, no. 6, pp. 165-184, 2022. Crossref, https://doi.org/10.14445/22315381/IJETT-V70I6P220

[21] P. K.,Sethy, N. K., Barpanda, A. K.Rath, & S. K. Behera, "Deep Feature Based Rice Leaf Disease Identification using Support Vector Machine," *Computers and Electronics in Agriculture*, vol. 175, pp. 105527, 2020.

[22] M,TakieddineSeddik, O.Kadri, C.Bouarouguene, &H.Brahimi, "Detection of Flooding Attack on OBS Network Using ant Colony Optimization and Machine Learning," *Computing and Systems*, vol. 25, no. 2, 2021.

[23] İ., Kahramanoğlu, E. D., Ülker, &S.Ülker, "Integrating Support Vector Machine (SVM) Technique and Contact Imaging for Fast Estimation of the Leaf Chlorophyll Contents of Strawberry Plants," *International Journal of Engineering Trends and Technology*, vol. 69, no. 3, pp. 23-28, 2021. Crossref, https://doi.org/ 10.14445/22315381/IJETT-V69I3P205.

[24] A.K.Singh, Dr. B.Chourasia, Dr. N.Raghuwanshi, Dr. K. Raju, "BPSO based Feature Selection for Rice Plant Leaf Disease Detection with Random Forest Classifier," *International Journal of Engineering Trends and Technology*, vol. 69, no. 4, pp. 34-43, 2021. Crossref, https://doi.org/10.14445/22315381/IJETT-V69I4P206

[25] B. K.Hatuwal, A.Shakya, & B.Joshi, "Plant Leaf Disease Recognition using Random Forest, KNN, SVM and CNN," *Polibits*, vol. 62, pp. 13-19, 2020.

[26] H.,Qi, Y.Liang, Q.Ding, &J.Zou, "Automatic Identification of Peanut-Leaf Diseases Based on Stack Ensemble," *Applied Sciences*, vol. 11, no. 4, 2021.

[27] Remya R S, Ebin Antony, "Water Quality Analysis of Different Water Sources in Kerala, India," *SSRG International Journal of Agriculture & Environmental Science*, vol. 9, no. 4, pp. 6-11, 2022. Crossref, https://doi.org/10.14445/23942568/IJAES-V9I4P102

[28] JAWani, S. Sharma, M. M., S. Ahmed, S. Sharma, and S. Singh, "Machine Learning and Deep Learning Based Computational Techniques in Automatic Agricultural Diseases Detection: Methodologies, Applications, and Challenges," *Archives of Computational Methods in Engineering*, pp. 1-37, 2021.

[29] D. K.Behera, M. Das, S Swetanisha, J.Nayak, S. Vimal, and B.Naik, "Follower Link Prediction using the Xgboost Classification Model with Multiple Graph Features," *Wireless Personal Communications*, pp. 1-20, 2021.

[30] S. S Patil, H.A.Dinesha, "Modified Weight Optimized XG Boost (MWO-XGB) for Concept Drift and Data Imbalance Problems in the Online Environment," *International Journal of Engineering Trends and Technology*, vol. 70, no. 6, pp. 308-316, 2022. Crossref, https://doi.org/10.14445/22315381/IJETT-V70I6P232

[31] M.Panigrahi, D. K.Behera, K. C.Patra, "Epileptic Seizure Classification of Electroencephalogram Signals using Extreme Gradient Boosting Classifier," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 25, no. 2, pp. 884-891, 2022. DOI: 10.11591/ijeecs.v25.i2.pp884-891.

[32] AkankshaPandey, L S Maurya, "Career Prediction Classifiers based on Academic Performance and Skills using Machine Learning," *SSRG International Journal of Computer Science and Engineering*, vol. 9, no. 3, pp. 5-20, 2022. Crossref, https://doi.org/10.14445/23488387/IJCSE-V9I3P102

[33] T.Chen, & C.Guestrin, "XGBoost: A Scalable Tree Boosting System," In *International Conference on Knowledge Discovery Data Mining*, pp. 785-794, 2016.

[34] S. K. Maria, S. S. Taki, M.Mia, A. A. Biswas, A. Majumder, & F. Hasan, "Cauliflower Disease Recognition using Machine Learning and Transfer Learning," In *Smart Systems: Innovations in Computing*, pp. 359-375, 2022.

[35] D. K. Behera, M. Das, S. Swetanisha, P. K. Sethy, "Hybrid Model for Movierecommendation System using Content K-Nearestneighbors and Restrictedboltzmann Machine," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 1, pp. 445-452, 2021. Doi: 10.11591/ijeecs.v23.i1.

[36] Xie, Yunxin, Chenyang Zhu, Runshan Hu, and Zhengwei Zhu, "A Coarse-to-Fine Approach for Intelligent Logging Lithology Identification with Extremely Randomized Trees," *Mathematical Geosciences*, vol. 53, no. 5, pp. 859-876, 2021.