International Journal of Advanced Technology and Engineering Exploration, Vol 10(102)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2022.10100456

Research Article

An effective crop recommendation method using machine learning techniques

Disha Garg* **and Mansaf Alam**
Department of Computer Science, Jamia Millia Islamia, New Delhi, India

Received: 12-November-2022; Revised: 17-May-2023; Accepted: 19-May-2023

Abstract
The soil plays a vital role in agriculture, and soil testing serves as the initial step in determining the optimal nutrient levels for cultivating specific crops. Machine learning (ML) classification techniques can leverage soil nutrient data to recommend suitable crops. The Wrapper-PART-Grid approach, which incorporated crop recommendation data to suggest appropriate crops, was introduced in this paper. This hybrid method combined the grid search (GS) method for hyperparameter optimization, wrapper feature selection strategy, and the partial C4.5 decision tree (PART) classifier for crop recommendation. The proposed approach was compared with other ML techniques, including multilayer perceptron (MLP), instance-based learning with parameter k (IBk), C4.5 decision tree (CDT), and reduced error pruning (REP) tree. Evaluation metrics such as true positive rate, false positive rate, precision, recall, F1-score, root mean squared error (RMSE), and mean absolute error (MAE) were employed to assess these models. The suggested method demonstrated superior reliability, accuracy, and effectiveness compared to other ML models for crop advisory purposes. This method attained a remarkable accuracy rate of 99.31%, the highest among all the approaches considered. In this paper, a ML-based crop recommendation technique aimed at assisting farmers in enhancing their knowledge of cultivating appropriate crops. The technique not only seeks to reduce overall wastage but also aims to increase crop yield and improve crop quality.

Keywords
WEKA tool, PART, ML, Smart farming, Crop recommendation, Feature selection, IoT.

1.Introduction
One of the primary strengths of the national and international economy has recently been highlighted as agriculture [1]. There are a variety of crops, but the quality of crops, productivity, and yield forecast have all raised concerns for the future of agriculture [2]. Digital technology has reduced the need for manual labor in agriculture, leading to increased productivity, better living standards, and more people working in the field [3]. Nowadays agriculture has developed a lot in India. Precision agriculture has achieved better enhancements and is important in recommending crops. The recommendation of crops is dependent on various parameters. The first and most crucial phase in farming is the Prediction of soil properties like pH, humidity, temperature, nitrogen (N), phosphorus (P), and potassium (K). These are directly related to the geographical and climatic conditions of the area being utilized [4–6].

In recent years, there have been drastic climatic changes occurring because of global warming [5–8]. The selection of inappropriate crops has a tremendous impact on farmer's hopes and dreams because it uses up all available resources (such as the cost of seeds, fertilizers, etc. Using machine learning (ML) as a key technology, traditional farming can be reshaped. This research aims to introduce ML-based crop suggestion system for farmers, hoping to use this information to produce more productive and higher-quality crops with less waste.

ML recommends suitable crops using various mathematical or statistical methods. By employing these methods, we can advise the farmer on the most suitable crop to grow in his particular agricultural region, helping him to maximize his profits. To help farmers make informed decisions about what to grow, crops are classified according to the nutrients they contain. Classification is ML technique that has enormous potential for the farming sector. Different classifiers are currently available for this purpose [9]. Classification uses training data to categorize new

*Author for correspondence

observations. However, it is impossible to say which is best because it relies on the application and the dataset. Analyzing a collection of training data is initially required before employing a classification technique. The training data predict the relation between the features and the class label.

The novelty of the present study is to produce the Wrapper-PART-Grid method in this paper. The wrapper algorithm selects appropriate features from the collected data, and the Partial C4.5 decision tree (PART) algorithm is used for classifying crops in the proposed prediction technique. The wrapper method uses the grid search (GS) algorithm to examine the combination of all feasible features and choose the subset that performs best for a given ML algorithm, known as Wrapper-PART-Grid algorithm. The objective of this study is to suggest optimal crops using input variables such as soil pH, humidity, temperature, nitrogen (N), phosphorus (P), and potassium (K) levels. Then, based on the predicted future yields of different crops, including rice, kidney beans, maize, chickpeas, pomegranate, pigeon peas, moth beans, black gramme, lentil, banana, mango, grapes, watermelon, mungbean, muskmelon, apples, oranges, papayas, coconuts, cotton, jute, and coffee, the most suitable crop is suggested using various ML model.

The crop recommendation dataset was utilized for the experiment. The experiment is divided into two main sections. Firstly, feature selection is performed to find the best features because it is well-recognized that different features can have varying effects. Then, we assess our approach using the selected features on different ML models after applying hyperparameter tuning. Finally, the results were compared using standard metrics, i.e., accuracy, precision, recall, F1-score, root mean squared error (RMSE), mean absolute error (MAE), and confusion matrix. This approach performed better than conventional ML methods.

The main contribution are as follows:

- An efficient system for agricultural crop recommendation was proposed, utilizing ML techniques.
- The Wrapper-PART-Grid method was introduced for classifying agricultural data to provide crop recommendations.
- To optimize the models for crop recommendation, the optimal parameters were identified using grid hyperparameter optimization.
- Experiments were conducted to evaluate the effectiveness of the method and compare the results with other approaches.

An overview of the related work on the topic is provided in Section 2. Section 3 includes the methodology, dataset preparation, preprocessing, the proposed method, feature selection, data analysis, k-fold cross-validation, and hyperparameter optimization. Section 4 presents the experimental study and result analysis. Section 5 is dedicated to the discussion of the results and their interpretation. Finally, in Section 6, the paper concludes.

2.Related literature

Previous literature offered numerous works that may be used to predict crops for the user. However, most of the study is not focused on various soil factors. This makes it necessary to enhance the effectiveness of crop prediction and recommendation systems so that they can match the soil characteristics and climate circumstances in a better way. A model was used to analyze the sufficient amounts of soil nutrients, including nitrogen, potassium, and phosphorus, and advise the crops that should be grown in the future. The crops in [10] were predicted using a neural network, and the accuracy was 89.88%. This paper predicts suitable crops, but crop rotation has not been thoroughly studied in this study. The study has suggested a technique to help farmers choose crops by considering all the variables, including soil type, sowing season, and geographic location. The suggested method considers soil properties like soil type, pH value, and nutrient concentration, as well as climatic factors like rainfall, temperature, and geographic location in terms of the state when recommending a suitable crop to the user. Various ML algorithms were used, but the results are not promising.

Similarly, a Naive Bayes algorithm incorporating the soil's temperature, humidity, and moisture as crucial variables were suggested for crop recommendation [11]. By utilizing ML, one of the most cutting-edge technologies in crop prediction, this research helps beginner farmers with a method that directs them to sow good crops. Furthermore, a supervised learning method called Naive Bayes suggests how to do it. The prediction accuracy of these models must be increased. In order to analyze the many soil properties and recommend the crop for cultivation, another ML approach was proposed [12]. They used k-nearest neighbor (KNN) algorithms, but prediction was based only on soil properties.

Random forest (RF) method was used to predict crop yields in the agricultural sector [13]. The RF method provides the optimal crop production model by considering the fewest number of models possible. The results indicate that crop production prediction is beneficial in the agricultural sector.

A winter wheat prediction model was proposed by estimating the characteristics of the soil using online soil spectroscopy and a prototype sensor [14]. The model used A self-organizing map with supervised Kohonen networks, XY-fused networks, and artificial neural networks based on counter-propagation. Even though the technique yields valuable data, studying the parameters related to the soil will not be sufficient to maximize crop productivity. Crop prediction depends on various variables, so feature selection is crucial. In order to predict crops utilizing different classifiers using soil attributes and environmental data, such as rainfall, season, texture, and temperature, a comparative evaluation of several feature selection approaches was conducted [15].

Suresh et al. developed a system for crop classification based on specific data. Increased precision and productivity were attained by utilizing a support vector machine (SVM). The sample dataset for location data and the sample dataset for crop data were the two datasets that were the target of this investigation. With this proposed approach, specific crops, including rice, black gram, maize, carrot, and radish, were advised based on the availability of the specific nutrients, i.e., N, P, K, and pH [16].

Kulkarni et al. [17] suggested a technique to accurately recommend the best crop based on the kind and features of the soil, such as the average rainfall and surface temperature. The ML algorithms used by this suggested system included linear SVM, RF, and Naive Bayes. This crop recommendation algorithm classified the input soil dataset into the recommended crop types, Kharif and Rabi. Applying the suggested approach produced a 99.91% accuracy rate.

The study in [18] accurately compares many ML algorithms to determine the crop's recommended yield, with an overall improvement over multiple other techniques of 3.6%. The resultant work assists agronomists in making the proper crop selections for farming. Furthermore, the crops' output will increase exponentially. As a result, increasing India's income in the process.

Different factors like N, P, K, pH, temperature, humidity, and rainfall to advise the crops were discussed [19]. The dataset consists of 2200 instances and eight features. The best model is created by utilizing ML algorithms in Waikato environment for knowledge analysis (WEKA). The ML algorithms chosen for classification are decision tree classifiers, multilayer perceptron, and rule-based classifiers. They have not evaluated the feature importance in this study.

Priya and Yuvaraj [20], deep learning algorithms like artificial neural network (ANN) are used to produce precise crops at the appropriate times. By providing inputs like moisture, temperature, pH, and humidity utilizing a sensor network and the Internet of Things, a deep neural network and graphical user interface are used to forecast crops. Farmers can choose crops to cultivate with the help of crop ideas.

In [21], internet of things (IoT) and ML system were suggested, which uses sensors to allow soil testing. It is based on measuring and observing soil properties. This method reduces the likelihood of soil deterioration and supports crop vitality. This system uses many sensors to monitor temperature, humidity, soil moisture, pH, and nitrogen, phosphorus and potassium (NPK) nutrients of the soil. These sensors include soil temperature, soil moisture, pH, and others. They have considered all the features and have not analyzed feature importance in this study. Also, hyperparameter optimization is not applied to the input parameters.

In [22], a recommendation system using an ensemble model with majority voting methods employing random trees, chi-squared automatic interaction detection (CHAID), KNN and Naive Bayes as learners to suggest a good crop based on soil data with high specific accuracy and efficacy was suggested. However, there was no result comparison or analysis in this study, and there was no feature importance evaluation.

In [23], the best crop prediction model that can assist farmers in selecting the right crop to produce based on local climate factors and soil nutrient levels was identified. This article contrasts two widely used criteria, Gini and Entropy, for algorithms like KNN, decision tree, and RF classifier. Findings show that RF has the best accuracy. Further, features should be analyzed to determine the most effective features to recommend crops.

AgroConsultant was introduced, as a smart system designed to help Indian farmers choose the best crops for their regions [24]. During the planting season, his farm's location, soil, and climatic elements like temperature and rainfall are crucial. In the future, crop rotations can be predicted.

In [25], a recommendation system is proposed that uses Arduino microcontrollers to collect data on the surrounding environment, ML techniques like Naive Bayes (Multinomial) and SVM, K-means clustering, and natural language processing (sentiment analysis) to make recommendations about what to plant. The neural network achieved the highest accuracy (98.8%) among all algorithms.

In [26], a multiclass soil fertilizer recommendation system for paddy fields was developed. In addition, the SVM parameters are tuned using various optimization techniques, such as the genetic algorithm and particle swarm optimization.

Using a preliminary set on a fuzzy approximation space and a neural network, the authors of [27] could estimate the crop's suitability in the Vellore District.
In [28], the most optimal crops for the current climate are predicted. Given the variables mentioned above, the study presented here gives farmers a more accurate idea of what crops to put where in their fields. An overview of some significant studies on various prediction models is given in *Table 1*.

Table 1 An overview of significant studies

Methods used	Advantage	Limitations	Accuracy	Reference
Neural Network	Helps in identifying suitable crops	Do not consider environmental factors	91%	[25]
Gradient Descent	Considers soil factors	No comparison of results with other classifiers	97%	[15]
Neural Network	Predicts suitable crops	There is not a thorough study on crop rotation available	89.88%	[10]
Naïve Bayes classifier	Uses environmental factors	No detailed result analysis is given	97%	[11]
Neural networks	Regular feedback is taken from the farmers	No comparison of results with other classifiers	95%	[26]
Extreme learning machine	Improved classification result	No comparison of results with other classifiers	--	[28]
Supervised self-organizing maps	Provides better results	Do not use the climate to predict yields or other variables.	81.65%	[14]
KNN classifier	Displays solid efficiency	Only soil properties are used for predicting crops	---	[7]
Fuzzy approximation	Improved classification accuracy	Do not consider crop predecessors	93.2%	[13]
Regression-based ensemble	Ideally suited for primary crops	No evaluation of alternative classifier models	94.78%.	[14]
Ensemble model	Provides improved Prediction using random forest and XGBoost	Compares poorly to other classifiers	96.69%	[29]
Majority voting scheme	This approach helps agronomists in selecting the best crop for their fields.	Feature analysis is not done	Overall improvement 3.6%	[18]
Multilayer Perceptron	Based on current environmental parameters, the smart module will provide irrigation and yield recommendations for the crop	Preprocessing and feature analysis is not done	98.2273%	[20]

Based on our literature analysis, most crop suggestion and Prediction methods use ML techniques such as decision trees, SVM, ANN, RF, logistic regression, KNN, and others [20]. The findings of these researches are needed to improve because very few studies have concentrated on determining the significance of the traits for crop recommendation. The main difficulties are finding high-quality publically available datasets, selecting the best features, and choosing the best algorithms. The literature study found that current comparisons of artificial intelligence (AI) algorithms for crop

recommendation are still lacking in obtaining reliable results.

3.Proposed methodology

The concepts and materials utilized for this experiment are described to make the proposed methodology more easily readable and clear.

3.1Proposed system

In the proposed system, the dataset is processed, and features are chosen. After choosing the relevant features, these were given as input to the ML models. In order to improve the model's effectiveness, the GS. performs parameter tuning. In order to build our ML model, we used various ML classification algorithms, such as IBk, multilayer perceptron, C4.5 decision tree (CDT), reduced error pruning (REP) tree, and partial decision tree (PART) algorithms. The best features are extracted through hyperparameter optimization. After the models had been built, a performance assessment of these models was done using performance metrics. The block diagram of the proposed system is illustrated in *Figure 1*, and a detailed flow diagram is shown in *Figure 2*.

3.2Dataset preparation and preprocessing

A crop recommendation dataset was used with 2200 records and seven parameters (N, P, K, Temperature, Humidity, pH, and Rainfall). The required soil content was determined for each crop to understand the data's nature better. Cross-validation was carried out after splitting the dataset into a training set and a validation set. Data was obtained from Kaggle [30]. *Table 2* gives the summary of the dataset used in this work. This dataset was chosen to train the model because it has parameters crucial for crop suggestion, such as humidity, temperature, rainfall, pH, nitrogen, phosphorous, and potassium requirement ratio. Temperature, humidity, rainfall, nitrogen, potassium, and phosphorus values are specific to each crop. The attributes in the crop recommendation dataset do not have any empty fields. After confirming that there are no missing values, the data type of the attributes (int64) is determined, and labels are listed.

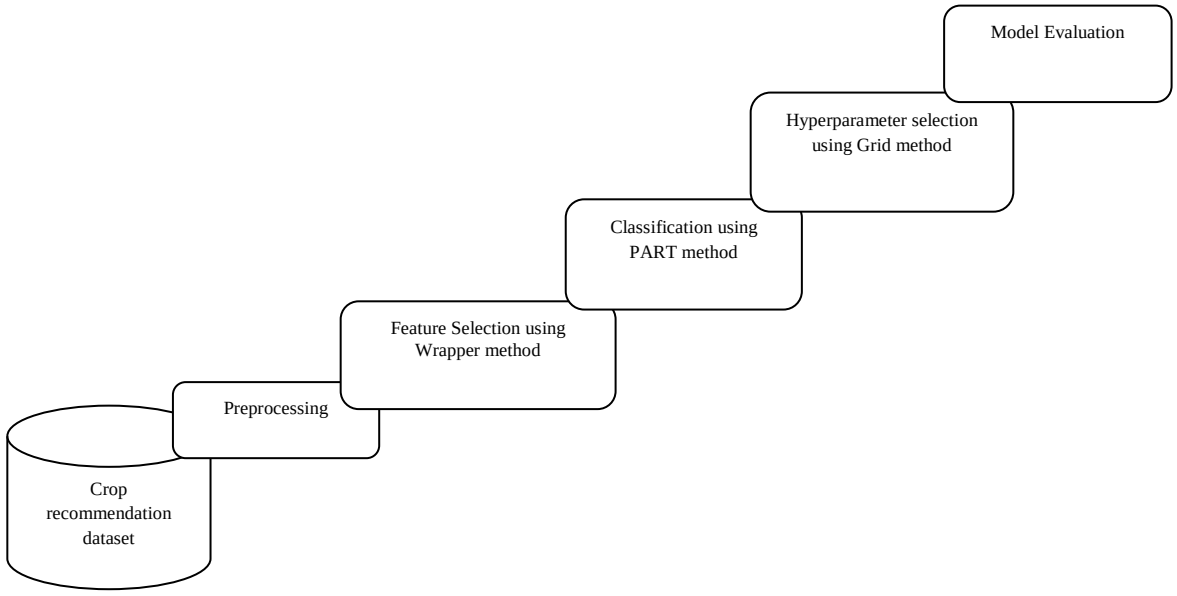


Figure 1 Schematic of the developed methodology

Table 2 Dataset description

Parameters	Crop to be recommended (Label)
N is the ratio of Nitrogen content in soil - kg/ha **P** is the ratio of Phosphorous content in soil - kg/ha **K** is the ratio of Potassium content in soil - kg/ha **Temperature** is the temperature in degree Celsius **Humidity** is the relative humidity in % **pH** value of the soil **Rainfall** is in mm	Rice, maize, jute, cotton, coconut, papaya, orange, apple, muskmelon, watermelon, grapes, mango, banana, pomegranate, lentil, blackgram, mungbean, mothbeans, pigeonpeas, kidneybeans, chickpea, coffee

The data must be perceptually prepared before applying ML models to analyze the experimental study. The input features are normalized this way because ML models cannot effectively train and test the non-uniform distribution of real-world farming data collected by sensors. According to the dataset, attributes like N, P, and K values of soil play a significant role from a biological point of view because these are the primary macronutrients for crops. These macronutrients' primary contributions can generally be divided into the following categories:

N—Nitrogen is primarily in charge of the plant leaves growing.

P—Phosphorus is essential for growing roots, flowers, and fruits.

K—Potassium performs the overall functions of the plant efficiently.

3.3Feature selection

Feature selection is an essential preprocessing step that resolves the issues with large dimensionality in many ML applications. First, a subset of features from the available data must be chosen to use a learning algorithm. Feature selection selects the most significant features from the initial complete feature set and removes the irrelevant, redundant, and noisy ones based on an assessment criterion, narrowing the feature set to those most significant or pertinent to the ML model [29].

Different features, i.e., N, P, K, rainfall, temperature, humidity, and pH, can be selected to find and suggest the best crop. The dependent variable in this experiment is the name of the various crops. Our proposed method considers N, P, K, temperature, humidity, pH, and rainfall-independent variables. In this phase, various feature selection approaches, i.e., filter methods and wrapper methods such as principal component analysis (PCA), correlation analysis, and information gain (IG), are applied to the dataset. Both methods were used to identify the most beneficial indicators for the agricultural system. By analyzing and choosing features independently of any learning algorithm, filter techniques rely on the features of the datasets to assess their importance [31]. Wrapper approach evaluates all potential feature combinations and chooses the one that produces the best outcome for a particular ML technique. Wrapper techniques choose fewer features to maximize the effectiveness of the learning process [32]. The potential for these strategies to be generalized is thus limited. A simple nonparametric method called PCA is used to extract the most crucial information from a set of redundant or noisy data. PCA uses an orthogonal transformation to turn samples of correlated variables into samples of linearly uncorrelated features. The degree of feature redundancy is considered while searching for feature subsets using correlation-based feature selection. The objective of the evaluation technique is to identify subsets of features that are individually highly correlated with the class but have low inter-correlation. IG calculates the difference in Entropy between the presence and absence of a feature. The difficulty of determining the significance of a feature inside a feature space is addressed here by using more generic techniques, such as the measurement of informational entropy [32].

Additionally, filtering methods give each feature a score before selecting the features having the highest scores [33]. It shows how closely and differently each feature matches the output labels. In this study, the effective environment indicators are chosen using a wrapper selection strategy. The comparison of the wrapper feature selection technique to the other feature selection strategies is demonstrated in *Table 3*.

Table 3 Number of selected features

Algorithm for feature selection	Selected features count	Selected features
PCA	6	Temperature, pH, P, k, humidity, N
Correlation	7	Temperature, pH, N, P, k, humidity, rainfall
IG	7	Temperature, pH, N, P, k, humidity, rainfall
Wrapper	5	K, N, P, humidity, rainfall

3.4Resampling

After appropriate feature selection, resampling is done. A few classes (also known as majority classes) frequently occupy the majority of instances in real-world data, whereas many other classes (also known as minority classes) have few instances. It is called a class-imbalanced classification problem. A common technique for balancing class distributions is resampling [34]. It consists of removing samples from the majority class (under-sampling) and adding more examples from the minority class (over-sampling). Resampling alters class distributions using two well-known techniques known as cross-validation and bootstrapping.

Several models are fitted to a portion of data using the resampling approach known as cross-validation,

and the model is then tested on a different subset of data. While trying to make accurate predictions, resampling throughout the training phase was crucial since it made it possible to determine which algorithms generalized best depending on our data. Also, it considerably uses the process of hyperparameter tuning, which involves modifying specific parameters of algorithms to improve outcomes [35]. The commonly used variations on cross-validation are train/test split, leave one out cross-validation (LOOCV), k-fold cross-validation, etc. LOOCV divides the samples n times, where n is the sample count. Although it is similar to k-fold cross-validation, the main distinction is that n different data splits are carried out. Simple cross-validation uses well-known k values (5 and 10) to reduce complexity [35]. Train-test split typically divides the dataset into training and test data in an 80:20 ratio and mimics how a model would perform on new and unseen data. In other words, the model would be trained using 80% of the training data and evaluated using 20% of the test data for which we already know the real truth. Then, using 20% of the data, we compared this ground truth with the model prediction. We then check how well our model would work with hypothetical data. It is the initial method of model evaluation [36]. However, the train-test split has disadvantages. Because this method reduces the quantity of the train data and does not use all our observations for testing, it creates bias. To solve this issue, we periodically split the data into training and testing using a method known as cross-validation (CV) [37]. As a result, the authors lessen the bias that the train-test split introduced. For this experiment, the k-fold CV approach was used. Each fold out of the training set was taken to create a model using the other folds, and then conduct testing on the excluded data. It is referred to as the k-fold CV method [38].

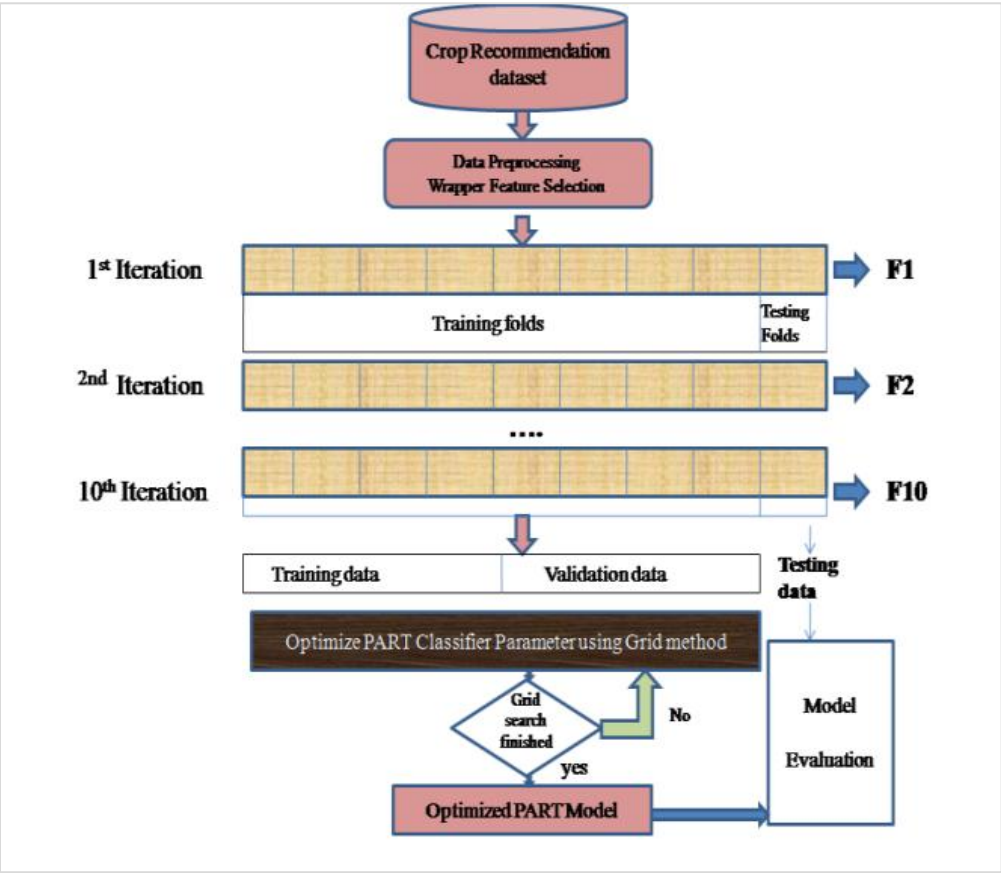


Figure 2 Workflow of the Proposed Wrapper-PART-Grid based crop recommendation system

3.5Classification algorithms

This study applied different ML-based classifiers on the crop recommendation dataset to evaluate overall performance and identify the best classifier for crop prediction. We were mainly interested in multiclass classifiers, which is why the following classifiers were selected randomly: multilayer perceptron,

Instance-based learning with parameter k (IBk), CDT, REP tree, and PART.

- *IBk*: WEKA uses the KNN method with its IBk algorithm. This paper used The IBk approach with k = 1 and k = 3 [39].
- *Multilayer perceptron:* A multilayer perceptron has one or more hidden layers whose neurons are called hidden neurons. This algorithm can be used for non-separable problems [40].
- *CDT:* The decision tree is built using the J48 method, from its root down to its leaf nodes. Starting at the tree's root and progressing through it until we reach a leaf node, which offers the classification of the instance, we may get the class label for a test item from a decision tree [41].
- *REP Tree:* A decision tree is built through IG and pruned using reduced-error pruning [41].
- *PART:* This technique generates rules by repeatedly building partial decision trees from data collection. Because of this, the algorithm is known as PART. PART is C4.5's extended version [42].

PART algorithm outperformed the others in terms of several different metrics. Witten et al. [43] proposed a separate-and-conquer rule learner. The algorithm produces decision trees, which are ordered sets of rules. The item is given the category of the first matching rule when a new set of data is compared to each rule in the list. Each iteration of the PART classifier creates a partial CDT, with the best leaf being a rule. The method combines rule learning with C4.5 and RIPPER.

Using training vectors Ai ∈ Rn, i=1... l and a label vector B∈ Rl, a decision tree recursively splits the space to group samples with similar labels.

Let D as a representation of the data at node m. Then, data is divided into Subsets $D_{LEFT(\theta)}$ and $D_{RIGHT(\theta)}$ for each candidate split with the formula θ = (j, tm) with feature j and threshold t_m as shown in Equation 1 and Equation 2:

$$D_{LEFT\ (\theta)} = ((A, B)\ |Aj<t_m) \quad (1)$$

$$D_{RIGHT(\theta)} = (D|D_{LEFT\ (\theta)}) \quad (2)$$

Depending on the problem being solved, Impurity function I is used to calculating the impurity at m (Classification or regression) as illustrated in Equation 3:

$$G\ (D, \theta) = (n_{LEFT}/N_m)\ I\ (D_{\ LEFT}\ (\theta)) + (n_{RIGHT}\ /N_m)\ I\ (D_{\ RIGHT}\ (\theta)) \quad (3)$$

Choose parameters to reduce the impurity in Equation 4:

$$\theta^* = \operatorname{argmin}_\theta G\ (D, \theta) \quad (4)$$

Recurse until the maximum allowable depth is reached for subsets $D_{LEFT\ (\theta^*)}$ and $D_{RIGHT\ (\theta^*)}$

Nm < min samples or $N_m = 1$

Values 0, 1… K-1 is assigned to the classification result for node m, which represents the region R_m with the observations as Equation 5:

$$P_{mk}=1/N_m \sum A_{i\ \epsilon}\ R_m\ I\ (B_i=K) \quad (5)$$

Be the proportion of class k observations in node m. The standard measure of impurity is Gini, as shown in Equation 6:

$$I\ (A_m) = \Sigma_k P_{mk}\ (1-P_{mk}) \quad (6)$$

Furthermore, Entropy, as shown in Equation 7:

$$I\ (A_m) = -P_{mk}\ \log P_{mk} \quad (7)$$

Moreover, misclassification is shown in Equation 8:

$$I\ (A_m) = 1-MAX\ (P_{mk}) \quad (8)$$

Here, A_m represents the training data at node m

3.6Metrics for comparative analysis

The efficiency of numerous supervised ML algorithms was analyzed and compared. The following are crucial parameters employed in this phase:

$$\text{Accuracy} = \frac{(TP+TN)}{(TN+FP+TP+FN)} \quad (9)$$

$$\text{Precison} = \frac{(TP)}{(TP+FP)} \quad (10)$$

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (11)$$

$$\text{F1 score} = \frac{2\times TP}{(2*TP+FP+FN)} \quad (12)$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \quad (13)$$

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \quad (14)$$

Accuracy is a measurement of how closely a predicted value corresponds to the actual value, as determined by the % of cases that were adequately identified. The ratio of correctly classified instances true positives (TP) and true negatives (TN) over the total predictions, including TP, TN, and wrong predictions, false positives (FP) and false negatives (FN) is known as accuracy Equation 9. Precision evaluates how accurate examples with positive labels are Equation 10. How many instances of the positive class were correctly identified, or how precisely positive examples were classified, is measured by recall Equation 11. The harmony and balance are measured by the f-measure Equation 12 [44]. The amount of misclassifications or errors in the model's

Prediction is measured using MAE Equation 13. When MAE values are comparable, the RMSE rate determines which classification method is superior Equation 14. Finally, the similarity level between two or more variables is evaluated using Cohen's Kappa. Equation 15 can be used to express the equation resulting from Cohen's Kappa evaluation as follows:

$$K = (P0 - Pe)/ (1\text{-}Pe) \quad (15)$$

P0 is the total diagonal proportion of the observation frequency, Pe is the total marginal proportion of the observation frequency, and k is the kappa coefficient value. The Cohen's kappa coefficient's value can be understood in terms of the degree of agreement: poor ≤ 0.20; fair = 0.21–0.40; moderate = 0.41–0.60; good = 0.61–0.80; very good = 0.81–1.00. The values of the kappa statistic are above .90, indicating good.

Determining which model will provide the fastest results is essential, so researchers calculated the time in seconds taken by each algorithm. This value represents the time required to train the model. Decision trees and other common ML approaches exhibit a bias in favor of the majority class and tend to neglect the minority class. They frequently misclassify the minority class relative to the majority class because they tend to forecast the majority class exclusively. The confusion matrix is also used to evaluate how well a classification algorithm is performing. The confusion matrix provides a comparison between actual and expected values. It is utilized to improve ML models. N is the number of classes or outputs, and N is the size of the confusion matrix. We obtain a 2×2 confusion matrix for two classes. We obtain a 3×3 confusion matrix for three classes. The confusion matrix, which displays each class's accurate and inaccurate predictions, may be used to assess the outcomes. The first row's first column shows how many classes "True" were accurately predicted, whereas the second row shows how many classes "True" were incorrectly predicted. All class "False" items in the second row were expected to be class "Yes." Therefore, the higher the diagonal values of the confusion matrix, the better the correct Prediction [45].

4.Experimental study and result analysis

4.1 Experimental environment

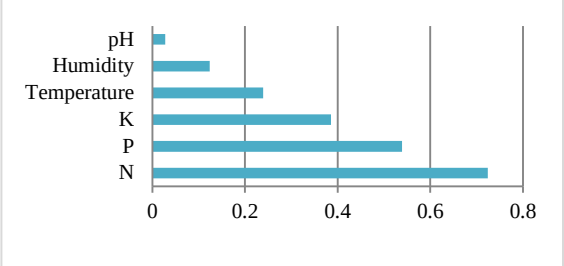
In this experimental investigation, the ML method for crop recommendation is implemented using WEKA. All research communities working on supervised and unsupervised learning approaches can use the open-source WEKA tool. This tool works well with ML approaches with Java platform implementation [46]. Furthermore, experiments involving ML techniques written in Python are implemented using WEKA. WEKA on Windows 10, equipped with an Intel Core i7-8665U CPU@ 4.80 GHz processor (8.00 GB RAM).

Various classification approaches have been used to select a crop, including IBk, multilayer perceptron, D.T., REP TREE, and PART. The wrapper algorithm selects appropriate features and the most beneficial environmental indicators for the PART classification algorithm. The wrapper method uses the GS. algorithm to examine the combination of all feasible features and choose the subset that performs best for a given ML algorithm. The PART algorithm is used for classifying crops in the proposed prediction technique, "Wrapper-PART-Grid."

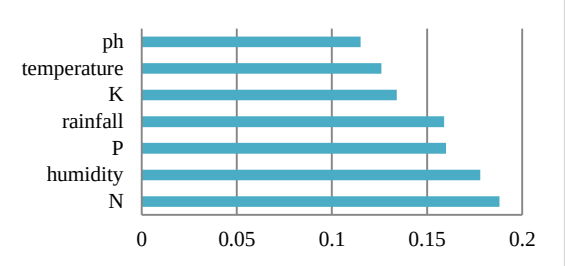
4.2 Results analysis

The IG ranking, correlation, PCA, and wrapper ranking filters are the four types used in this study. These filters were applied to the dataset to determine which feature combination is more important for classification models, as demonstrated in *Figure 3*. In accordance with the ranks, the wrapper ranking filter chose fewer attributes, and it discovered that the five most significant attributes are rainfall, humidity, N, P, and K. Undesirable features (Temperature and pH) were removed based on the returned ranking of the features to maximize the performance of the models by updating the dataset. The wrapper method has selected minimum and valuable features. The outcome demonstrates that (as shown in *Figure 4*) filter methods are less reliable than wrapper feature selection techniques as they identified the relevant features only.

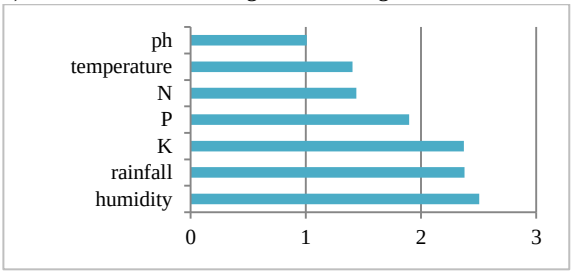
In the selected dataset, 2200 instances are available. Suppose k = 10, 2200/10 = 220 observations would be in each fold. K-fold CV determines test accuracy by using fold-1 (220 samples) as the testing set and k-1 (9 folds) as the training set. The method is repeated k times, or ten times if k = 10. Every time, a distinct collection of observations is used as a validation/test set. K-test accuracy predictions are produced as a result of this method and then averaged [47].



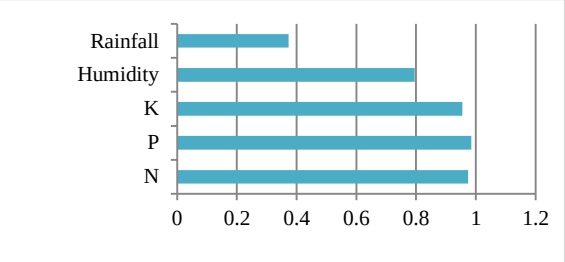
a)Feature evaluation using PCA ranking filter



b) Feature evaluation using correlation ranking filter



c) Feature evaluation using Information Gain ranking filter



d) Feature evaluation using wrapper filter ranking

Figure 3 Feature evaluation

4.2.1 Hyperparameter tuning

This section discusses the outcomes of the modeling-related experiments, including the hyperparameter tuning and the parameters set for each experiment. The ideal configuration used throughout the modeling phase may significantly improve the performance of an algorithm. The comparative analysis of the algorithms that were created is also included in this part. This study was done to decide which supervised ML method would be most effective in crop recommendation. All the studies employed 10-fold cross-validation and a batch size of 100 to assess algorithm performance.

Choosing a set of ideal hyper-parameters is known as hyper-parameter tuning. Before beginning the ML task, the model value of the hyper-parameter is fixed. In ML approaches, the hyper-parameter adjustment has a significant impact. First, the data is protected from the model parameters. Then, tuning of the hyper-parameters is done to achieve the optimum fit. Given the complexity of the problem, GS and random search methods are utilized to find the optimum hyperparameter. The accuracy of the ML classifier is improved using this strategy. In the proposed method, GS hyperparameter optimization is used. GS identifies the ideal hyper-parameters for a model or those that produce the most "correct" predictions. GS examines every possible set of hyper-parameter combinations. When using GS, the user defines a finite set of values, and the system evaluates the cartesian product of those values. GS cannot fully utilize the productive areas on its own [48]. GS algorithm is based on brute force. By doing so, a thorough search for a specific subset of the hyperparameter space is made. Use a different algorithm if the search space is too big. Random search is faster but does not always guarantee the most significant outcome [49].

A three-hidden-layered ANN with ten hidden units in each layer made up the multilayer Perceptron (MLP) classifier. Experimental decisions were made on the number of levels and hidden units in each layer. Rectified linear unit (ReLU) served as the activation function of the hidden layer. The hyperparameter setting of MLP is shown in *Table 4:*

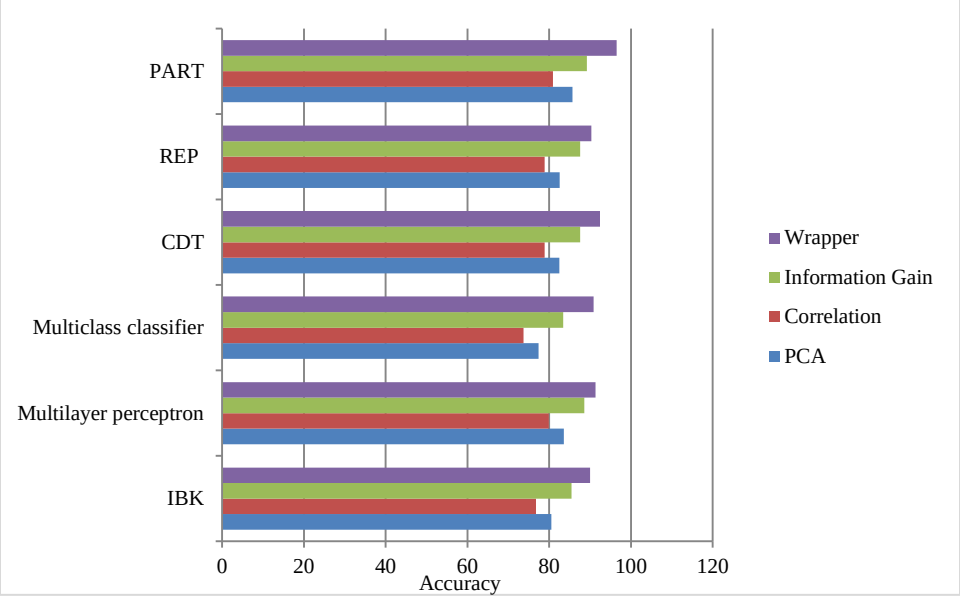


Figure 4 Accuracy of the classification algorithm using various feature selection techniques for selected features

Table 4 Multilayer Perceptron hyperparameter optimization results

Training	Number of hidden layers	Activation function For hidden layers	Initial learning rate	Accuracy%
1	3	ReLU	.001	**98.2**
2	6	Tanh	.01	98
3	5	ReLU	.05	97.21

It is one of the most important factors when creating a neural network. Choosing the optimal learning rate might be challenging if it is too low. As a result, the training process may be slowed down. Nevertheless, the model may not be optimized appropriately if the learning rate is too high. How successfully the network model learns the training dataset will depend on the activation function used for the hidden layer. Hidden layers are only needed in artificial neural networks when non-linear data separation is necessary. The highest accuracy is achieved as 98.2 for training 1.

The IBk or KNN algorithm's hyperparameter tuning is done by selecting the number of neighbors and the distance function in *Table 5*. In order to avoid either overfitting or underfitting, several values of k must be considered while defining it. Larger values of k may result in solid bias and low variance, whereas smaller values of k may have high variance but low bias. The distance measure makes finding the closest train data points with known classes easier. The best accuracy of 98.1% is achieved for training 2.

Table 6 presents the outcomes obtained by employing WEKA's confidence factor and using unpruned choices (True/False) for the PART method. For PART, four experimental trainings were conducted. The unpruned parameter was set to true for the first two trainings, and the corresponding confidence factors were 0.25 and 0.50. In this experiment, the model performed 99.3% and 99%, respectively. However, when the unpruned parameter was set to false in the most recent two trainings, the model's accuracy reached 98.32% for the 0.25 confidence factor and 97.21% for the 0.50 confidence factor. Therefore, the unpruned option was set to True to show that no pruning is performed.

In *Table 7*, the parameter was used as the criterion and maximum depth. This criterion sets the standard by which the impurity of a split is evaluated. Although "Gini" is the default parameter for measuring impurity, "entropy" is another option. It is decided to keep the criterion as a Gini index and maximum depth as a minimum (5) and achieved an accuracy of 98.4%. One of the reasons for overfitting in decision trees is allowing the tree to grow too deep, resulting in a more complicated model due to the increased number of splits and the more data captured. Bag size hyperparameter values of 100, 40, and 20 are shown in *Table 8*. The accuracy was

97.4% and 96.8% for 100 and 75 bag sizes, respectively, and 96.32% for the 50 bag size. The default bag size of 100 was considered in WEKA.

Table 5 IBk hyperparameter optimization results

Training	Neighbors (K)	Distance metric	Accuracy%
1	3	Euclidian	**98.3**
2	5	Manhattan	98.1
3	4	Euclidian	97.7

Table 6 PART hyperparameter optimization results

Training	Confidence factor	Unpruned	Accuracy %
1	.25	True	**99.3**
2	.50	True	99
3	.25	False	98.32
4	.50	False	98.21

Table 7 CDT hyperparameter optimization results

Training	Criterion	Max depth	Accuracy%
1	Gini index	6	97.21
2	Entropy	8	96.3
3	Gini index	5	**98.4**

Table 8 REP hyperparameter optimization results

Training	Bag size	Accuracy
1	20	96.32
2	40	96.8
3	100	**97.4**

5.Discussion

All algorithms were used to recommend the suitable crop on the crop recommendation dataset to compare the efficiency of various approaches. The preprocessing of the dataset involves removing unnecessary attributes that do not add value to the result. The Wrapper-PART-Grid technique has proven to be highly effective compared to alternative methods. Various factors, i.e., recall, accuracy, precision, and F1 measure, supported the analysis. In order to compare the effectiveness of the chosen features with the other ways, the wrapper feature selection technique is also examined. The wrapper feature selection approach discovers the fewest useful features among the different filter selection strategies. As can be seen in *Figure 5*, the findings demonstrate that this methodology has selected less number of features as compared to other feature selection techniques. Each feature is assessed using its similarity data to the output labels as part of feature selection methods. Since it estimates all feasible feature combinations and selects the combination set that yields the maximum accuracy when applied to various ML models, as illustrated in *Figure 4*. The wrapper selection approach is effective in the situation of the high similarity data of each attribute as opposed to filter selection strategies.

Using the PART algorithm led to the highest recall (0.993). In other words, PART algorithms will recommend something 99% of the time accurately but 1% of the time incorrectly. The PART algorithm achieved the highest precision and f1-score (0.993), indicating that the algorithm is accurate and that its high precision score is related to its low false positive rate. According to *Table 9*, PART had the lowest RMSE of 0.0249 after hyperparameter tuning, giving the most precise result. The study was further analyzed using kappa statistics and times to construct the model. *Table 9* shows a composite chart of these metrics. The kappa statistic was used to judge how well the model performed. Results showed that the PART algorithm was the best-performing algorithm with a kappa value of 0.9929 and MAE (0.0007). The kappa statistic was used to judge how well the model performed compared to the actual labels in the dataset. The kappa scores of the MLP, IBk, and CDT are all relatively close to one another (.9814, .9824, and.9833, respectively), whereas the kappa score of the REP tree is the lowest of all models (.9729). All the created models had scored over 0.81, which is close to 1, indicating that their interpretations agreed almost perfectly. If the value is 1, then all of the other classifiers agree with each other perfectly. The

comparison chart of kappa statics, MAE, and time taken by each model is given in *Figure 6.*

The PART algorithm has the maximum accuracy among all selected ML models, as seen in *Table 10*. However, it can be challenging to identify which class (positive or negative) the models predict when their accuracy score is low. Therefore, assessing a model accurately using accuracy alone is not always possible. To clarify this, we calculated precision, recall, and f1-score for each model, and then, the models were compared using these calculated metrics to determine precisely where one model outperforms the other.

MAE was considered to assess the discrepancy between the predicted and actual classes. This allows it to compare the predicted labels of the samples to the actual values in the dataset and measure the correctness of the constructed model. It is found that the model with the lowest MAE was the most successful. With an MAE of 0.0007, the PART classifier was the most accurate. A lower score indicates less likelihood of misclassification during Prediction. The results for MLP, IBk, CDT, and PART were.004, .0026, .0076, and.003, respectively. *Table 10* shows that before and after hyperparameter optimization results are also best in the case of the PART algorithm. Hyperparameters were tuned while using several methods to create a precise predictive model. Many alternative models with potentially different outcomes may result from optimizing the parameters.

Finally, the time the PART method took to construct the model was considered, as this is additional information supplied by WEKA following the 10-fold cross-validation. The findings indicated that the PART method was the most efficient in training time for the classifier model. During the 10-fold cross-validation process, the time taken to build the model was recorded in seconds, and the PART algorithm took significantly less time in training (.01 seconds). The findings indicated that IBk needs even more training time. Training for MLP took .04 s, IBk .16 s, CDT .051 s, and REP tree .06 s. These are the top-performing algorithms at the exact moment.

Table 9 Results of Kappa statistics

Algorithms	Kappa statistic	MAE	Time taken in seconds
Multilayer perceptron	0.9814	0.004	0.04
IBk	0.9824	0.0026	0.16
CDT	0.9833	0.0076	.051
REP Tree	0.9729	0.003	.06
PART	0.9929	0.0007	0.01

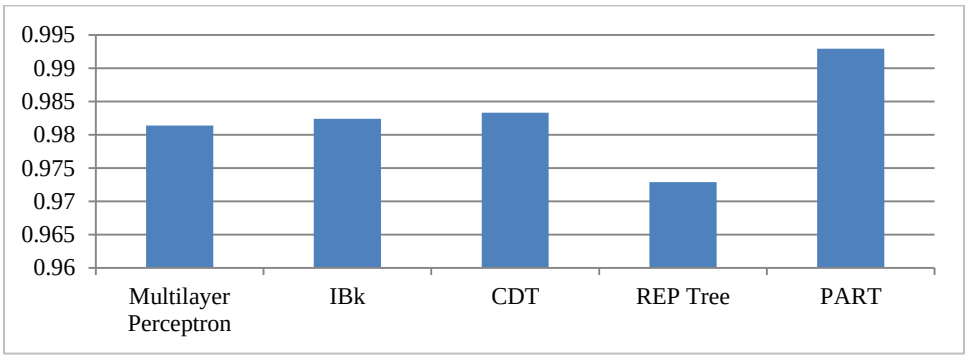
```
  a   b   c   d   e   f   g   h   i   j   k   l   m   n   o   p   q   r   s   t   u   v   <-- classified as
 88   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  12   0 |   a = rice
  0  99   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0 |   b = maize
  0   2  98   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 |   c = chickpea
  0   0   0  99   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 |   d = kidneybeans
  0   0   0   0 100   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 |   e = pigeonpeas
  0   0   0   0   0  96   0   3   1   0   0   0   0   0   0   0   0   0   0   0   0   0 |   f = mothbeans
  0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 |   g = mungbean
  0   1   0   0   0   1   0  98   0   0   0   0   0   0   0   0   0   0   0   0   0   0 |   h = blackgram
  0   0   0   0   0   2   0   1  97   0   0   0   0   0   0   0   0   0   0   0   0   0 |   i = lentil
  0   0   0   0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0   0   0   0 |   j = pomegranate
  0   0   0   0   0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0   0   0 |   k = banana
  0   0   0   0   0   0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0   0 |   l = mango
  0   0   0   0   0   0   0   0   0   0   0   0 100   0   0   0   0   0   0   0   0   0 |   m = grapes
  0   0   0   0   0   0   0   0   0   0   0   0   0  99   1   0   0   0   0   0   0   0 |   n = watermelon
  0   0   0   0   0   0   0   0   0   0   0   0   0   0 100   0   0   0   0   0   0   0 |   o = muskmelon
  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 100   0   0   0   0   0   0 |   p = apple
  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 100   0   0   0   0   0 |   q = orange
  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 100   0   0   0   0 |   r = papaya
  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 100   0   0   0 |   s = coconut
  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 100   0   0 |   t = cotton
  2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  98   0 |   u = jute
  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 100 |   v = coffee
```

Figure 5 Confusion matrix (PART)

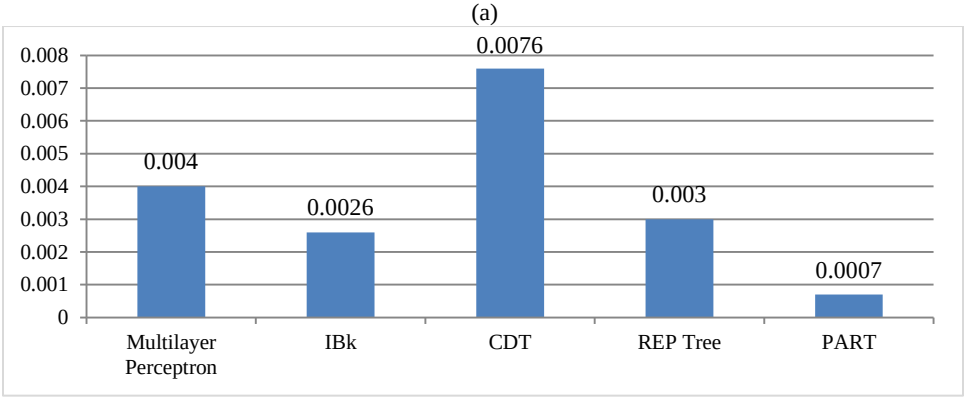
Table 10 Performance of implemented classifiers

Models	Hyperparameter optimization status	TP	FP	Precision	Recall	F1-score	RMSE
Multilayer Perceptron	Before	0.963	0.002	0.963	0.963	0.963	0.0809
	After	0.982	0.001	0.983	0.982	0.982	0.035

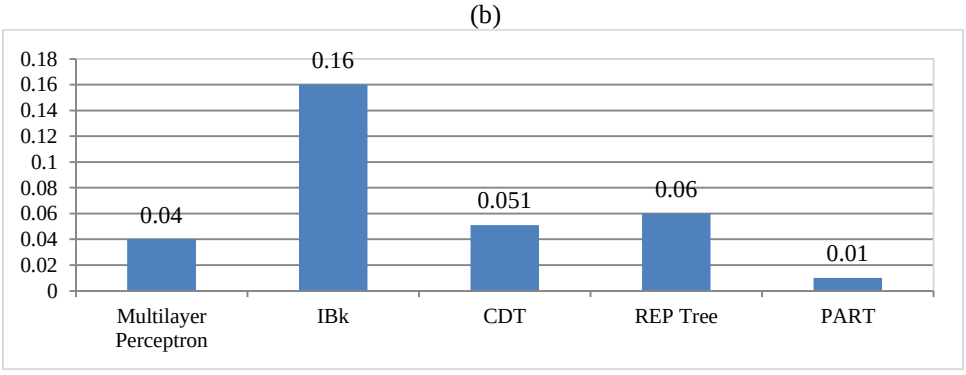
IBk	Before	0.982	0.001	0.982	0.982	0.982	0.0405
	After	0.983	0.001	0.984	0.983	0.983	0.0346
CDT	Before	0.975	0.001	0.976	0.975	0.975	0.0451
	After	0.984	0.001	0.985	0.984	0.984	0.0405
REP Tree	Before	0.966	0.002	0.968	0.966	0.966	0.051
	After	0.974	0.001	0.975	0.974	0.974	0.0454
PART	Before	0.991	0.000	0.991	0.991	0.991	0.028
	After	0.993	0.000	0.993	0.993	0.993	0.0249



(a)



(b)



(c)

Figure 6 Comparison based on (a) kappa statistics, (b) MAE, (c) Time taken in seconds

Another way of evaluation is the confusion matrix. Only a few errors were found in the confusion matrix for the model trained using the proposed method. The diagonal elements of the confusion matrix indicate how often the Prediction was accurate. *Figure 5* shows the confusion matrix for the proposed method after hyperparameter tuning. Clearly, out of 2200 instances, 27 were misclassified for different classes, and 98.77% were correctly classified. It was the best result as compared to others. A complete list of abbreviations is shown in *Appendix I.*

6. Conclusion

In this paper, Wrapper-PART-Grid was proposed as a prediction technique for decision-making systems in the domain of crop recommendations. The proposed approach utilized wrapper feature selections, GS hyperparameter optimization, and the PART algorithm. The most informative features from the crop recommendation dataset were selected by the wrapper method based on the results of other feature selection methods. The accuracy of each method was evaluated after selecting the optimal parameters for each model. By tuning hyperparameters using the grid optimization method, an impressive accuracy of 99.31% was achieved by the PART algorithm.

The findings of this research and the developed crop recommendation model have the potential to be integrated into a farmer's decision-making system. As a result, farmers may become more inclined to seek crop recommendations during soil testing, leading to a reduction in crop losses. The utilization of clustering for crop classification in classifiers is expected to enhance accuracy in the future. Despite several positive aspects of this study, there are also certain limitations. Only five models were examined in this research, and exploring various machine learning models for this task could be beneficial. In the future, the application of a deep learning-based computer vision system can be explored to enhance productivity in the smart farming sector.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

Author's contribution statement

Disha Garg: Conceptualization, investigation, writing-original draft, editing, data collection, analysis, and interpretation of results. **Mansaf Alam:** Study conception, design, supervision, investigation.

References

[1] AlZu'bi S, Hawashin B, Mujahed M, Jararweh Y, Gupta BB. An efficient employment of internet of multimedia things in smart and future agriculture. Multimedia Tools and Applications. 2019; 78:29581-605.

[2] Rezk NG, Hemdan EE, Attia AF, El-Sayed A, El-Rashidy MA. An efficient IoT based smart farming system using machine learning algorithms. Multimedia Tools and Applications. 2021; 80:773-97.

[3] Ansari M, Ali SA, Alam M. Internet of things (IoT) fusion with cloud computing: current research and future direction. International Journal of Advanced Technology and Engineering Exploration. 2022; 9(97):1812-45.

[4] Treboux J, Genoud D. High precision agriculture: an application of improved machine-learning algorithms. In 6th SWISS conference on data science (SDS) 2019 (pp. 103-8). IEEE.

[5] Sharma A, Jain A, Gupta P, Chowdary V. Machine learning applications for precision agriculture: a comprehensive review. IEEE Access. 2020; 9:4843-73.

[6] Thilakarathne NN, Yassin H, Bakar MS, Abas PE. Internet of things in smart agriculture: challenges, opportunities and future directions. In Asia-pacific conference on computer science and data engineering 2021 (pp. 1-9). IEEE.

[7] Lawal ZK, Yassin H, Zakari RY. Flood prediction using machine learning models: a case study of Kebbi state Nigeria. In Asia-pacific conference on computer science and data engineering 2021 (pp. 1-6). IEEE.

[8] Lawal ZK, Yassin H, Zakari RY. Stock market prediction using supervised machine learning techniques: an overview. In Asia-pacific conference on computer science and data engineering 2020 (pp. 1-6). IEEE.

[9] Durai SK, Shamili MD. Smart farming using machine learning and deep learning techniques. Decision Analytics Journal. 2022.

[10] Priyadharshini A, Chakraborty S, Kumar A, Pooniwala OR. Intelligent crop recommendation system using machine learning. In 5th international conference on computing methodologies and communication 2021 (pp. 843-8). IEEE.

[11] Kalimuthu M, Vaishnavi P, Kishore M. Crop prediction using machine learning. In third international conference on smart systems and inventive technology 2020 (pp. 926-32). IEEE.

[12] Mariappan AK, Madhumitha C, Nishitha P, Nivedhitha S. Crop recommendation system through soil analysis using classification in machine learning. International Journal of Advanced Science and Technology. 2020; 29(3):12738-47.

[13] Kumar YJ, Spandana V, Vaishnavi VS, Neha K, Devi VG. Supervised machine learning approach for crop yield prediction in agriculture sector. In international conference on communication and electronics systems 2020 (pp. 736-41). IEEE.

[14] Pantazi XE, Moshou D, Alexandridis T, Whetton RL, Mouazen AM. Wheat yield prediction using machine learning and advanced sensing techniques. Computers and Electronics in Agriculture. 2016; 121:57-65.

[15] Anguraj K, Thiyaneswaran B, Megashree G, Shri JP, Navya S, Jayanthi J. Crop recommendation on analyzing soil using machine learning. Turkish Journal of Computer and Mathematics Education. 2021; 12(6):1784-91.

[16] Suresh G, Kumar AS, Lekashri S, Manikandan R, Head CO. Efficient crop yield recommendation system using machine learning for digital farming. International Journal of Modern Agriculture. 2021; 10(1):906-14.

[17] Kulkarni NH, Srinivasan GN, Sagar BM, Cauvery NK. Improving crop productivity through a crop recommendation system using ensembling technique. In 3rd international conference on computational systems and information technology for sustainable solutions 2018 (pp. 114-9). IEEE.
[18] Garanayak M, Sahu G, Mohanty SN, Jagadev AK. Agricultural recommendation system for crops using different machine learning regression methods. International Journal of Agricultural and Environmental Information Systems. 2021; 12(1):1-20.
[19] Bakthavatchalam K, Karthik B, Thiruvengadam V, Muthal S, Jose D, Kotecha K, et al. IoT framework for measurement and precision agriculture: predicting the crop using machine learning algorithms. Technologies. 2022; 10(1).
[20] Priya PK, Yuvaraj N. An IoT based gradient descent approach for precision crop suggestion using MLP. In journal of physics: conference series 2019 (p. 012038). IOP Publishing.
[21] Gosai D, Raval C, Nayak R, Jayswal H, Patel A. Crop recommendation system using machine learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 2021: 554-69.
[22] Reddy DA, Dadore B, Watekar A. Crop recommendation system to maximize crop yield in ramtek region using machine learning. International Journal of Scientific Research in Science and Technology. 2019; 6(1):485-9.
[23] Rao MS, Singh A, Reddy NS, Acharya DU. Crop prediction using machine learning. In Journal of Physics: Conference Series 2022 (p. 012033). IOP Publishing.
[24] Doshi Z, Nadkarni S, Agrawal R, Shah N. AgroConsultant: intelligent crop recommendation system using machine learning algorithms. In fourth international conference on computing communication control and automation 2018 (pp. 1-6). IEEE.
[25] Bandara P, Weerasooriya T, Ruchirawya T, Nanayakkara W, Dimantha M, Pabasara M. Crop recommendation system. International Journal of Computer Applications. 2020; 175(22):22-5.
[26] Suchithra MS, Pai ML. Improving the performance of sigmoid kernels in multiclass SVM using optimization techniques for agricultural fertilizer recommendation system. In soft computing systems: second international conference, ICSCS 2018, Kollam, India, 2018 (pp. 857-68). Springer Singapore.
[27] Anitha A, Acharjya DP. Crop suitability prediction in vellore district using rough set on fuzzy approximation space and neural network. Neural Computing and Applications. 2018; 30:3633-50.
[28] Ashok T, Suresh Varma P. Crop prediction based on environmental factors using machine learning ensemble algorithms. In proceedings of intelligent computing and innovation on data science 2019 (pp. 581-94). Singapore: Springer Singapore.
[29] Bouchlaghem Y, Akhiat Y, Amjad S. Feature Selection: a review and comparative study. In E3S web of conferences 2022 (p. 01046). EDP Sciences.
[30] https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset. Accessed 13 April 2013.
[31] Kanyongo W, Ezugwu AE. Feature selection and importance of predictors of non-communicable diseases medication adherence from machine learning research perspectives. Informatics in Medicine Unlocked. 2023.
[32] Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques in machine learning. In science and information conference 2014 (pp. 372-8). IEEE.
[33] Mafarja MM, Mirjalili S. Hybrid binary ant lion optimizer with rough set and approximate entropy reducts for feature selection. Soft Computing. 2019; 23(15):6249-65.
[34] Verma A. Evaluation of classification algorithms with solutions to class imbalance problem on bank marketing dataset using WEKA. International Research Journal of Engineering and Technology. 2019; 5(13):54-60.
[35] Christias P, Mocanu M. A machine learning framework for olive farms profit prediction. Water. 2021; 13(23).
[36] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. Journal of machine Learning Research. 2011; 12:2825-30.
[37] Villavicencio CN, Macrohon JJ, Inbaraj XA, Jeng JH, Hsieh JG. Covid-19 prediction applying supervised machine learning algorithms with comparative analysis using Weka. Algorithms. 2021; 14(7).
[38] Smith TC, Frank E. Introducing machine learning concepts with WEKA. Statistical genomics: Methods and Protocols. 2016:353-78.
[39] Reynolds K, Kontostathis A, Edwards L. Using machine learning to detect cyberbullying. In 10th international conference on machine learning and applications and workshops 2011 (pp. 241-4). IEEE.
[40] Amin MN, Habib MA. Comparison of different classification techniques using WEKA for hematological data. American Journal of Engineering Research. 2015; 4(3):55-61.
[41] Sultana J, Jilani AK. Predicting breast cancer using logistic regression and multi-class classifiers. International Journal of Engineering & Technology. 2018; 7(4.20):22-6.
[42] Mazid MM, Ali AS, Tickle KS. Input space reduction for rule based classification. WSEAS Transactions on Information Science and Applications. 2010; 7(6):749-59.
[43] Witten IH, Frank E, Hall MA, Pal CJ, DATA M. Practical machine learning tools and techniques. In Data Mining 2005.
[44] Garg D, Alam M. Integration of convolutional neural networks and recurrent neural networks for foliar disease classification in apple trees. International

Journal of Advanced Computer Science and Applications. 2022; 13(4):357-67.
[45] Armah GK, Luo G, Qin K. A deep analysis of the precision formula for imbalanced class distribution. International Journal of Machine Learning and Computing. 2014; 4(5):417-22.
[46] https:// www.cs.waika to.ac.nzml/weka/. Accessed 13 April 2013.
[47] Nie Y, De Santis L, Carratù M, O'Nils M, Sommella P, Lundgren J. Deep melanoma classification with k-fold cross-validation for process optimization. In international symposium on medical measurements and applications (MeMeA) 2020 (pp. 1-6). IEEE.
[48] Belete DM, Huchaiah MD. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. International Journal of Computers and Applications. 2022; 44(9):875-86.
[49] Anggoro DA, Mukti SS. Performance comparison of grid search and random search methods for hyperparameter tuning in extreme gradient boosting algorithm to predict chronic kidney failure. International Journal of Intelligent Engineering and Systems. 2021; 14(6):198-207.

Disha Garg is currently pursuing Ph.D. in the Department of Computer Science, Faculty of Natural Sciences, Jamia Millia Islamia New Delhi-110025. Her research interests are Big Data Analytics, Machine Learning and Deep Learning. She has been presented several articles in international conferences and published several research articles in reputed International Journals. She is working in Agriculture Area using IoT and Data Analytics.
Email: disha.garg3399@gmail.com

Prof. Mansaf Alam is currently serving as a Professor in the Department of Computer Science, Faculty of Natural Sciences, Jamia Millia Islamia, located in New Delhi-110025. He holds the position of Young Faculty Research Fellow, DeitY, Govt. of India, and serves as the Editor-in-Chief of the Journal of Applied Information Science. Prof. Alam has a notable publication record with numerous research articles published in International Journals and Proceedings at conferences by prestigious publishers such as IEEE, Springer, Elsevier Science, and ACM. His research interests encompass various areas including AI, Big Data Analytics, Machine Learning & Deep Learning, Cloud Computing, and Data Mining. Prof. Alam is actively involved in the academic community as a reviewer for renowned international journals, including Information Science published by Elsevier Science. He also serves as a member of the program committee for several esteemed international conferences. Additionally, he is a valued member of the Editorial Board for reputable International Journals in the field of Computer Sciences. Prof. Alam has authored three books: "Digital Logic Design" published by PHI, "Concepts of Multimedia" by Arihant, and "Internet of Things: Concepts and Applications" published by Springer. He has also contributed to the books "Big Data Analytics: Applications in Business and Marketing" and "Big Data Analytics: Digital Marketing and Decision Making" by Taylor and Francis, as well as "Extended reality for Healthcare System: Recent Advances in Contemporary Research" by Elsevier, UK. Recently, Prof. Alam achieved an international patent (Australian) for his work on "An AI Based Smart Dustbin," highlighting his innovative contributions in the field.
Email: malam2@jmi.ac.in

Appendix I

S. No.	Abbreviation	Description
1	AI	Artificial Intelligence
2	ANN	Artificial Neural Network
3	CV	Cross-Validation
4	CDT	C4.5 Decision Tree
5	CHAID	Chi-Squared Automatic Interaction Detection
6	GS	Grid Search
7	KNN	K-Nearest Neighbor
8	FN	False Negatives
9	FP	False Negatives
10	IBk	Instance-Based Learning with Parameter k
11	IG	Information Gain
12	LOOCV	Leave One out Cross Validation
13	MAE	Mean Absolute Error
14	ML	Machine Learning
15	MLP	Multilayer Perceptron
16	N	Nitrogen
17	NPK	Nitrogen, Phosphorus and Potassium
18	P	Phosphorus
19	PART	Partial C4.5 Decision Tree
20	PCA	Principal Component Analysis
21	K	Potassium
22	KNN	K-Nearest Neighbor
23	RF	Random Forest
24	ReLU	Rectified Linear Unit
25	REP	Reduced Error Pruning
26	RMSE	Root Mean Squared Error
27	SVM	Support Vector Machine
28	TN	True Negatives
29	TP	True Positives
30	WEKA	Waikato Environment for Knowledge Analysis