Indonesian Journal of Electrical Engineering and Computer Science
Vol. 38, No. 3, June 2025, pp. 1840~1850
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v38.i3.pp1840-1850

Crop prediction using an enhanced stacked ensemble machine learning model

D. Madhu Sudhan Reddy, N. Usha Rani
Department of Computer Science and Engineering, Sri Venkateswara University College of Engineering, Tirupati, India

Article Info

Article history:

Received Apr 3, 2024
Revised Nov 28, 2024
Accepted Feb 27, 2025

Keywords:

Crop prediction
Decision tree
Machine learning
Multi-layer perceptron
Random forest
Stacked ensemble
XGBoost

ABSTRACT

In India, agriculture is a major sector that fulfils the population's food requirements and significantly contributes to the gross domestic product (GDP). The careful selection of crops is fundamental to maximizing agricultural yield, thereby elevating the economic vitality of the farming community. Precision agriculture (PA) leverages weather and soil data to inform crop selection strategies. Conventional machine learning (ML) models such as decision trees (DT), support vector classifier, K-nearest neighbors (KNN), and extreme gradient boost (XGBoost) have been deployed to predict the best crop. However, these model's efficiency is suboptimal in the current circumstances. The enhanced stacked ensemble ML model is a sophisticated meta-model that addresses these limitations. It harnesses the predictive power of individual ML models, stratified in a layered architecture to improve the prediction accuracy. This advanced model has demonstrated a commendable accuracy rate of 93.1% prediction by taking input of 12 soil parameters such as Nitrogen, Phosphorus, Potassium, and weather parameters such as temperature and rainfall, substantially outperforming the accuracies achieved by the individual contributing models. The efficacy of the proposed meta-model in crop selection based on agronomic parameters signifies a substantial advancement, fortifying the economic resilience of India's agriculture.

Corresponding Author:

D. Madhu Sudhan Reddy
Department of Computer Science and Engineering, Sri Venkateswara University College of Engineering
Tirupati, Andhra Pradesh, India
Email: madhu.dagada@gmail.com

1. INTRODUCTION

The agriculture is leading sector of India as 60% of the population is directly and indirectly involved in the development by contributing considerably to the nation's gross domestic product (GDP), employment, and food security. Agriculture has been a leading area of the Indian economy for centuries, supporting the livelihoods of the vast population of approximately 1.42 billion. In terms of economic importance, agriculture contributes around 16-17% to India's GDP [1]. Farming is a step-by-step process that starts from preparation of soil, selection of a crop, sowing, adding manure and fertilizers, irrigation, harvesting, and storage. Farmers usually follow traditional methods to select crops based on previous experience and similar crops of surrounding farmers. These methods are unable to produce better results every time.

Artificial intelligence (AI) enabled farm cultivation, which helps farmers to make perfect decisions about crop selection, disease prediction, and pest detection [2]. Recently, farmers-initiated data-driven strategies such as precision agriculture (PA), which uses AI-driven methods to increase crop yields by selecting suitable crops and supporting the nation's ecological farming growth. Machine learning (ML) is a sub-area of AI. The underlying application of ML [3] in the present study is the prediction of the most

suitable crops for cultivation. The core concept of ML is, to develop a model in such a way that it learns from experiences and improves performance. Various ML applications have been introduced, encompassing auto-irrigation systems for agri-farms [4], drones to analyze agricultural land [5], monitoring systems for crops [6], PA [7], and animal identification, among others. Consequently, this approach proves highly advantageous for agricultural practitioners.

A framework for crop recommendations [8] was created using the ensemble method of ML. The optimal crop is recommended by the ensemble technique based on the properties and nature of the soil with an accuracy of 99%. The ensemble model uses random forest (RF), Naïve Bayes (NB), and linear support vector machines (SVM) as base classifiers, which are conventional ML techniques. The dataset contained samples of surface temperature and annual rainfall as well as chemical and physical characteristics of the soil. The best algorithms for crop categorization were found by evaluating and comparing the output of several classification algorithms [9]. Additionally, they examined the impact of such algorithms on crop prediction and offered improved crop-related tactics. Lastly, they recommended enhancing the estimation and response time of the existing methods. Despite of high accuracy of the system, very few models were considered for the ensemble approach.

A majority voting procedure followed by an ensemble approach, NB, K-nearest neighbors (KNN), and RF as base learners, was suggested by a recommended system [10] to crop for site-specific parameters to recommend a crop with high efficiency. The recommended crop is based on the crop yield estimation model [11], which assessed the ANN-GWO (artificial neural networks with grey wolf optimizer)'s efficiency for crop yield with a root mean square error (RMSE) of 3.19, and mean absolute error (MAE) of 26.65. This research's main objective was to develop a module that would help farmers choose the best crop for their region. However, it is a complex method for farmers to use for crop recommendation. These methods were recommended based on yield prediction of individual crops.

An automated crop recommendation website [12] was created, utilizing datasets that offer comprehensive records of various area characteristics, development specifics, and soil parameters. Depending on the parameters in the dataset, their system might suggest crops. Crop projections covered every type of crop grown in the US and were not restricted to any one crop species. The dataset included data on all crops in every province at the district level, totaling over 2.5 lakh documents. The results demonstrated the effectiveness of ML techniques. With an accuracy of 93.2%, RF outperformed the other classifiers. The crop is suggested by a simple and better mobile application with a graphic user interface (GUI) integrated with the model, which helps to suggest crops [13] based on input parameters of weather and soil data. It also took into account crop cultivation expenses and the location, time, and source of irrigation.

Several ML methods [14] use soil data about the area to forecast a suitable crop for a recommendation. Several ML approaches are used for soil classification, such as bagged trees, weighted KNN, and SVM models with gaussian kernel assistance. Accuracy can be increased since the crop was chosen by analyzing the quantities of soil rather than soil types. Performance metrics such as accuracy and F1-scores of a few ML algorithms [15], including decision tree (DT), SVM, NB, RF, logistic regression (LR), and extreme gradient boosting (XGBoost) evaluated, which use soil data to propose crops, XGBoost performed better than the other models.

Farmers can choose crops by taking into consideration several factors such as geographic location, soil type, and planting season by using a crop-suggested system [16]. In addition, models like LR, NB, KNN with cross-validation, KNN, DT, and neural network (NN) are taken into consideration in PA, which concentrates on site-specific crop management. At 89.88%, the NN had provided a more accurate result. Nevertheless, NN implementation is a challenging process.

Three steps weight calculation, categorization, and prediction make up the crop selection method [17], which was developed. There were 27 input criteria in total, which were broken down into 7 major categories: facilities, soil risk, input, season, water, and support. The initial stage involved utilizing the rough set methodology to assess the relative weights of each main criterion's sub-criteria, and then applying Shannon's entropy to determine the relative weights of the main criteria themselves. VIKOR (visekriterijumska optimizacija i kompromisno resenje) was used to determine the ranking index of the primary criteria because it is an effective method for sorting the alternatives and a multicriteria optimization and compromise solution. Understanding this model will take more expertise due to its complexity.

A crop is recommended by the suggestion system [18] uses pattern matching techniques to enable farmers to choose the best crop for the sowing area and season. Farmers get benefits from it as a result since their net profit will increase. The system is capable of recommending a variety of crops that are most advantageous to producers in their decision-making process. This is accomplished by analyzing a dataset that primarily comprises five criteria: rainfall, soil moisture, temperature, slope, and humidity data values that are associated with horticulture. When soil parameters differ from one farm to another, then the pattern-matching technique may not be appropriate to consider in the model.

The crop selection method (CSM) [19] proposed to solve the crop selection problem to maximize the yield of crops in a season. It led to maximum economic improvement for the nation. Soil characteristics are ignored by the method in the crop selection process, though it is an important parameter. To predict the best crop(s) for the area, a comparative [20] based analysis of several wrapper feature selection methods along with ML classification techniques was conducted. Based on the results of the trial, the recursive feature elimination technique in conjunction with the adaptive bagging classifier outperforms the other analytical approaches. The accuracy of this approach can be increased by adjusting the hyper-parameters of the ML models.

The deep learning technique (DLT) based crop-specific recommender system [21] by considering historical crop and climate data. ACO-IDCNN-LSTM, a hybrid technique combining ant colony optimization (ACO) with deep convolution neural networks (DCNN) and long short-term memory (LSTM), has been proposed for crop prediction in DLT, LSTM, and DCCN networks. High accuracy levels, typically 95.1%, are often achieved by DCNNs. There are additional layers and NN operations involved in this process. Implementing CNN and LSTM is therefore highly complex. Additionally, there is a high ACO convergence rate.

A study methodology combining machine learning and data balance was put out [22] for crop recommendation. 14 ML models are tested using Kaggle data, and boosting (Cboost) obtains the highest accuracy (99.15%), F-measure (0.9916), and precision (0.9918). Gaussian Naïve Bayes (GNB) does well in Matthews correlation coefficient (MCC) and receiver operating characteristic (ROC) (0.9569). Most classifiers took into account a small number of factors when suggesting crops. A crop recommendation system (CRS) [23] for Maharashtra that improves farmer production by utilizing data from 2001–2022. By using DL and ML, such as RF for 92% accuracy and LSTM for weather forecasting, the CRS enhances agricultural efficiency by recommending the best crops based on local conditions. This model is limited to a relatively small number of crops. By making informed judgements regarding irrigation, planting, and harvesting, the ML prediction model [24] in agriculture improves crop production. The model emphasizes the potential of integrating internet of things (IoT) data and online resources to enhance accuracy, attaining a classification accuracy of up to 99.59% using algorithms such as Bayes Net classifier.

A crop recommendation system [25] was proposed to assist producers in making informed decisions by utilizing ML. The system has the potential to increase crop output and reduce costs in the face of challenges such as population growth by predicting agricultural yields and suggesting optimum crop management practices in the context of algorithms such as DT, NB, and RF. The application of data analytics techniques, such as LR with NN, was employed to forecast crop prices [26], taking into account factors such as the area harvested and planted. The study determined that XGBoost was the most effective technique for price prediction.

Utilizing a variety of visualization tools, a model [27] incorporated mobile applications and ML to assist farmers in identifying the most effective conditions for planting, harvesting, and fertilizing crops. This model can also be modified to provide fertilizer recommendations. A regression-based ML system [28] that employs NB classifiers to forecast fertilizer usage and crop yield for crops in Mysore based on soil nutrients has demonstrated high accuracy for wheat, ragi, and paddy. These models can also be improved to create user-friendly applications specifically designed to meet the requirements of producers.

Existing solutions are developed using synthetic data to model design for crop recommendation. These models are unable to consider other soil parameters like Copper and Sulphur. Few solutions utilized DL methods for crop recommendation. Particularly in developing nations, it is imperative to customize recommendations for small-scale and subsistence farmers. Significant computational resources are necessary for numerous sophisticated models, particularly those that involve deep learning. Research could investigate methods to decrease the energy consumption of these models, particularly in developing countries with restricted access to high-performance computation.

Considering all gaps in the existing literature, a proposed model needs to be less complex, more accurate and consider all soil and weather parameters. The key contributions of this study have been listed below:

- Propose an enhanced stacked ensemble model for crop prediction with high accuracy by comparing six ML models, multi-layer perceptron (MLP), XGBoost, KNN, DT, and SVM.
- Seven different crops have been classified for the prediction, based on input soil parameters such as Nitrogen, Phosphorus, Potassium, pH, Manganese, Organic Carbon, Zinc, Electrical Conductivity, Iron, Boron, Copper, Sulphur, and weather parameters such as rainfall, and temperature.

The paper has been structured as follows: section 2 analyses the literature review of crop prediction or recommendation systems. Section 3 explains the details of the proposed model along with other ML models with analysis. Section 4 discusses about model's performance evaluation results and analysis. Section 5 concludes the research paper.

2. METHOD

The basic approach of ML is categorized into three broad types, based on the nature of the learning paradigm. These are supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the predictive machine learns from patterns and relationships among features of data. It further classifies into two types based on the target/output value of the model, i.e. classification and regression. Examples of supervised learning are DT classifier, RF classifier, KNN classifier, and MLP.

ML models are learned or experienced by taking training and testing on a given dataset. There is a limitation in individual ML models that tend to perform poorly, due to the occurrence of high bias. An alternative solution is to combine individual models into either parallel or sequential. Combining multiple models can happen in three ways, namely bagging, boosting, and stacked. In bagging, the same ML model can be considered parallelly for intermediate prediction, and the final prediction is evaluated based on the major voting of these intermediate predictions. In boosting, the same ML model takes in sequence so that incorrect prediction of the first training model forwarded to the next training model to make the combined model to be strong in prediction.

Stacking means combining multiple base models and making a meta-model. It combines the prediction of multiple ML models to create a more robust predictive model. It leverages the diversity among individual models to improve overall performance, accuracy, and generalization. Ensemble methods are widely used in various ML tasks, including regression, classification, and anomaly detection. The basic structure of stacked ensemble learning is represented in Figure 1. It is a four-step process, initially multiple base learners (Classifier 1, Classifier 2, ..., Classifier n) train on the dataset. Next, by using the prediction outputs of each learner to form a new dataset. Later a meta-model train on the newly formed dataset. At last, the meta-model produces the final prediction value.

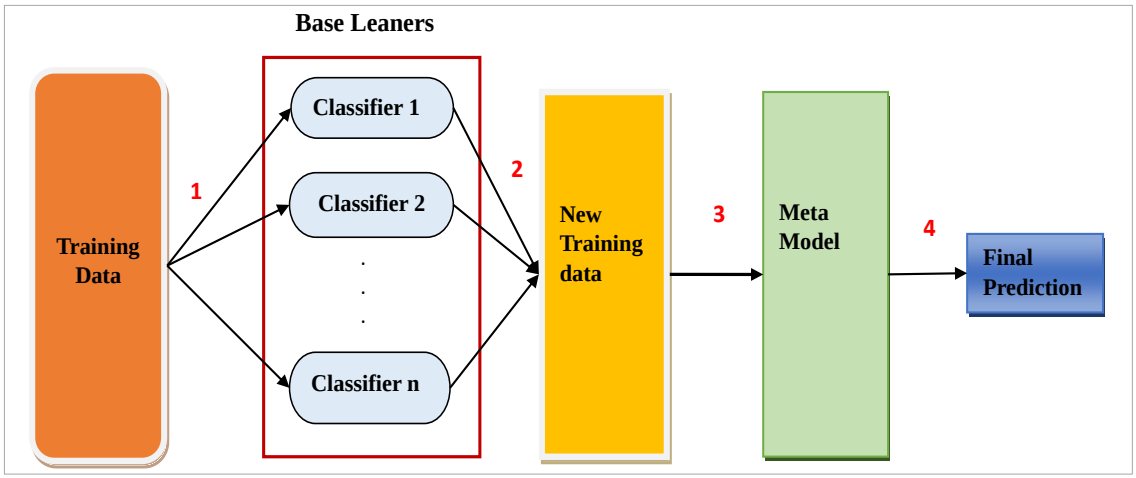

Figure 1. Simplified structure of stacked ensemble learning

2.1. Dataset description

In this work, the dataset that has been utilized for prediction is taken from a website [29]. This data is related to five districts named Anantapur, Chittoor, Kadapa, Kurnool, and SPSR Nellore of Andhra Pradesh, which is a state in south India, which contains instances of 14 input parameters, 12 out of 14 parameters are soil parameters. It is categorized into macronutrients such as Nitrogen, Phosphorous, Potassium, Sulphur, Calcium, and Magnesium and micronutrients such as Iron, Boron, Zinc, and Copper and 2 are climate parameters such as rainfall and temperature. These nutrients are very important to grow a healthy crop.

2.2. Pre-processing

The dataset contains a few outliers and missing values, which may be sensitive to a few ML models like SVM, LR, KNN, and DT. As it is related to classification, the mean imputation method is used to make missing values into suitable values. Features of the dataset are in different ranges, so each data point of features needs to scale in the same range. To make scale data points, the first mean of the column vector X is calculated, next the standard deviation of X, and calculate new scale value by using (1).

$$X_{new} = \frac{X_i - X_{mean}}{Standard\ Deviation} \quad (1)$$

2.3. Data splitting

In this work, the dataset is split into "train-tests" with different sizes to find better training and testing of the model and get better accuracy. Here, 70%-30% means 70% of total data is used for training models and the remaining 30% is kept for testing the models. The total number of instances is 315,344, the training 70% size is 220,740 instances, and the test 30% size is 94,604 instances.

2.4. Classification algorithms

ML models are used for either classification or regression. The majority of ML models perform both prediction, such as classification and regression. This paper mainly focuses on the classification work and models to make a clear analysis of the study. This research paper is concerned with classification models.

2.4.1. Decision tree

DT [30] are a popular classification technique that uses top-down procedure to create tree structure classifiers using given data. The ID3 algorithm, based on entropy, is used to calculate information gain, determining which attribute to be as root and internal node in the tree to split further. An expansion, the C4.5 algorithm, is based on ID3 and includes features like predicting continuous values and handling missing values. The DT is created by selecting the highest IG feature from the dataset and splitting it into sub-trees. This process is repeated until all features are covered in the DT.

2.4.2. Multi-layer perceptron

A MLP [31] is a type of ANN-based non-linear model that falls under the category of feedforward NNs. It consists of multiple layers of nodes (neurons) arranged layer by layer structure, including an input layer, one or more hidden layers, and an output layer. MLPs are widely used for supervised learning tasks such as regression and classification. The prediction capability of MLP comes from by maintenance of multi-level layers of neuron networks.

The basic structure of MLP is represented in Figure 2, X_1, X_2, X_3, and X_4 are input data values, which are assigned to neurons of the input layer along the bias. In each layer, neurons perform operations such as the summation of weights and activation functions. In the output layer, the SoftMax function generates the final predicted result Y based on probability. MLPs can make flexibility and complex model relationships in data, making powerful tools for various applications.

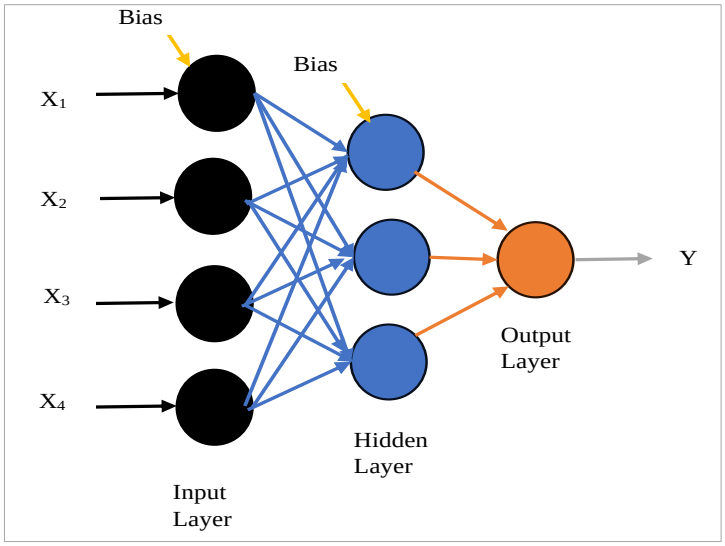

Figure 2. Multi-layer perceptron

2.4.3. Support vector machines

A SVM [32] is a supervised ML algorithm that is generally applied in high-dimensional spaces for applications of classification and regression. It is a binary classifier that categorizes data variables into either

class 0 or class 1. The hyperplane of the SVM is selected to optimize linear separation between two-class data sets of two-dimensional space points. The objective of generalization is to identify an n-dimensional hyperplane that optimizes the separation of data points from their potential classes. Data points that are closest to the hyperplane and have the minimum distance are referred to as support vectors. The foundation for data point separation calculations is a kernel function, which includes linear, polynomial, gaussian, sigmoid, and radial basis function (RBF) functions. The efficiency and fluidity of class separation are regulated by these functions, and the hyperparameters may be adjusted to induce overfitting or underfitting.

2.4.4. K-nearest neighbours

KNN [33] works based on the principle of 'k' nearest labels or values to datapoint. It is useful either for classification or regression. For classification, it considers the nearest 'k' values takes the majority voting label and gives it as output. For regression, it considers the nearest 'k' values average as output. To find the nearest 'k' values, Euclidian or Manhattan distance measure will be used. The formula for Euclidian distance 'd' of two points (X_1, X_2) and (Y_1, Y_2) is:

$$d = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \quad (2)$$

2.4.5. Extreme gradient boost classifier

The XGBoost [34], [35] extension for gradient-boosted DT. It is a popular and skillfully executed approach, represented as a DT, in gradient-boosted trees. It employs a technique that builds on the sequence of weak learners. To construct the XGBoost tree, start with finding the residual of DT-1, a weak tree, is given to DT-2, another weak tree, to reduce the overall residue. This process is continued until the final tree, n. Every XGBoost tree model lowers the residual from the tree model that came before it, in contrast to RF. Similar to the derivation of first-order for error information that the traditional gradient boosted DT (GBDT) employed. Cost functions are performed by XGBoost using both first- and second-order derivatives. The configurable cost function is additionally enabled by the XGBoost tool. To accurately predict a target variable, it combines the predictions of simpler models and multiple weak trees.

2.5. Proposed enhanced stacked ensemble learning

Enhanced stacked ensemble learning is an advanced form of ensemble learning based on stack generalization [36] that combines the strengths of individual models with additional enhancements to improve stacked model performance and robustness.

Algorithm 1. Algorithm for stacked ensemble machine learning

Input: Training dataset D= (X,Y), where X ∈ set of input features and Y ∈ output labels. A set of base learners 'B'= {DT, XGBoost, KNN, SVM, MLP}, Meta learner: Random Forest model.
Output: Predict a crop for a given input.

1: The dataset 'D' is divided into 'k' (Example 'k'=5) fold partitions, denoted as D={D_1, D_2, D_3,…, D_k}. It helps in obtaining predictions for the training set without overfitting.
2: **for** b=1 to B **do**
3: Train base classifier using D_{i-1} folds as the training set.
4: **end for**
5: Create a new training set for the meta-learner.
6: **for** b=1 to B **do**
7: Use fold D_i as the test set for prediction by the base classifier.
8: **end for**
9: Aggregate the predictions from all k folds to form a new dataset D'={(X', Y)}, where X_i' is a vector of predictions from each base learner for the j^{th} vector sample. Y_j is the true output label for the j^{th} sample.
10: Train the meta-learner RF on the new dataset D' with true output labels Y_i as the target.
11: To make a final prediction on a new, unseen test sample X_{test}:
12: Obtain predictions from each base learner B_i on X_{test}.
13: Combine these predictions to form a new feature vector X_{test}' for the meta-learner.
14: Predict the final output using the meta-learner model with X_{test}'.

Stacking, in general, involves training multiple individual base learners parallelly, combining their predictions and use to train meta learner, often referred to as a meta-model or aggregator. In Figure 3, DT, gradient boosting, KNN, support vector classifier, and MLP are base learners and RF is a meta learner. The final prediction will be given by the meta-learner. For this research study, enhanced stacked ensemble learning has different ML algorithms some models are base learners at the initial level, and prediction results of these base learners are considered as input parameters for the next level meta learner for training and cross-validation.

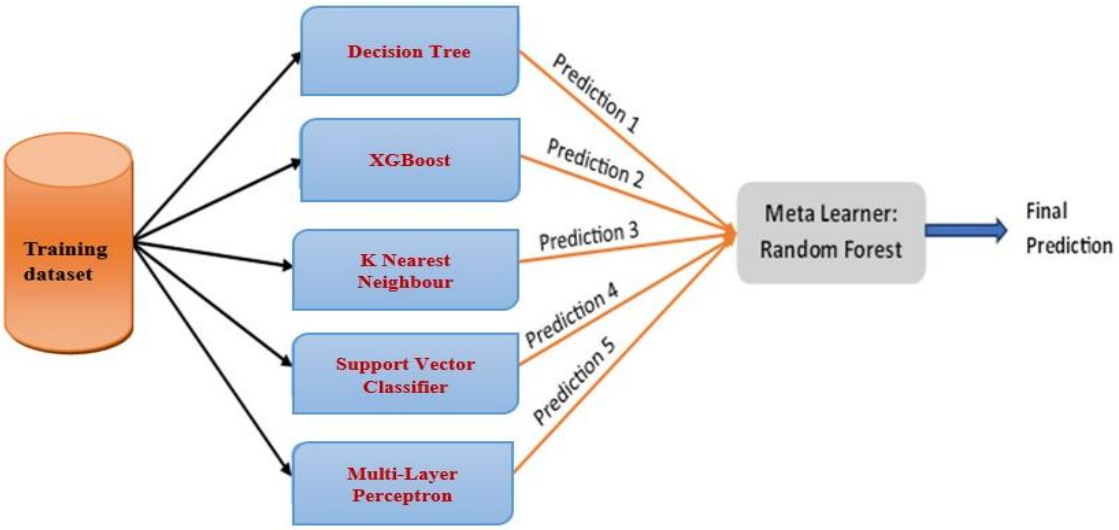

Figure 3. Structure of the proposed enhanced stacked ensemble model

3. RESULTS AND DISCUSSION

This research study investigated existing and traditional ML models such as DT, XGBoosting, KNN, and SVM, which have not comprehensively incorporated with soil and weather data for crop recommendation. Additionally, ensemble models have scope to improve performance in terms of accuracy and F1-score. Accuracy measures overall correctness, precision evaluates the quality of positive predictions, recall evaluates sensitivity to positive instances, and F1-score balances precision and recall. It is essential to comprehend these metrics to conduct a thorough evaluation and optimization of models in ML applications.

A model generates the appropriate number of predictions by analyzing the observed values, which is the essence of accuracy. The defined values are evaluated to determine whether they are true or false. A measurement of accuracy is illustrated in (3). It is assessed based on true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. Here TP means a correct prediction that an outcome is positive, TN means a correct prediction that an outcome is negative, FP means an incorrect prediction that an outcome is positive, and FN means an incorrect prediction that an outcome is negative.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (3)$$

Precision is a term that is used to assess the sensitivity and efficacy of a classification model. TP and FP statements are employed to quantify it. This classifier generates a positive probability result, which is computed by the values specified in (4).

$$Precision = \frac{TP}{(TP+FP)} \quad (4)$$

Recall refers to the scenario in which classification outcomes are deemed bad based on the classifier's probability assessment. It is assessed by genuine positive and false negative statements. The (5) illustrates the computation of recall.

$$Recall = \frac{TP}{(TP+FN)} \quad (5)$$

The F1-score is a value that is utilized in the process of calculating prediction performance. Recall and accuracy are both weighted and averaged together to determine the F1-score. The accuracy and recall are the metrics that are used to evaluate it. The computation of the F1-score is displayed in the (6).

$$F1\ Score = \frac{2*Precision*Recall}{(Precision+Recall)} \quad (6)$$

In Figure 4, different ML models and the proposed enhanced stacked ensemble models are compared with respective performance using accuracy. This clearly states that the enhanced stacked ensemble model outperforms remaining ML models such as DT, XGBoost, KNN, MLP, and SVC. When there are imbalances in the classes of a dataset, the F1-score is a more useful metric than accuracy. An improved metric to assess different ML models along with stacked ensemble learning is the F1-score. In Figure 5, F1-score comparison of different ML models. It is covey that stacked ensemble learning outperforms than remaining ML models SVC, MLP, KNN, XB, and DT. Future research can integrate the method with web or mobile applications effectively used by farmers.

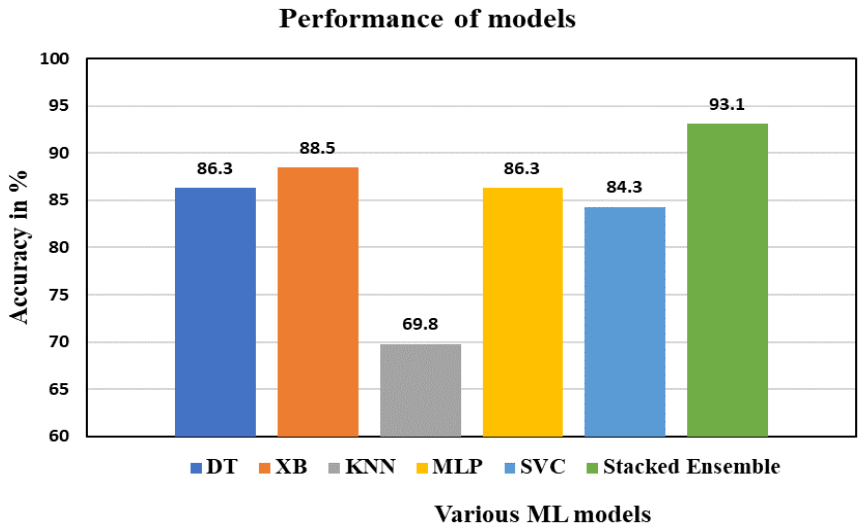

Figure 4. Accuracy comparison of different ML models

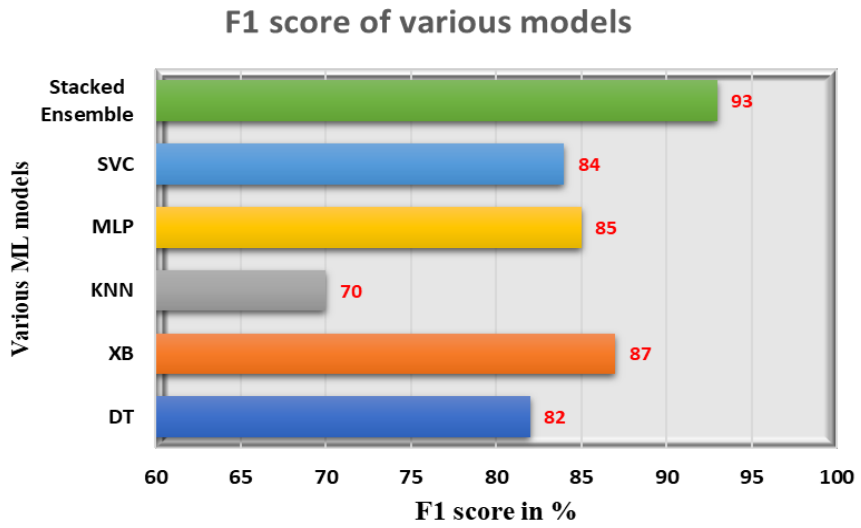

Figure 5. F1 score comparison of different ML models

In this research work, ML models such as DT, XB, MLP, and SVC perform better than KNN with above 84% accuracy. As the dataset has more dimensions, KNN is unable to handle it properly. SVC produces 84.30% accuracy by forming multiple planes in such a way as to classify data efficiently. MLP has given the accuracy at 86.3% by training data non-linearly with adjusted weights and bias. The accuracy of DT is 86.3% given by taking the high IG feature as root to split tree. A boosting technique, XGBoost maintains an accuracy of 88.5% which is better than the accuracy of DT. The XGBoost tree was constructed in such a way that residuals of a DT were reduced level by level. The proposed enhanced stacked ensemble model gives better performance at 90-10 train test size with 93.10% accuracy.

The individual ML model just creates a relationship function in an attempt to map input towards output. But stacked ensembled learning takes things a step further by determining the connection between each ensembled model's prediction result on out-of-sample data and the actual value. As stacked ensembled are designed to be more robust than average boosting models or individual models, they typically produce better predictive performance. There are instances where even small gains in prediction accuracy have a significant impact on the business situation.

This study considered and applied to five districts of Andhra Pradesh, future research can apply to the entire state as diverse soil and weather conditions. Future studies may explore integrating models with IoT devices to get real-time soil and weather data for crop recommendation. By utilizing a rich set of soil and climate data, the model significantly improves the precision of crop recommendations, thereby promising substantial benefits for the agricultural sector in terms of yield optimization and economic stability.

4. CONCLUSION

Enhancing productivity in PA has required the right crop-selection method by considering soil and weather data. Traditional ML methods like DT and XGBoost have made strides in predicting crop viability, but still, these fall short under complex, real-world conditions. The enhanced stacked ensemble ML model is an innovative approach that unites the strengths of various models into a singular, potential meta-model. This model's layered architecture significantly elevates prediction accuracy to an impressive 93.1%, by analyzing 12 soil parameters, including essential nutrients N, P, and K, alongside climatic elements such as temperature and rainfall. The outstanding performance of the proposed ensemble model marks a transformative leap in PA, promising to revolutionize crop selection processes. By doing so, it empowers India's agricultural sector with the tools for not just survival, but also for thriving economically, guiding in a new era of agronomic intelligence and economic sustainability.

ACKNOWLEDGMENTS

This work was supported by the Department of Computer Science and Engineering at Sri Venkateswara University. We sincerely appreciate their invaluable assistance.

FUNDING INFORMATION

The authors declare that they have no known competing financial interests.

AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
D. Madhu Sudhan Reddy	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
N. Usha Rani		✓			✓	✓			✓	✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data supporting this study's findings are available from a government website. Restrictions apply to the availability of these data, which were used under license for this study. Data are available at https://soilhealth.dac.gov.in with the permission of the concern team. Dataset was obtained upon request for research experiments from the government website https://soilhealth.dac.gov.in.

REFERENCES

[1] A. Gulati and R. Juneja, "Transforming Indian agriculture," in *Indian agriculture towards*, 2022, pp. 9–37.

[2] D. Crevier, *AI: the tumultuous history of the search for artificial intelligence*. United States: Basic Books, Inc., 1993.

[3] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[4] A. Vij, S. Vijendra, A. Jain, S. Bajaj, A. Bassi, and A. Sharma, "IoT and machine learning approaches for automation of farm irrigation system," *Procedia Computer Science*, vol. 167, pp. 1250–1257, 2020, doi: 10.1016/j.procs.2020.03.440.

[5] L. El Hoummaidi, A. Larabi, and K. Alam, "Using unmanned aerial systems and deep learning for agriculture mapping in Dubai," *Heliyon*, vol. 7, no. 10, p. e08154, 2021, doi: 10.1016/j.heliyon.2021.e08154.

[6] L. Benos, A. C. Tagarakis, G. Dolias, R. Berruto, D. Kateris, and D. Bochtis, "Machine learning in agriculture: a comprehensive updated review," *Sensors*, vol. 21, no. 11, p. 3758, May 2021, doi: 10.3390/s21113758.

[7] V. Hakkim, E. Joseph, A. Gokul, and K. Mufeedha, "Precision farming: the future of Indian agriculture," *Journal of Applied Biology and Biotechnology*, vol. 4, no. 6, pp. 068–072, 2016, doi: 10.7324/jabb.2016.40609.

[8] N. H. Kulkarni, G. N. Srinivasan, B. M. Sagar, and N. K. Cauvery, "Improving crop productivity through a crop recommendation system using ensembling technique," in *Proceedings 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions, CSITSS 2018*, 2018, pp. 114–119, doi: 10.1109/CSITSS.2018.8768790.

[9] A. A. Mahule, "Hybrid method for improving accuracy of crop-type detection using machine learning," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 2, pp. 2284–2288, 2020, doi: 10.30534/ijatcse/2020/209922020.

[10] K. Palanivel and C. Surianarayanan, "An Approach for prediction of crop yield using machine learning and big data techniques," *International Journal of Computer Engineering and Technology*, vol. 10, no. 3, 2019, doi: 10.34218/ijcet.10.3.2019.013.

[11] S. Nosratabadi, K. Szell, B. Beszedes, F. Imre, S. Ardabili, and A. Mosavi, "Comparative analysis of ANN-ICA and ANN-GWO for crop yield prediction," 2020, doi: 10.1109/RIVF48685.2020.9140786.

[12] P. Patil, V. Panpatil, and S. Kokate, "Crop prediction system using machine learning algorithms," *International Research Journal of Engineering and Technology (IRJET)*, vol. 7, no. 2, pp. 748–753, 2020.

[13] N. Usha Rani and G. Gowthami, "Smart crop suggester," in *Advances in Computational and Bio-Engineering: Proceeding of the International Conference on Computational and Bio Engineering*, 2020, pp. 401–413.

[14] S. A. Z. Rahman, K. C. Mitra, and S. M. M. Islam, "Soil classification using machine learning methods and crop suggestion based on soil series," in *2018 21st International Conference of Computer and Information Technology (ICCIT)*, Dec. 2018, pp. 1–4, doi: 10.1109/ICCITECHN.2018.8631943.

[15] D. M. S. Reddy and U. R. Neerugatti, "A comparative analysis of machine learning models for crop recommendation in India," *Revue d'Intelligence Artificielle*, vol. 37, no. 4, pp. 1181–1190, Aug. 2023, doi: 10.18280/ria.370430.

[16] A. Priyadharshini, S. Chakraborty, A. Kumar, and O. R. Pooniwala, "Intelligent crop recommendation system using machine learning," in *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, 2021, pp. 843–848, doi: 10.1109/ICCMC51019.2021.9418375.

[17] N. Deepa and K. Ganesan, "Multi-class classification using hybrid soft decision model for agriculture crop selection," *Neural Computing and Applications*, vol. 30, no. 4, pp. 1025–1038, Aug. 2018, doi: 10.1007/s00521-016-2749-y.

[18] Anjana, A. K. K, A. Sana, B. A. Bhat, S. Kumar, and N. Bhat, "An efficient algorithm for predicting crop using historical data and pattern matching technique," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 294–298, Nov. 2021, doi: 10.1016/j.gltp.2021.08.060.

[19] R. Kumar, M. P. Singh, P. Kumar, and J. P. Singh, "Crop selection method to maximize crop yield rate using machine learning technique," in *2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, May 2015, pp. 138–145, doi: 10.1109/ICSTM.2015.7225403.

[20] A. Suruliandi, G. Mariammal, and S. P. Raja, "Crop prediction based on soil and environmental characteristics using feature selection techniques," *Mathematical and Computer Modelling of Dynamical Systems*, vol. 27, no. 1, pp. 117–140, Jan. 2021, doi: 10.1080/13873954.2021.1882505.

[21] R. R. Mythili K., "Crop Recommendation for better crop yield for precision agriculture using ant colony optimization with deep learning method," *Annals of the Romanian Society for Cell Biology*, vol. 25, pp. 4783–4794, 2021, [Online]. Available: https://www.annalsofrscb.ro/index.php/journal/article/view/3024.

[22] S. K. Apat, J. Mishra, K. S. Raju, and N. Padhy, "An artificial intelligence-based crop recommendation system using machine learning," *Journal of Scientific and Industrial Research*, vol. 82, no. 5, pp. 558–567, May 2023, doi: 10.56042/jsir.v82i05.1092.

[23] Y. Mahale *et al.*, "Crop recommendation and forecasting system for Maharashtra using machine learning with LSTM: a novel expectation-maximization technique," *Discover Sustainability*, vol. 5, no. 1, 2024, doi: 10.1007/s43621-024-00292-5.

[24] E. Elbasi *et al.*, "Crop prediction model using machine learning algorithms," *Applied Sciences (Switzerland)*, vol. 13, no. 16, p. 9288, 2023, doi: 10.3390/app13169288.

[25] P. Rawat, M. Bajaj, S. Vats, and V. Sharma, "An analysis of crop recommendation systems employing diverse machine learning methodologies," *Proceedings - IEEE International Conference on Device Intelligence, Computing and Communication Technologies, DICCT 2023*, pp. 619–624, 2023, doi: 10.1109/DICCT56244.2023.10110085.

[26] P. Samuel, B. Sahithi, T. Saheli, D. Ramanika, and N. A. Kumar, "Crop price prediction system using machine learning algorithms," *Quest Journals Journal of Software Engineering and Simulation*, vol. 6, no. 1, pp. 14–20, 2020.

[27] T. Gupta, S. Maggu, and B. Kapoor, "Crop prediction using machine learning," *Iconic Research And Engineering Journals*, vol. 6, no. 9, pp. 279–284, 2023.

[28] C. Chandana and G. Parthasarathy, "Efficient machine learning regression algorithm using Naïve Bayes classifier for crop yield prediction and optimal utilization of fertilizer," *International Journal of Performability Engineering*, vol. 18, no. 1, pp. 47–55, 2022, doi: 10.23940/ijpe.22.01.p6.4755.

[29] "Soil health card scheme." https://soilhealth.dac.gov.in (accessed Sep. 08, 2024).

[30] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, "A comparative study of decision tree ID3 and C4.5," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 2, 2014, doi: 10.14569/specialissue.2014.040203.

[31] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 19–27, 1958.

[32] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003*, 2003, pp. 986–996.

[33] F. Murtagh, "Multilayer perceptrons for classification and regression," *Neurocomputing*, vol. 2, no. 5–6, pp. 183–197, Jul. 1991, doi: 10.1016/0925-2312(91)90023-5.
[34] M. Chen, Q. Liu, S. Chen, Y. Liu, C.-H. Zhang, and R. Liu, "XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system," *IEEE Access*, vol. 7, pp. 13149–13158, 2019, doi: 10.1109/ACCESS.2019.2893448.
[35] K. L. Ong, C. P. Lee, H. S. Lim, and K. M. Lim, "Speech emotion recognition with light gradient boosting decision trees machine," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 4, pp. 4020–4028, Aug. 2023, doi: 10.11591/ijece.v13i4.pp4020-4028.
[36] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, Jan. 1992, doi: 10.1016/S0893-6080(05)80023-1.

BIOGRAPHIES OF AUTHORS

D. Madhu Sudhan Reddy has received the B.Tech. degree in computer science and engineering from the Alfa College of Engineering Technology, the M.Tech. degree in information technology from Guru Nanak Engineering College, Hyderabad. He is pursuing a full-time Ph.D. in computer science and engineering at Sri Venkateswara University. Earlier he worked as an assistant professor at multiple premier colleges of Hyderabad. He has a total of 10 years of experience in teaching. He published several research papers in Scopus-indexed journals. He is interested in research areas of machine learning, deep learning, artificial intelligence, and algorithms. He can be contacted at email: madhu.dagada@gmail.com.

Dr. N. Usha Rani has been working as an associate professor at, the Department of Computer Science and Engineering, Sri Venkateswara University College of Engineering, Tirupati, Andhra Pradesh. She has 14 years of teaching experience. Her areas of interest are machine learning, deep learning, fuzzy logic, genetic algorithms data mining, and big data analytics. She has been awarded Ph.D. degree in the School of Computer and Information Sciences, University of Hyderabad, Hyderabad, Telangana, India. She did M.Tech in the specialization of artificial intelligence at the University of Hyderabad, Hyderabad, Telangana, India. She published several papers in reputed national and international journals with high-impact factor. She attended and presented many research papers in various national and international conferences. She guided the number of B.Tech students in their project work. She has been giving research guidance to M.Tech and Ph.D. students. She is serving as a member of BOS (UG&PG), CSE, SVUCE, Tirupati. She can be contacted at email: usharani.ur@gmail.com.