SN Computer Science (2025) 6:726
https://doi.org/10.1007/s42979-025-04261-1

ORIGINAL RESEARCH

Precision Agriculture Using a Two-Tier ML Model: Integrating aKNCN Soil Classification with ELM-mBOA Yield Prediction

Awad Bin Naeem[1] · Biswaranjan Senapati[2] · Jawad Rasheed[3,4,5] · Fazeel Abid[6] · Shtwai Alsubai[7]

Received: 12 January 2025 / Accepted: 27 July 2025

Abstract

Accurate crop yield prediction is critical for sustainable agricultural planning and resource optimization, especially amid increasing food demand and climate variability. This study proposes a novel two-tiered machine learning (ML) framework that integrates IoT-based soil data with advanced classification and regression models to enhance prediction accuracy. In the first tier, an Adaptive k-Nearest Centroid Neighbour (aKNCN) classifier evaluates soil quality based on key nutrient metrics. The second tier utilizes an Extreme Learning Machine (ELM) optimized via the modified Butterfly Optimization Algorithm (mBOA) to forecast crop yields, incorporating both soil quality and agro-environmental factors. The system is trained and validated on a publicly available Indian crop production dataset containing 10,000 samples across major crops (wheat, maize, rice), with features including soil moisture, temperature, and rainfall. Feature selection is performed using Correlation-Based Feature Selection (CBFA) and Variance Inflation Factor (VIF) methods to reduce noise and multicollinearity. Experimental results demonstrate that the proposed aKNCN-ELM-mBOA model significantly outperforms traditional ML models—such as Support Vector Machine (SVM), Artificial Neural Network (ANN), Gradient Boosting (GB), and Random Forest (RF)—in terms of error metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and R^2. The model achieves a notably low RMSE of 0.301 and MAPE of 3.932, alongside a high R^2 score of 0.817, indicating strong generalization. This approach underscores the potential of hybrid ML systems, enriched by IoT-driven data and robust optimization, to drive precision agriculture and informed decision-making. Future work may involve time series forecasting and scaling the model with real-time sensor data for broader deployment.

Keywords Crop yield prediction · Machine learning · Soil quality classification · Smart farming systems

✉ Jawad Rasheed
jawad.rasheed@izu.edu.tr

[1] Department of Computer Science, National College of Business Administration and Economics, Multan, Pakistan

[2] Department of Computer Science and Data Science, Parker Hannifin Corp, Chicago, IL 60503, USA

[3] Department of Computer Engineering, Istanbul Sabahattin Zaim University, Istanbul 34303, Turkey

[4] Department of Software Engineering, Istanbul Nisantasi University, Istanbul 34398, Turkey

[5] Applied Science Research Center, Applied Science Private University, Amman, Jordan

[6] Department of Computer Science and Information Technology, University of Lahore, Lahore, Pakistan

[7] Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, P.O. Box 151, Al-Kharj 11942, Saudi Arabia

Introduction

Artificial intelligence (AI), in the form of machine learning (ML), is gaining increasing use in various sectors, including agriculture [1, 2]. Several variables, including soil type, environmental conditions, weather, fertilizer application, and seed variety, make it challenging to forecast crop production accurately. By finding trends and connections between these elements, ML may enhance yield prediction. To train models for anticipated outcomes, a variety of attributes and historical data are combined to generate prediction models during the training phase [3–6]. ML models can be either predictive or descriptive, depending on the issues and problems the research raises. A comprehensive analysis of the literature was done to find out whether ML might be used to estimate agricultural production. A good SLR study should include a

Published online: 06 August 2025

thorough description of its methods, be repeatable, and be accessible to other researchers [7–9]. Robots, drones, and remote sensors have all been significantly impacted by modern Internet of Things (IoT) technologies. More than half of the world's food demand is expected to be satisfied by 2025, despite increased productivity, which requires smart agricultural systems capable of anticipating and monitoring crop performance. Farmers can boost production by utilizing the IoT to predict agrarian yields [10–12].

Agriculture plays a crucial role in supporting human life and the global economy. With the increasing demand for food driven by population growth, improving crop productivity has become more important than ever. Recent advances in data science and ML have opened new possibilities for precision agriculture, especially in predicting crop yields based on various environmental and agronomic factors [13, 14]. ML techniques have shown significant potential in uncovering complex patterns within large datasets—patterns that traditional statistical models might miss. These technologies can help optimize resource use, reduce waste, and enable more informed decision-making by farmers and policymakers. In this study, we propose a two-tiered ML approach to enhance crop yield prediction accuracy. The model combines Adaptive K-Nearest Centroid Networks (aKNCN) with several ML classifiers, including Random Forest (RF), Gradient Boosting (GB), Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Extreme Learning Machines (ELM). The primary objective of this work is to assess the performance of the proposed two-tiered model in comparison to traditional methods. The findings aim to support the development of intelligent agricultural systems that promote sustainable farming practices.

AI enhances agricultural practices and reduces the need for human labor, despite the high costs associated with technology. The suggested approach involves two classification models with optimal weight selection for crop production prediction. This study examines the effect of categorization and yield prediction within IoT-based smart farming systems. The primary research question examines how these techniques can enhance prediction accuracy in such systems. This paper tackles the challenge of crop yield prediction using a two-tiered ML approach. We explore how advanced models can be integrated to boost prediction accuracy in agriculture. The following is the contribution of the paper:

- The proposed method ensures the effective implementation of an IoT-based agricultural system for the deployment of a crop production forecasting model.
- This task includes pre-processing, feature selection (FS), and classification. FS approaches are used for feature selection and data pre-processing. Then, to improve crop output projections, a two-tier ML-based IoT smart agricultural system is proposed. ML-based classification is used to categorize soil samples into multiple categories, considering the features of the soil dataset.
- One proposes using the ELM model to forecast agricultural output. After being carefully assessed utilizing a range of experimental results and ML performances, the recommended technique for predicting crop output is.

Literature Review

The ability to have a greater variety of food products in the same land area due to population growth is the primary driver behind the adoption of AGRI-IoT. At an annual growth rate of 83 million, the world's population reached 7.6 billion in 2017. The global population is expected to reach 8.6 billion by 2030 and 9.8 billion by 2050. Recent years have seen a persistent management of population growth issues via agricultural innovation. Farmers have always been able to do more with less because they employ fertilizer in genetically modified crops. Artificial and deep neural networks are the most commonly utilized models in agriculture [15]. An ANN with input, hidden, and output layers may approximate a node-link structure via the application of bias and weight optimization [16]. Deep learning (DL), a subset of ML, uses an intensive learning process in a deep network to anticipate agricultural output based on fluctuations in input data layout [17]. These deep learning methods may also produce a probability model utilizing field data. In addition to this benefit, DL techniques provide information on crop performance over a wide range of environmental variations [18].

One of the most popular applications of AI is reinforcement learning. This essential set of techniques is required for both lowering logic for dynamic programming and creating ML models for decision-making sequences [19]. Another ML technology that could enhance neural network training for agricultural output prediction is called Extreme Learning Machine (ELM). It speeds up the learning process and yields superior outcomes. However, these approaches also have some shortcomings: restricted sustainability, high computing costs, excessive complexity, and imprecise forecasting [20]. This work proposes addressing these challenges by utilizing a more effective ML-based agricultural production forecast model. The primary reason for India's motivation is its agricultural sector, which is a key economic resource. The challenges associated with costly and labor-intensive traditional agricultural planting management can be addressed through the use of IoT-based real-time detection, intelligent crop growth management, and standardization of agricultural planting modes. Numerous yield approaches, both empirical and mathematical, have been tested on various crops. These models are challenging to apply in many sectors because they require crop and soil data. Numerous satellite-based remote sensing methods have been developed for yield modelling purposes. However, these technologies are insufficient to provide small farms with sufficient geographic information for crop optimization. Advances in ML models have gained significant importance recently since they allow researchers to understand and solve difficult problems. Among other ML approaches, decision trees, SVM, and ANN have been used in agricultural production forecasting. With ELM and aKNCN, this research predicts crop output and produces better results. Pakistan's economy is largely centered on the agricultural industry, which is highly competitive globally. Emerging nations are enhancing their efficiency by leveraging emerging technologies such as IoT and LPWA. While a great deal of research has been conducted on the use of IoT technology in the Agri-4 food supply chain, little is known about its suitability for primary agriculture. The results are more difficult to apply to Pakistan's goals, as it appears that large American farms are the primary focus of most agricultural research in the United States. Consequently, as farms become more environmentally friendly, their profit margins decrease. Developing more broadly accessible AGRI-IoT technologies that enable more effective resource utilization and improved agricultural waste management could be a timely solution. Projected crop output is approached using ML and proportional sensing. The experiment's goal was to gather the data required for agricultural yield control [21]. Proximal sensing has been utilized to collect data on crop characteristics and soil conditions relevant to potato tubers. Using a large dataset, prediction accuracy was assessed. Support vector regression (SVR), k-nearest neighbours (KNN), linear regression, and elastic net ML techniques were used to anticipate agricultural production. R^2, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) were used to evaluate performance. Given that it incorporates more information than other networks, KNN performed poorly in predicting agricultural productivity. demonstrated an ML and Internet of Things-driven smart agriculture system. WPART projected drought and agricultural production by combining the Wrapping and PART techniques. The wrapper feature selection technique assisted in selecting the best features for further categorization, which is one of the two most important stages in the prediction method, alongside feature selection [22]. PART, a partial decision tree approach, was used to complete the prediction and classification task. The WPART exams included accuracy, precision, sensitivity, and the F1 score. Soybeans, sugarcane, jowar, and bajra were the ingredients of the experiment. Due to the dataset containing a large number of samples with incorrect labels, erroneous predictions were frequently made. Agricultural inputs, soil nutrients, irrigated areas, and soil quality will all be used to estimate future crop yields. developed a deep reinforcement learning-based ML approach for the creation of an Internet of Things smart agriculture system. A combination of cloud computing and AI was used to forecast and categorize agricultural produce. When possible, this research tried to minimize resource usage to enhance food production. One may demonstrate the Markov decision-making process by using a hierarchical Bayesian multi-task reinforcement learning approach. Following that, policy distillation was used to assess the Q-value regression function [23]. With the use of AI and cloud computing, agricultural output was forecast and categorized. The primary goal of this project was to produce more food with fewer resources. The Markov decision process was modelled using a multi-task reinforcement learning technique based on hierarchical Bayesian methodologies. After policy distillation, one may examine the regression function of the Q-value.

Computational complexity was noted as one of this approach's primary drawbacks, nonetheless. Furthermore, neither the ability to adapt to changing circumstances nor human-level skill in performing challenging tasks was achieved. By building an incremental model and utilizing transfer learning strategies, this endeavour aims to enhance performance efficiency in the future. Created a deep learning methodology to forecast agricultural production. The primary objectives of the work were to estimate biomass, identify crops and weeds, and forecast crop yield. The yield of wheat and barley crops was predicted using a Convolutional Neural Network (CNN) model that included feature extraction, training, hyperparameter tuning, and regularization. The simulation study's assessment criteria are MAE and MAPE. However, the suggested CNN performs badly with a large dataset.

Materials and Methods

Thus, soil elements like nitrogen, phosphorus, and potassium have a significant impact on farming, as do air temperature and crop rotation. ML algorithms are essential tools for calculating agricultural production in decision-making, much like selecting which crops to plant. Various ML methods are used to facilitate the forecasting of agricultural output. Pre-processing, FS, and a two-tier model make up the recommended smart agricultural framework for crop output forecasting. It is recommended to utilize AkNCN, an improved version of KNCN. When compared to other algorithms, ELM is often a powerful model that utilizes faster learning processes, enhances performance, and reduces training error. In this work, two classification strategies are proposed to improve system accuracy and yield outcomes that surpass those of existing models. Figure 1 illustrates the complete process pipeline, beginning with data preprocessing and feature selection (CBFA and VIF), followed by tiered classification using aKNCN for soil quality and ELM-mBOA for yield estimation, integrating IoT-sensed features to enhance predictive accuracy. Data noise is first eliminated through pre-processing, and then feature selection methods, such as CBFA and VIF, are employed to refine the dataset. Finally, two-tier classification schemes are used. The first layer determines soil quality based on nutrients collected by IoT sensors using the proposed aKNCN model. The second layer then utilizes ELM-mBOA to predict crop yield; mBOA aids in selecting the optimal weight, thereby improving accuracy. This model enhances the system by minimizing error values and raising system accuracy.

Preprocessing

We utilised the publicly available *District-wise Season-wise* Crop Production Statistics dataset provided by the Government of India. This dataset comprises detailed records on crop production across Indian districts, segmented by crop type, region, and season, spanning multiple years. It is available at: https://www.data.gov.in/cata

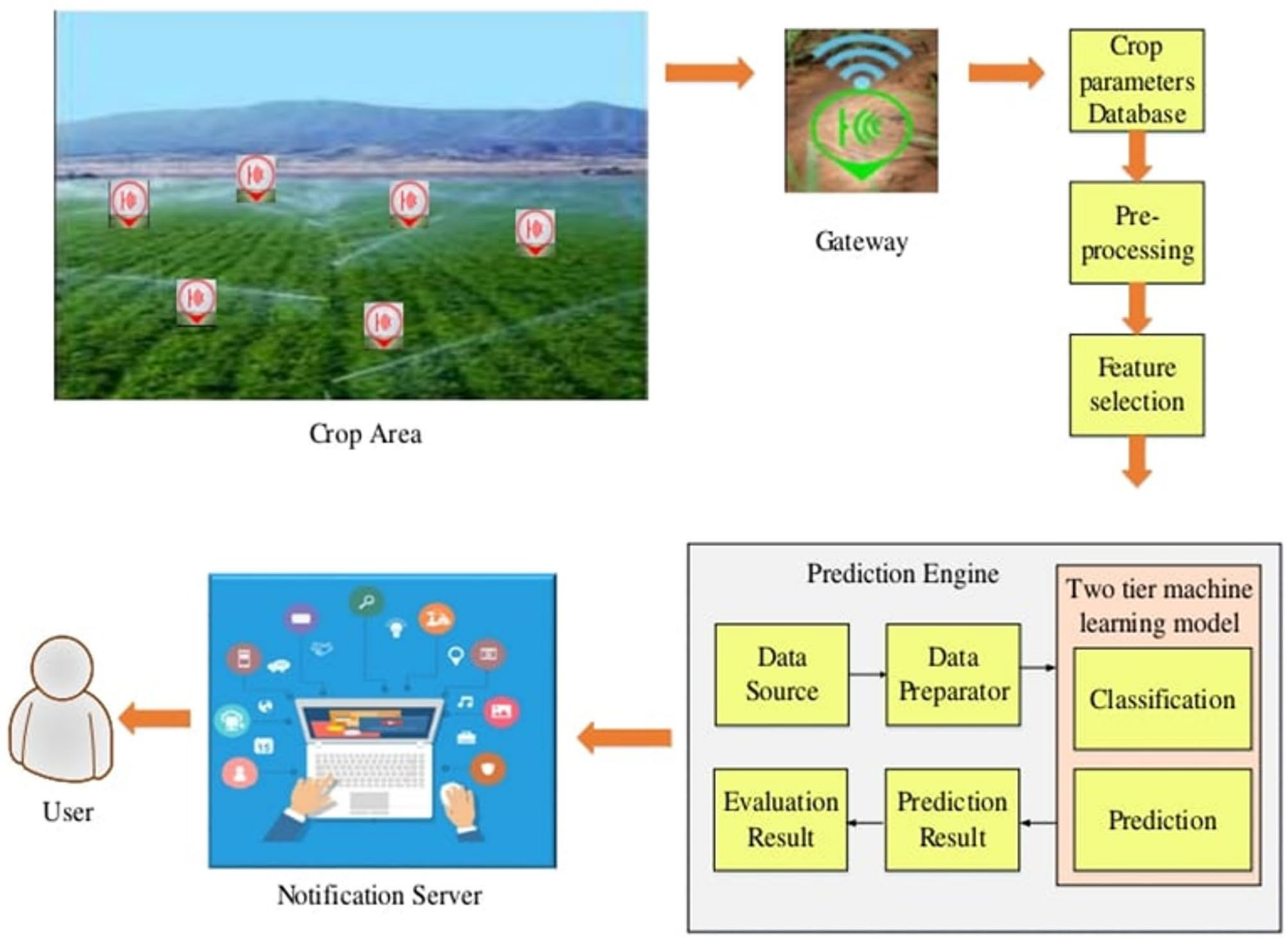

Fig. 1 Workflow diagram of the proposed two-tier machine learning framework for crop yield prediction

log/district-wise-season-wise-crop-production-statistics-0. This dataset including environmental and crop-specific features relevant to yield prediction. Table 1 presents a breakdown of the input dataset used for model training and evaluation, detailing the types of features (e.g., categorical or numeric), the overall sample size (1500), and the proportional distribution across major crop types (wheat, maize, rice), which provides context for data balance and diversity. The data distribution is relatively balanced, though wheat samples slightly outnumber others, which was accounted for during model training by applying stratified sampling techniques.

Preprocessing steps included handling missing values, normalizing continuous variables, and encoding categorical attributes. The final dataset was divided into a training set (80%) and a testing set (20%) for model evaluation. This dataset serves as the foundation of the proposed two-tiered ML approach, enabling models to learn from both agro-climatic conditions and management practices to predict crop yield with improved accuracy. The comprehensive dataset allows models to understand the complex relationships between environmental factors and crop yield, thereby contributing to robust and generalizable predictions.

The initial step after gathering data from various sources is pre-processing. Because ML cannot handle noisy data, pre-processing is necessary. Errors and outliers make up noisy data. Before categorization, the data must be pre-processed to fill in missing values, eliminate unnecessary data, extract functionality, and maintain the proper data range. In this work, category data in text format is converted to numerical data using the label encoder technique and the isnull approach to identify null values.

Feature Selection

ML is a computer learning paradigm that is driven by statistical value prediction. Using the FS model, important components highly linked to crop production are identified. The main argument in favour of FS is that it expedites the training process of the ML method, reduces model complexity, and simplifies interpretation. Selecting the right subset also helps reduce overfitting and improve system accuracy. For feature sets of typical size, the method's classification time takes precedence over its computation time. However, when working with large datasets, feature selection is crucial. FS might utilize statistical techniques such as filtering, embedding, and wrappers. When using filter approaches, cross-validation performance is subordinated to the inherent properties of features derived from univariate statistics. Compared to wrapper approaches, these methods operate more quickly and computationally efficiently. Filters enable higher-dimensional data to be handled more computationally effectively. Thus, filter-based FS methods such as the VIF and CBFA algorithms are used in this work. CBFA selects the ideal feature set, which is mostly related to yield. The multicollinearity of independent characteristics is validated using VIF. In this way, all multicollinear features are removed.

Table 1 Summary of dataset attributes, including feature types, sample size, and crop category distribution

Attribute	Description	Data Type	Crop Distribution
Crop Type	Wheat, Maize, Rice	Categorical	Wheat 40%, Maize 35%, Rice 25%
Soil Moisture	Percentage (%)	Numeric	-
Temperature	°C	Numeric	-
Rainfall	Mm	Numeric	-
Yield	kg/ha	Numeric	

Correlation-based Feature Selection Algorithm (CBFA)

In CBFS, features are arranged in order of correlation heuristic. This feature is intended for a set of attributes that are not associated with one another but have a strong connection across classes. They must be removed because irrelevant traits have a significant correlation with other features but a weak association with different categories. Acceptance of a feature is determined by how successfully it separates classes in situations where other features have not been able to do so in the past. In the CBFS method, features are ranked based on a correlation heuristic, which prioritizes attributes that exhibit a strong correlation with the target class while minimizing their correlation with one another. This method aims to identify features that uniquely contribute to class discrimination, even when other features fail to do so. Irrelevant or redundant features often exhibit a high correlation with different features but a weak association with the target variable. Such features must be removed to reduce noise and enhance model generalization.

Let N be the number of selected features. The CBFS score is computed based on the average feature-class correlation, $\overline{r_{cf}}$, and the average inter-feature correlation $\overline{r_{ff}}$. The heuristic merit M_s of a feature subset S is calculated as:

$$M_n = \frac{N.\overline{r_{cf}}}{\sqrt{N + N\left(N-1\right).\overline{r_{ff}}}} \tag{1}$$

Where:

- $\overline{r_{cf}}$ is the average correlation between each feature and the target class.
- $\overline{r_{ff}}$is the average intercorrelation between feature pairs.

A higher M_s indicates a more optimal subset, consisting of features that are highly predictive of the target variable and minimally redundant.

Variance Inflation Factor Algorithm (VIF)

The VIF approach is used in least squares regression analysis to calculate multicollinearity. It estimates the degree to which a certain regression coefficient's variance is increased by collinearity. The VIF model is used to identify and reduce characteristics that are independent yet correlated with each other. This method is fast since it uses the predictor's one-pass search. Furthermore, this method is computationally efficient and prevents overfitting while assessing each predictor in the model.

The Variance Inflation Factor (VIF) is used in least squares regression analysis to detect multicollinearity among features. It quantifies the degree to which the variance of a regression coefficient is inflated due to the linear relationships among predictors.

To calculate the VIF for a given independent variable Xj, it is regressed against all other independent variables. The VIF is then defined as:

$$VIF_j = \frac{1}{1 - \boldsymbol{R_j^2}} \tag{2}$$

Where *R_j^2 is the coefficient of determination from regressing XjX_jXj on the remaining variables.*

Interpretation:

- If R_j^2=0, meaning no multicollinearity, then VIF_j =1.
- A VIF value greater than 10 typically indicates significant multicollinearity and suggests that the feature should be removed.

This method is computationally efficient, requires only a one-pass regression per feature, and helps prevent overfitting by ensuring that only orthogonal (non-redundant) features are retained.

Tier 1-Classification

The aKNCN is used in this work to classify soils based on various factors. The proposed KNCN simultaneously enhances the KNCN classification performance and addresses the limitations of the prior KNCN. KNCN, a non-parametric classifier, uses the centroid distance. This indicates that the nearest neighbours of the test samples must meet two conditions: they must be located near the test samples, and the distribution of their nearest neighbours must be symmetrical within the test samples. But it's not easy to find neighbours on a property that meets these requirements. Despite its amazing accuracy, KNCN has a slow categorization speed. Consequently, aKNCN is designed to alter the nearest centroid neighbour for each input sample appropriately, enhancing classification precision and reducing classification time. The aKNCN has two attributes, as given below.

Property 1

The aKNCN approach only reaches a stable searching phase when the distance to the closest centroid, the product of the multiplier k above a predefined threshold, using the first centroid, zinc, 1, to describe the neighborhood's test distance, is achieved.

$$\boldsymbol{d}(\boldsymbol{y}, \boldsymbol{z^s}) > \boldsymbol{kX}\boldsymbol{d}(\boldsymbol{y}, \boldsymbol{z_{ncn}}, 1) \tag{3}$$

The symbol d(y, z s) represents the lowest centroid distance between test samples. This is the original centroid distance; the multiplier product is greater than or equal to one.

$$\boldsymbol{d}(\boldsymbol{y},\ zinc, 1) \tag{4}$$

Property 2

The aKNCN approach only succeeds in the searching phase when the whole sample class, Mi, is among the j closest neighbours and the total number of samples per class to compete, class, is less than Mi-1. The following describes the quality:

$$(\therefore MCM(x) \therefore \{mcmy \mid mcmy \therefore V\}jIisj) \qquad ((5))$$

Results

Table 2 evaluates the predictive performance of several ML models—including aKNCN-ELM-mBOA, ELM-BOA, ANN, SVM, RF, and GB—across standard metrics: MAE, MSE, RMSE, MSLE, MedAE, MAPE, R^2, and EVS in predicting crop yield. SVM, ANN, ELM, gradient boost, and RF are expected to be state-of-the-art methods that rely on squared logarithmic error prediction

Table 2 Comparative analysis of model performance across different algorithms based on error and accuracy metrics

Approaches	Error Measures							
	MAE	MSE	RMSE	MSLE	R^2	EVS	MedAE	MAPE
aKNCN -ELM-more	0.064	0.091	0.301	0.011	0.817	0.818	0.0504	3.932
aKNCN- ELM-BOA	0.067	0.095	0.309	0.011	0.806	0.808	0.0529	3.871
aKNCN -ELM	0.097	0.134	0.366	0.016	0.730	0.731	0.078	5.565
aKNC _ANN	0.130	0.181	0.426	0.022	0.636	0.636	0.105	7.809
aKNC-SVM	0.165	0.230	0.480	0.028	0.538	0.538	0.132	9.685
aKNC-GB	0.231	0.320	0.566	0.039	0.359	0.359	0.187	13.71
aKNC-RF	0.293	0.413	0.642	0.050	0.174	0.17	0.233	17.4
aKNCN -ELM-mBOA	0.064	0.091	0.301	0.011	0.817	0.818	0.0504	3.932
aKNCN- ELM-BOA	0.067	0.095	0.309	0.011	0.806	0.808	0.0529	3.871

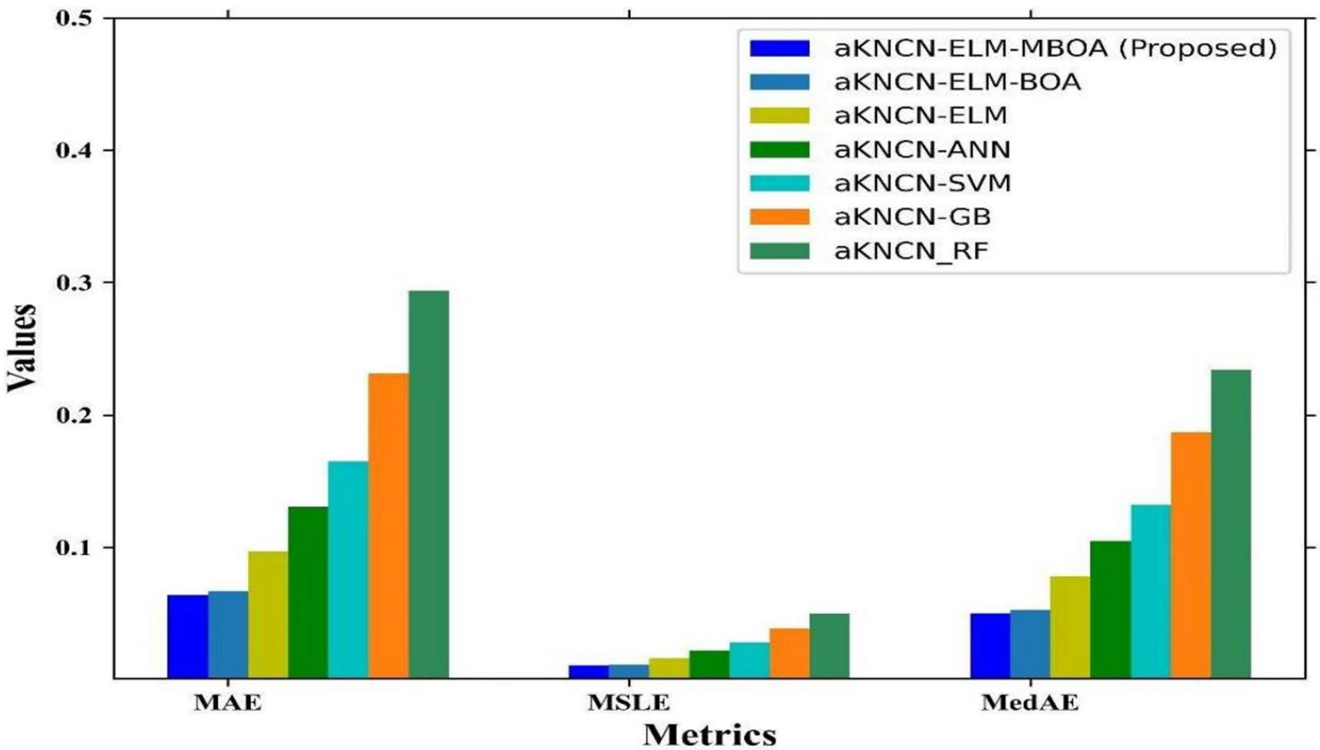

Fig. 2 Error metric comparison of competing models in predicting crop yield using the proposed two-tier ML approach

or MSLE. Using these current methods, a KNCN-ELM-mBOA is simulated, and error metrics are employed to evaluate the prediction efficacy of each technique. It is evident from Table 2 that the MAE, MSE, RMSE, MSLE, MedAE, and MAPE values generated by the proposed aKNCN-ELM-mBOA were lower than those obtained by the existing approaches. R^2 and EVS values produced by the proposed aKNCN-ELM-mBOA were higher than those of the existing techniques.

Figure 2 visualizes key error metrics—MAE, MSLE, and MedAE—for each model evaluated. The aKNCN-ELM-mBOA demonstrates consistently lower error values, validating its improved performance over existing methods. The suggested approach yielded lower error values for MAE, MSLE, and MedAE compared to previous methods, as illustrated in Fig. 2. Similarly, the proposed model enhances MedAE and MSLE scores. aKNCN-ELM-mBOA has an MAE of 0.064; aKNCN-ELM has an MAE of 0.067; and aKNCN-ELM has an MAE of 0.130, 0.165, 0.231, 0.233, and 0.293. Generally speaking, if the prediction system has a low error rate, the model is considered successful. As a result, the suggested method for forecasting crop yield is practical.

Figure 3 highlights the performance of various algorithms, with the aKNCN-ELM-mBOA achieving the lowest RMSE and highest R^2, confirming its superior ability to model complex, non-linear relationships in agricultural yield prediction. The recommended model outperformed other methods in producing lower error levels for both RMSE and MSE, as illustrated in Fig. 3. Nevertheless,

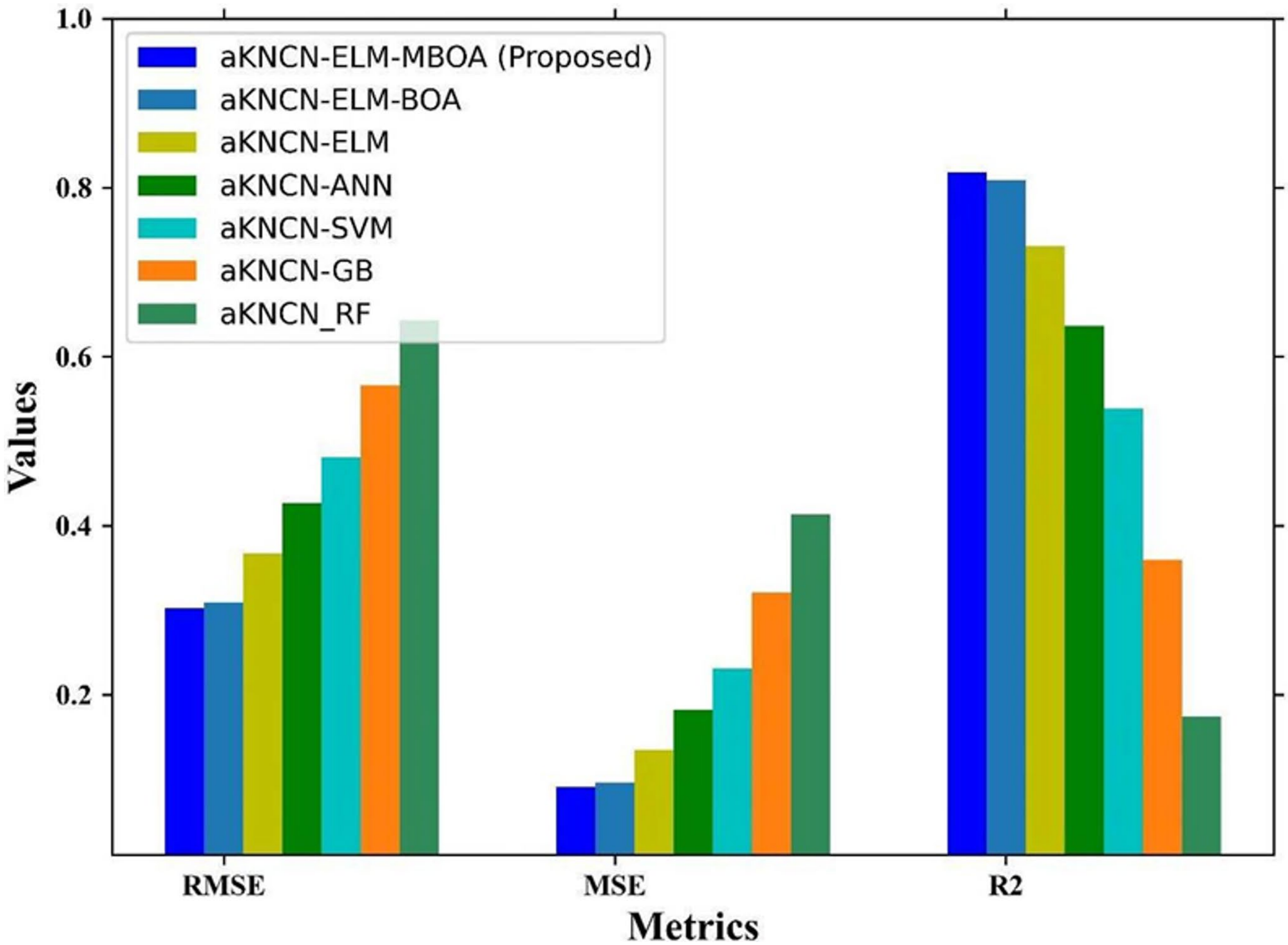

Fig. 3 Comparative performance of machine learning models based on RMSE, MSE, and R² scores

the recommended strategy outperforms existing strategies in terms of R^2 values. The R2 values for the following models are 0.806, 0.730, 0.636, 0.538, 0.359, and 0.174 for aKNCN-ELM-BOA, aKNCN-ELM, aKNC-ANN, aKNC-SVM, aKNC-GB, and aKNC-RF, respectively. The R2 value of the suggested model is 0.817. Furthermore, the recommended model's RMSE value of 0.301 indicates a lower error rate than the other classification methods. As a result, the proposed paradigm worked in every scenario.

Figure 4 compares the Mean Absolute Percentage Error (MAPE) and Explained Variance Score (EVS) across models. The aKNCN-ELM-mBOA achieved the lowest MAPE and highest EVS, emphasizing its enhanced predictive reliability and variance explanation capability. With respect to the results presented in Fig. 4, the recommended strategy yielded a MAPE value that was lower than that of any other strategy. Nevertheless, the proposed method yields an EVS value that is higher than that of existing methods. Thus, the MAPE values of aKNCN-RF, aKNCN-GB, aKNCN-ANN, aKNCN-SVM, and aKNCN-ELM (Extreme Learning Machine) are 17.40, 13.71, 7.80, 9.68, and 5.56, respectively. The comparable MAPE values for aKNCN-ELM-mBOA and aKNCN-ELM-BOA are 3.932 and 3.871, respectively. More individuals selecting the optimal path of action leads to better outcomes when using our recommended technique.

Figure 5 illustrates the close match between predicted values and actual data. The aKNCN-ELM-mBOA model shows superior alignment with real observations, confirming its robustness and practical utility in real-world agricultural forecasting. The approach is considered successful in predicting crop yields, as the forecasted results are nearly identical to the actual results. It displays both expected figures and actual data obtained from various methods. The proposed aKNCN-ELM-mBOA predicted the graph's terminus more accurately than previous methods. The recommended aKNCN-ELM-mBOA was more precise than other methods and closely matched real data. AKNCN-GB provided lower accuracy compared to other techniques. As a result, the proposed aKNCN-ELM-mBOA outperforms existing techniques in effectiveness.

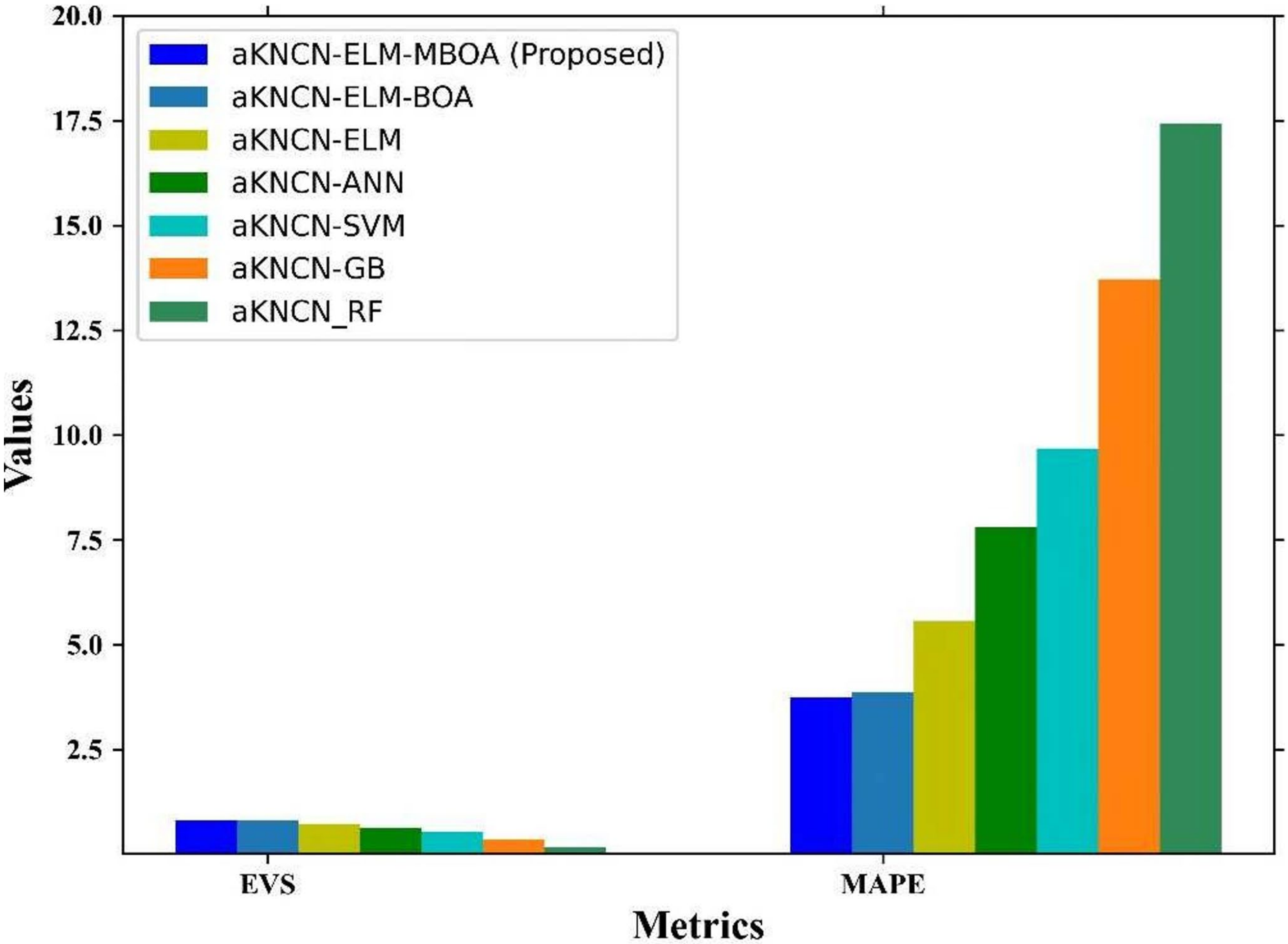

Fig. 4 Comparison of MAPE and EVS scores for yield prediction models

For the SVM, a radial basis function (RBF) kernel was used with a regularization parameter of C = 1.0. The ANN model consisted of three hidden layers, each with 128, 64, and 32 neurons, respectively. The ELM model used 500 hidden neurons with a linear activation function.

The aKNCN-ELM-mBOA model outperforms other techniques primarily due to: (1) efficient hyperparameter tuning via the mBOA, enabling exploration of optimal model parameters; (2) effective feature selection using CBFA and VIF, which reduces noise and multicollinearity; and (3) the adaptive architecture of the aKNCN combined with ELM, which captures nonlinear dependencies in the data better than traditional models. The superior performance of the aKNCN-ELM-mBOA model can be attributed to multiple factors:

- Advanced hyperparameter tuning via mBOA effectively optimizes the model parameters, preventing overfitting and improving generalization.
- The aKNCN architecture allows for better nonlinear feature representation compared to traditional ELM models.
- The two-tiered feature selection reduces noise and multicollinearity, providing the model with the most relevant inputs.

Conclusions and Future Work

This study predicts agricultural production using a two-tier ML approach called aKNCN and ELM-mBOA. The proposed aKNCN model is employed in the first layer to assess soil quality based on data collected from Internet of Things devices about nutrients. The soil quality score from the ELM model, along with other factors such as temperature and precipitation that influence crop production, are used to estimate yield in the second layer. The mBOA helps optimize the hyperparameters of the ELM prediction model to increase accuracy. Python is the programming language used

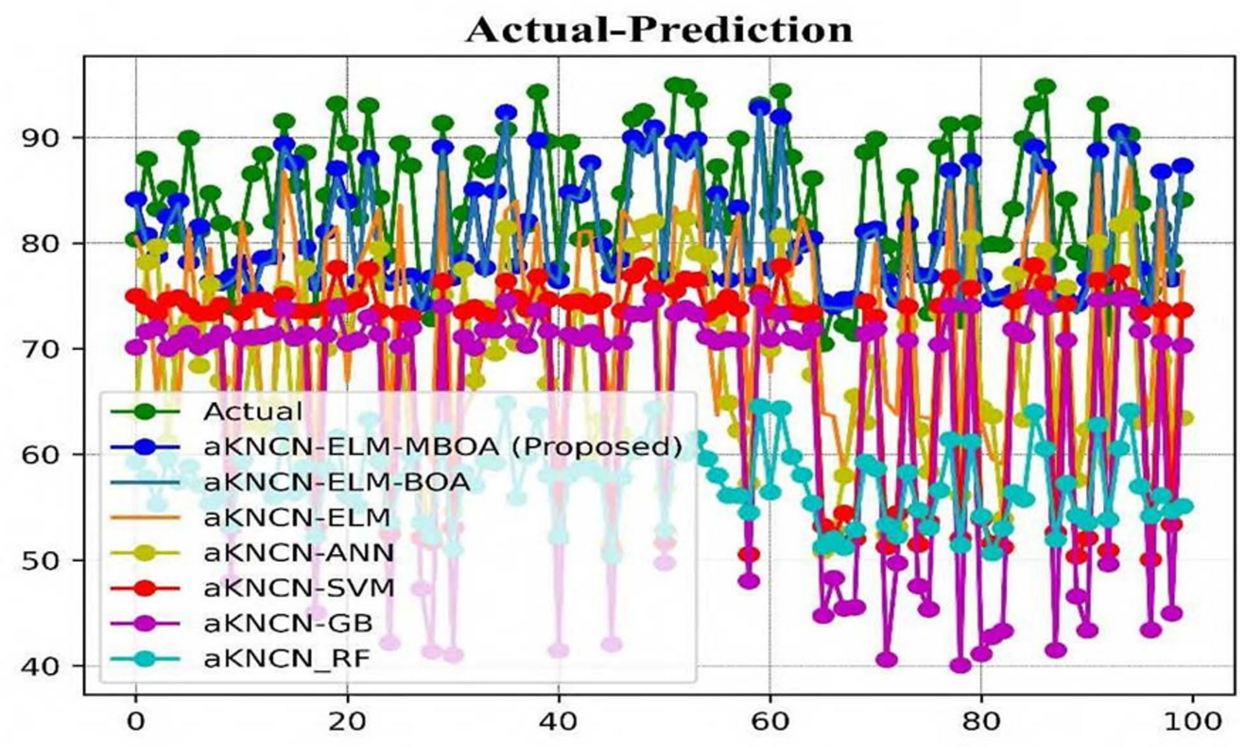

Fig. 5 Accuracy comparison between predicted and actual crop yields across machine learning models

to develop this system. The performance of the proposed prediction model is evaluated with a soil dataset. Compared to existing classification methods, this approach yields better results in terms of accuracy, RMSE, R^2, MSE, MedAE, MAE, MAPE, and EVS. The RMSE and MAE of the aKNCN ELM-mBOA model are 0.301 and 0.064, respectively. Future time series analysis will facilitate the prediction of future values. Incorporating various factors such as soil nutrients, soil quality, irrigated areas, and agricultural locations could enhance the study's scope and improve system accuracy. Additionally, deep learning-based smart agriculture, which utilizes the Internet of Things, may further enhance output quality. In future, incorporating a variety of criteria, such as soil nutrients, soil quality, irrigated regions, and agricultural locations, can expand the study's scope and improve the system's accuracy. To enhance output quality, deep learning-based smart agriculture may also be implemented through the Internet of Things system.

Funding This research received no external funding.

Data Availability The dataset used in this study is publicly available and can be accessed from the Open Government Data (OGD) Platform India at the following URL: https://www.data.gov.in/catalog/district-wise-season-wise-crop-production-statistics-0.

Declarations

Conflict of Interest The authors declare no conflict of interest.

References

1. Lee E, et al. Causes of food allergy according to age and severity: a recent 10-year retrospective study from a single tertiary hospital. Allergy Asthma & Respiratory Disease; 2020.
2. Gibril MBA, et al. Deep convolutional neural network for large-scale date palm tree mapping from UAV-based images. Remote Sens. 2021;13(14):2787.
3. Xiao Q, Li W, Kai Y, Chen P, Zhang J, Wang B. Occurrence prediction of pests and diseases in cotton based on weather factors by long short-term memory network. BMC Bioinformatics. 2019;20(25):1–15.
4. Soomro AM, Naeem AB, Shahzad K, Madni AM. Armado Dacalap Del mundo, Muhammad sajid, & Muhammad Ashad baloch. Forecasting cotton whitefly population using deep learning. J Comput Biomedical Inf. 2022;4(01):64–76. https://doi.org/10.56979/401/2022/67
5. Naeem AB, Senapati B, Chauhan AS, Kumar S, Orosco Gavilan JC, Abdel-Rehim F. W. M. Deep learning models for cotton leaf disease detection with VGG-16| international journal of intelligent systems and applications in engineering. Deep learning models for cotton leaf disease detection with VGG-16|. Int J Intell Syst Appl Eng. 2023, February 17. https://www.ijisae.org/index.php/IJISAE/article/view/2710

6. Kataria SK, Pal RK, Kumar V, Singh P. Population dynamics of whitefly, bemisia tabaci (Gennadius), as influenced by weather conditions infesting Bt cotton hybrid. J Agrometeorology. 2019;21(4):504–9.
7. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Ippodrino R. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Translational Med. 2020;18(1):1–9.
8. Kumar, S., Jain, A., Shukla, A. P., Singh, S., Raja, R., Rani, S.,… Masud, M. A comparative analysis of machine learning algorithms for detection of organic and nonorganic cotton diseases. Mathematical Problems in Engineering. 2021;2021.
9. Yang J, et al. Development and validation of a deep learning model for detection of allergic reactions using safety event reports across hospitals. JAMA Netw Open. 2020;3(11):e2022836–2022836.
10. Salman A et al. Leaf classification and identification using Canny Edge Detector and SVM classifier. 2017 International Conference on Inventive Systems and Control (ICISC), 2017:1–4.
11. Filippi P, Han SY, Bishop TF. On crop yield modelling, predicting, and forecasting and addressing the common issues in published studies. Precision Agric. 2025;26:8. https://doi.org/10.1007/s11119-024-10212-2
12. Ahmad HS et al. Improving water use efficiency through reduced irrigation for sustainable cotton production, sustainability. 2021;13(7):4044.
13. Screpnik CR, Zamudio E, Gimenez LI. Artificial intelligence in agriculture: a systematic review of crop yield prediction and optimization, in *IEEE Access*. 2025;13:70691–70697. https://doi.org/10.1109/ACCESS.2025.3560631
14. Chen Z, Gallie DR. The ascorbic acid redox state controls guard cell signalling and stomatal movement, (in eng). Plant Cell. 2004;16(5):1143-62. https://doi.org/10.1105/tpc.021584
15. Racchi ML, et al. Genetic characterization of Libyan date palm resources by microsatellite markers. 3 Biotech. 2014;4(1):21–32.
16. Abbasi-Oshaghi E, Mirzaei F, Farahani F, Khodadadi I, Tayebinia H. Diagnosis and treatment of coronavirus disease 2019 (COVID-19): laboratory, PCR, and chest CT imaging findings. Int J Surg. 2020;79:143–53.
17. Dhaka VS et al. A survey of deep convolutional neural networks applied for prediction of plant leaf diseases. Sensors. 2021;21(14):4749. [Online]. Available: https://www.mdpi.com/1424-8220/21/14/4749
18. Halken S, et al. EAACI guideline: preventing the development of food allergy in infants and young children (2020 update). Pediatr Allergy Immunol. 2021;32(5):843–58.
19. Li T, Zhao H, Wang S, Yang C, Huang B. Attack and defense strategy of distribution network cyber-physical system considering EV source-charge bidirectionality. Electronics. 2021;10(23):2973. [Online]. Available: https://www.mdpi.com/2079-9292/10/23/2973
20. Chen, W., Hasegawa, D. K., Kaur, N., Kliot, A., Pinheiro, P. V., Luan, J.,… Fei,Z. The draft genome of whitefly Bemisia tabaci MEAM1, a global crop pest,provides novel insights into virus transmission, host adaptation, and insecticide resistance. BMC Biology. 2016;14(1):1–15.
21. Sufyan Tahir M et al. Jan., Transformation and evaluation of broad-spectrum insect and weedicide resistant genes in Gossypium arboretum (Desi Cotton), GM Crops Food. 2021;12(1):292–302.
22. Zhi J, Qiu T, Bai X, Xia M, Chen Z, Zhou J. Effects of nitrogen conservation measures on the nitrogen uptake by cotton plants and nitrogen residual in soil profile in extremely arid areas of Xinjiang, China. Processes. 2022;10(2):353. [Online]. Available: https://www.mdpi.com/2227-9717/10/2/353
23. Senapati B, Talburt JR, Bin Naeem A, Batthula VJR. Transfer learning based models for food detection using ResNet-50, 2023 IEEE International Conference on Electro Information Technology (eIT), Romeoville, IL, USA, 2023, pp. 224–229. https://doi.org/10.1109/eIT57321.2023.10187288