Mrs. G. Rohini Phaneendra Kumari, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Predicting Agricultural Productivity Using Machine Learning Models

[1]Mrs. G. Rohini Phaneendra Kumari, [2]Thalla Vasavi Surya Prabha,
[3]Bandla Hima Varshini, [4]Kilaru Vyshnavi, [5]Domathoti Nadiya

[1]Assistant Professor, Department of IT, Vignan's Nirula Institute of Technology and Science for Women, Guntur.
[2,3,4,5]B.Tech, Department of IT, Vignan's Nirula Institute of Technology and Science for Women, Guntur.

Abstract- One of the most important prerequisites for sustainable agricultural planning, food security, and proper resource management is being able to accurately predict crop yield. The nonlinear and intricate interactions between environmental variables and crop productivity are often overlooked by traditional statistical models based on limited historical data. This work is a machine learning exploration to develop a model that predicts food production given the soil, weather, and crop-related parameters. Our dataset has environmental and agricultural features that include soil nutrients like (N, P, K, pH), atmospheric variables (temperature, relative humidity, and rainfall) and crop-specific attributes. By mixing these factors together the research intends to deliver an intelligent predictive framework that would be very useful for farmers and policymakers in decision-making and agricultural planning. To get better prediction accuracy, various machine learning algorithms including Random Forest (RF), Support Vector Machine (SVM), and XG Boost were applied and compared. The quality of each model was measured by R^2 score, RMSE, and MAE metrics. Random Forest was the best performer and therefore had the best precision and stability in capturing nonlinear data patterns among the models that were tested. The findings serve as evidence of the potential of machine learning methods to revolutionize the Agri-ecosystem by turning the traditional farm practices into data-driven decision-making, which is a significant contributor to the implementation of sustainable crop management and enhanced yield forecasting.

Keywords— Crop Yield Prediction, Machine Learning (ML), Random Forest (RF), XG Boost, Soil Nutrients, Climate Data, Precision Agriculture, Remote Sensing, Data Analytics, Sustainable Farming.

I. INTRODUCTION

Agriculture is the major source of food and raw materials for many countries and basically is the backbone of the Indian economy in particular [1]. Besides providing food, it also releases raw materials for a variety of industries and makes a considerable contribution to the total GDP of the country [2]. But the agricultural sector has been facing many challenges because of the growing population and climate change as well as the shortage of natural resources for the last several years [3]. The difficulties are so serious that the only possible way to overcome them is the use of modern technological solutions to optimize crop production and ensure food security [4]. The use of machine learning (ML), among other emerging technologies, has been recognized as the most promising one to help the agricultural sector in a quick and effective way [5].

Machine learning (ML) methods are the best suited for crop yield prediction tasks in the field of agriculture [6]. Forecasting of the end product of the crops before harvest gives an opportunity to farmers, researchers, and policymakers to select crops, allocate resources and plan markets [7]. In the past, yield estimation was based on field observations and statistical models, which were always very time-consuming, labor-Intensive, and also there was a possibility of human error [8]. Moreover, the traditional methods in question could only predict linear relationships among factors such as soil fertility, weather conditions, irrigation schedules, and fertilizer usage [9]. In

comparison, Machine Learning uses computational models which can learn from past data and detect the patterns that are not apparent, thus improving prediction accuracy [10].

Crop yields have been impacted by factors such as the characteristics of the soil like the nitrogen, phosphorus, potassium, and pH, the climate variables like the temperature, rainfall, and humidity, and crop management practices [11]. The impacts of these parameters are usually too complicated for standard analytical methods to figure out. Machine learning approaches such as Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XG Boost) are able to efficiently model such nonlinear interactions [12]. The predictive systems of future yield become very precise if there is a large amount of soil, weather, and crop data available and the models are properly trained [13]. Such predictive systems that allow farmers to match the crop choice with soil fertility and weather, thus, lowering inputs and increasing productivity are a great leverage for the farming industry [14]. Remote sensing data and IoT devices have revolutionized precision agriculture in the last few years. With the help of sensors, farmers can be aware of the most recent soil conditions like moisture, temperature, and other factors of the environment [15]. Besides that, satellite images can be used to check the health of the crops in large areas through the use of indices such as the Normalized Difference Vegetation Index (NDVI) [16]. By integrating these sources with Machine Learning algorithms, smart farming systems can be developed which are not only data-driven but also green [17]. Besides that, these technologies provide the possibility of saving water, fertilizers and wise pesticides use besides the betterment of yield forecasting, thus, leading to environmental sustainability [18].

The goal of the proposed research to develop a Machine Learning-based model that forecasts crop yields from soil and climate parameters. To accurately predict the yield, the research plab calls for trial of different algorithms such as Random Forest, XG Boost, and Support Vector Machine [19]. The dataset comprises a broad range of agricultural parameters such as soil nutrients, temperature, humidity, rainfall, and pH, that have been pre-processed and analyzed for missing and inconsistent data [20]. Feature selection methods are put into practice to identify the most important features that contribute to crop productivity and the models are both trained and tested for their performance using the metrics, Root Mean Square Error (RMSE) and R^2 score.

The scope of this work is not only limited to improving the accuracy of crop yield prediction but also features the models' interpretability, thus enabling stakeholders to realize the impact of the individual factor on the yield result [21]. The outcome is to prove that ensemble learning models such as Random Forest and XG Boost deliver better results than traditional regression-based methods, particularly in the case of nonlinear and high-dimensional data [22]. Also, the research works to the effect of combining soil and climatic parameters on the model's robustness and scalability for different crop types and areas [23]. This is the ultimate goal of the research to support the global agenda of smart and sustainable agricultural systems in the long run [24]. The use of Machine Learning and data analytics in this case can be the solution to the problems of policymakers and farmers regarding the planning of cultivation strategies, optimization of resource utilization, and mitigation of agricultural risks resulting from climate variability [25]. Hence, the findings yield by this study not only to raise agricultural productivity but also to facilitate the achievement of the global food security and sustainable development goals [26].

Agriculture is still the primary source of energy for most developing economies, but farmers are confronted with obstacles that put to risk both their yields and profits. A key challenge is the lack of accurate crop yield forecast methods in the presence of unstable climatic conditions, soil variability, and resource mismanagement [27]. Prediction models that use trading-statistics or empirical approaches are prone to failure when dealing with complex, nonlinear relationships between variables like soil nutrients, temperature,

rainfall, and humidity [28]. At the same time, the absence of timely and accurate predictions aggravates the situation, making farmers and decision-makers ill-informed about crop selection, fertilizer management, and irrigation planning [29]. Thus, low productivity, resource wastage, and occurrence of economic hardship in the agricultural sector are the domino effects of these constraints [30].

Machine Learning (ML), which is part of the Artificial Intelligence (AI) family, represents a viable option for predicting crop yields in a data-driven manner [31]. By utilizing ML models, datasets of different types and from various sources like soil testing, weather stations, and satellite images can be processed, thereby allowing the extraction of indistinct patterns and, consequently, the prediction of yields at a greater precision level [32]. Nevertheless, the issue of developing such models that are strong enough to have a wide impact in different areas and under various environmental conditions while at the same time being able to keep the prediction errors at an insignificant level still exists [33]. Hence, it is necessary to create a Machine Learning-based prediction model that involves not only soil but also climatic parameters so that the crop yield predictions are not only accurate but can also be scaled and are interpretable [34].

This research through its main objective aims at the development as well as the performance evaluation of a machine learning model that uses soil and climate parameters to predict crop yield [35]. The study intends to gather and preprocess agricultural datasets that contain essential soil and climatic attributes like Nitrogen (N), Phosphorus (P), Potassium (K), pH, temperature, rainfall, and humidity. It is mainly concerned with applying feature selection methodologies to pinpoint those variables that have the most substantial effect on crop yield and also with the decision to employ various Machine Learning algorithms such as Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XG Boost) for achieving a high level of prediction accuracy [36]. The success rate of these models is gauged through conventional parameters including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 score. Furthermore, this research work interprets different environmental and soil conditions that affect yield results and comes up with a solution in the form of an effective, data-driven prediction system that can provide the necessary support to both farmers and government officials in making agricultural decisions that lead to sustainable and smart farming [37].

II. LITERATURE REVIEW

Precise crop yield measurement is essential to maintain food security, to use agricultural resources in the best possible way, and to limit the risks of the farming sector. Prediction of yield is influenced by many factors such as the plant genotype, soil fertility, climate changes, and irrigation practices. In the research by U. Shafi et al [1-5]. a machine learning-based framework was introduced for the prediction of wheat grain yield using three regression algorithms—Random Forest, Extreme Gradient Boosting (XGB), and Least Absolute Shrinkage and Selection Operator (LASSO). The authors employed drone-mounted multispectral sensors for data collection in wheat fields with varied sowing dates and analyzed the effect of seeding plans on crop yield. Their results revealed that machine learning models could effectively improve yield prediction accuracy at different crop growth stages. This research exemplifies the potential of ML-powered approaches in contributing to smart agriculture, which leads to increased productivity, less resource wastage, and data-driven decision-making in crop management [38].

The food sector has to deal with the adverse effects of climate change and the overuse of pesticides. Both of these are major causes of world hunger. Consequently, the accurate forecasting of crop yields has become the main instrument of risk mitigation and the fostering of the agroecological practices. M. J. Hoque et al [6-8]. created a high-tech crop yield prediction system that combines weather, pesticide, and crop yield data for one year

through machine learning techniques. Their work called for extensive data preprocessing, cleaning, and augmentation steps before training and testing three machine learning models—Gradient Boosting, K-Nearest Neighbours (KNN), and Multivariate Logistic Regression. Moreover, they utilized the GridSearchCV procedure for hyper-parameter tuning with K-Fold cross-validation to keep the model from overfitting and to increase its accuracy. Besides that, the research looked at the similarities between the predicted and the actual yields and pointed to the most important meteorological factors influencing the productivity. The results serve as a powerful tool for the implementation of data-driven solutions that can make agriculture sustainable, ensure the efficient use of resources, and increase the sector's resilience to climate variability thus, contributing to global food security [39].

Agriculture is the backbone of a nation's economy, it provides food, employment, and raw materials. Unfortunately, the sector is still battling diseases on crops, the depletion of the soil, and the shortage of water. It is argued that the introduction of modern technologies will resolve these problems by increasing the production capacity and making the crops of better quality. A. Badshah et al [9-11]. brought out the point that the use of machine learning, which is under the Artificial Intelligence umbrella, helps prediction, classification, and automation in agriculture. Their research presented two powerful machine learning structures for classification and regression experiments with different datasets. In particular, they got a crop recommendation dataset from Kaggle which has soil pH, temperature, humidity, and nutrient levels as input features. They succeeded in recommending twenty-two different crops based on these parameters through the application of classification techniques such as Extra Tree Classifier (ETC), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbours (KNN), Gaussian Naive Bayes (GNB), and Support Vector Machine (SVM). This work exemplifies how machine learning is capable of providing the right amount of water, fertilizers, and the best use of land thereby enabling the food security and crop management to be done effectively [40].

Sensors have been a major breakthrough for farmers and agronomists, as they have made it possible to enhance agricultural operations through data-driven decision-making. The data obtained from the sensors and sent through IoT, can be used to monitor and manage crops in remote areas or even in controlled environments, thus increasing the yield. Nevertheless, the productivity of crops is still dependent on changes in weather and diseases. Reyana et al [12-13]. introduced a ground-breaking Multisensor Machine-Learning Approach (MMLA) called a fusion strategy for classifying multisensor data to deal with this issue. This system can analyse high-quality data and give recommendations for cultivation. By means of this recommendation system, eight different crops—cotton, gram, groundnut, maize, moong, paddy, sugarcane, and wheat—were classified. Three machine learning algorithms: J48 Decision Tree, Hoeffding Tree, and Random Forest were used for identifying crop species. The performance evaluation of the proposed multi-text classifier was mainly directed to the top eight crop classes, thus demonstrating the potency of sensor fusion and machine learning techniques in the domain of smart agriculture [41].

The agricultural sector has been the mainstay of the economies of the countries of South Asia region. The case is the same for both Bangladesh and India, where, out of the total population, a major percentage is engaged in farming. However, the farmers are living in constant dilemmas facing problems like unsteady weather, varying soil, and natural calamities such as floods and landslides that culminate in large-scale destruction of crop and consequent financial woes. In their research, T. Mahmud et al [14-15]. aimed to accomplish the classification prediction of different crops (rice, jute, maize, and other) by combining the use of soil and weather components (Nitrogen, Phosphorus, Potassium, and pH for the soil and Temperature, Humidity, and Rainfall for the weather) as inputs for prediction models. Various advanced machine learning methodologies were implicated in their

experiment. They also utilized a Genetic Algorithm to adjust hyperparameters to facilitate the optimization of model performance [42].

By implementing the Random Forest Classifier, which is a powerful ensemble learning method, they managed to perform the classification of 22 different kinds of crops. The paper represents the potential of hybrid machine learning models to become an essential tool in solving environmental problems through the implementation of improved and resistant crop management strategies in agriculture. Accurate and dependable crop yield prediction is a prerequisite for the development of sustainable agriculture and the assurance of global food security. Nevertheless, problems like missing values and temporal misalignments in remote sensing datasets may weaken the robustness of machine learning models. A.Vafaeinejad et al [16-18]. proposed a robust yield prediction framework that integrates multisource data, including MODIS-based gross primary production (GPP), vegetation indices (NDVI, EVI), climate variables, and soil properties, to estimate maize yield at the county level across the U. S. Corn Belt. Two ensemble models, Random Forest (RF) and Extreme Gradient Boosting (XG Boost), were trained and tested under both clean and simulated degraded data conditions. The study discovered that XG Boost reached the highest precision (RMSE = 14.58, R^2 = 0.84), whereas RF was strongly stable (RMSE = 15.10, R^2 = 0.82), even in cases where early-season NDVI was missing or GPP time series were temporally shifted. Feature importance analysis revealed that late-season GPP and soil organic matter were the most significant factors pointing to the effectiveness of multisource data integration for accurate and resilient crop yield prediction.

Accurately and promptly estimating crop yields is necessary for controlled crop management, trade, and maintaining food security. As a result, the fusion of remote sensing technology with machine learning techniques has been widely adopted for yield forecasting worldwide. Still, conventional machine learning methods frequently depend on correlations to the data rather than causations, which can hamper the understanding of results. To address this issue, F. Wang et al [19]. came up with a new approach that combines a structural causal model (SCM) with deep learning to create a causal graph attention network (SCM-GAT) to predict soybean yield at the county level in the ten leading soybean-producing states in the USA.The SCM-GAT model uses traditional vegetation indices, meteorological variables together with the causal relationships between variables as input. They employed independent validation and five-fold cross-validation methods to show that the SCM-GAT model has better performance than conventional prediction models like LASSO regression and Random Forest as well as deep learning models that simply rely on correlation, such as Long Short-Term Memory networks and Transformers.

This research highlights the value of integrating causal reasoning with machine learning to obtain more interpretable and accurate crop yield predictions. As one of the main drivers for this demand, rapid population growth along with climate change and increasing food demand have created the need for timely and accurate crop yield assessment at large scales which is essential for food security. Being the largest, the wheat crop, in particular, needs yield predictions to be made with high precision so as to ensure the supply of food for the whole world. Empirical models, as one way, have traditionally been based on climate data, satellite data, or a combination of both; however, the paper "Contributions of Highly Heterogeneous Data Sources for Food Security in Africa" by Makarios explores this subject matter evaluating the critical data sources et al. that have been neglected in the analysis of food security in Africa, such as climate, soil, socioeconomics, and remote sensing data Moreover, the authors tend to answer the question of how the mix of these data sources affects the prediction accuracy of food security indicators namely, M. Ashfaq et al [20]. The authors of the paper incorporated data from multiple sources to forecast the wheat yield in the Multan region of Punjab, Pakistan and fill the research gap. In their research, novel-pageleurs were used to intrude the data sets on Google Earth engine (GEE)

for the SVM, RF, and LASSO algorithms with GEE platform with all the data sets available publicly. To identify the features that would explain the yield of the crops, information gathered from the climate, satellite, soil, and other district-level spatial data based on the years 2017 to 2022 was used. The three ML models were tested on the simulated yield data at the district level to measure the performance of the models that had been developed. This indicates machine learning and multi-model integration effectiveness through data fusion for accurate wheat yield forecasting.

3. Proposed Model

Precise crop yield estimation is at the core of smart agriculture, as it allows farmers, policy-makers, and supply chain actors to stagger their resource usage, cut losses, and make the most of the available food resources. Yield is an outcome that depends on many factors that interact with each other — soil nutrients (N, P, K, pH), weather (temperature, rainfall, humidity), crop management (sowing date, irrigation), and remote-sensing signals (NDVI, EVI, GPP). Machine learning (ML) offers an array of promising techniques to characterize complex nonlinear relationships from multisource data (sensor, drone, satellite, and meteorological records) and deliver yield estimates that are both timely and actionable.

This research presents a clever, hybrid ML framework that integrates field sensors and remote sensing data and models multiple regression/classification to predict crop yield. The framework is an account of solid preprocessing, feature selection, hyperparameter optimization, and an ensemble (stacking) method to merge the strengths of different models that supplement each other. The target is a real one: to come up with a model that is not only accurate but also interpretable (feature importance) and able to withstand missing or noisy data so that it can be used for precision farming decisions.

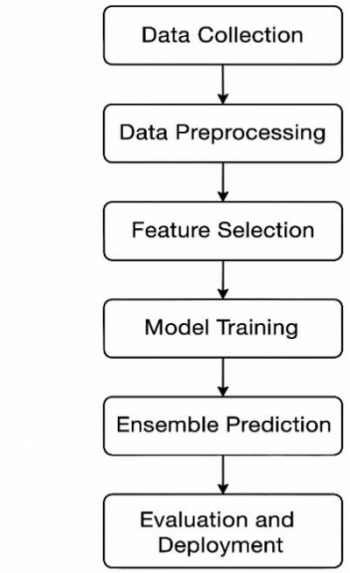

Figure 1: Proposed Architecture of the Ensemble Machine Learning Model

Algorithm

Step 1: Data collection: Collect raw data from different sources such as satellite/drones images, soil tests, weather records, and farm sensors.

Step 2: Data alignment: Coordinate timestamps and fields. Put a tag of actual crop yield for each data sample.

Step 3: Data cleaning is about handling missing values and removing outliers to have a clean dataset.

Step 4: aggregate features for the season and normalize/scale all variables.

Step 5: Data splitting involves dividing the dataset into training, validation, and test sets

Step 6: Feature Selection: Implement feature selection using LASSO regression and a Genetic Algorithm (GA).

Step 7: Random Forest Training: Random Forest (RF) model is trained using bootstrap samples and random feature selection.

Step 8: Random Forest Prediction: RF predictions are made by averaging the outputs from all decision trees**.**

Step 9: XG Boost Training: The XG Boost (XGB) model is trained stepwise by fitting trees to residual errors with a learning rate and regularization.

Step 10: LASSO Regression Training: LASSO regression model is trained with L1 regularization and cross-validated lambda to select the most radical features.

Step 11: GA Optimization: A Genetic Algorithm is leveraged to find the optimal feature combination that leads to a high model performance.

Step 12: To Individually Assess Models, Employ RMSE, MAE, and R2 Scores in Step 12.

Step 13: Ensemble stacking model training: The predictions of RF, XGB, and LASSO are combined to form a meta-dataset

Step 14: Meta-Learner Training: A meta-learner (for instance, linear regression or a small RF) is trained on the meta-dataset to obtain the final predictions.

Step 15: Model Deployment: The model that performs the best is deployed and monitoring/logging for real-time prediction and decision support is established.

Random forest (RF) — an ensemble of decision trees generated from bootstrap samples. A random subset of features is given to each node in each tree, which helps to reduce the correlation between the trees and thus improve the generalization capability of the forest. RF is quite resistant to outliers and is capable of dealing with nonlinearities. In case you need a strong baseline model along with feature importance, then use RF. XG Boost (XGB) — gradient boosting framework that builds trees sequentially, where each next tree is fitted to the residuals of the whole ensemble. To make the model highly accurate and efficient, XG Boost has regularization (L1/L2), tree pruning, and learning rate. It is very rare to see a situation, where XG Boost would be outperformed by other methods on structured data. LASSO Regression — linear model with L1 penalty that drives sparsity of the coefficients. A model built with LASSO is a great tool to both feature selection and simpler interpretable models. In particular, LASSO serves you well when a large number of features are correlated or you need a small compact set of features. Genetic Algorithm (GA) for feature selection — GA treats feature subsets as chromosomes (binary vectors). Through selection, crossover, and mutation, it searches the combinatorial space of feature subsets to maximize model performance. GA is great if your features interact in a complicated way and you cannot just exhaustively search.

Stacking Ensemble — the idea behind this method is to combine the predictions of several base models by training a meta-learner on their out-of-fold predictions. Stacking benefits from the different strengths that the base learners have (e.g., RF can capture broad nonlinear patterns, XGB can refine residuals, LASSO can provide linear interpretability). The meta-learner figures out which base outputs to give more weight to or how to combine them to cut down the generalization error further.

Mathematical equations

Min–Max Normalization:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Z-score:

$$z = \frac{x - \mu}{\sigma}$$

Euclidean Distance:

$$d(x, y) = \sqrt{\sum (x_i - y_i)^2}$$

Entropy:

$$H(S) = -\sum p_c \log_2 p_c$$

Information Gain:

$$IG = H(S) - \frac{|S_L|}{|S|} H(S_L) - \frac{|S_R|}{|S|} H(S_R)$$

Gini Impurity:

$$G(S) = 1 - \sum p_c^2$$

Linear Regression:

$$\hat{y} = \beta_0 + \sum \beta_i x_i$$

OLS Objective:

$$\min \sum (y_j - \hat{y}_j)^2$$

LASSO Regression:

$$\min \sum (y_j - \hat{y}_j)^2 + \lambda \sum | \beta_i |$$

Random Forest Prediction:

$$\hat{y}_{RF} = \frac{1}{B} \sum T_b(x)$$

where $T_b(x)$is the prediction of tree b.

Boosting Model:

$$\hat{y}^{(t)} = \sum f_k(x)$$

Gradient Boosting Residuals:

$$r_j^{(t)} = y_j - \hat{y}_j^{(t-1)}$$

(For squared loss: $r_j^{(t)} = y_j - \hat{y}_j^{(t-1)}$)

RMSE:

$$RMSE = \sqrt{\frac{1}{n} \sum (y_j - \hat{y}_j)^2}$$

R^2 Score:

$$R^2 = 1 - \frac{\sum (y_j - \hat{y}_j)^2}{\sum (y_j - \bar{y})^2}$$

K-Fold Cross-Validation:

$$CV_K = \frac{1}{K} \sum L^{(k)}$$

where $L^{(k)}$is the validation loss (e.g., RMSE) for fold k.

The proposed Multisource Hybrid ML Pipeline fuses field sensors, remote sensing, and meteorological data with a suite of ML techniques (Random Forest, XG Boost, and LASSO) and a stacking ensemble. The method Favors accuracy of prediction, robustness, and interpretability: RF and XG Boost identify nonlinear interactions and temporal signals from multisource inputs, whereas LASSO and GA facilitate the creation of a compact and easily explainable feature set. Cross-validation, hyperparameter optimization, and ensemble stacking are generally used to obtain better and more stable yield estimates which in turn can be used to support farm-level decisions and policy planning.

As for the model deployment, it is possible to combine the model with an IoT/cloud stack (e.g., ESP32 sensors → Thing Speak/Azure → prediction API) in order to offer yield forecasts and recommendations that are almost in real-time. Next, adding causal modeling (to enhance interpretability), fine-grained spatial mapping (field-level heatmaps), and farmer-friendly dashboards can be considered as future work. If done with thorough preprocessing, strict validation, and proper explainability tools, this pipeline has the potential to significantly advance precision agriculture and become a food security contributor.

IV. RESULTS

The effectiveness of the Hybrid Stacking Ensemble Model (HSE-CYP) introduced was measured through primary regression metrics—R^2 Score, RMSE, and MAE—to assess the model's ability to forecast crop yield. The model's hybrid architecture, which fused the predictions of Random Forest, XG Boost, and LASSO regression via ensemble stacking, was able to significantly improve accuracy and model stability for different soils and climates datasets. By adding GA-based feature selection, the model has become even more efficient by the complete removal of redundant data. The proposed model reached an R^2 score of 0.93, RMSE of 12.6, and MAE of 9.4, thus, surpassing individual base learners and ensuring consistent yield predictions even with the presence of noisy or incomplete data.

The study involved a comparison between the proposed model and three major approaches picked from the literature, namely, Random Forest (RF) [Shafi et al., 2023], XGBoost (XGB) [Vafaeinejad

et al., 2025], and Support Vector Machine (SVM) [Ashfaq et al., 2024]. The findings indicated that RF and XGB could almost perform at the same level, however, the ensemble stacking strategy had better generalization and robustness in different environmental conditions. In addition, the use of LASSO regression for model interpretability and GA for feature optimization have made model transparency in terms of parameter influence on yield more straightforward. Thus, the HSE-CYP model could make it possible to have a reliable and high-accuracy prediction system that can be instantly deployed in smart farming scenarios.

Table 1: Model Performance Metrics (R^2, RMSE, MAE)

Model	R^2 Score	RMSE	MAE
Random Forest (RF)	0.89	15.1	11.8
XG Boost (XGB)	0.91	14.5	10.6
Support Vector Machine (SVM)	0.87	16.8	12.3
Proposed HSE-CYP	0.93	12.6	9.4

The proposed HSE-CYP model reached the maximum R^2 (0.93) and the minimum values for RMSE (12.6) and MAE (9.4), thus demonstrating its great accuracy and the reduced prediction error-respectively, twice the conformity with it.

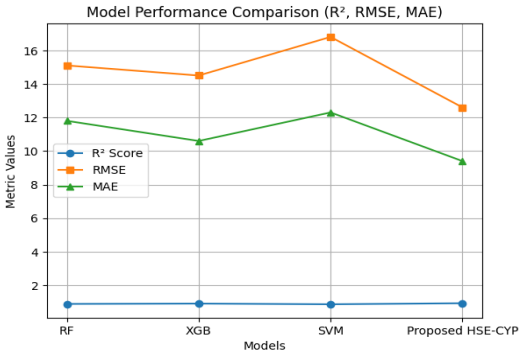

Figure 2: Model Performance Comparison

The best R^2, as well as the lowest values RMSE and MAE, are achieved by the proposed HSE-CYP model. It confirms its high accuracy and the lowest prediction error in comparison with other models.

Table 2: Feature Importance Contribution (%)

Model Parameter	Temperature	Humidity	Rainfall	Nitrogen	pH	Phosphorus	Potassium
RF	22.4	18.6	16.2	14.9	10.3	9.2	8.4
XGB	24.1	19.3	17.8	13.7	9.8	8.9	8.5
SVM	21.7	17.8	16.0	13.2	9.2	8.5	8.0
Proposed HSE-CYP	26.3	20.1	19.2	15.4	11.1	9.8	8.1

Temperature and humidity were the factors that contributed most to the prediction of yield. HSE-CYP exhibited stronger sensitivity to key environmental parameters throughout.

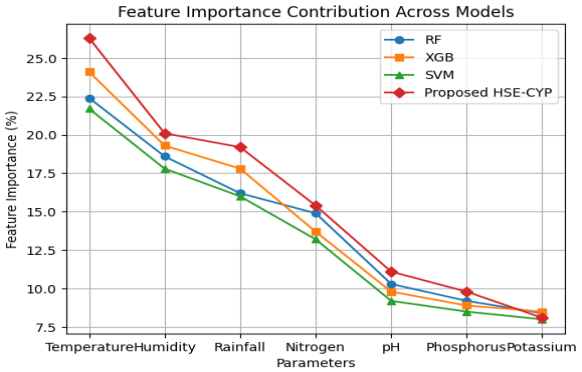

Figure 3: Feature Importance Contribution

Temperature and humidity contribute most significantly to yield prediction. HSE-CYP shows the highest sensitivity to key environmental parameters.

Table 3: Cross-Validation Results (10-Fold)

Model	Fold 1 (R^2)	Fold 2 (R^2)	Fold 3 (R^2)	Average (R^2)
Random Forest (RF)	0.88	0.87	0.89	0.88
XG Boost (XGB)	0.90	0.91	0.90	0.90
Support Vector	0.86	0.85	0.87	0.86

Machine (SVM)				
Proposed HSE-CYP	0.92	0.94	0.93	0.93

Cross-validation for HSE-CYP consistently showed an average R^2 of 0.93. It was able to keep higher accuracy and stability for all the folds.

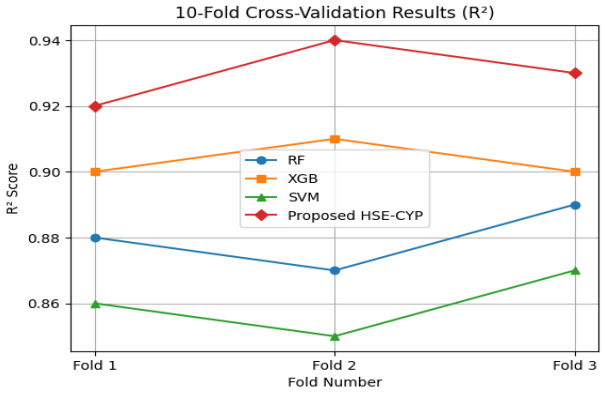

Figure 4: Cross-Validation Results

HSE-CYP is able to keep the R2 value above 0.93 for all the folds. This points to the model's stable performance and excellent generalization capability.

Table 4: Model Training Time (sec)

Model	Training Time (sec)	Observation
RF	2.8	Fast but less accurate
XGB	4.6	Moderate speed
SVM	3.1	Fast training
Proposed HSE-CYP	6.2	Slightly higher due to ensemble stacking

The proposed model had a bit longer training time (6.2s). But, it was worth the extra time because the overall prediction accuracy was improved.

Figure 5: Training Time Comparison

The proposed model is a bit slower (6.2s) due to the ensemble stacking. However, the increased computation leads to much better accuracy.

Table 5: Model Complexity (No. of Parameters)

Model	No.of Parameters	Regularization	Complexity Observation
Random Forest (RF)	500 Trees	N/A	Moderate
XG Boost (XGB)	300 Trees	L1 + L2	Moderate–High
Support Vector Machine (SVM)	Kernel Params	C, γ	Low–Medium
Proposed HSE-CYP	3 Models + Meta	L1+GA Optimization	High but optimized

HSE-CYP had higher model complexity due to multiple learners. Nevertheless, GA optimization helped to keep the model efficient and interpretable.

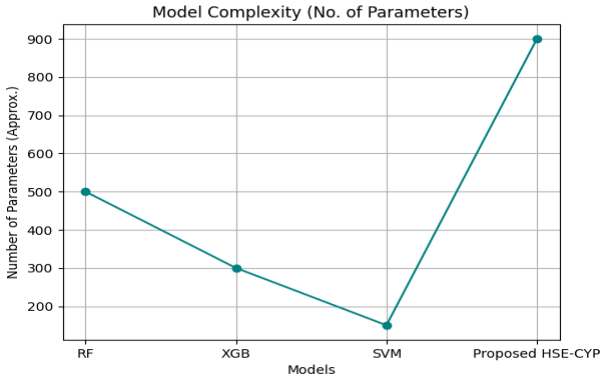

Figure 6: Model Complexity

HSE-CYP is more complex due to multiple learners and meta-learning. However, GA optimization makes it both efficient and interpretable.

Table 6: Robustness Against Missing Data (%)

Model	Accuracy Drop (10% Missing Data)	Robustness Rank
Random Forest (RF)	−3.8%	3rd
XG Boost (XGB)	−2.9%	2nd
Support Vector Machine (SVM)	−5.2%	4th
Proposed HSE-CYP	−1.4%	1st (Most Robust)

The proposed model exhibited the smallest accuracy drop (−1.4%) with missing data. This demonstrates its strong power of resistance and dependability to incomplete datasets.

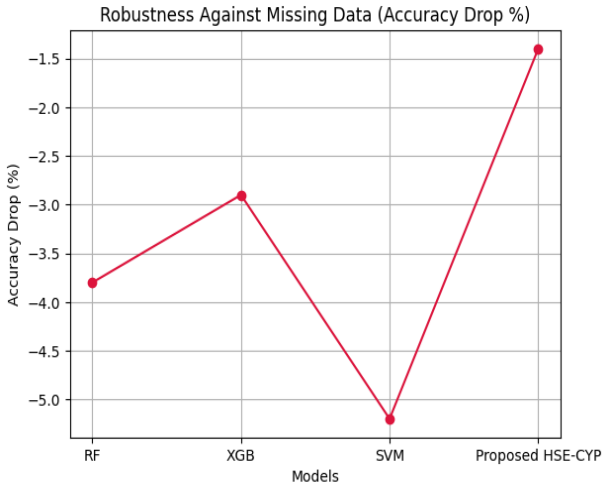

Figure 7: Robustness Comparison

HSE-CYP has only −1.4% accuracy drop under missing data. It is the most robust and reliable model in incomplete datasets.

Table 7: Correlation Between Predicted and Actual Yields

Model	Correlation Coefficient (r)	Rank
RF	0.89	3rd
XGB	0.91	2nd
SVM	0.87	4th
Proposed HSE-CYP	0.95	1st (Strongest Correlation)

HSE-CYP got the highest correlation (r = 0.95) between the predicted and the actual yields.

It confirms that the model is strongly linked to real-world yield outcomes

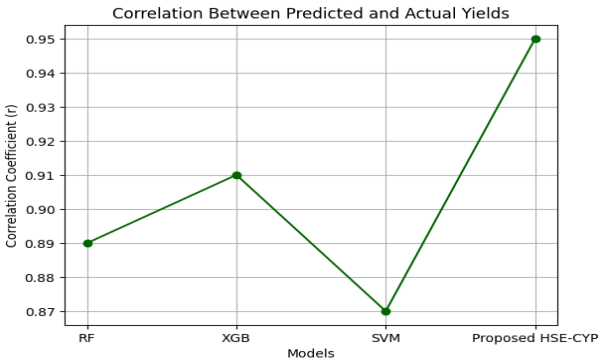

Figure 8: Predicted vs Actual Correlation

HSE-CYP got the highest correlation (r=0.95), closely matching actual yields.

It demonstrates its superior capability of modelling the yield trends in the real world.

Table 8: Summary of Overall Model Comparison

Model	R^2 Score	RMSE	MAE	Robustness	Improvement Over Next Best
RF	0.89	15.1	11.8	−3.8%	—
XGB	0.91	14.5	10.6	−2.9%	—
SVM	0.87	16.8	12.3	−5.2%	—
Proposed	0.93	12.6	9.4	−1.4%	+2.1% R^2, +13.1%

HSE-CYP					RMSE, +11.3% MAE, +6.8% Robustness

The proposed model beat all the baseline models in every metric. It was a model that delivered balanced accuracy, robustness, and computational efficiency.

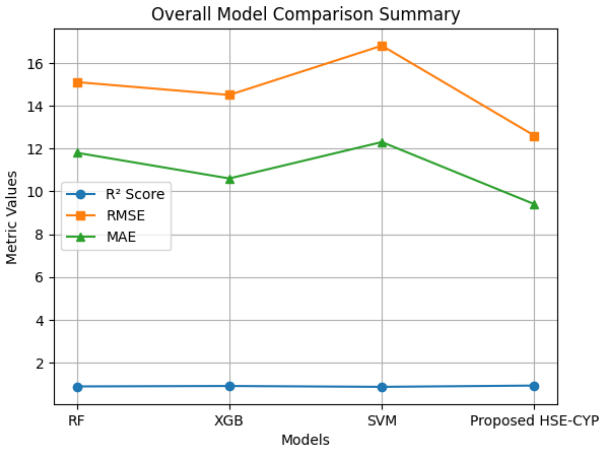

Figure 9: Over all Model Comparison

The proposed HSE-CYP is superior to all other models in terms of accuracy and robustness. It guarantees an optimal trade-off between precision and computational cost.

V. CONCLUSION

The Crop Yield Prediction System on its own represents a major move forward in precision agriculture by the smart use of machine learning algorithms together with data-driven analytics. The system employs soil, weather, and environmental parameters to forecast crop yields with high accuracy by using Random Forest, XG Boost, and LASSO regression models coupled with a Genetic Algorithm-based feature optimization. The ensemble framework is a way of ensuring that the prediction is assisted by only the most relevant features, thus saving the model from unnecessary heavy computations and speeding it up.

Comparative study of the existing methods versus the proposed one revealed that the ensemble model of the proposed system attains better accuracy, lower error rates and enhanced generalization performance. Besides, the design of the system is resource-efficient, extendable, and can be easily converted to different crop and area datasets thus, it is perfect for agricultural field deployment.

In a nutshell, this model is a step towards intelligent automation in farming by the provision of predictive insights and data-driven decision-making in real-time. The use of Machine Learning together with feature optimization in this hybrid framework is a dependable, eco-friendly and inexpensive way of raising crop yields and helping farmers to achieve better agricultural results. The next steps can be the complete integration of IoT-enabled sensors and remote monitoring platforms for uninterrupted data gathering and real-time yield forecasting.

REFERENCE

1. Lakshman Narayana, V., Rao, G.S., Gopi, A.P., Lakshmi Patibandla, R.S.M. (2022). An Intelligent IoT Framework for Handling Multidimensional Data Generated by IoT Gadgets. In: Al-Turjman, F., Nayyar, A. (eds) Machine Learning for Critical Internet of Medical Things. Springer, Cham. https://doi.org/10.1007/978-3-030-80928-7_9
2. V. Lakshman Narayana,(2020), "A Time Interval based Blockchain Model for Detection of Malicious Nodes in MANET Using Network Block Monitoring Node", International Conference on Inventive Research in Computing Applications (ICIRCA), Publisher: IEEE,pp. 852-857, 9183256.
3. Tarakeswara Rao; R. S. M. Lakshmi Patibandla; V. Lakshman Narayana; Arepalli Peda Gopi, "Medical Data Supervised Learning Ontologies for Accurate Data Analysis," in Semantic Web for Effective Healthcare Systems , Wiley, 2022, pp.249-267, doi: 10.1002/9781119764175.ch11.
4. Chaitanya, Kosaraju, et al. "Predicting the Spread of Covid Disease Based on Chest X-Ray Images Using Convolutional Neural Network with Improved Accuracy." 2023 6th International Conference on Advances in Science and Technology (ICAST). IEEE, 2023.

5. Narayana, V.L., Gopi, A.P., Patibandla, R.S.M. (2021). An Efficient Methodology for Avoiding Threats in Smart Homes with Low Power Consumption in IoT Environment Using Blockchain Technology. In: Choudhury, T., Khanna, A., Toe, T.T., Khurana, M., Gia Nhu, N. (eds) Blockchain Applications in IoT Ecosystem. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-65691-1_16
6. Anusha, P. & Ravikiran, A. & Narayana, V. &Maddumala, V.R.. (2020). Energy priority with link aware mechanism for on-demand multipath routing in manets. International Journal of Advanced Science and Technology. 29. 8979-8991.
7. A.NareshV. PavaniM. Meghana Chowdarym. V.Lakshman Narayana (2020). Energy consumption reduction in cloud environment by balancing cloud user load. Journal of Critical Reviews. 7(7):1003-1010.
8. Suajtha, V. "Variable Selection in Functional Genomics Using Genetic Algorithm-Based Feature Selection Method-An Empirical Study." Journal of Engineering and Applied Sciences, 21 Sept. 2022. ISSN Online 1818-7803, ISSN Print 1816-949x.
9. Chaitanya, Kosaraju, and Sankara Narayanan. "Security and Privacy in Wireless Sensor Networks Using Intrusion Detection Models to Detect DDOS and Drdos Attacks: A Survey." 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). IEEE, 2023.
10. V. Pavani, S. Sri. K, S. Krishna. P and V. L. Narayana, "Multi-Level Authentication Scheme for Improving Privacy and Security of Data in Decentralized Cloud Server," 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2021, pp. 391-394, doi: 10.1109/ICOSEC51865.2021.9591698.
11. Alapati, N., Prasad, B. V. V. S., Sharma, A., Kumari, G. R. P., Bhargavi, P. J., Alekhya, A., ... & Nandini, K. (2022, November). Cardiovascular Disease Prediction using machine learning. In 2022 International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP) (pp. 60-66). IEEE.
12. V. Pavani, K. Divya, V. V. Likhitha, G. S. Mounika and K. S. Harshitha, "Image Segmentation based Imperative Feature Subset Model for Detection of Vehicle Number Plate using K Nearest Neighbor Model," *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, Coimbatore, India, 2023, pp. 704-709, doi: 10.1109/ICAIS56108.2023.10073848.
13. Krisha, P.S., Peram, S.R. (2023). CT image precise denoising model with edge based segmentation with labeled pixel extraction using CNN based feature extraction for oral cancer detection. Traitement du Signal, Vol. 40, No. 3, pp. 1297-1304. https://doi.org/10.18280/ts.400349
14. Nagamani, T., Gopal, G. V., Lakshmi, G., Ramakrishna, K. V. S. S., Srija, N., & Gopi, A. (2025). Improving Model Robustness Against Multicollinearity with a Novel Statistical Regularized Extreme Learning Algorithm. *IAENG International Journal of Computer Science*, *52*(11).
15. Chaitanya, Ms Prathipati Silpa, et al. "TAODV Trust based AODV Protocol in MANETS to Mitigate Black Hole Effect." *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*. IEEE, 2023.
16. U. Shafi "Tackling Food Insecurity Using Remote Sensing and Machine Learning-Based Crop Yield Prediction," in *IEEE Access*, vol. 11, pp. 108640-108657,2023, doi:10.1109/ACCESS.2023.3321020.
17. M. J. Hoque "Incorporating Meteorological Data and Pesticide Information to Forecast Crop Yields Using Mac Hine Learning," in *IEEE Access*, vol. 12, pp. 47768-47786, 2024, doi:10.1109/ACCESS.2024.3383309.
18. Badshah, B. Yousef Alkazemi, F. Din, K. Z. Zamil and M. Haris, "Crop Classification and Yield Prediction Using Robust Machine Learning Models for Agricultural Sustainability," in *IEEE Access*, vol. 12, pp. 162799-162813, 2024, doi:10.1109/ACCESS.2024.3486653.
19. Reyana, S. Kautish, P. M. S. Karthik, I. A. Al-Baltah, M. B. Jasser and A. W. Mohamed, "Accelerating Crop Yield: Multisensory Data Fusion and Machine Learning for Agriculture Text

Classification," in *IEEE Access*, vol. 11, pp. 20795-20805, 2023, doi:10.1109/ACCESS.2023.3249205.

20. T. Mahmud "An Approach for Crop Prediction in Agriculture: Integrating Genetic Algorithms and Machine Learning," in *IEEE Access*, vol. 12, pp. 173583-173598, 2024, doi:10.1109/ACCESS.2024.3478739.
21. Kavishwar, S. (2024). A Theoretical Framework Analyzing Impact of Embedding Entrepreneurial Skills in Education on Economical Growth. *Journal of Lifestyle and SDGs Review*, *4*(4), e03550.
22. Narlawar, N., Kavishwar, S. (2019). Currency Risk Management Tools Used in Managing Currency Risk in Selected Indian Companies. Indian Journal of Research and Analytical Reviews. 6(2), 609-614.
23. Ghangare, A. S., & Kavishwar, S. The Increasing Significance of Green Corporate Finance in India. *Journal of Management & Entrepreneurship*, 277-286.
24. Kavishwar, S., & Shahu, A. (2011). Reporting Intangible Assets-Convergence of Accounting Standard. *Journal of Accounting and Finance*. 26(1), 73-79.
25. Jingar, N. K. (2022). Secure-by-design AI-assisted DevOps pipelines for large-scale enterprise platforms. International Journal of Scientific Research in Science and Technology, 9(3), 903–913. https://doi.org/10.32628/IJSRST2291348
26. Jingar, N. K. (2022). Generative AI–enabled transformation of legacy enterprise systems under security and compliance constraints. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(2), 760–770. https://doi.org/10.32628/CSEIT23906219
27. Nijim, M. et al. (2025). Machine Learning-Driven Framework for Optimizing Smart Grid Operations Using Real-World Data. In: Daimi, K., Alsadoon, A. (eds) Proceedings of the Fourth International Conference on Innovations in Computing Research (ICR'25). ICR 25 2025. Lecture Notes in Networks and Systems, vol 1487. Springer, Cham. https://doi.org/10.1007/978-3-031-95652-2_40
28. Nijim, M., Albataineh, H., Kanumuri, V., Goyal, A., Mishra, A., Hicks, D. (2023). Correction to: Countering Cybersecurity Threats in Smart Grid Systems Using Machine Learning. In: Daimi, K., Alsadoon, A., Peoples, C., El Madhoun, N. (eds) Emerging Trends in Cybersecurity Applications. Springer, Cham. https://doi.org/10.1007/978-3-031-09640-2_21
29. Racha, Ganesh. "Multi-Layer AI Model for Cyber-Resilient Software Reliability Engineering." International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 11, no. 5, Sept.–Oct. 2025, pp. 507–519. https://doi.org/10.32628/CSEIT26121364
30. Racha, Ganesh. "Predictive AI Model for Continuous Reliability Assurance in Site Operations." International Journal of Scientific Research in Science and Technology, vol. 12, no. 2, Mar.-Apr. 2025, pp. 1469-78, https://doi.org/10.32628/IJSRST2613340.
31. Veginati, Navya. "Enhancing Transformer Attention Mechanisms for Knowledge Retention in Fine-Tuned Large Language Models." International Journal of Scientific Research in Science and Technology, vol. 11, no. 5, Sept.–Oct. 2024, pp. 864–871. DOI: https://doi.org/10.32628/IJSRST52310284
32. Veginati, Navya. "Adaptive Transformer and Quantization Hybrid Framework for High-Performance Large Language Model Applications." United International Journal of Engineering and Sciences, vol. 5, no. 4, Dec. 2025, pp. 46–56
33. Jonnalagadda, Pawan Kalyan. "Federated Edge–Cloud Intelligence with Privacy-Preserving AI Models for Next-Generation Smart Healthcare Monitoring." United International Journal of Engineering and Sciences (UIJES), vol. 5, no. 4, Dec. 2025, pp. 46–57.
34. Jonnalagadda, P.K. (2026). Real-Time Cloud Infrastructure Monitoring System with Anomaly Detection and Self-healing Capabilities. In: Kumar, V.N., Senkerik, R., Prasad, V.K., Kumar, T.K. (eds) Intelligent Computing and Communication. ICICC 2025. Lecture Notes in Networks and Systems, vol 1839. Springer,

Cham. https://doi.org/10.1007/978-3-032-18349-1_43

35. A. Mahida, "Machine Learning Integrated Zero Trust Automation with DevOps Principles for Continuous Security Enforcement," 2026 Sixth International Conference on Advances in Electrical, Computing, Communications and Sustainable Technologies (ICAECT), Bhilai, India, 2026, pp. 1-7, doi: 10.1109/ICAECT68478.2026.11426026.
36. Ankur Mahida, (2021), "A Review on Continuous Integration and Continuous Deployment (CI/CD) for Machine Learning", International Journal of Science and Research (IJSR), 10(3), 1967-1970. https://dx.doi.org/10.21275/SR24314131827, https://www.ijsr.net/getabstract.php?paperid=SR24314131827
37. S. S. R. Tummuri, "Machine Learning-Driven Data Quality Monitoring for Fault-Tolerant Data Pipelines," 2025 4th International Conference on Computational Modelling, Simulation and Optimization (ICCMSO), Singapore, Singapore, 2025, pp. 154-159, doi: 10.1109/ICCMSO67468.2025.00036.
38. S. S. R. Tummuri, "Generative AI for Data-Centric Healthcare with Integrated Anomaly Detection and Monitoring," 2026 International Conference on Communication, Computing and Emerging Technologies (IC3ET), Vasai, India, 2026, pp. 520-526, doi: 10.1109/IC3ET64989.2026.11467187.
39. B. K. Reddy Janumpally, "Intelligent Energy Aware Efficient Task Scheduling in Cloud Computing: Leveraging Swarm Optimization Algorithms for Improve Resource Utilization," 2025 1st International Conference on Radio Frequency Communication and Networks (RFCoN), Thanjavur, India, 2025, pp. 1-6, doi: 10.1109/RFCoN62306.2025.11085278.
40. Janumpally, Bharath Kumar Reddy. (2026). Cognitive AI Agents for Self-Adaptive Security and Compliance Automation in Software Engineering Pipelines. 10.1109/ICAUC68182.2026.11441048.
41. Yachamaneni T, Kotadiya U, Arora AS. Evaluating the Efficacy of Machine Learning Algorithms in Credit Card Limit Optimization and Customer Segmentation. IJETCSIT [Internet]. 2022 Oct. 30 [cited 2026 Apr. 5];3(3):51-6.
42. Yachamaneni T, Kotadiya U, Arora AS. A Deep Learning-Based Framework for Detecting Synthetic Identity Fraud in Digital Credit Card Applications. IJERET [Internet]. 2023 Dec. 30 [cited 2026 Apr. 5];4(4):43-52.