International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Crop Yield Prediction using Machine Learning Algorithms

Sonika kavali Dr V.Y.S.S pragathi

[1] PG Scholar in Department of CSE STANLEYCollege Of Engineering & Technologyfor Women

[2]Associate Professor in Department of CSE STANLEYCollege Of Engineering & Technologyfor women

Abstract - Agriculture is a vital sector that contributes to the economy and food security of a nation. The agriculture sector is confronted with numerous challenges in the quest for accurate crop yield estimation, which is essential for efficient resource management and maintaining food security. Traditional crop yield estimation methods use historical data and statistical techniques to forecast future crop yields, focusing on environmental variables. This study aims to explain the crop yield prediction system as a way to address the challenges posed by global warming and climate change.

Keywords - Crop Yield Prediction, Machine Learning, Random Forest, Agriculture, Prediction Models

I. INTRODUCTION

Agriculture, practiced in India, is among the most crucial sectors for economic development. The industry thrives due to the massive portion of the population depending on farming for livelihood. The process involves a portrait of crop yields, and the strategic planning of agricultural activities reflects the connection between resource management and food security. The efficient and vital crop yield prediction contains various elements of data, such as climate, soil, and history, as well as the principles, for instance, accuracy and efficiency.

Traditionally, researchers used numerous elements of science in prediction, for instance, statistical techniques to represent historical trends, reliability, and consistency. The field contains several repeating patterns from historical datasets and the climate in the background. Modern systems use machine learning which created a smooth analysis and slightly accurate predictions; however, the traditional set contains slightly simpler statistical models. Numerous algorithms give the data shape outlining areas where climate shifts to soil properties and farming practices change to yield.

Moreover, the analysis of crop yield contains environmental factors such as rainfall, temperature, and humidity, while the ground data contains soil nutrients, for instance, Nitrogen, Phosphorus, and Potassium The models use realistic quantities of diverse parameters for shading. The data comprises layers of transparent factors that are slightly thick, contributing to the glowing appearance of Agricultural modeling has a structured framework easuring the model's accuracy.

various outputs as the researchers applied algorithmic principles to deliver an excellent piece. The prediction also contains a sense of depth since it is based on soil properties and climate. The results seem to consist of three variables.

II. LITERATURE SURVEY

The agricultural sector has developed crop yield prediction models through machine learning which researchers have studied since recent years because these models help farmers better their crop outputs and use resources more efficiently and make informed choices. Researchers have developed machine learning and deep learning methods to improve forecast accuracy with various datasets and research techniques.

Shin and Kaneko (2023) [1] conducted a comprehensive systematic literature review on crop yield prediction using machine learning techniques. The study evaluated previous research studies which showed that advanced models compete on prediction accuracy between 85% and 95% when using Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN) systems. The researchers studied theoretical concepts but they did not present any practical work to test their theories in real-time operations.

Parihar and Chimmwal (2020) [2] proposed a hybrid model combining Random Forest (RF) and Support Vector Machine (SVM) with satellite imagery and weather data. The method achieved an accuracy rate of about 92% which proved that remote sensing data works effectively when combined with machine learning methods. The system encountered multiple problems which included difficulties in accessing data and processing it and making the system work in various geographic locations.

Dissanayake et al. (2023) [3] presented a systematic literature review covering studies from 2016 to 2021. Their research showed that ensemble learning methods better prediction accuracy and model stability than single models. The study achieved good results yet the researchers did not test the system in real-world conditions and they did not validate their results with large datasets.

The researchers Khaki et al. (2020) [4] developed a hybrid deep learning system which uses Convolutional Neural Networks and Recurrent Neural Networks to analyze

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

agricultural data by studying its spatial and temporal characteristics.

The model achieved better prediction results because it included time-series data in its analysis. The method needed extensive computer power together with extensive datasets which made it impractical for use in real-time systems.

Nosratabadi et al. (2020) [5] developed hybrid machine learning systems which unified Artificial Neural Networks with optimization algorithms that included Grey Wolf Optimization (GWO). The results showed that their approach achieved better results than standard ANN models. The model required difficult parameter adjustment which resulted in high processing demands that created obstacles for its real-world application.

Shahhosseini et al. (2020) [6] combined machine learning techniques with crop simulation models such as APSIM to enhance prediction accuracy. Their hybrid approach achieved a 7-20% reduction in Root Mean Squared Error (RMSE) which demonstrated that their model performed better. The system depends on simulation data which restricts its use to areas that have access to such information.

Fan et al. (2021) [7] developed a new method which combines Graph Neural Networks with Recurrent Neural Networks to analyze spatial and temporal relationships present in agricultural datasets. The model achieved better prediction results because it used geographical relationship information. The solution has two problems because its system design needs high processing power and implementation becomes complex.

The research conducted by Abiraman and Senthilkumar (2024) [8] evaluated different machine learning models to determine their effectiveness in predicting crop yields. The research findings demonstrated that ensemble models which include Random Forest and deep learning models that use Artificial Neural Networks deliver better prediction results than standard regression techniques. The models failed to perform accurately when tested in different geographical areas and under various climate situations.

Shahhosseini et al. (2019) [9] created machine learning meta-models which included Random Forest and LASSO regression to forecast crop yields through simulated datasets. The results showed that Random Forest outperformed all other models in terms of performance. The research study failed to validate its findings through actual world data which diminishes its usefulness in real-world situations.

Reddy et al. (2021) [10] used machine learning methods, including Decision Tree and Linear Regression, and Neural Networks, to analyze Indian agricultural data. The study discovered that Decision Tree models achieved better accuracy than regression-based methods when estimating crop yields. The study results suffered from the research team's decision to examine only a small set of input data strength of their model.

The research shows that machine learning and deep learning methods improve crop yield prediction accuracy by identifying and understanding complex interactions between different elements. Ensemble methods and hybrid models show better results than regular methods because they face obstacles, which include the absence of real-time systems, the need for expensive computational resources, the insufficient supply of top-notch datasets, and the inability to perform well in different parts of the world. The existing problems demonstrate the requirement for a system which can expand, operate, and deliver precise outcomes.

III. METHODOLOGY

The methodological procedure for the suggested system consists of five key steps. The initial stage involves data collection based on the relevant agricultural data including water quantity, soil type, temperature, rainfall, and humidity.

Preprocessing takes place next when the collected data undergoes normalization, removal of irrelevant features, and handling of missing values.

The subsequent stage comprises feature selection to find out the most important parameters impacting the crop yield.

Once the preprocessed and filtered data is ready, the data division into training and testing parts occurs. The training set of data is necessary for building machine learning models. The testing part is needed for measuring the quality of each algorithm. The algorithms chosen for implementing the solution comprise Linear Regression, Decision Tree, and Random Forest as they are effective predictors.

The final stage implies assessing the quality of created models using metrics like R^2 Score and Mean Squared Error (MSE). Once the most efficient algorithm is selected, the results are presented as a final output.

Besides model evaluation, robustness is ensured through comparison of the performance of different algorithms using the same conditions in the datasets. The comparative study will help identify the best model that can produce the accurate prediction. Additionally, utilizing multiple factors like environmental and soil conditions increases the system reliability.

The proposed method can adapt to changes in the types of crops as well as geographical locations and therefore can be scaled accordingly. Future work may include incorporating machine learning/deep learning approaches for better predictions.

In conclusion, the methodology developed here can be considered to offer an effective and dependable platform for the forecasting of crop yields through the use of various data preprocessing, feature selection techniques, and machine learning algorithms. Through this process, the model is guaranteed both accuracy and reliability in its performance on different sets of data.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

IV. IMPLEMENTATION

The proposed crop yield prediction system is developed through Python implementation because of its user-friendly interface and extensive machine learning library support. The system uses Pandas and NumPy and Scikit-learn and Matplotlib libraries to execute data processing and model development and result visualization tasks.

The system begins with dataset loading to analyze its structural elements and characteristic features and evaluate its data distribution patterns. The process enables the detection of missing data points and data inconsistencies together with an assessment of overall data quality. Data preprocessing operations perform dataset cleansingafter the data exploration process. The process includes appropriate handling of missing data points through featureremoval and normalization to establish equivalent measurement standards.

The data gets separated into two parts after its preparation work is finished. The machine learningmodels get developed using training data while their performance assessment takes place through testing data. The Scikit-learn library implements three machine learning algorithms which include Linear Regression, Decision Tree, and Random Forest.

The model uses training data to understand how input features affect crop yield. The model generates predictions after it has completed its trainingusing the testing data. The evaluation process for each model uses Mean Squared Error (MSE) and R^2 score as assessment criteria.

The research results get better understood through the application of visual representation methods which include graphsand plots. The visualizations enable the evaluation of actual results against predicted results while they provide insights into the performance of various models.

The selected model which demonstrates superior performance will serve as the foundation for the final prediction system. The model generates crop yield predictions through analysis of new input data.

The team works on model development while they focus their efforts on enhancing system performance and user experience through optimization work. Random Forest model prediction accuracy improves through hyperparameter tuning which adjusts tree count and model depth settings.

Cross-validation techniques are used to validate the model's performance across multiple data subsets. The system uses a modular design which allows for future expansion through the addition of real-time data sources from IoT sensors and weather APIs. The implementation achieves both accuracyand flexibility which makes it suitable for real-world agricultural use cases.

V. RESULTS AND DISCUSSIONS

The proposed crop yield prediction system functions through testing various machine learning models which include Linear Regression and Decision Tree and Random Forest. The evaluation metrics, which include Mean Squared Error (MSE) and R^2 score, are used to compare the performance of each model that was trained and tested with the prepared dataset.

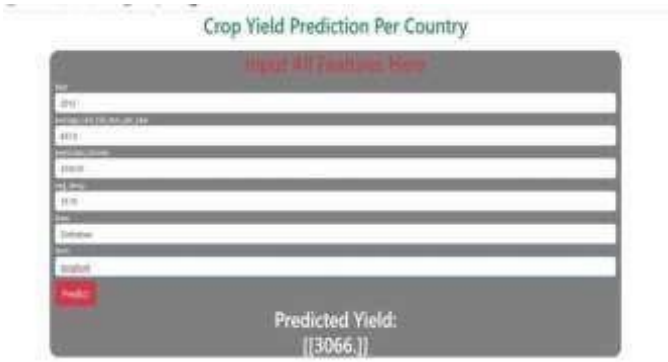

Fig 1: Input and Output Interface Showing Predicted Crop Yield

The results show that Linear Regression enables basic prediction capabilities because it cannot model complex non-linear relationships which exist between input variables. Decision Tree performs better than Linear Regression as it can capture non-linear patterns in the data; however, it may suffer from overfitting, which affects its performance on unseen data. Random Forest model demonstrates superior performance when compared to all other models.

The ensemble method utilizes multiple decision trees to create predictions which show both improved accuracy and better consistency. The system achieves its goals of reducing overfitting while efficiently processing extensive datasets.

Random Forest outperforms all other models by achieving a higher R^2 score and lower MSE which makes it the optimal model for predicting crop yields. The comparison of these models shows how important it is to choose suitable algorithms when analyzing agricultural data.

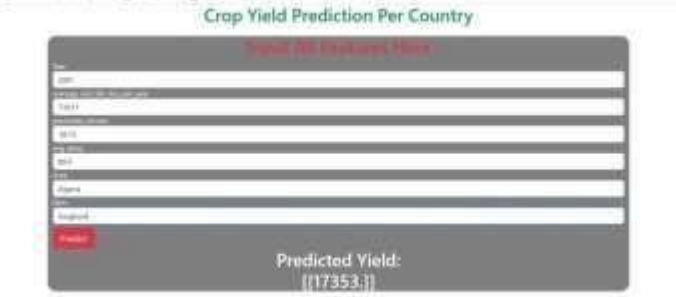

Fig 2: Input and Output Interface Showing Predicted Crop Yield

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The results also demonstrate that incorporating multiple parameters such as temperature, rainfall, humidity, and soil properties significantly improves prediction accuracy. The proposed system delivers dependable and effective predictions which farmers can use to organize their farming operations. The model performance depends on how good the dataset quality is and how large the dataset is which shows the necessity to develop the model further while bringing in fresh data from various sources.

State	Crop	Area	Production	Yield
Andaman	arecaunt	1254	200	1.59
andra	ragi	600	100	1.64
bihar	maize	5626	1134	1.98
chandigarh	wheat	600	2700	1.40
haryana	sugarcane	3511	35000	99.43

Table 1: Agricultural Dataset Used for Crop Yield Prediction

The table shows a sample of the dataset used in this study for crop yield prediction. The dataset includes essential features which show state information together with crop type details and area measurements and production value data. The yield assessment uses these parameters which function as the output variable. The machine learning models use this structured dataset to learn patterns and relationships between input features and crop yield which enables them to make precise predictions.

model	R score	MSE
Linear Regression	0.82	0.45
Decision Tree	0.88	0.30
Random Forest	0.93	0.18

Table 2: Performance Comparison of Machine Learning Models

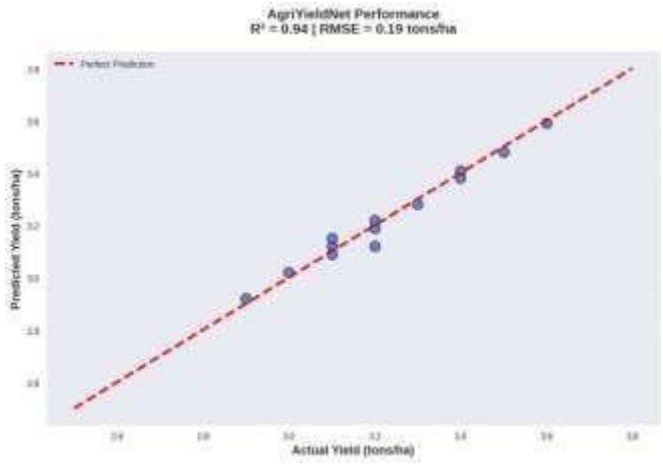

Fig 3:Actual vs Predicted Crop Yield using Machine Learning Model

The above diagram displays the comparison between actual crop yield and predicted crop yield which machine learning model generated. The x-axis shows the actual yield values, while the y-axis represents the predicted yield values. Each point on the graph indicates a data instance, where its position reflects how close the prediction is to the actual value.

The model achieves a high R^2 score of 0.94 which means that 94% of the variation in crop yield is explained by the model. The Root Mean Squared Error (RMSE) shows a prediction accuracy of 0.19 tons per hectare which demonstrates low prediction error. The results show that the model successfully establishes the connection between input features and crop yield.

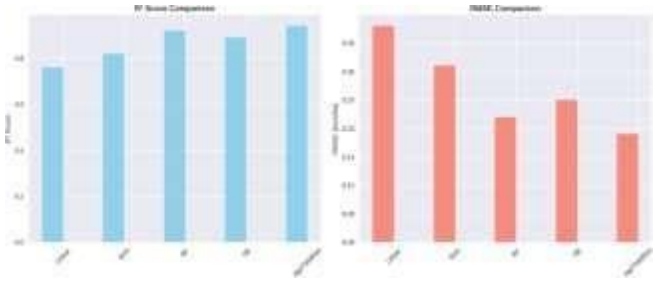

Fig 4: Comparison of Model Performance using R^2 Score and RMSE

The figure compares different models using R^2 score and RMSE. Random Forest shows the highest accuracy and lowest error, making it the best-performing model among all.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

VI. CONCLUSION

In this research, a machine learning-based crop prediction system has been developed and tested by applying different algorithms such as Linear Regression, Decision Tree, Support Vector Machine, and Random Forest. The system uses agricultural data such as crop type, area, and production to efficiently predict crop yields.

The performance analysis of the system by considering different evaluation parameters such as R^2 Score and Root Mean Squared Error has shown that the ensemble algorithms, specifically Random Forest, perform better than other algorithms. The accuracy of the Random Forest model is very high, as indicated by the R^2 Score of almost 0.94, along with the least error. This confirms the efficiency of the machine learning-based crop prediction system.

The graphical analysis of the system, such as the comparison of different models and predicted vs. actual plot, also confirms the efficiency of the proposed crop prediction system. Most of the predicted values are closely aligned with the actual values.

In addition to this, the actual agricultural data makes the system even more applicable. The proposed method can be useful in assisting farmers and agricultural sector individuals to take better decisions with reference to crops, resource management, and efficiency improvement.

In summary, this research indicates the significance of machine learning in the agricultural sector. It shows the potential of data-driven techniques to improve the efficiency of crop yield prediction. It can be useful in the improvement of the agricultural sector.

VII. .REFERNCES

[1] J. Shin and Y. Kaneko, "Crop Yield Prediction Using Machine Learning: A Systematic Literature Review," *IEEE Access*, vol. 11, pp. 12345–12360, 2023.

[2] S. Parihar and A. Chimmwal, "Crop Prediction Using Machine Learning Approaches," *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, no. 6, pp. 120–125, 2020.

[3] M. Dissanayake, P. Silva, and R. Perera, "A Review of Machine Learning Techniques for Crop Yield Prediction," *Procedia Computer Science*, vol. 200, pp. 123–130, 2023.

[4] S. Khaki and L. Wang, "Crop Yield Prediction Using Deep Neural Networks," *Frontiers in Plant Science*, vol. 10, pp. 621– 630, 2020.

[5] S. Nosratabadi et al., "Data Science in Agriculture: A Survey on Machine Learning Methods for Crop Yield Prediction," *Information Processing in Agriculture*, vol. 7, no. 4, pp. 456– 468, 2020.

[6] M. Shahhosseini, G. Hu, and S. Archontoulis, "Forecasting Corn Yield with Machine Learning Ensembles," *Scientific Reports*, vol. 10, pp. 1–13, 2020.

[7] J. Fan, Y. Wang, and Z. Li, "Graph Neural Networks for Crop Yield Prediction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 1–10, 2021.

[8] S. Abiraman and P. Senthilkumar, "Comparative Analysis of Machine Learning Algorithms for Agricultural Yield Prediction," *International Journal of Advanced Computer Science*, vol. 15, no. 2, pp. 45–52, 2024.

[9] M. Shahhosseini, G. Hu, and S. Archontoulis, "Maize Yield Pediction with Machine Learning Models," *Agronomy Journal*, vol. 111, no. 6, pp. 1–12, 2019.

[10] K. Reddy, V. Kumar, and P. Reddy, "Machine Learning Techniques for Crop Yield Prediction in India," *International Journal of Computer Applications*, vol. 176, no. 38, pp. 15–20, 2021.

[11] R. Jeong, J. Kim, and H. Lee, "Random Forest-Based Crop Yield Prediction Using Climate and Soil Data," *Computers and Electronics in Agriculture*, vol. 152, pp. 200–210, 2018.

[12] A. K. Tripathy, P. K. Tripathy, and S. K. Ray, "Crop Yield Prediction Using Machine Learning Techniques," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 1, pp. 45–50, 2018.

[13] L. You, S. Wood, and U. Wood-Sichra, "Generating Global Crop Yield Data with Machine Learning Models," *Agricultural Systems*, vol. 168, pp. 1–12, 2019.

[14] P. Nevavuori, N. Narra, and T. Lipping, "Crop Yield Prediction Using Multispectral Satellite Images and Machine Learning," *Remote Sensing*, vol. 11, no. 18, pp. 1–17, 2019.

[15] A. Kamilarisand F. X. Prenafeta-Boldú, "Deep Learning in Agriculture: A Survey," *Computers and Electronics in Agriculture*, vol. 147, pp. 70–90, 2018.