Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 05, May-2025

A Review of Crop Yield Prediction and Recommendation using Machine Learning

Sandeep Chitalkar
Dr. D. Y. Patil Institute of Technology
Savitribai Phule Pune University
Pune, India

Utkarsh Takmoge
Dr. D. Y. Patil Institute of Technology
Savitribai Phule Pune University
Pune, India

Om Gade
Dr. D. Y. Patil Institute of Technology
Savitribai Phule Pune University
Pune, India

Vishal Thakare
Dr. D. Y. Patil Institute of Technology
Savitribai Phule Pune University
Pune, India

Gunjan Potdukhe
Dr. D. Y. Patil Institute of Technology
Savitribai Phule Pune University
Pune, India

***Abstract*—This article delves into the recent techniques in Machine Learning for developing recommendation systems and forecasting agricultural production. It focuses on popular models like Gradient Boosting, Random Forest, and Long Short-Term Memory (LSTM) neural networks. It also looks at how combining different data sources, building hybrid models, and using tools like remote sensing and feature selection can make predictions more accurate. Traditional methods often can't handle the huge and complex data that modern farming produces. For food security and agricultural efficiency, crop production prediction is crucial. Machine learning analyzes different types of data like soil properties, satellite images, patterns in weather, and crop management practices—more effectively offers a strong solution. By comparing different ML models, the study shows the unique strengths each one offers and suggests a hybrid approach that merges Random Forest and LSTM to improve prediction accuracy even more. It also identifies areas where future research could make these models more user-friendly, precise, and adaptable for real-time application in precision farming.**

***Keywords*—Machine Learning (ML), Crop Yield Prediction, Agricultural Production Forecasting, Recommendation Systems, Gradient Boosting, Random Forest, Long Short-Term Memory (LSTM) Neural Networks, Hybrid Models, Remote Sensing, Feature Selection, Data Integration, Food Security, Precision Farming, Soil Properties, Weather Patterns, Satellite Imagery, Crop Management Practices, Prediction Accuracy, Real-time Application, Agricultural Efficiency**

I. INTRODUCTION

With changing climate and geographical changes, the importance of forecasting the crop yields is necessary for the quality production of food in a sustainable manner. All of these forecasts are vital for not having the risks, ensuring the steady supply and efficient use of resources. Nonlinear and complex data patterns that we often observe in agricultural data are challenging for the statistical and mathematical approaches to handle. Hence, deep learning and machine learning techniques are crucial factors that can evaluate and generate precise forecasts to guide choices. These machine learning techniques can significantly help in giving the desire outputs. Machine learning models that are rigorously trained over data are advanced in processing complex agricultural data from multiple resources, including satellites, weather data patterns, and soil and land properties. The primary models for the prediction of the yield prediction and recommendation will be tested and evaluated based on their effectiveness and their future uses.

II. MOTIVATION

With rising population demands and fluctuations in climate globally, the food quality has become a topic of interest in global communities, providing a good quality of food via effective and sustainable farming has emerged as worldwide imperative. Traditional farming methods typically depend on past trends, intuition and broad recommendations, which are inadequate for addressing the fluctuations of contemporary agricultural settings. With changing weather conditions, limited access to practical information and soil quality further creates challenges for effective decision-making on the farm. The Motivation for this review comes by addressing the transition to data centric agricultural solutions that helps in boosting the productivity and optimizing the use of resources Our drive originates from connecting theoretical ML advancements with their real-world applications. We seek to emphasize models that are not just precise but also understandable, scal- able, and usable in practical agricultural systems. In the end, this paper aims to facilitate the transition of agriculture into a more intelligent, adaptable field by incorporating sophisticated computational technologies

III. LITERATURE SURVEY

Table I summarizes research conducted in recent crop yield predictions using machine learning techniques, showcasing the key advantages and limitations of each approach. This analysis helps in identifying existing gaps and potential directions for future work in precision agriculture.

TABLE I
LITERATURE SURVEY TABLE

Paper Name	Pros	Cons
B. Bischke, et al., 2023, **P. Helber, et al. 2023**	Focuses on an operational approach to yield modeling at both field and subfield levels, making it practical for real world application	Lacks detailed experimental validation for model's better Performance in various agricultural settings various Agricultural settings
N. Yamsani et al. 2023	Utilizes deep reinforcement learning with a modified reward function for Yield prediction, offering a novel approach to handling diverse environmental factors.	The approach requires significant computational resources for model training and model implementation.
U.B.A. et al. 2023	Introduces precision agriculture using machine learning for targeted resource optimization, promoting sustainable farming practices.	Primarily based on theoretical models, with limited practical field trials.

A. Soil Data Analysis

Healthy soils ensure nutrient availability, water retention, and root development, which are important factors for robust and accurate crop yield predictions. The accuracy of the machine learning model can be significantly boosted by the characteristics of soils, such as pH levels, moisture, and organic matter content [1]. The machine learning models, like random forest and support vector machine, have proven efficient and effective, particularly for analyzing soil and agricultural data, which is beneficial for crop yield prediction [2]. Many research and outcomes have been observed that say that integrating the soil data increases the accuracy of yield prediction by up to 20 %.

Moreover, in recent times there has been an increasing spike in developing the ML models that predict the crop's suitability based on soil conditions as it influences the potential yield predictions [3]. For example, Zang et al. (2022) developed an RF-based model that considered nitrogen levels and moisture of the soil to predict paddy levels. Making the accurate yield predictions is crucial step of soil properties, especially when it comes to constantly changing environmental conditions and this approach has helped in an increase this accuracy by 18% [3]. Soil development and health monitoring field using the IoT technology, data on soil and microbial activity further enhances the predictive capabilities of ML models.

B. Weather-Driven Yield Models

To some extent the prediction of a Crop yield is dependent on the weather forecast. Numerous studies and research have predicted crop outcomes using weather data, such as temperature, humidity, and rainfall. Two well-known models that make good use of this essential components are Random Forest (RF) and Gradient Boosting Machines (GBMs) [4]. Several of these models have achieved accuracy rates of 90% or more in tests conducted in a variety of climates [5].

For example, a study by Li et al in 2021 predicted seasonal maize yields by combining crop yield data with weather data from different seasons. The model displayed its ability to manage climate swings by performing well even in areas with different climate conditions [7]. The timing of weather data is also very important, time-series analysis has been used in recent models to capture the influence of weather patterns over time [9].

C. Remote Sensing and Satellite Data

Crop yield scenario has completely changed after use of the satellite technologies. Taking detailed pictures of the agricultural crops, one can observe factors such as the growth stage and health and the surrounding environment more closely [10]. Deep learning techniques like Convolutional Neural Networks, which process the image in layers, and Long Short-Term Memory networks are proven effective to analyze these images in detail, providing accuracy for prediction of crop yield [11]. One of the examples, a study by Chen et al, combined weather forecasting data with images from the satellites to predict the yield of rice crop. Model used by them were CNN and LSTM approach, which gave exceptionally better results than the traditional models such as linear regression model, which showed a clear improvement in accuracy [15]. Using a hybrid model leverages the advantages of CNN for spatial data and LSTM for tracking the changes over a period of time, making it a suitable choice for agricultural data. Additionally, models using indices like Enhanced Vegetation Index and Normalized Difference Vegetation Index give real-time and precise information about the crops condition and lead to more reliable crop yield predictions [17].

IV. RESEARCH GAP AND CHALLENGES

The major drawback of the agricultural dataset is that they lack complete information, inconsistent standards and discrepancies. The datasets that are used nowadays are the outdated and they failed to provide the real time updates for training and making timely decisions [4]. Moreover, the label data from small farms also limits the creation on universal models. Also, the sharing of such crucial data and ownership forms an obstacle to data sharing and collaborative efforts.

While significant advancements have been done in this field to apply machine learning to predict crop yield and develop recommendation systems, research gaps and challenges still hinder their practical scalability and acceptance [7]. challenges including issues such as the quality of data, methods of modeling, infrastructure and regulatory frameworks.

Challenges Associated with Infrastructure and Execution The significant computational demands of ensemble and hybrid models render them impractical in resource-constrained regions, especially in developing nations. Delays are observed in producing the real-time forecasts due to limited processing power and lack of edge infrastructure and cloud. Difficulties in Engaging Farmers Numerous recommendation systems utilizing machine learning often overlook the diverse socio-economic backgrounds and educational levels of farmers [10].

Farmers sometimes don't have the continuous supply to work steadily with electronic tools, so the deployment of the ML system won't work on the field. Some of these gadgets do not support the local language of the farmers, which make them difficult to use. The disparity in digital accessibility, especially in rural areas, indicates that farmers might find it challenging to access or understand the results generated by these models.

V. COMPARATIVE ANALYSIS

A. Random Forest (RF)

To predict the agricultural yields one of the widely used machine learning technique is Random Forest which is based on ensemble learning. It processes the data by creating multiple decision trees during the training period and then averaging their prediction results. This method or algorithm is known for its high accuracy. It helps to prevent overfitting and effectively manages large and complex datasets. Random Forest is beneficial when working with structured data like weather conditions, crop management practices, and soil parameters [8] [16]. A key feature of Random Forest is its ability to assess the importance of different feature variables, making it easier to identify which factors have the most influence on crop yields, enhancing the accuracy.

Although having a few advantages, Random Forest additionally has a few barriers on the subject of managing sequential information, along with time-collection climate styles that contain beyond and destiny information points [5]. While Random Forest is commonly much less susceptible to overfitting in comparison to different device mastering models, it could have demanding situations even as managing noisy or imbalanced datasets.

B. Support Vector Machines (SVM)

Support Vector Machine (SVM) is a common machine-learning technique for classification and prediction tasks, especially when working with small to medium-sized datasets. In the context of yield prediction, SVMs are used frequently to calculate yields based on patterns in weather, soil conditions, and crop management approaches. While SVMs perform well with smaller datasets, they may struggle with larger datasets with more features.

One of the primary benefits of SVMs is their ability to prevent overfitting, particularly when proper kernel functions are used. However, SVMs can be costly when it comes to computational resources, especially for dealing with large datasets, and may struggle to capture the complicated non- linear correlations that are frequently observed in crop produc- tion prediction [5]. Moreover, SVMs need careful parameter tweaking to avoid overfitting or underfitting, which can be a time-consuming and challenging process.

C. Convolutional Neural Networks and Long Short-Term Memory Networks

Deep learning models, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have proven extremely effective at processing satellite photos. CNNs are excellent at identifying spatial patterns in images, whereas LSTMs excel at understanding data that varies over time, such as crop growth. CNN and LSTM combined in hybrid models have given the results that are proven to be extremely effective in predicting crop yields using satellite images and weather data patterns. These models can process both spatial and temporal data. However, these models demand a significant amount of computational power and training data, which can be a constraint [5]. Furthermore, applying these models in resource-constrained environments, such as small farms or underdeveloped countries, might be difficult due to the high computing and data needs. Table 1 Represent the comparative analysis or previously used models

TABLE II
ML Models and Accuracy in Crop Yield Prediction

ML Model	Data Source	Accuracy (%)
Random Forest	Weather, Soil	92
SVM	Weather, Soil, Crop Management	85
CNN-LSTM	Remote Sensing, Satellite	88

VI. PROPOSED METHOD

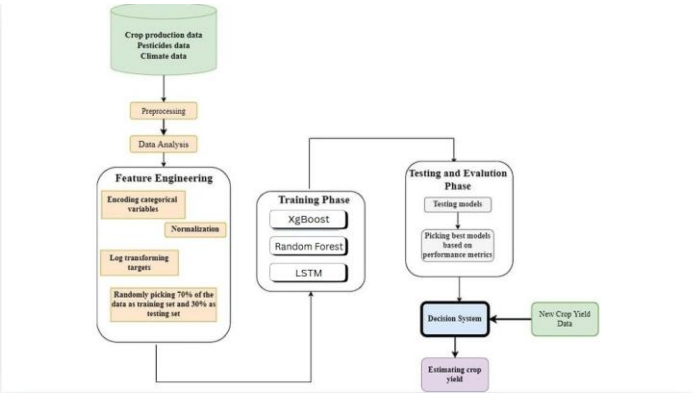

Fig. 1. System Architecture for Crop Yield Prediction

This article presents a structured novel hybrid model that combines the benefits of Random Forest, Long Short-Term Memory (LSTM) networks, and Extreme Gradient Boosting (XGBoost). Random Forest is well known for its outstanding performance when working with large datasets and for capturing static variables like soil characteristics and crop management strategies. LSTM is a powerful deep-learning algorithm that is particularly good at modeling dynamic patterns in time-series data, past weather and yield trends. With the help of XGBoost, a scalable and incredibly efficient gradient-boosting algorithm, we intend to greatly improve the model's prediction capabilities, especially when working with structured data. Together, these three algorithms offer a robust and accurate framework for forecasting agricultural output.

A. Data Preprocessing and Integration

The first step in this approach is data preparation, which involves gathering a wide variety of data related to agricultural factors from different sources. This includes data from sources like soil sensors placed in farms, crop management systems and weather stations, as well as unstructured data from satellite pictures. Every data point is standardized before being included to the model so that consistency and great quality across all sources are assured. Additionally, we use imputation techniques to handle any missing values. This is crucial for training since XGBoost is sensitive to data gaps and quality.

B. Feature Selection and Dimensionality Reduction

It works by eliminating the less important features gradually which do not have a big impact in final result, leaving behind only the ones that have the most significant impact and importance on predictions of crop yield. This approach increases the accuracy of the model and helps in reducing computational complexity. XGBoost well-suited to benefit from this process since it has its own feature importance mechanism, which can further validate the features chosen by RFE.

Further lower the number of dimensions by the technique called Principal Component Analysis (PCA), specially in case of large and complicated datasets which has several charac- teristics like soil data, satellite data and weather patterns etc. PCA helps to simplify the dataset without any loss of crucial information, making it easier for all three models including RF, LSTM, and XGBoost—to handle and perform optimally. XGBoost, in particular works better when working with struc- tured data that has a reduced number of dimensions, making this step even more crucial for improving its performance

C. Model Training and Fusion

The Random Forest model is trained using the structured data, which consist the information on soil conditions, weather trends, and crop management practices. The LSTM model, which focuses on weather trends over time, is trained using time-series data. In order to effectively handle structured data, we utilize XGBoost's capabilities, particularly when it comes to identifying intricate correlations between different factors, such soil characteristics and historical yield data. XGBoost improves the accuracy of short-term agricultural yield estimates by learning from high-dimensional structured data.

Each model is cross-validated to fine-tune its parameters and ensure that we get the best performance out of the three. The primary goal of RF and XGBoost is to generate precise short-term forecasts using soil and crop management data, whereas LSTM is designed to identify long-term trends, especially those impacted by weather fluctuations. The final crop yield projection is produced by combining the outputs of each model after they have all been trained using the data.

The key advantage of using this hybrid model is the ability to adjust the weight of each model's prediction depending on the specific context. For example, if we're predicting crop yields for a vicinity with substantially special soil conditions, the contributions of the RF and XGBoost fashions can be given extra weight due to the fact they're higher applicable to fixing those varieties of issues. However, the long-time period climate traits can have a larger impact on the output, LSTM's manufacturing is probably given extra weight. This flexibility in changing the model's outputs guarantees that we get the maximum correct and dependable yield predictions.

D. Real-Time Deployment and Cloud Integration

To enable real-time crop yield estimates, we suggest integrating our hybrid model with cloud platforms. The cloud's scalability allows us to process large amounts of data and provide up-to-date predictions, helping farmers make informed decisions quickly. In addition, by deploying the model on IoT devices in the field, we can collect real-time data like soil moisture and temperature. Farmers can get the maximum updated meteorological records way to forecasts which are constantly up to date through facts. The performance of XGBoost in processing dependent facts substantially improves the accuracy of those real-time forecasts. Cloud computing, IoT facts, and our hybrid version provide a powerful, real-time method to assist farmers efficiently manipulate their plants and maximize manufacturing projections.

VII. CONCLUSION

This paper concludes via way of means of summarizing numerous gadget studying techniques for predicting agricultural output and growing smart advice systems. It highlights how crucial it's miles to get the data from different sources, such as climate patterns, soil traits and satellite tv for pc imagery, that allows you to enhance forecast accuracy. The suggested hybrid model combines Random Forest and Long Short-Term Memory networks and XGBoost to provide a reliable and flexible method for real-time agricultural yield forecasting. Future research will focus on optimizing the model's deployment in resource-constrained contexts and enhancing prediction interpretability through advanced explainability techniques.

REFERENCES

[1] P. Helber, B. Bischke, P. Habelitz, C. Sanchez, D. K. Pathak, M. Miranda Lorenz, H. Najjar, F. Mena, J. Siddamsetty, D. Arenas, M. Vollmer, M. Charfuelan Oliva, M. Nuske, and A. Dengel, ”Crop yield prediction: An operational approach to crop yield modeling on field and sub-field level with machine learning models,” in Proc. 2023 IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), Pasadena, CA, USA, 2023, pp. 4092-4094, doi: 10.1109/IGARSS46784.2023.10283302.

[2] M. Rashid, B. S. Bari, Y. Yusup, M. A. Kamaruddin, and N. Khan, ”A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction,” IEEE Access, vol. 9, pp. 63406-63439, Apr. 2021, doi: 10.1109/ACCESS.2021.3075159.

[3] D. Jayanarayana Reddy and M. Rudra Kumar, ”Crop yield prediction using machine learning algorithm,” in Proc. IEEE 2021 5th Int. Conf. Intelligent Computing Control Syst. (ICICCS), Madurai, India, May 2021, pp. 1466-1470, doi: 10.1109/ICICCS51141.2021.9432236.

[4] S. Author, ”Machine learning techniques in crop recommendation based on soil and crop yield prediction system – Review,” in Proc. IEEE 2022 Int. Conf. Artificial Intelligence Data Eng. (AIDE), Karkala, India, Dec. 2022, pp. 230-235, doi: 10.1109/AIDE57180.2022.10078849.

[5] S. V. Gaikar, M. S. Zambare, and A. D. Shaligram, ”A systematic review of the soil fertility monitoring and organic farming techniques for an improved crop yield,” in Proc. 2023 IEEE Third Int. Conf. Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, Feb. 2023, pp. 1-8, doi: 10.1109/ICAIS56108.2023.10073868.

[6] K. P. K. Devan, S. B., U. S. P., and V. S., ”Crop Yield Prediction and Fertilizer Recommendation System Using Hybrid Machine Learning Algorithms,” in Proc. 2023 IEEE 12th Int. Conf. Commun. Syst. Netw. Technol. (CSNT), Bhopal, India, Apr. 2023, pp. 171-175, doi: 10.1109/CSNT57126.2023.10134662.

[7] A. Lakshmanarao, M. Naveen Kumar, K. S. V. Ratnakar, and Y. Satwika, ”Crop yield prediction using regression models in machine learning,” in Proc. IEEE 2023 2nd Int. Conf. Appl. Artif. Intell. Comput. (ICAAIC), Surampalem, India, May 2023, pp. 423-426, doi: 10.1109/ICAAIC56838.2023.10141462.

[8] D. Elavarasan and P. M. Durairaj Vincent, ”Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications,” IEEE Access, vol. 8, pp. 86886–86901, 2020, doi: 10.1109/ACCESS.2020.2992480.

[9] N. Yamsani, R. Vijayarangan, V. Thirumurugan, G. M. Ramadan, and H. M. Al-Jawahry, ”Deep reinforcement learning with modified reward function for crop yield prediction,” in Proc. IEEE 2023 Int. Conf. Ambient Intell., Knowl. Informatics Ind. Electron. (AIKIIE), Warangal, India, Nov. 2023, pp. 1-5, doi: 10.1109/AIKIIE60097.2023.10390279.

[10] A. Reyana, S. Kautish, P. M. Sharan Karthik, I. A. Al-Baltah, M. B. Jasser, and A. W. Mohamed, ”Accelerating crop yield: Multisensor data fusion and machine learning for agriculture text classification,” IEEE Access, vol. 11, pp. 20795-20805, Feb. 2023, doi: 10.1109/ACCESS.2023.3249205.

[11] P. G, T. V. R, A. Pushpalatha, and P. Kavitha Rani, ”Effective crop yield prediction using gradient boosting to improve agricultural outcomes,” in Proc. IEEE 2023 Int. Conf. Networking and Commun. (ICNWC), Chennai, India, Apr. 2023, pp. 1-6, doi: 10.1109/ICNWC57852.2023.10127269.

[12] U. B. A., S. K. N., B. D. Shetty, S. Patil, K. Dullu, and S. Neeraj, ”Machine learning in precision agriculture,” in Proc. IEEE 2023 4th Int. Conf. Communication, Computing and Industry 6.0 (C216), Bangalore, India, Dec. 2023, pp. 1-6, doi: 10.1109/C2I659362.2023.10431196.

[13] S. R. Gopi and M. Karthikeyan, ”Effectiveness of crop recommendation and yield prediction using hybrid moth flame optimization with machine learning,” Engineering, Technology and Applied Science Research journal, vol. 13, no. 4, pp. 11360-11365, 2023, doi: 10.48084/etasr.6092.

[14] K. Lohitha Reddy and A. P. Siva Kumar, ”Machine learning techniques for weather based crop yield prediction,” in Proc. IEEE 2023 Third Int. Conf. Artificial Intelligence Smart Energy (ICAIS), Anantapur, India, Feb. 2023, pp. 1263-1268, doi: 10.1109/ICAIS56108.2023.10073740.

[15] C. Kiruthiga and K. Dharmarajan, ”Machine learning in soil borne diseases, soil data analysis crop yielding: A review,” in Proc. IEEE 2023 Int. Conf. Intelligent Innovative Technol. Comput., Electr. Electron. (IITCEE), Bengaluru, India, Jan. 2023, pp. 702-706, doi: 10.1109/IITCEE57236.2023.10091016.

[16] P. Sharma, P. Dadheech, N. Aneja, and S. Aneja, ”Predicting agriculture yields based on machine learning using regression and deep learning,” IEEE Access, vol. 11, pp. 111255-111264, Oct. 2023, doi: 10.1109/ACCESS.2023.3321861.

[17] U. Shafi, R. Mumtaz, Z. Anwar, M. M. Ajmal, M. A. Khan, Z. Mahmood, M. Qamar, and H. M. Jhanzab, ”Tackling food insecurity using remote sensing and machine learning-based crop yield prediction,” IEEE Access, vol. 11, pp. 108640-108657, Sep. 2023, doi: 10.1109/ACCESS.2023.3321020.