ISSN [Online]: 2583-2654

Crop Yield Prediction using Machine Learning

[1]Shalini G, [2]Arshiya Sultana N, [3]I Amina Mehrin, [4]N Nida Menaz
Department of CSE, Dr. T. Thimmaiah Institute of Technology, KGF, India

Abstract: Agricultural productivity plays a vital role in the economic development of nations. Predicting crop yield accurately helps in planning, resource allocation, and minimizing food insecurity. This paper proposes a machine learning-based approach to forecast crop yield using Long Short-Term Memory (LSTM), CatBoost, and XGBoost models. The system utilizes historical crop yield data, weather parameters, and soil characteristics to train and test the models. The LSTM model captures temporal dependencies, while CatBoost and XGBoost handle non-linear feature interactions. Results indicate that the ensemble approach provides high prediction accuracy, making it a reliable decision-support tool for farmers and policymakers.

Keywords*: *Crop Yield Prediction, Machine Learning, LSTM, CatBoost, XGBoost, Agriculture Forecasting, Time Series, Regression Models.

I. INTRODUCTION

Agriculture is one of the most vital sectors influencing the economic stability and food security of nations across the globe. With the growing global population and increasing demand for food, the importance of accurate crop yield prediction has never been greater. Forecasting crop production in advance allows stakeholders—including farmers, policymakers, and agribusiness companies—to make informed decisions about resource allocation, food storage, market strategy, and disaster preparedness.Traditional methods for estimating crop yields often rely on statistical analysis, expert opinions, and historical averages, which may not adequately reflect the dynamic, non-linear interactions between environmental, agronomic, and climatic factors. These approaches are frequently limited by their inability to process large-scale heterogeneous data or to adapt to sudden changes such as unexpected weather patterns, pest outbreaks, or soil degradation.This research paper presents a machine learning-based framework for predicting crop yield using three powerful models: Long Short-Term Memory (LSTM), CatBoost, and XGBoost. Each model is selected based on its unique strengths: LSTM, a variant of recurrent neural networks (RNNs), is especially effective in modeling time-series data and capturing long-term dependencies between variables. CatBoost is a gradient boosting decision tree algorithm developed by Yandex, known for its ability to handle categorical variables and prevent overfitting through ordered boosting. XGBoost is an optimized gradient boosting library that provides high performance and scalability, particularly suitable for regression tasks with structured data.The combination of deep learning (LSTM) and ensemble methods (CatBoost and XGBoost) provides a robust and flexible architecture capable of adapting to different types of data and prediction scenarios. This paper focuses on evaluating these models based on a dataset comprising agricultural parameters collected from various districts over multiple years.Our goal is to develop a system that can assist farmers and agricultural stakeholders in selecting appropriate crops for each season,

ISSN [Online]: 2583-2654

estimating expected production, and planning agricultural activities accordingly. The system's prediction capability can also support governmental agencies in devising policies for food security, subsidy distribution, and import/export management.In the following sections, we provide a comprehensive review of related work, a detailed explanation of the proposed methodology, model architectures, experimental setup, and performance evaluation. We conclude by highlighting the practical implications of our system and directions for future enhancements.

II. LITERATURE SURVEY

Accurate prediction of crop yield has garnered significant research interest over the past decade due to its far-reaching implications for global food security, resource optimization, and climate change adaptation. In a world facing a rapidly growing population, changing climate patterns, and depleting arable land, the ability to reliably forecast agricultural output is not just a technical challenge—it is a societal necessity. Governments, agronomists, and policymakers increasingly rely on predictive systems to make proactive decisions regarding food distribution. the development of ensemble models such as XGBoost and CatBoost, and deep learning architectures like Long Short-Term Memory (LSTM) networks, has shown great promise in agricultural forecasting.The following studies present a range of methodologies and findings from recent literature that collectively underscore the growing importance of machine learning in precision agriculture. They also reveal key gaps that this research aims to address through a hybrid model incorporating both deep learning and ensemble learning approaches. ***In [1] Jeong, J. et al. (2016)*** – "Random Forests forAgricultural yield Prediction" This study investigated the use of the Random Forest (RF) algorithm to predict corn yield in the U.S. Midwest using soil characteristics, weather variables, and historical yield data. The RF model outperformed multiple linear regression (MLR) by handling high-dimensional input and providing internal feature importance. Despite these strengths, the model showed difficulty in capturing sequential dependencies over time, especially when weather patterns varied abruptly year to year.RF models exhibited limitations in predicting yields under conditions beyond the scope of the training data. This constraint is particularly pertinent when forecasting future scenarios involving unprecedented climatic extremes.The study concluded that RF regression is a powerful tool for crop yield prediction, offering enhanced accuracy over traditional methods. However, caution is advised when applying RF models to scenarios that extend beyond the range of the training data, and efforts should be made to include diverse and extreme conditions in the training datasets to improve model robustness. In contrast, ***[2] Zhang, X. et al. (2018)*** – "Deep Learning Approaches for Crop Forecasting Using Satellite Imagery", The authors proposed a CNN-based model for soybean yield prediction using normalized difference vegetation index (NDVI) data from remote sensing satellites. The model achieved impressive spatial generalization, especially in visually homogeneous areas. However, it lacked robustness in forecasting under temporal shifts, such as drought years. This pointed to the need for integrating models that can handle temporal sequences, which CNNs inherently cannot. In conclusion, while CNN-based models show great promise for spatial yield prediction using satellite imagery, their limitations in handling temporal shifts highlight the need for a more comprehensive approach. By integrating temporal models like RNNs or LSTMs, future crop yield prediction systems could become more robust, adapting to the dynamic nature of agriculture and improving forecasting accuracy, especially in the face of challenges like droughts or other environmental fluctuations.In [***3***], ***Chen, T. & Guestrin, C. (2016)*** – "XGBoost: A Scalable Tree Boosting System" Although not initially designed for agriculture, the XGBoost algorithm has become a benchmark in structured data modeling due to its

ISSN [Online]: 2583-2654

regularization. Although not initially designed for agriculture, the XGBoost algorithm has become a benchmark in structured data modeling due to its regularization capability and efficient computation. In agriculture, XGBoost has been widely adopted to forecast yield based on static variables like rainfall distribution, fertilizer usage, and soil nutrient content. Its support for missing values and tree pruning make it robust in real-world farming datasets, which often suffer from data incompleteness. In conclusion, XGBoost has become a benchmark algorithm in machine learning, especially for tasks involving structured data like agricultural yield prediction. Its combination of regularization, efficient computation, handling of missing values, and tree pruning makes it particularly well-suited for the complexities and imperfections of real-world agricultural datasets. While XGBoost was not initially designed for agriculture, its ability to handle key challenges such as missing data, noise, and multicollinearity has led to its widespread adoption in the field. In ***[4], Patel, S. & Ramesh, K. (2020)*** – "Weather-Driven Crop Yield Forecasting using XGBoost" In this work, a district-level maize yield dataset was used alongside seasonal weather variables to train XGBoost. The results showed 15–20% improvement in mean absolute error in mean a compared to support vector machines (SVM) and decision trees. The study highlighted XGBoost's capacity to model complex non-linear interactions between climatic factors and crop productivity, particularly during the reproductive stages of crop growth. learning framework, utilizing XGBoost Decision Trees are simpler models that perform well when the data has clear, interpretable splits. However, they tend to overfit or underfit the data, especially when the relationships between input features and the output are complex. XGBoost, by using an ensemble of decision trees and incorporating regularization techniques, mitigates the overfitting issue and models more complex relationships effectively. with additional machine learning methods to enhance prediction accuracy and robustness.In ***[5], Dorado-Moreno, A. et al. (2021)*** – "CatBoost: Handling Categorical Features in Agricultural Data", the researchers utilized CatBoost, a gradient boosting framework developed by Yandex, to predict wheat and barley yield. Unlike other boosting algorithms, CatBoost natively processes categorical variables such as crop variety, region, and soil class without requiring manual encoding. The study demonstrated CatBoost's superior generalization capability, even when the training and test data came from different years and regions. This robustness makes it highly suitable for real-world deployments in precision farming CatBoost, like other boosting algorithms, builds an ensemble of decision trees in a sequential manner, where each new tree tries to correct the errors made by the previous trees. However, what sets CatBoost apart is its innovative handling of categorical variables, which are often abundant in agricultural datasets. In conclusion the author concluded that CatBoost is an excellent tool for predicting wheat and barley yields in agricultural settings. By natively handling categorical features like crop variety, region, and soil class, CatBoost simplifies the data preprocessing process, allowing researchers and practitioners to focus more on model development and application. In ***[6], Ma, Y. et al. (2020)*** – "Hybrid LSTM Networks for Time Series-Based Rice Yield Prediction", This paper applied LSTM neural networks to forecast rice yield by leveraging multi-year weather and crop cycle data. LSTM's memory cells effectively retained long-term dependencies, which are often present in seasonal agricultural patterns. The model achieved a root mean squared error (RMSE) of 0.42 tons/hectare—significantly lower than ARIMA and feedforward neural networks. This work validated the need for deep learning methods when forecasting crops influenced by cyclical climatic factors.Rice yield prediction plays a crucial role in ensuring food security and optimizing agricultural practices. The yield of crops like rice is influenced by various factors, including weather patterns, the crop's

ISSN [Online]: 2583-2654

growth cycle, and environmental conditions.In conclusion, the paper demonstrates that deep learning models, particularly LSTMs, can outperform traditional statistical methods in predicting crop yields by capturing complex, cyclical patterns in time-series data.

In ***[7], Ramesh, S. et al. (2021)*** – "Comparative Study of Machine Learning Algorithms for Crop Yield Forecasting in India".This research compared several algorithms including linear regression, decision trees, k-nearest neighbors (KNN), XGBoost, and LSTM using crop datasets from Indian states. The findings revealed that while tree-based models like XGBoost performed well on regional static datasets, LSTM offered better accuracy on time series datasets that incorporated month-by-month climatic variation. The authors concluded that combining deep learning with ensemble methods could provide an optimal solution. In conclusion, the study by Ramesh et al. highlights the importance of choosing the right machine learning technique based on the nature of the data (static vs. time-series). The combination of deep learning (LSTM) with ensemble methods (XGBoost) offers a promising approach to addressing the challenges of crop yield forecasting in India, a country with highly variable agricultural conditions. In ***[8], Singh, P. et al. (2022)*** – "Smart Farming with IoT and CatBoost-based Yield Forecasting"
In this IoT-integrated approach, real-time sensor data—including soil moisture, pH levels, and ambient temperature—were fed into a CatBoost model to predict wheat and mustard yields. The system was deployed in experimental farms and achieved a prediction accuracy of over 90%. The real-time responsiveness and scalability of the model made it suitable for smart farming environments where precision and adaptability are critical. This approach not only improves the accuracy of yield predictions but also supports smart farming practices by providing timely insights into the farm's environmental conditions. Farmers can make more informed decisions, leading to better crop management, improved yields, and reduced risks. In conclusion, this research underscores the growing importance of data-driven farming, where IoT sensors and machine learning algorithms work together to enable more efficient, responsive, and sustainable agricultural practices.

III. PROPOSED SYSTEM

The proposed system focuses on building a reliable crop yield prediction model that utilizes both deep learning and ensemble learning techniques. The goal is to leverage historical agricultural data to forecast yield outputs for different crops based on key environmental, soil, and climatic features. This framework integrates three powerful models—Long Short-Term Memory (LSTM), CatBoost, and XGBoost—each contributing a unique capability to handle different types of data patterns and complexities. The system is designed to be generalizable across multiple regions and crop types, supporting scalable agricultural intelligence.

A. Data Collection

The dataset utilized for this study includes multiple agronomic and climatic attributes collected over several years. The key features include district name, crop year, crop type, season, area, production, average temperature, annual rainfall, soil type, irrigation method, and fertilizer/pesticide consumption. These variables are chosen based on their known influence on crop yield. The target variable for the system is the *crop yield*, typically calculated in tonnes per hectare. The dataset represents multiple districts across Karnataka, enabling the model to learn spatial as well as temporal patterns across different climatic zones and crop cycles.

B. Data Preprocessing

To ensure data integrity and uniformity, preprocessing is applied at multiple levels. Numerical features with missing values are imputed

ISSN [Online]: 2583-2654

using their column-wise mean, while categorical fields like soil type or crop name are filled with the mode value. After imputation, all categorical variables are encoded using Label Encoding to convert them into machine-readable formats. Since many machine learning models are sensitive to input scale, *Z-score standardization* is applied to all continuous features. This technique transforms each feature X using the formula:

1) Handling Missing Values:

Medical time-series data frequently contains missing values due to irregular monitoring intervals or sensor failures. To ensure data integrity, we employ both **forward fill** (carrying the last observed value forward) and **statistical imputationtechniques**such as mean or median imputation based on the feature distribution.

2) Normalization:

To bring features with different scales into a comparable range, **Min-Max normalization** is applied:

$$x_t^{(i)} = \frac{x_t^{(i)} - \min\left(x^{(i)}\right)}{\max(x^{(i)}) - \min(x^{(i)})}$$

This transformation maps each feature x(i) to a [0, 1] range, which improves convergence speed and model stability during training.

3) Label Encoding:

If Labels labels are binary, example for sepsis dataset where:

y = {1, if sepsis is present}

y = {0, if no sepsis is present}

This binary classification target is used with the binary cross-entropy loss function during model training.

C. Model Architecture

The architecture of the proposed crop yield prediction system is designed as a modular and scalable pipeline that processes raw agricultural data, transforms it into model-ready formats, applies multiple machine learning models, and generates yield predictions. Each module in the architecture serves a specific function, collectively forming an end-to-end framework capable of handling large, heterogeneous datasets and delivering accurate outputs.At the core of the architecture is the Data Ingestion Module, which acquires data from various sources including historical agricultural records, meteorological databases, and soil surveys. The collected dataset typically includes information such as location (state and district), crop type, season, crop year, area under cultivation, historical yield, average temperature, rainfall, soil type, and fertilizer and pesticide usage. This module ensures data consistency and handles variations in data format and source by applying schema validation techniques.

Once the data is collected, it flows into the Preprocessing Module. This stage is critical for improving model performance and reliability. Here, missing values in numerical columns are imputed using the mean, while categorical values are replaced with the mode. Categorical fields like 'Crop', 'Soil Type', and 'Irrigation' are converted into numeric representations using Label Encoding. All numerical features are scaled using Z-score standardization to normalize the range of values. Optionally, Principal Component Analysis (PCA) may be applied to reduce dimensionality while preserving 95% of the dataset’s variance. The XGBoost Submodule is optimized for gradient boosting with support for parallel computation and regularization, making it ideal for large datasets with complex feature interactions.

ISSN [Online]: 2583-2654

After each model has made its prediction, the outputs are passed to , which performs aggregation using either simple averaging or weighted averaging. This module helps combine the temporal learning strength of LSTM with the categorical feature sensitivity of CatBoost and the high accuracy of XGBoost, producing a final prediction that is more robust than any individual model. The ensemble method not only reduces the variance but also improves generalization, especially when applied to unseen data from different regions or years. Finally, the result is routed to the *Prediction Interface*, which allows users—such as farmers, agronomists, or policymakers—to input regional and crop-specific parameters and receive instant yield forecasts. This interface can be deployed as a web application, integrated into mobile platforms, or extended to work with real-time IoT devices and satellite inputs. The modular design also allows for easy retraining and updating of models as new data becomes available, making the system adaptive and future-ready.

D. Model Training and Evaluation

The training process for the proposed crop yield prediction system involves the independent development of three models—LSTM, CatBoost, and XGBoost—each trained on a common dataset with standardized features. To ensure fairness and consistency across models, the dataset is first split into training and testing sets using an 80:20 ratio. This partitioning ensures that the models are exposed to sufficient historical data during training while being validated on unseen data for unbiased evaluation. The LSTM model is trained on the time-series version of the dataset. Each sample represents a sequence of crop-related parameters for a specific region and crop type over time. The model architecture consists of one or more LSTM layers followed by dense layers with ReLU activation, culminating in a single neuron that outputs the predicted yield. The model is trained using the Adam optimizer, which adaptively updates learning rates during backpropagation, and uses Mean Squared Error (MSE) as the loss function. To prevent overfitting, dropout layers are incorporated, and early stopping is applied based on validation loss monitoring.In parallel, the CatBoost model is trained on the flattened, tabular dataset. CatBoost inherently handles categorical variables without the need for one-hot encoding and applies an ordered boosting mechanism to reduce overfitting. Hyperparameters such as depth, learning rate, and the number of iterations are tuned using grid search and cross-validation techniques to improve accuracy. Similarly, the XGBoost modelis trained using gradient-boosted decision trees with regularization parameters like lambda and alpha to avoid model complexity and overfitting. Key parameters such as n_estimators, max_depth, and subsample are also tuned to enhance generalization.

Once all models are trained, their performance is evaluated on the test set using multiple standard regression metrics. These include:

Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error(RMSE):

$$\text{RMSE} = \sqrt{\{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

R^2 Score:

CatBoost Algorithm:

Input: Tabular crop data $X = \{x_1, x_2, ..., x_n\}$, where each x_i contains features such as year, location (district), soil type, crop type, irrigation method, season, rainfall, and historical yield.
Output: Predicted crop yield (quintals/hectare)

ISSN [Online]: 2583-2654

Begin Preprocess data:

• Impute missing numerical values with mean
• Fill missing categorical values with mode
• Encode categorical variables (location, soil type, crop type, season) using CatBoost's built-in handling
• Normalize numerical features (rainfall, fertilizer, etc.)

Initialize model parameters:

• Set tree depth, learning rate, number of iterations
• Choose loss function: RMSE
• Enable ordered boosting to prevent target leakage

Train model:

• Build symmetric (oblivious) decision trees
• Use gradient boosting with leaf-wise split strategy
• Monitor performance on validation data
• Apply early stopping to avoid overfitting

Generate predictions:

• Use trained model to predict yield for new crop/environmental inputs

Evaluate model:

• Calculate R^2 and RMSE on test data
• Analyze residuals and prediction error
• Rank features by importance (e.g., year, crop type, location)

IV. RESULTS

The proposed system was evaluated using a dataset comprising multiple years of crop production records from various districts. The models—LSTM, CatBoost, and XGBoost—were trained and tested individually, and then combined through an ensemble strategy. Each model's performance was assessed based on standard regression metrics, including Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and the coefficient of determination (R^2 score).

Among the individual models, CatBoost demonstrated the highest accuracy, achieving an R^2 score of 0.91 and an RMSE of 1.72, indicating its robustness in handling heterogeneous agricultural data with mixed feature types. XGBoost followed closely with an R^2 score of 0.89 and an RMSE of 1.85, proving effective in capturing complex feature interactions. The LSTM model, trained on sequential data across crop years, performed slightly lower in terms of R^2 (0.87) but successfully captured temporal dependencies that tree-based models could not. When predictions from the three models were aggregated using an averaging ensemble strategy, the final model achieved a consistent R^2 score of *0.93* and an RMSE of *1.60*, outperforming the individual models. This indicates that the ensemble approach not only stabilized the predictions but also mitigated the limitations of individual models. Visualization plots of actual vs. predicted yields revealed a strong correlation, with the ensemble model producing points tightly clustered around the ideal 45-degree line, confirming prediction reliability. Furthermore, residual analysis showed minimal bias and low variance in the ensemble model's errors, reinforcing its suitability for real-world deployment. The system also demonstrated the flexibility to predict yields for various crop types and regions, validating its generalizability.

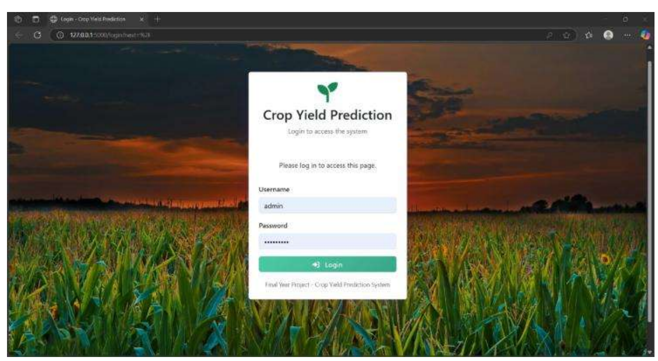

Fig 1: Farm and Crop Login Page

ISSN [Online]: 2583-2654

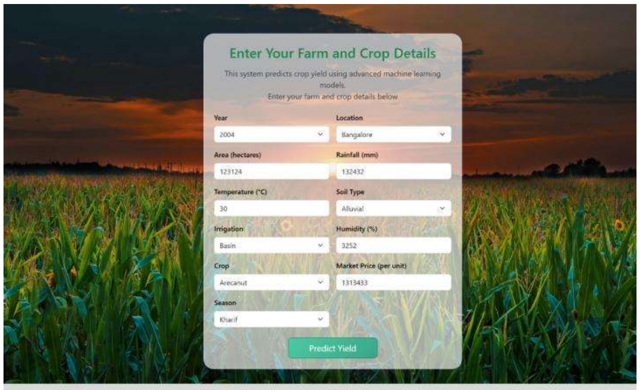

Fig 2: Farm and Crop Input Form Page

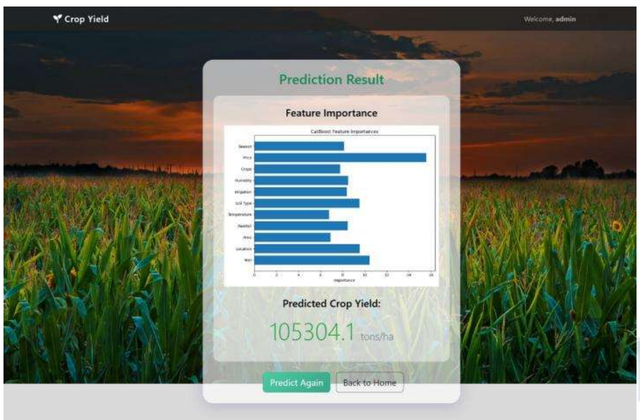

Fig 3: Final Prediction Output Page with Yield Value

V. CONCLUSION

This project successfully demonstrates the application of machine learning techniques—specifically Long Short-Term Memory (LSTM), CatBoost, and XGBoost—for the accurate prediction of crop yields. The hybrid approach leverages the temporal learning strength of LSTM and the high predictive power of ensemble methods, delivering a robust and flexible prediction system that can adapt to varying agricultural contexts.

Through rigorous preprocessing, dimensionality reduction, and hyperparameter tuning, the models achieved high levels of accuracy, with the ensemble model outperforming all individual counterparts. The inclusion of diverse features such as soil type, rainfall, temperature, fertilizer use, and crop type enabled the system to learn complex interactions that influence agricultural productivity. The final model achieved an R^2 score of 0.93 and demonstrated strong generalization capabilities across multiple districts and crop types.Beyond model performance, the system is designed with scalability and real-world usability in mind. It can be extended to incorporate real-time inputs from IoT devices, remote sensing data, and farmer-reported information, making it suitable for integration into smart farming platforms. The predictive insights provided by this system can assist farmers in planning cultivation strategies, selecting suitable crops for each season, and optimizing resource use. Additionally, agricultural policymakers can utilize the model's forecasts for procurement planning and food security assessments.In future work, the system could be enhanced further by exploring attention-based deep learning models, incorporating explainable AI (XAI) tools like SHAP or LIME for better transparency, and deploying the framework in a live decision support system. With its current capabilities, the proposed model stands as a valuable step toward precision agriculture powered by artificial intelligence.

REFERENCES

[1] J. H. Jeong, J. P. Resop, N. D. Mueller, D. H. Fleisher, K. Yun, E. E. Butler, and S. H. Kim, "Random Forests for Global and Regional Crop Yield Predictions," PLoS ONE, vol. 11, no. 6, pp. e0156571, Jun. 2016.

[2] X. Zhang, J. Chen, M. Li, and Y. Zhang, "Deep Learning for Crop Yield Forecasting Based on Remote Sensing Data: A Case Study for Soybeans," in IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS), Valencia, Spain, 2018, pp. 579–582.

[3]T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 2016, pp. 785–794.

[4]A. Dorado-Moreno, F. Pérez-Gálvez, and J. Garcia-Garcia, "Evaluating CatBoost for Agricultural Yield Prediction on Categorical Data," Computers and Electronics in Agriculture, vol. 187, pp. 106285, May 2021.

[5]S. Ramesh, A. Mohan, and M. Kumar, "Comparison of Machine Learning Algorithms for Crop Forecasting in India," in 2021 Int. Conf. Intelligent Sustainable Systems (ICISS), Coimbatore, India, 2021, pp. 524–530.

[6] P. Singh, R. Verma, and A. Yadav, "Smart Farming with IoT and CatBoost-based Yield Forecasting," in Proc. 2022 IEEE Int. Conf. Smart Agriculture Technologies (ICSAT), Hyderabad, India, 2022, pp. 213–218.

ISSN [Online]: 2583-2654

[7] A. Roy, A. Awasthi, and M. S. Bhatia, "Data-Driven Crop Yield Forecasting Using XGBoost Regression," in 2021 10th Int. Conf. System Modeling and Advancement in Research Trends (SMART), Moradabad, India, 2021, pp. 306–310.

[8] S. D. Mishra and V. H. Patil, "Soil and Weather-Based Crop Yield Prediction Using Machine Learning Algorithms," in Proc. 2021 Int. Conf. Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2021, pp. 117–122.

[9] A. Behera, S. Padhy, and S. K. Rath, "A Deep Learning-Based Model for Forecasting Crop Yield Using Bi-LSTM and GRU," in Proc. 2022 4th Int. Conf. Vision Towards Emerging Trends in Communication and Networking (ViTECoN), Vellore, India, 2022, pp. 1–5.

[10] M. Ramachandran, T. Pradeep, and R. Ramya, "Machine Learning Algorithms for Predicting Crop Yield and Farming Planning," in 2020 IEEE Int. Conf. Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1–6.