



Data-Driven Approach for Crop Yield Prediction Using Machine Learning Techniques

Janhavi Jawalkar ¹, Janhavi Jari ², Tejaswini Ghule ³, Arpita Solanke ⁴, Atharv Sune ⁵,
V B Gadicha ⁶

Research Scholars, Computer Science and Engineering, P. R. Pote Patil College of Engineering & Management,
Amravati, India^{1,2,3,4,5}

Head of Department, Computer Science and Engineering, P. R. Pote Patil College of Engineering & Management,
Amravati, Maharashtra, India⁶

Abstract: Agriculture plays a vital role in food security and economic development, making accurate crop yield prediction important for effective farm planning and resource management. This study proposes a data-driven web-based agricultural decision-support system that integrates crop yield prediction, crop recommendation, real-time weather information, plant disease detection, farm analytics, and government scheme assistance. Machine learning algorithms such as Random Forest, XGBoost, and Gradient Boosting are used for crop yield prediction, while a Random Forest classifier is used for crop recommendation based on soil and weather parameters. The system further provides agronomic explanations, spray dosage calculation, multilingual support, voice input, and WhatsApp-based information sharing to improve accessibility and usability for farmers. The proposed KrushiMitra platform combines predictive analysis and practical agricultural support features within a single farmer-friendly interface.

Keywords: Crop Yield Prediction, Machine Learning, Random Forest, Precision Agriculture, Agriculture Decision Support.

I. INTRODUCTION

Crop yield is influenced by several interconnected factors, including weather conditions, soil properties, crop type, geographical location, and farming practices. Parameters such as rainfall, temperature, humidity, soil pH, nitrogen (N), phosphorus (P), potassium (K), fertilizer usage, and irrigation directly affect crop growth and final production. Therefore, reliable crop-yield prediction requires the analysis of relevant agricultural, environmental, and climatic data rather than relying only on traditional estimation methods.

The need for accurate crop yield prediction has increased due to changing weather conditions, inappropriate crop selection, inefficient utilization of agricultural resources, and variations in soil and climatic conditions. Farmers often depend on experience and conventional estimation techniques to determine suitable crops, expected production, and appropriate farm practices. Conventional methods may not adequately consider the complex relationships among multiple environmental and agricultural factors, which can lead to reduced productivity, resource wastage, and financial losses. A data-driven decision-support framework can address these limitations by providing quantitative information for crop selection, yield estimation, farm analysis, and agricultural planning.

Machine learning (ML) provides an effective approach for identifying complex relationships between agricultural parameters and crop yield. Previous studies have investigated various machine learning techniques, including Random Forest, regression, decision trees, neural networks, and ensemble learning methods for crop yield prediction [3]. Recent research has also explored advanced techniques such as XGBoost, stacking, and other ensemble models to improve prediction performance [9]. However, the effectiveness of machine learning techniques can vary depending on crop type, geographical region, dataset characteristics, feature selection, and environmental conditions. Consequently, comparative evaluation of suitable machine learning models is important for identifying reliable prediction techniques for agricultural datasets.

The integration of multiple agricultural and environmental data sources can further improve crop-yield forecasting and decision support. Weather and pesticide-related information have been incorporated into yield forecasting models [5], while climate variables and Normalized Difference Vegetation Index (NDVI) data have been applied to wheat-yield prediction [11]. Other research has investigated multimodal and deep-learning techniques that combine agricultural and remote-sensing information [12][14]. Although such techniques can provide promising prediction results, their implementation may require large datasets, extensive computational resources, and specialized infrastructure. These requirements can create challenges for practical deployment in resource-constrained agricultural environments.



The proposed research aims to develop a data-driven agricultural decision-support framework using machine learning techniques and integrated agricultural information. Random Forest, XGBoost, and Gradient Boosting algorithms are employed for crop yield prediction, along with a machine-learning-based crop recommendation component using soil and weather parameters. The proposed framework further incorporates agronomic explanations, real-time weather information, farm analytics and visualization, plant disease detection, spray dosage calculation, and government scheme information. A multilingual web application named KrushiMitra, supporting English, Hindi, and Marathi, is designed to present agricultural information in an accessible and farmer-friendly manner. Voice input and WhatsApp-based advisory sharing are also incorporated to improve ease of interaction and information access. The integration of predictive analysis, crop recommendation, disease identification, farm analytics, agricultural advisory, and communication features provides a broader decision-support mechanism for crop planning, crop management, and agricultural resource utilization.

The research is aligned with the following objectives:

- (1) To collect, preprocess, and integrate historical agricultural data containing weather parameters, soil nutrients (N, P, K, and pH), fertilizer usage, irrigation, crop type, geographical location, and yield records.
- (2) To analyse agricultural and environmental factors using exploratory data analysis and feature-selection techniques to identify the parameters that have the greatest influence on crop yield.
- (3) To develop and optimize machine learning models, particularly Random Forest, XGBoost, and Gradient Boosting, for accurate crop yield prediction.
- (4) To develop a crop recommendation component that suggests suitable crops based on soil properties, weather conditions, geographical information, and predicted yield.
- (5) To integrate real-time weather information through a weather API to provide location-specific weather conditions and support weather-aware agricultural recommendations.
- (6) To develop a multilingual and farmer-friendly web application supporting English, Hindi, and Marathi for providing crop yield predictions, crop recommendations, weather information, farm analytics, graphical visualizations, and downloadable reports.
- (7) To evaluate and compare the developed machine learning models using train-test splitting, 10-fold cross-validation, and performance metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2 Score, and model training time.
- (8) To extend the decision-support framework with integrated agricultural assistance including plant disease identification, agronomic explanations, farm-level and regional analytics, spray pump dosage calculation, government scheme and subsidy information, and advisory sharing.

The scope of the research covers agricultural data collection, data preprocessing, exploratory analysis, feature selection, machine learning-based crop yield prediction, crop recommendation, real-time weather integration, plant disease identification, agronomic explanation, farm and regional analytics, agricultural input calculation, visualization, and report generation. The proposed crop yield prediction and recommendation framework is intended to function as a decision-support tool that provides data-driven insights for farmers and agricultural stakeholders. The multilingual web interface improves accessibility by presenting agricultural predictions, recommendations, and supporting information in English, Hindi, and Marathi, while the integration of weather and farm-level information enables more contextual agricultural analysis. The research primarily focuses on developing and evaluating machine learning-based predictive models and integrating practical agricultural decision-support capabilities using available agricultural and environmental data. The system is intended to support, rather than replace, expert agricultural knowledge and field-level decision-making.

II. LITERATURE REVIEW

Crop yield prediction has been studied using different machine learning (ML) techniques and combinations of agricultural, soil, weather, and remote-sensing data. Earlier research shows that prediction performance varies with the crop, region, input features, dataset size, and evaluation method [3]. Therefore, no single ML model can be considered universally best for all agricultural datasets.

Screpnik *et al.* reviewed artificial intelligence techniques used in agriculture, including Random Forest, XGBoost, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), and highlighted the need for practical AI-based decision-support systems [4]. A broader systematic review also reported the growing use of machine learning and deep-learning methods for crop yield prediction while showing considerable variation in datasets and experimental approaches [9].

Tree-based ensemble models are particularly suitable for structured agricultural data. Random Forest combines the predictions of multiple decision trees to improve robustness, while Gradient Boosting builds models sequentially to reduce prediction errors. XGBoost further improves the boosting approach through efficient training and regularization. Previous crop-yield studies have evaluated these and other ML techniques on agricultural datasets [2][7][8]. The 2025



ICDICI study also compared different ML algorithms for crop-yield prediction [15]. These findings support the need to compare multiple models using the same dataset and evaluation procedure before selecting a suitable model.

TABLE I: Comparison of existing studies on crop-yield prediction

Study	Method	Data/application	Key contribution	Limitation or design implication
Sundararajan <i>et al.</i> , 2025 [1]	Data-driven ML	Agricultural and environmental data	Used several agricultural inputs to predict crop yield.	A complete farmer-oriented web system has not yet been developed.
Ranjani <i>et al.</i> , 2022 [2]	Machine learning	Crop yield data	Showed the use of ML methods for yield prediction.	Wider regional testing and decision-support features are required.
van Klompenburg <i>et al.</i> , 2020 [3]	Systematic review	ML and deep-learning studies	The common models, crops, and input features are summarized.	The review does not provide any deployed applications.
Screpnik <i>et al.</i> , 2025 [4]	AI systematic review	Yield prediction and optimization	Reviewed RF, XGBoost, SVM, KNN, and related AI methods.	Practical integrated systems remain important.
Hoque <i>et al.</i> , 2024 [5]	ML regression	Weather and pesticide data	This study demonstrated the value of combining weather and pesticide information.	The Findings depend on the selected dataset and inputs.
Badshah <i>et al.</i> , 2024 [10]	Classification and regression	Crop classification and yield	Combined crop classification and yield prediction.	A farmer-oriented multilingual dashboard was not the main focus.
Ashfaq <i>et al.</i> , 2024 [11]	ML with data fusion	Wheat climate and NDVI data	Combined climate and vegetation information for wheat prediction.	This work is crop- and region-specific.
Waqar <i>et al.</i> , 2025 [13]	Stacking ensemble	Real and synthetic agricultural data	Stability is improved by combining models and synthetic data.	The model structure is more complex for simple deployment.
Sarkar <i>et al.</i> , 2024 [12]	Meta-transformer and temporal GNN	Combined agricultural data	Advanced models were used for complex agricultural information.	High data and computing needs may limit lightweight use.
MirhoseiniNejad <i>et al.</i> , 2024 [14]	ConvLSTM-ViT	Satellite and earth-observation data	Used time-based satellite information for yield prediction.	It requires remote-sensing data and stronger hardware.

Recent research indicates three major directions in crop yield prediction. First, combining multiple data sources can provide a more comprehensive representation of crop conditions. Hoque *et al.* incorporated weather and pesticide information [5], while Ashfaq *et al.* combined climate and NDVI data for wheat-yield prediction [11]. Second, crop classification and yield prediction can be integrated into a single agricultural solution, as demonstrated by Badshah *et al.* [10]. Third, advanced multimodal and remote-sensing approaches can provide detailed predictions but generally require greater data-processing and computational resources [12][14]. Research on sustainable agriculture further demonstrates the increasing role of ML-based crop yield prediction [6].

The reviewed studies provide a strong foundation for selecting suitable ML techniques and agricultural parameters. However, their reported performance values are not directly used as benchmarks for the proposed system because the datasets, geographical regions, crops, and experimental settings are different.

A. RESEARCH GAP IDENTIFIED

As per literature survey it is observed that existing crop-yield prediction studies mainly focus on model development and accuracy, while many are limited to specific crops, regions, or datasets. Advanced approaches may also require large datasets and higher computational resources. There is a practical gap in integrating yield prediction, weather information,



crop recommendation, plant disease detection, agronomic explanations, farm analytics, spray dosage calculation, government scheme information, visualization, reports, and regional-language support into one farmer-friendly system. The proposed work addresses this gap by comparing Random Forest, XGBoost, and Gradient Boosting and integrating the selected model into a multilingual web-based decision-support platform.

III. PROPOSED SYSTEM WORKING

The proposed crop yield prediction system follows a data-driven approach for estimating crop yield using agricultural and environmental parameters. Agricultural, soil, weather, fertilizer, irrigation, crop, location, and yield data are collected and preprocessed before training three machine learning models: Random Forest, XGBoost, and Gradient Boosting. All models are trained and evaluated using the same dataset and evaluation procedure. The best-performing model is integrated into a web application that provides crop-yield prediction, crop recommendations, weather information, graphical analysis, disease detection, agricultural advisory support, and report generation.

A. DATA COLLECTION AND PREPROCESSING

The agricultural and environmental data are collected from the Kaggle Crop Yield Prediction Dataset, government sources, research repositories, and a weather API. The main inputs include rainfall, temperature, humidity, soil pH, N-P-K nutrients, fertilizer, irrigation, crop type, location, and season, with crop yield as the target. The data is cleaned by handling missing values, duplicates, and inconsistent values. Categorical features are converted into numerical form, and feature analysis is performed to identify important factors affecting yield.

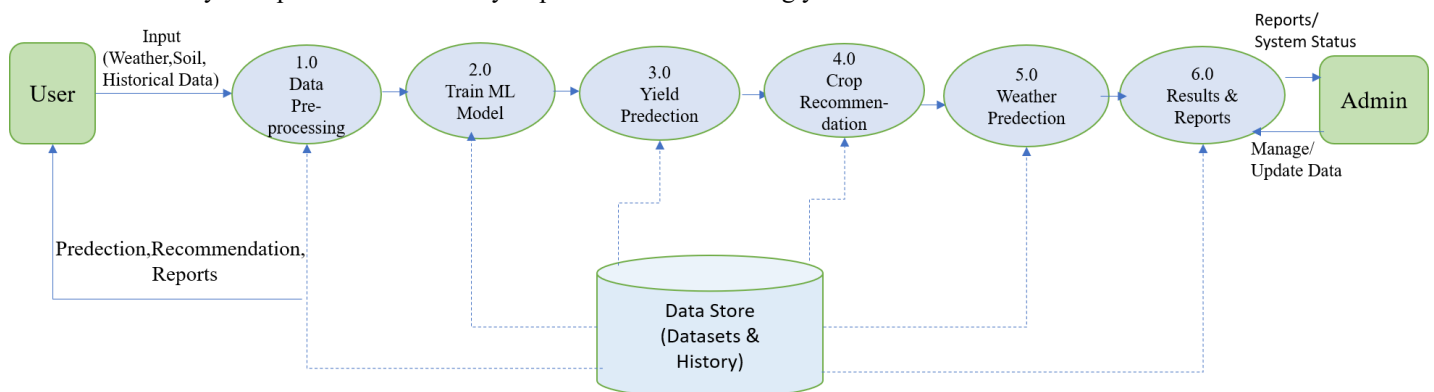


Fig. 1. Proposed crop yield prediction and decision-support architecture

B. PREPROCESSING AND FEATURE ENGINEERING

The collected data is preprocessed by converting categorical variables such as crop type, soil type, irrigation, season, and region into numerical form. Missing and abnormal values are handled appropriately, while numerical scaling is applied only when required. Feature analysis is performed to identify important relationships between agricultural factors and crop yield. Relevant features such as seasonal weather conditions may also be derived from the available data to improve model performance.

C. MODEL DEVELOPMENT AND TRAINING

The prepared dataset is divided into 80% training and 20% testing data. Random Forest, XGBoost, and Gradient Boosting algorithms are trained using the same training data. Hyperparameters are tuned using the training set, while 10-fold cross-validation is used to examine model stability and reliability. For crop recommendation, a Random Forest classifier uses soil and weather parameters such as N, P, K, temperature, humidity, pH, and rainfall to recommend a suitable crop. The recommendation is also supported by an agronomic explanation describing the main soil and environmental factors behind the suggested crop. Random Forest combines predictions from multiple decision trees, while Gradient Boosting and XGBoost sequentially build trees to reduce prediction errors. The models are compared using MAE, MSE, RMSE, and R^2 , along with training time. The model providing reliable performance is selected for deployment.

D. MODEL EVALUATION

The performance of the developed regression models is evaluated using the following measures:



$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (1)$$

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

$$RMSE = \sqrt{MSE}, \quad (3)$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}. \quad (4)$$

MAE represents the average difference between the actual and predicted crop yield, while MSE measures the squared prediction error. RMSE gives higher importance to larger errors, making it useful for identifying significant prediction deviations. The R^2 score indicates how effectively the model explains variations in crop yield. Training time is also considered to compare the computational efficiency of the models. The final 20% testing data is kept separate during model selection and is used only for the final evaluation of the selected model.

E. WEB APPLICATION AND DECISION SUPPORT

The selected machine learning model is integrated into a Flask-based web application. The frontend is developed using React.js, Tailwind CSS, HTML, CSS, JavaScript, while MySQL is used for data storage. Pandas, NumPy, Scikit-learn, and XGBoost support data processing and model development. The application allows users to enter farm and environmental information and obtain predicted crop yield. It also provides crop recommendations, current location-based weather information, historical trends, charts, and downloadable reports. Additional decision-support features include plant disease detection through leaf images, spray-pump dosage calculation, government scheme and MahaDBT subsidy assistance, and agricultural guidance. The system can also provide the information through English, Hindi, and Marathi, with voice input and WhatsApp sharing to improve accessibility and communication.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed models were evaluated using classification accuracy for crop recommendation and MAE, MSE, RMSE, R^2 score, cross-validation, and training time for crop yield prediction. For crop recommendation, Random Forest achieved an accuracy of 99.55%, followed by XGBoost (98.86%), SVM (98.41%), Decision Tree (97.95%), and KNN (97.73%). Based on these results, Random Forest was selected for crop recommendation.

The final Random Forest model for crop yield prediction achieved an MAE of 0.861, an RMSE of 1.061, and an R^2 score of 0.202. The developed models were integrated into KrushiMitra to provide crop yield prediction and crop recommendations.

KrushiMitra also provides weather information, historical records, farm analytics, disease assistance, spray calculation, government scheme information, visualizations, downloadable reports, voice input, and an AI chatbot. Its multilingual interface supports English, Hindi, and Marathi, improving accessibility for farmers and other agricultural users.

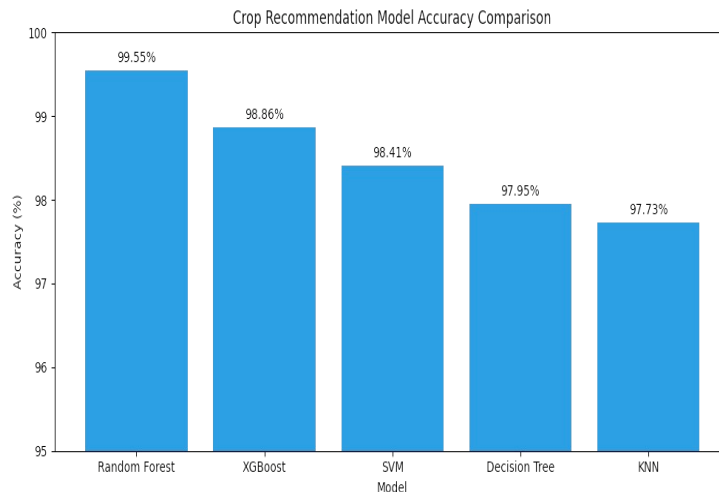


Fig. 2. Crop Recommendation Model Accuracy

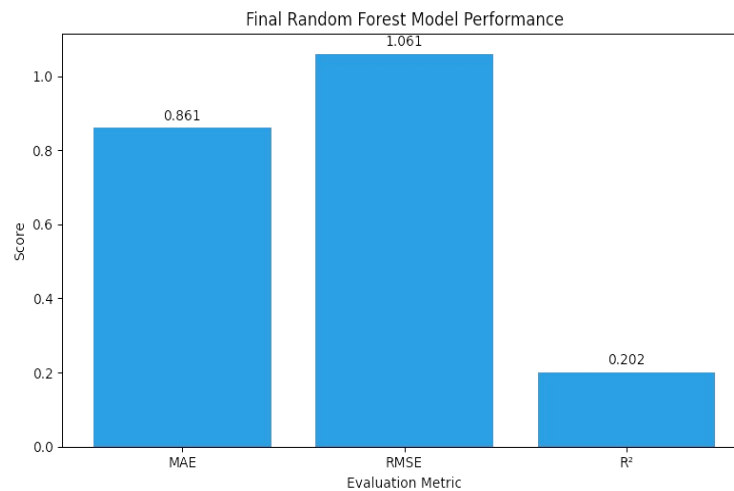


Fig. 3. Final Random Forest Performance

V. CHALLENGES AND LIMITATIONS

- 1) *Data Quality and Regional Suitability*: Differences in data units, locations, time periods, missing values, and outdated records may affect prediction accuracy.
- 2) *Weather API Dependency*: API failures or unavailable weather data may affect the delivery of real-time weather information.
- 3) *Data Security and Privacy*: Proper database security and protection of agricultural data are important considerations for web-based deployment.
- 4) *Limited Recommendation Factors*: Crop recommendations depend on available input data and may not fully consider market demand, local farming practices, or other field-specific conditions.

VI. CONCLUSION

Crop yield prediction systems have shown their significance in modern agriculture by supporting farmers in better crop planning and resource management. The proposed system uses Random Forest, XGBoost, and Gradient Boosting with agricultural and weather data to estimate crop yield. It further combines crop recommendation, weather information, farm analytics, disease-related assistance, spray dosage calculation, government scheme information, charts, reports, and multilingual support in a single web-based platform. However, proper data quality, model evaluation, and regional testing are necessary to ensure reliable predictions. With accurate data, systematic testing, and user feedback, the proposed system can provide a practical and integrated solution for data-driven agricultural planning and farmer support.

REFERENCES

- [1] S. Sundararajan, N. Belagalla, O. I. Al Mrayat, M. Saqib, S. Malhotra, and J. P. Jayabharathi, "Data-driven approach for crop yield prediction using machine learning techniques," in Proc. 2025 Int. Conf. Intelligent Computing and Smart Communication Systems (ICICSCS), IEEE, 2025, doi: 10.1109/ICICSCS65874.2025.11032031.
- [2] J. Ranjani, V. K. G. Kalaiselvi, A. Sheela, D. S. Deepika, and G. Janaki, "Crop yield prediction using machine learning," in Proc. IEEE Int. Conf. Smart Technologies and Systems for Next Generation Computing (ICSTSN), 2022, doi: 10.1109/ICSTSN53084.2022.9711853.
- [3] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Computers and Electronics in Agriculture*, vol. 177, Art. no. 105709, 2020, doi: 10.1016/j.compag.2020.105709.
- [4] C. R. Screpnik, E. Zamudio, and L. I. Gimenez, "Artificial intelligence in agriculture: A systematic review of crop yield prediction and optimization," *IEEE Access*, vol. 13, pp. 70691–70697, 2025, doi: 10.1109/ACCESS.2025.3560631.
- [5] M. J. Hoque, S. Islam, J. Uddin, A. Samad, B. Sainz de Abajo, D. L. Ramírez Vargas, and I. Ashraf, "Incorporating meteorological data and pesticide information to forecast crop yields using machine learning," *IEEE Access*, vol. 12, pp. 47768–47786, 2024, doi: 10.1109/ACCESS.2024.3383309.



- [6] "Crop yield prediction: Leveraging machine learning for sustainable agriculture," in Proc. 2024 2nd Int. Conf. Self Sustainable Artificial Intelligence Systems (ICSSAS), IEEE, 2024, doi: 10.1109/ICSSAS64001.2024.10760847.
- [7] "Significance of machine learning in crop yield prediction," in Proc. 2024 2nd World Conf. Communication and Computing (WCONF), IEEE, 2024, doi: 10.1109/WCONF61366.2024.10692141.
- [8] "Crop yield prediction based on machine learning techniques," *IEEE Conference Proceedings*, 2023, Art. no. 10245132.
- [9] "Crop yield prediction using machine learning: An extensive and systematic literature review," *Smart Agricultural Technology*, Elsevier, 2024, Art. no. S2772375524003228.
- [10] A. Badshah, B. Y. Alkazemi, F. Din, K. Z. Zamli, and M. Haris, "Crop classification and yield prediction using robust machine learning models for agricultural sustainability," *IEEE Access*, vol. 12, pp. 162799–162813, 2024, doi: 10.1109/ACCESS.2024.3486653.
- [11] M. Ashfaq, I. Khan, A. Alzahrani, M. U. Tariq, H. Khan, and A. Ghani, "Accurate wheat yield prediction using machine learning and climate–NDVI data fusion," *IEEE Access*, vol. 12, pp. 40947–40961, 2024, doi: 10.1109/ACCESS.2024.3376735.
- [12] S. Sarkar *et al.*, "Crop yield prediction using multimodal meta-transformer and temporal graph neural networks," *IEEE Transactions on AgriFood Electronics*, vol. 2, no. 2, pp. 545–553, 2024, doi: 10.1109/TAFE.2024.3438330.
- [13] M. Waqar, Y. W. Kim, and Y. C. Byun, "A stacking ensemble framework leveraging synthetic data for accurate and stable crop yield forecasting," *IEEE Access*, vol. 13, pp. 136909–136926, 2025, doi: 10.1109/ACCESS.2025.3591802.
- [14] S. M. MirhoseiniNejad, D. Abbasi-Moghadam, and A. Sharifi, "ConvLSTM–ViT: A deep neural network for crop yield prediction using earth observations and remotely sensed data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 17489–17502, 2024, doi: 10.1109/JSTARS.2024.3464411.
- [15] "Comparative performance evaluation of machine learning algorithms for crop yield prediction," in Proc. 2025 Int. Conf. Data Intelligence and Computing Innovations (ICDICI), IEEE, 2025, doi: 10.1109/ICDICI66477.2025.11135411.
- [16] "An approach towards cotton quality determination by using SVM algorithm," in Proc. 2nd World Congress on Optimization & Automation Techniques in Engineering (WCOATE-2022), G. H. Rasoni University, Amravati, India, Mar. 11–12, 2022.