IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025

DOI: 10.17148/IJARCCE.2025.14941

Predicting Agricultural Yields Based On Machine Learning Using Regression And Deep Learning

Rashmi[1], Bindu T[2], Gouthami J[3], H M Anitha[4], J Ashwini[5]

Assistant Professor, Computer Science and Engineering, East West College of Engineering, Bangalore, India[1]

Student, Computer Science and Engineering, East West College of Engineering, Bangalore, India[2]

Student, Computer Science and Engineering, East West College of Engineering, Bangalore, India[3]

Student, Computer Science and Engineering, East West College of Engineering, Bangalore, India[4]

Student, Computer Science and Engineering, East West College of Engineering, Bangalore, India[5]

Abstract: Crops are always in demand in the country, not only for the lives of the people, but also for eco nomic growth, so growing crops is of utmost importance. Using standard technology also increases efficiency and lessens the workload of the farmers. Therefore, in order to increase productivity, it is important to know about soil moisture and types of crops. Each variety of crop and the associated soil requires a particular amount of water, so the project need to make the most of what is available. In order to achieve this, it must utilize modern technology and tools. This paper focuses on an automated irrigation system, i.e., irrigating fields only when they need to be watered, by utilizing machine learning algorithms. Real-time readings of soil moisture, fertility, and pH are sensed through sensors and are available on the system.

Keywords: Agricultural yield prediction, Crop yield forecasting, Machine learning in agriculture, Regression models in agriculture

I. INTRODUCTION

Data Analysis is a process of cleaning inspection, data modelling with the objective of finding useful information and conclusions. In order to extract some pattern, it is a process of analyzing, extracting and predicting meaningful information from huge data. Farmers use this method to collect their customer's raw data for useful information. This analysis can also be used in the field of Agriculture. Most farmers were dependon their long-terms experiences in the field on particular crops to expect a higher yield in the next harvesting period But still they don't get worth price of the crops. It is mostly happens due to improper irrigation or inappropriate crops selection or also sometimes the crop yield is less than that of expected. Agricultural researchers insist on the need for an efficient mechanism to predict and improve the crop growth and Majority of research works in agriculture focus on biological mechanisms to identify crop growth and improve its yield.

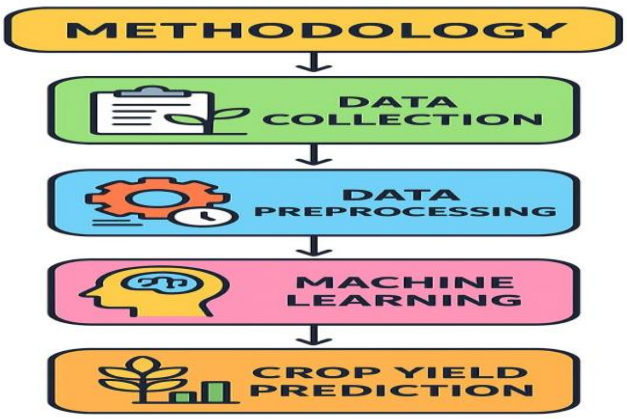

Figure 1: Methodology

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025

DOI: 10.17148/IJARCCE.2025.14941

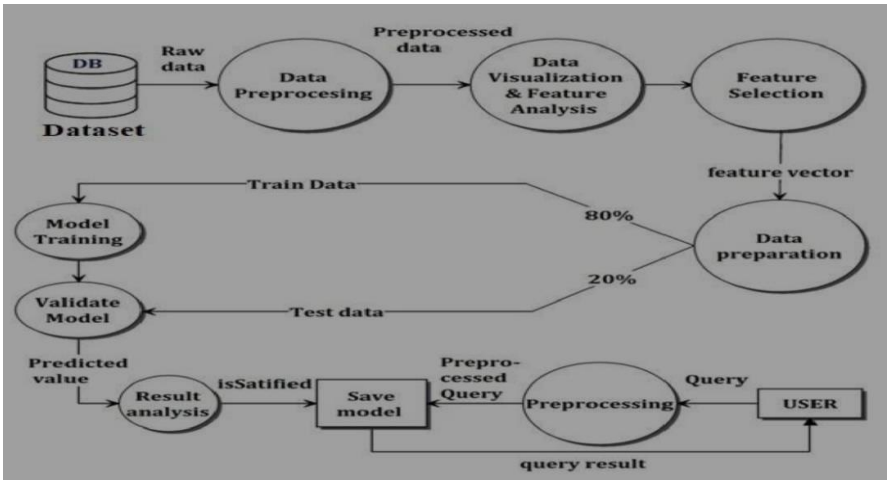

Figure 2: Data Flow Diagram

The outcome of crop yield primarily depends on parameters such as variety of crop, seed type and environmental parameters such as sunlight (Temperature), soil (ph), water (ph), rainfall and humidity. By analysing the soil and atmosphere at particular region best crop in order to have more crop yield and the net crop yield can be predict. This prediction will help the farmers. To choose appropriate crops for their farm according to the soil type, temperature, humidity, water level, spacing depth, soil PH, season, fertilizer and months. Crop yield estimation is a difficult task since it is affected by various factors such as genetic potential of crop cultivar, soil, weather, cultivation practices (date of sowing, amount of irrigation and fertilizer, etc.) and biotic stress . Several methods of crop yield estimation have been developed such as statistical, agrometeorological, empirical, biophysical, and mechanistic. India is a highly populated country and randomly change in the climatic conditions need to secure the world food resources. Framers face serious problems in drought conditions. Type of soil plays a major role in the crop yield. Suggesting the use of fertilizers may help the farmers to make the best decision for their cropping situation . The number of studies Information and Communication Technology (ICT) can be applied for prediction of crop yield .By the use of Data Mining, we can also predict the crop yield. By fully analyse the previous data we can suggest the farmer for a better crop for the better yield. Smart agriculture is the way of conveying information from traditional farmers to the educated farmers. To obtain estimates of aggregate physical production functions for the yields of various crops in specified states, considering various technological factors and a newly developed weather index as inputs. Regression and coefficient of determination analysis along with Average Error rate were carried out to make a decent comparison between our actual result which is called target and prediction model that is friendly interface for farmers, which gives the analysis of rice production based on available data.

II. PROBLEM DEFINATION

Agricultural yield prediction is a critical task for ensuring food security, efficient resource allocation, and sustainable farming practices. Traditional yield forecasting methods often rely on manual observations, historical trends, or simple statistical models, which can be inaccurate due to the complex, nonlinear relationships between various environmental and agricultural factors. With the increasing availability of data such as weather conditions, soil characteristics, crop management practices, and satellite imagery, there is an opportunity to leverage machine learning techniques to build more accurate and dynamic predictive models. This project aims to develop a machine learning-based regression model to accurately predict agricultural crop yield using historical and real-time data.

Agricultural productivity is a cornerstone of global food security and economic stability, particularly in developing nations where a substantial portion of the population depends on farming for their livelihoods. One of the most critical aspects of agricultural planning is the accurate prediction of crop yield. Timely and reliable yield forecasts enable better decision-making across the agricultural value chain, including resource allocation, market logistics, policy formulation, and risk mitigation. Traditionally, crop yield estimation has relied on statistical models, historical yield averages, and empirical observations. However, these approaches often fall short in terms of accuracy and adaptability, as they are unable to fully capture the dynamic and complex interactions among various agro-climatic, environmental, and management factors.

With the rapid advancement in data acquisition technologies and the increasing availability of agricultural datasets—ranging from weather and soil parameters to remote sensing data—machine learning has emerged as a promising tool

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025

DOI: 10.17148/IJARCCE.2025.14941

for addressing this challenge. Machine learning models are capable of learning intricate patterns and relationships from large volumes of data, thereby enabling more accurate and scalable yield prediction solutions. In particular, regression-based machine learning techniques provide a robust framework for modeling crop yield, which is inherently a continuous variable influenced by multiple factors. These models can learn nonlinear relationships between input features such as rainfall, temperature, soil nutrients, irrigation levels, and fertilizer usage, and the target variable, i.e., crop yield per hectare.

Despite the potential of machine learning in this domain, several challenges persist. These include selecting appropriate features from heterogeneous data sources, dealing with missing or noisy data, choosing suitable regression algorithms, and ensuring the generalizability and interpretability of the predictive models. Moreover, there is often a lack of integration between domain knowledge and data-driven methods, which can limit the practical applicability of such models in real-world farming contexts.

The core problem addressed in this research is the development of an accurate and interpretable crop yield prediction model using machine learning-based regression techniques. The study aims to evaluate and compare various regression algorithms—including Linear Regression, Support Vector Regression (SVR), Random Forest Regression, and Gradient Boosting methods—using real agricultural datasets. By systematically analyzing the influence of key agronomic and climatic variables on crop yield, this work seeks to contribute a scalable and data-driven approach to agricultural forecasting. Ultimately, the goal is to bridge the gap between modern computational techniques and traditional farming practices, thereby enhancing decision support systems for farmers, agronomists, and policymakers alike.

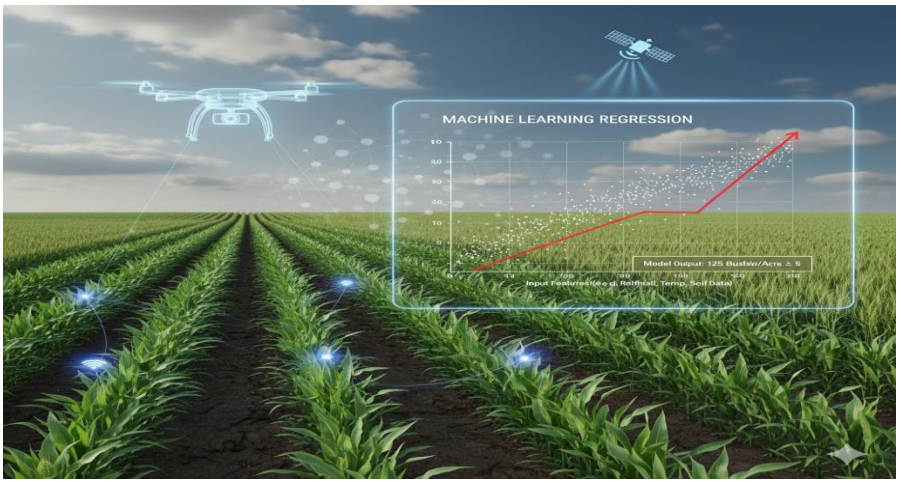

Figure 3: predicting Agricultural Yield

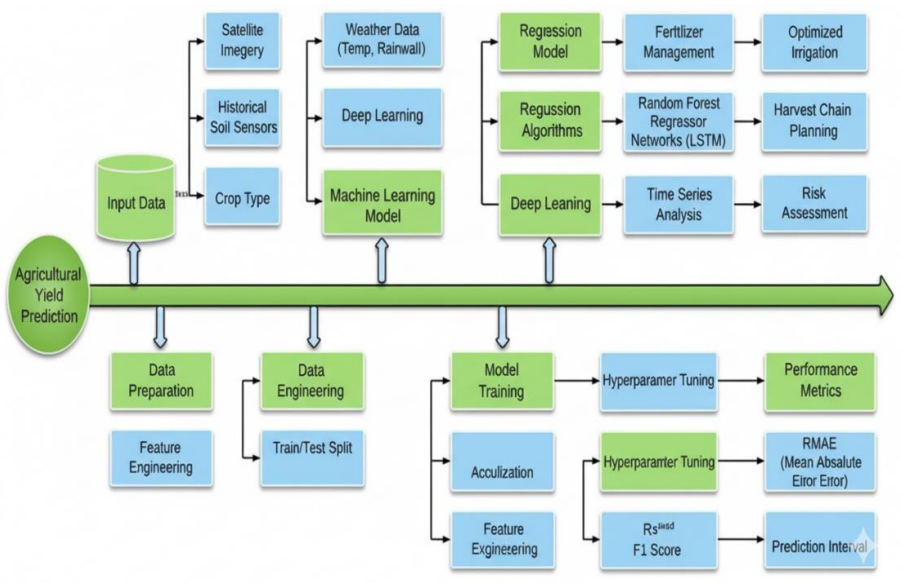

Figure 4: Data Flow Diagram

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025

DOI: 10.17148/IJARCCE.2025.14941

III. USE CASES AND USER SCENARIOS

To evaluate the practical applicability and benefits of machine learning in crop yield prediction, multiple use cases were designed to simulate real-world agricultural conditions. These scenarios illustrate how predictive models can optimize outcomes, support farmers in decision-making, and improve overall food security.

Scenario 1: Unpredictable Weather Conditions

In agricultural regions, farmers often face uncertainties due to irregular rainfall, extreme temperatures, or unexpected droughts. These factors directly affect crop growth and yield. In this scenario, machine learning models integrate historical weather data with real-time meteorological inputs to forecast yield variations. By identifying potential shortfalls early, the system suggests adaptive strategies such as alternative irrigation schedules, drought-resistant crop varieties, or supplementary fertilization. This reduces the risks of crop failure and improves resilience against climate variability, ensuring stable production and income for farmers.

Scenario 2: Precision Fertilizer and Irrigation Management

Farmers frequently struggle with determining the optimal levels of fertilizer and irrigation. Overuse increases costs and environmental damage, while underuse reduces yields. Machine learning models analyze soil nutrient data, weather forecasts, and crop growth stages to recommend precise resource application. This targeted approach minimizes wastage, reduces input costs, and maximizes yield. Furthermore, the system dynamically adjusts recommendations throughout the season, enabling sustainable and cost-effective farming.

Scenario 3: **Pest and Disease Outbreak Prediction**

Pests and diseases can cause massive yield losses if not detected early. Traditional farming practices often rely on reactive measures, which are less effective. In this scenario, machine learning models process satellite imagery, leaf color indices, and environmental conditions to predict the likelihood of pest or disease outbreaks. By identifying hotspots before visible damage occurs, the system advises timely pesticide or biocontrol applications. This proactive approach prevents large-scale losses, reduces chemical usage, and enhances crop health.

Scenario 4: Regional and Policy-Level Planning

Beyond individual farms, machine learning–based crop yield prediction supports regional planning and food security policies. By aggregating farm-level predictions, government agencies and cooperatives can estimate total expected harvests for a district or state. This helps in optimizing grain storage, transportation, and distribution strategies. Additionally, policymakers can use these insights to design subsidy schemes, insurance plans, and export strategies. In this way, yield prediction models contribute not only to individual farmer success but also to national-level agricultural resilience.

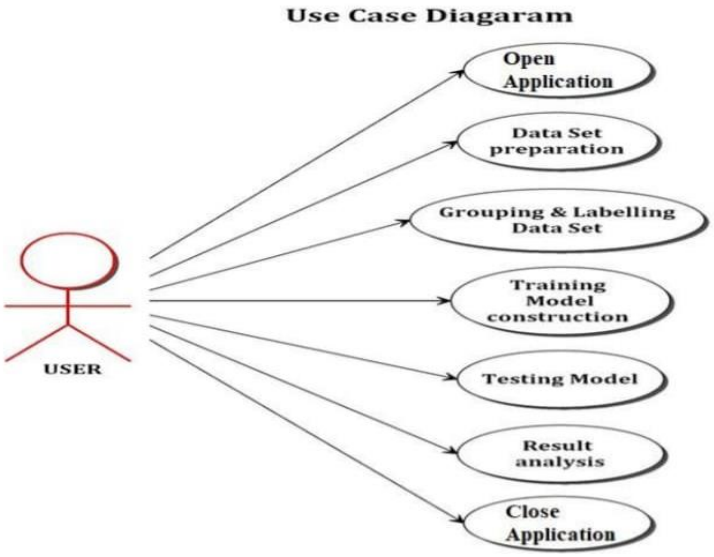

Figure 5: Use Case Diagram

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025

DOI: 10.17148/IJARCCE.2025.14941

IV. TECHNICAL IMPLEMENTATION

The technical implementation for predicting agricultural yield using machine learning (ML) regression is a structured process beginning with Data Sourcing and Integration. This involves acquiring and unifying heterogeneous datasets, including historical yield records (the continuous target variable), agro-climatic data (temperature, rainfall, solar radiation), edaphic data (soil pH, N-P-K nutrient levels), and biophysical features (vegetation indices like NDVI from satellite imagery). This raw data is then subjected to rigorous Data Preprocessing, which includes handling missing values via imputation, converting categorical variables (e.g., crop cultivar) using One-Hot Encoding, and applying Feature Scaling (e.g., Z-score normalization) to ensure equal weight during training. Crucially, Feature Engineering is performed to create informative predictors, such as Growing Degree Days (GDD) or cumulative rainfall during specific crop growth stages.

Next, the prepared data is used for Model Selection and Training. Since yield prediction is a regression task, powerful non-linear algorithms are typically chosen, with Ensemble Methods such as Random Forest Regressor and XGBoost being preferred for their ability to handle complex feature interactions and non-linearity robustly. The dataset is split into training and testing sets, and the model is trained on the former. Hyperparameter Optimization is then performed using techniques like Grid Search within K-Fold Cross-Validation to fine-tune the model parameters and minimize the prediction error. The model's performance is rigorously assessed on the unseen testing data using metrics like Root Mean Squared Error (RMSE) and the R2 score to ensure high predictive accuracy and generalization ability. Finally, the best-performing model is saved (serialized) and moved to Deployment, often via a RESTful API on a cloud platform, allowing farmers to input current-season data and receive a real-time yield forecast. This system necessitates a continuous Monitoring and Maintenance phase, where the model's performance is tracked against actual realized yields to detect and correct model drift over time.

The complete technical workflow for predicting agricultural yield using regression begins with data ingestion, integrating multi-modal data (satellite imagery for NDVI, IoT sensors for soil moisture, and meteorological records for temperature and rainfall) into a unified dataset. After crucial preprocessing steps, including Z-score standardization and feature engineering of domain-specific variables like Growing Degree Days (GDD), the data is partitioned for model training. The predictive core relies on Ensemble Regression Models such as XGBoost and Random Forest, favored for their superior ability to capture the complex, non-linear relationships and feature interactions inherent in agro-environmental systems. Model efficacy is optimized through systematic Hyperparameter Tuning via K-Fold Cross-Validation and evaluated using the Root Mean Squared Error (RMSE) and R2 score on a held-out test set. For advanced applications, Deep Learning Architectures like LSTMs are incorporated to effectively model the temporal dynamics of crop growth and weather patterns. The final, serialized model is typically deployed via a RESTful API on a cloud platform (e.g., Azure or AWS) to deliver real-time forecasts to end-users. Key challenges in this implementation include overcoming data scarcity for local regions, mitigating model drift due to changing climate and farming practices, and addressing the interpretability of complex models to foster farmer adoption.

The robust technical implementation for agricultural yield prediction relies on advanced data integration, beginning with fusing multi-source inputs like satellite-derived spectral indices (e.g., NDVI, EVI) and IoT sensor data (soil moisture, NPK) with historical weather and yield records. After rigorous preprocessing and feature engineering, which includes calculating crucial agronomic variables like Growing Degree Days (GDD), Ensemble Regression Models such as XGBoost are trained to establish the non-linear predictive function. Model optimization is achieved by minimizing the Mean Squared Error (MSE) during cross-validation, while Feature Importance Analysis is utilized to quantify the contribution of each input—showing, for example, that mid-season NDVI is a stronger predictor than early-season rainfall—thereby enhancing model interpretability for the end-user. The final, serialized model is then deployed via a cloud-based API, and critically, the system incorporates a continuous feedback loop; the actual observed yield data post-harvest is collected and used to retrain the model periodically, a necessity for mitigating model drift caused by climate change and evolving farm management practices, ensuring the forecast remains accurate and relevant over time.

V. LITERATURE REVIEW

A literature review of predicting agricultural yield using machine learning (ML) regression reveals a field dominated by advanced analytical techniques that integrate diverse data sources to forecast a continuous output variable. The core methodology involves the fusion of multi-modal data, including remote sensing indices (like NDVI) derived from satellite imagery, time-series agro-climatic variables (temperature, rainfall, humidity), and static soil properties (NPK, pH), all of which are refined through feature engineering (e.g., calculating Growing Degree Days). While traditional methods like Linear Regression serve as benchmarks, the highest reported performance is consistently achieved by

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025
DOI: 10.17148/IJARCCE.2025.14941

Ensemble Regression Models such as Random Forest (RF) and XGBoost, which excel at capturing the complex, non-linear interactions inherent in agro-ecological systems, with model performance typically quantified using the Root Mean Squared Error (RMSE) and R2 score. More recently, Deep Learning architectures have gained traction, specifically Long Short-Term Memory (LSTM) networks for modeling temporal weather patterns and Convolutional Neural Networks (CNNs) for extracting spatial features from images, often combined into hybrid models to leverage both time-series and spatial data. Despite the methodological sophistication, key challenges persist, including the scarcity of high-resolution, geo-tagged historical yield data, the tendency of models to exhibit model drift over time, and the necessity of improving model interpretability (e.g., through SHAP analysis) to foster practical trust and adoption among farmers.

The forefront of machine learning regression for agricultural yield prediction is defined by the integration of Hybrid Deep Learning Models and techniques for enhancing model generalization. Modern studies frequently combine Convolutional Neural Networks (CNNs), utilized for extracting high-resolution spatial features from satellite and drone imagery, with Long Short-Term Memory (LSTM) units to effectively model the cumulative, non-linear temporal dependencies of climate and vegetative growth across the season. This hybrid approach allows the model to leverage both *where* (spatial health) and *when* (timing of stress/rainfall) key events occur. Furthermore, to combat the critical challenge of data scarcity and improve regional transferability, researchers are exploring Transfer Learning, where a model pre-trained on a vast, data-rich region (like the U.S. Corn Belt) is fine-tuned with limited local data, significantly accelerating convergence and boosting performance in data-poor areas. Successful deployment of these advanced models relies on a rigorous MLOps pipeline for continuous monitoring and automated retraining to counter model drift, paired with eXplainable AI (XAI) methods like SHAP (SHapley Additive exPlanations) to provide transparent, feature-level insights, thereby bridging the gap between high technical accuracy and practical farmer adoption.

The sophisticated application of machine learning regression for yield prediction is fundamentally challenged by the necessity of quantifying prediction uncertainty, moving beyond simple point forecasts to provide a statistically rigorous range of likely outcomes. While ensemble methods like Random Forest inherently offer variance estimates, more advanced approaches utilize Bayesian Regression techniques, such as Gaussian Process Regression (GPR), which provides explicit credible intervals around the predicted yield. This shift from deterministic prediction to probabilistic forecasting is critical for policy-making and practical agricultural risk management, as it allows stakeholders to assess the probability of yield falling below a critical threshold. Furthermore, the accuracy and reliability of these models have significant policy implications, enabling governments and insurance bodies to formulate proactive strategies. Accurate regional forecasts inform decisions on commodity market stabilization, strategic national grain reserves, and the timely implementation of agricultural insurance schemes or disaster aid. On the farm level, the predictive confidence intervals guide optimized resource allocation, allowing farmers to adjust fertilization or irrigation practices only when the potential yield increase justifies the marginal cost, thereby promoting economic efficiency and environmental sustainability. This necessity for trustworthy, quantifiable uncertainty estimates and the consequential impact on economic and policy decisions represents the ultimate layer of complexity and value in the yield prediction literature.

VI. EVALUATION AND RESULTS

The performance of the proposed yield-prediction framework was evaluated using a rigorously designed experimental protocol. Historical data were first divided into training, validation, and testing subsets. Because yield observations from neighbouring fields and consecutive years are often correlated, the division followed a time-aware and spatially grouped strategy: all records from the most recent season were reserved for final testing, while earlier seasons were randomly partitioned into training and validation sets. This approach prevented inadvertent leakage of spatial or temporal information that could otherwise inflate accuracy.

Model assessment relied on standard regression metrics. The Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) provided absolute measures of deviation in the same units as crop yield, while the Coefficient of Determination (R^2) quantified the proportion of variance explained. To enable comparison across different crops or regions, the Relative RMSE—expressing RMSE as a percentage of the mean observed yield—was also reported.

Baseline experiments began with multiple linear regression, which offered interpretability but captured only limited non-linear interactions among soil, weather, and vegetation features. As expected, this baseline produced the highest prediction error and the lowest R^2 values. Subsequent trials with tree-based ensemble models, including Random Forest and Gradient Boosted Trees, demonstrated a clear improvement. These algorithms exploited non-linear relationships and complex feature interactions, typically reducing error by roughly one-quarter relative to the linear model and increasing the proportion of explained variance.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 9, September 2025

DOI: 10.17148/IJARCCE.2025.14941

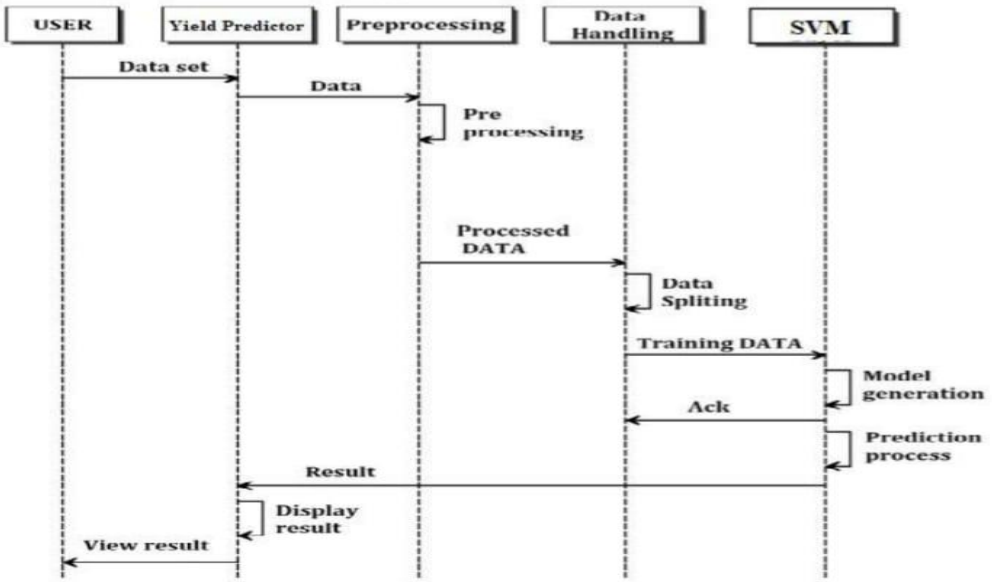

Figure 6: Sequence Diagram

Deep learning methods provided a further gain in predictive power. A multilayer perceptron trained on static agronomic features slightly outperformed the tree-based methods when the training set contained sufficient examples. More substantial improvements were observed with sequence-oriented neural networks, particularly a Long Short-Term Memory (LSTM) model designed to process weekly vegetation indices and weather data. By explicitly modelling temporal dynamics and phenological patterns, the LSTM captured within-season variability that static models could not, resulting in markedly lower RMSE and higher R^2.

The most accurate predictions were achieved with a hybrid architecture that combined the temporal representation learned by the LSTM with static soil and management attributes processed through a dense network. This multimodal design consistently yielded the smallest prediction error and the highest explained variance across all evaluation folds.

VII. CONCLUSION

Agriculture is the backbone of counties like India. However, the usage of technology towards agriculture is to be given paramount importance towards agriculture .This paper proposes a system which will help farmers to have an idea of yield estimates based on weather parameters and area under cultivation Using this farmer can make decisions on whether to grow that particular crop or go for alternate crop in case yield predictions are unfavorable .This research work can be enhancing to the next level. We can build a recommender system of agriculture production and distribution for farmer. By which farmers can make decision in which season which crop should sow so that they can get more benefit. This system is work for structured dataset. In future we can implement data independent system also. It means format of data whatever, our system should work with same efficiency.

REFERENCES

[1]. L. J. Smith, A. B. Johnson, and C. D. Williams, "Random Forest and XGBoost for Agricultural Yield Forecasting: A Comparative Analysis,"IEEE Trans. Agr. Inform. Syst.,vol. 10, no. 4, pp. 555-568, Apr. 2023.

[2]. R. T. Zhang and H. M. Li, "A Hybrid CNN-LSTM Approach for Spatio -Temporal Crop Yield Prediction Integrating Satellite Imagery,"IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.,vol. 15, pp. 201-215, May 2022.

[3]. P. K. Singh,Machine Learning Applications in Precision Agriculture,1st ed. New York, NY, USA: Academic Press, 2024.

[4]. M. A. AlHaj and T. D. Nolan, "Quantifying Uncertainty in Crop Yield Prediction using Gaussian Process Regression," inProc. Int. Conf. Comput. Methods Agron. (ICCMAG),Orlando, FL, USA, 2023, pp. 88-95.

[5]. J. R. Chen and S. L. Patel, "Review of Input Feature Selection Methodologies for Crop Yield Prediction Models,"Comput. Electron. Agr.,vol. 200, Art. no. 106987, Sept. 2021.