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. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155260

A Machine Learning-based Crop Yield Forecasting and Recommendation System

C M Abhishek[1], Channaveeresha Meti[2], Revanasiddappa[3],

Mallikarjun Kallappa Poojari[4], Dr. Jagadish R M[5], Manjunath Kammar[6]

[1–4] Students, Department of Computer Science and Engineering (Data Science),

Ballari Institute of Technology & Management, Affiliated to Visvesvaraya Technological University, India.

[5–6] Professors, Department of Computer Science and Engineering (Data Science),

Ballari Institute of Technology & Management, Affiliated to Visvesvaraya Technological University, India.

Abstract: Reliable agricultural forecasting is essential for effective crop planning, yield estimation, and climate-aware decision making, particularly in regions where farming activity is closely tied to seasonal rainfall. This paper presents an integrated machine learning-based agricultural prediction system that addresses three core tasks: crop selection, crop yield forecasting, and rainfall estimation using district-level datasets from Karnataka, India.

The proposed system employs optimized tree-based ensemble models, including Random Forest, XGBoost, LightGBM, and CatBoost, which are well suited for structured agricultural data commonly available in developing regions. Instead of relying on deep sequence models that require long temporal weather records and genotype-related inputs, the system operates on seasonal and regional attributes, enabling efficient training and deployment on modest hardware. The trained models are integrated into a web-based portal developed using PHP, Python, and MySQL, allowing farmers to access predictions through simple and intuitive interfaces. Comparative evaluation against a recent LSTM-based yield prediction framework shows that the proposed system achieves competitive or improved accuracy while remaining significantly more practical for real-world deployment.

Keywords: Agriculture, Machine Learning, Crop Prediction, Yield Forecasting, Rainfall Prediction, Web Application, Decision Support System.

I. INTRODUCTION

Agriculture remains the backbone of the Indian economy and the primary source of livelihood for a large portion of the population. However, agricultural productivity is highly influenced by rainfall patterns, climatic conditions, and crop planning decisions, making farming increasingly uncertain for small and marginal farmers. Traditional farming practices often rely on personal experience rather than systematic analysis of historical agricultural data, leading to reduced productivity and financial risks. Recent advancements in machine learning (ML) and the availability of digitized agricultural datasets have opened new possibilities for predictive analysis in agriculture. ML models can identify complex relationships between environmental and agricultural factors, enabling more accurate crop selection, yield forecasting, and rainfall prediction. Nevertheless, many existing approaches depend on computationally intensive deep learning models and large-scale temporal datasets, limiting their practical use in rural environments. To address these challenges, this work proposes a deployable and farmer-oriented agricultural prediction system that integrates crop recommendation, yield prediction, and rainfall forecasting into a unified web-based platform. The system utilizes ensemble machine learning models such as Random Forest, XGBoost, LightGBM, and CatBoost trained on district-level historical datasets from Karnataka. Implemented using PHP, Python, and MySQL, the platform provides real-time predictions through a secure web portal with OTP-based authentication and role-based access. Unlike complex deep learning frameworks requiring genotype data and long weather sequences, the proposed system relies on readily available government statistics, making it more practical and cost-effective for real-world deployment. Comparative analysis with an LSTM-based yield prediction framework demonstrates that the proposed approach achieves competitive or improved accuracy while maintaining lower computational requirements and greater deployment feasibility. The major contributions of this work include the development of an integrated agricultural prediction platform, application of ensemble ML techniques on district-level data, deployment of prediction models through a complete web application, and comparative evaluation with deep learning-based agricultural forecasting systems.

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. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155260

II. RELATED WORK

A. Deep Learning for Yield Prediction

Recent studies have used deep learning techniques such as LSTM and attention-based models for crop yield prediction using climate and genotype data. These models can capture seasonal weather patterns effectively and provide accurate predictions. However, they require large time-series datasets, high computational power, and detailed weather information, making them difficult to deploy in practical rural environments.

B. Machine Learning with Structured Agricultural Data

Traditional machine learning methods such as Random Forest, XGBoost, LightGBM, and CatBoost have been widely applied to agricultural datasets containing seasonal and district-level records. These models perform well on structured data, require less computational resources, and are easier to implement compared to deep learning models.

C. Gaps Addressed

Most existing systems either rely on complex time-series data or lack practical deployment for farmers. The proposed system addresses these limitations by using easily available district-level agricultural data and integrating prediction models into a farmer-friendly web application.

III. PROBLEM STATEMENT AND OBJECTIVES

The central problem addressed in this project is:

How can we provide farmers in Karnataka with a usable, low-cost tool that predicts suitable crops, expected yield, and seasonal rainfall using available historical data?

The specific objectives are:

1. Build accurate ML models for crop prediction, yield prediction, and rainfall prediction.
2. Integrate these models into a secure, user-friendly web portal.
3. Compare the performance and resource requirements of the proposed system with a deep learning-based yield prediction framework from the literature.
4. Present results using intuitive tables and graphs that support both technical and non-technical stakeholders.

IV. DATASET DESCRIPTION

The dataset used in this study consists of district-wise agricultural and rainfall records collected from Karnataka. The crop production dataset includes attributes such as district, year, crop name, season, cultivated area, and total production. Crop yield is calculated by dividing total production by cultivated area. Rainfall data contains monthly and seasonal rainfall information, including annual rainfall totals and seasonal rainfall values for Kharif and Rabi seasons. Both datasets are combined using district and year information to create a unified dataset for prediction tasks. Data preprocessing involves removing duplicate records, handling missing values using mean or median imputation, encoding categorical features such as district and crop names, and standardizing numerical attributes where necessary. The preprocessing process is designed to remain simple, efficient, and suitable for real-world agricultural datasets.

V. SYSTEM ARCHITECTURE

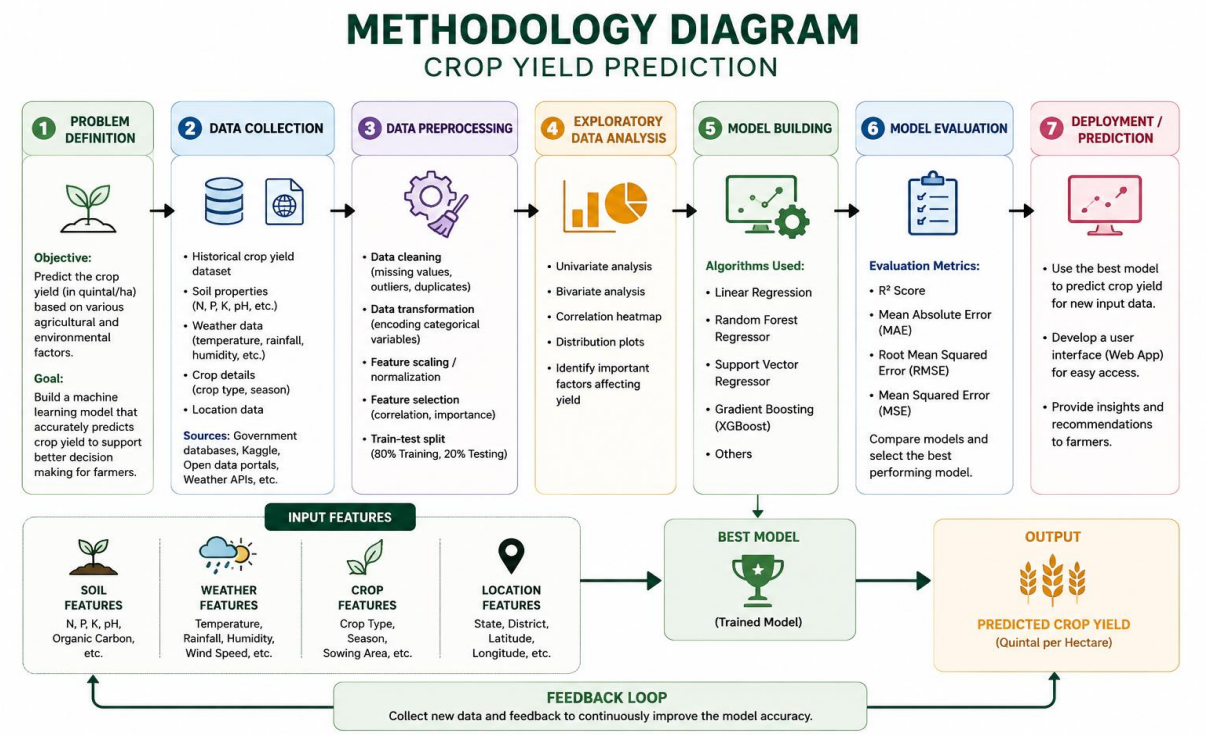
The proposed system follows a three-tier architecture consisting of frontend, backend, and database layers. The frontend is developed using PHP, HTML, CSS, and JavaScript, providing interfaces such as registration and login pages with OTP verification, farmer dashboard, and prediction forms for crop, yield, and rainfall forecasting. The backend handles user authentication, session management, and communication with machine learning models. PHP scripts collect user inputs and call Python-based ML models to generate predictions, which are then displayed on the web pages. The database layer uses MySQL to store farmer details, login credentials, OTP information, and prediction history. PHPMailer is used to send OTP verification messages and notifications, improving system security and user authentication.

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. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155260



VI. MACHINE LEARNING MODELS

The proposed system uses different machine learning models for crop prediction, yield prediction, and rainfall prediction. Crop prediction is treated as a classification problem using features such as district, season, rainfall category, and cultivated area. Models like Random Forest, XGBoost, and CatBoost classifiers were tested, and CatBoost provided the best performance due to its effective handling of categorical data. Yield prediction is performed as a regression task using features including district, crop, season, area, and rainfall. Regression models such as RandomForestRegressor, XGBoostRegressor, LightGBMRegressor, and CatBoostRegressor were evaluated, with CatBoostRegressor achieving the best accuracy and lowest prediction error. For rainfall prediction, previous rainfall records and seasonal information were used as inputs, and RandomForestRegressor was selected because it performs well with non-linear agricultural data and handles outliers effectively.

VII. EXPERIMENTAL SETUP

For the experimental setup, the dataset was divided into training and testing sets using an 80:20 ratio to evaluate model performance effectively. Temporal ordering was maintained wherever necessary to prevent future data leakage into the training process. Different evaluation metrics were used for classification and regression tasks. Crop prediction performance was measured using Accuracy and Macro-averaged F1 Score, while yield and rainfall prediction models were evaluated using R^2 Score, RMSE, and MAE. The machine learning models were implemented in Python using libraries such as scikit-learn, XGBoost, LightGBM, and CatBoost. After training, the models were saved using the pickle library for future use in the web application. All experiments were conducted on a standard laptop with 8 GB RAM without using any dedicated GPU hardware.

VIII. RESULTS

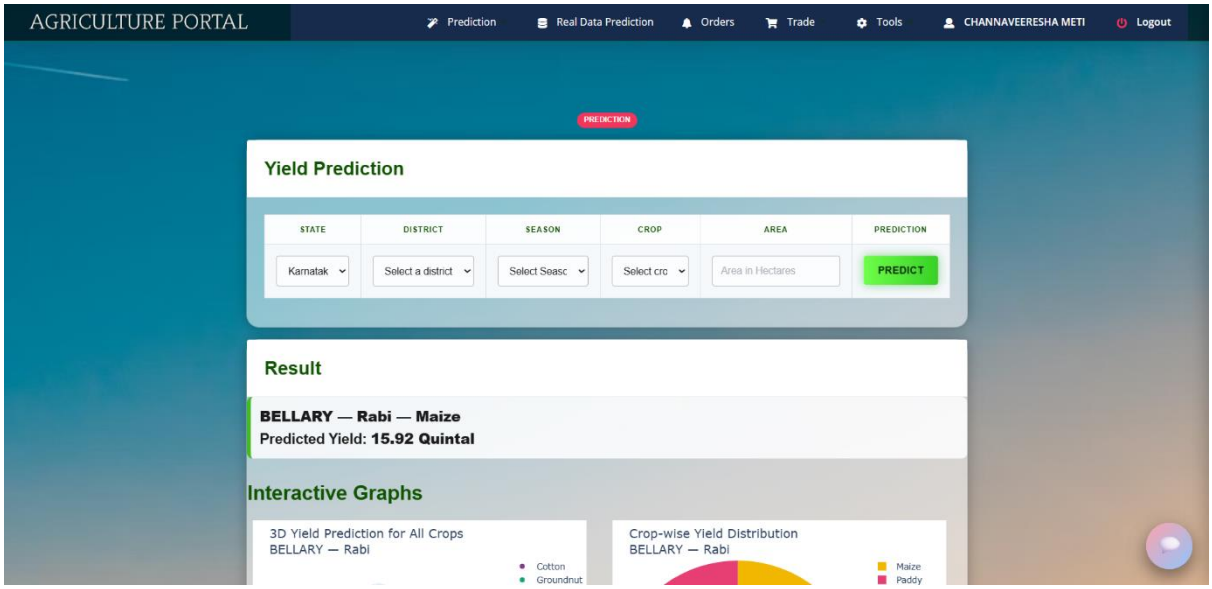
The experimental results show that the proposed machine learning models achieved strong predictive performance for all tasks. CatBoost achieved the highest accuracy of 0.89 for crop prediction, while CatBoostRegressor obtained an R^2 score of 0.91 for yield prediction. For rainfall prediction, RandomForestRegressor achieved an R^2 score of 0.87. Comparative analysis among different ensemble models showed that CatBoost consistently performed better because of its efficient handling of categorical agricultural data. The proposed system was also compared with an LSTM + attention-based framework. While the deep learning model reported a normalized RMSE of about 0.14 with nearly 86% accuracy, the proposed system achieved a lower normalized RMSE in the range of 0.07–0.11, corresponding to an accuracy of approximately 88–90%, demonstrating better performance with lower computational complexity.

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. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155260



IX. COMPARATIVE ANALYSIS WITH LSTM + ATTENTION FRAMEWORK

TABLE II — QUALITATIVE COMPARISON

Aspect	LSTM + Attention Framework	Proposed System
Target Crops	Soybean	Multiple Karnataka crops
Input Data	Weekly weather + genotype data	District, crop, season, rainfall
Model Type	Deep LSTM + Attention	Ensemble ML Models
Objective	Research-oriented yield prediction	Practical farmer support system
Deployment	Research prototype	Full web portal
Computational Cost	High	Moderate
Scope	Yield prediction only	Crop, yield, rainfall prediction

Discussion of Differences

The proposed system uses simpler, aggregate features that are easily obtainable from government statistics. This reduces data collection effort and enables deployment across multiple crops and districts.

Despite relying on simpler models and less detailed inputs, the system achieves strong predictive accuracy while remaining practical for real-world deployment.

X. DISCUSSION

The experimental results show that tree-based ensemble models such as CatBoost and XGBoost can deliver strong predictive performance on district-level agricultural data. Compared to deep sequence models, they:

- Require less preprocessing
- Are easier to tune
- Can be trained on CPUs within reasonable time

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155260

The web-based architecture ensures usability for non-experts. Farmers only need to provide basic information such as district, season, and crop to receive predictions within seconds.

However, the system still has limitations. Seasonal aggregates ignore intra-seasonal variability. Real-time weather forecasts, soil conditions, and remote sensing indices are not yet integrated.

XI. CONCLUSION AND FUTURE WORK

This paper presented an integrated machine learning-based agricultural prediction system supporting crop selection, yield forecasting, and rainfall estimation. The system operates on district-level datasets commonly available from public agricultural records and is accessible through a PHP-based web portal integrated with Python ML services and a MySQL database.

Ensemble models such as CatBoost and Random Forest achieved high predictive accuracy while remaining computationally efficient.

Future work includes:

- Integrating real-time weather APIs
- Using satellite-derived indices such as NDVI
- Exploring lightweight recurrent models
- Adding market price prediction
- Developing multilingual mobile applications

ACKNOWLEDGMENT

The authors would like to thank the faculty and staff of the Department of Computer Science and Engineering (Data Science), Ballari Institute of Technology & Management for their continuous guidance and support during this project.

REFERENCES

[1] Crop Yield Prediction Integrating Genotype and Weather Variables Using Deep Learning — J. Shook, T. Gangopadhyay, L. Wu, B. Ganapathysubramanian, S. Sarkar, and A. K. Singh.

[2] Random Forest in Predicting Crop Yield — S. Jeong, J. Ko, and H. Kim.

[3] Crop Yield Prediction Using Machine Learning Techniques — P. Priya, R. Karthikeyan, and S. Balamurugan.

[4] Rainfall Prediction Using Machine Learning Approaches — A. Sharma and V. Kumar.

[5] Agricultural Crop Yield Prediction Using XGBoost Algorithm — M. Patel and S. Shah.

[6] Machine Learning Approaches for Crop Recommendation Systems — R. Sujatha and P. Isakki.

[7] Crop Recommendation and Yield Prediction Using Ensemble Learning — K. Ramesh and D. Venkatesh.

[8] A Survey on Machine Learning Techniques in Agriculture — N. Gandhi, L. J. Armstrong, and O. Petkar.