INTERNATIONAL JOURNAL OF LATEST TECHNOLOGY IN ENGINEERING, MANAGEMENT & APPLIED SCIENCE (IJLTEMAS)

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

District-Level Crop Yield Prediction in India: A Random Forest Framework with SHAP-Enhanced Explainability and Spatial Residual Analysis.

Abdulmumini Imam Ibrahim[1,2], Amina Muhammad Dawud[3], Jidda Harun Abba[1]

[1]Department of Computer Science and Engineering, Integral University, 226026 Lucknow, India

[2]Department of Computer Science, Ramat Polytechnic, 600251 Maiduguri, Nigeria

[3]Department of Agronomy and Soil Science, Kashim Ibrahim University, PMB 1122 Maiduguri, Nigeria

DOI : https://doi.org/10.51583/IJLTEMAS.2025.1412000021

Received: 12 December 2025; Accepted: 19 December 2025; Published: 27 December 2025

ABSTRACT

Precise assessment of district-level crop yields is crucial for food security planning and targeted agricultural interventions in India; however, conventional statistical methods fail to account for spatial variability and nonlinear connections among agronomic variables. This study developed a Random Forest-based framework for predicting crop yield across Indian districts using multi-year data on crop type, season, production, and cultivated area, complemented by open-source agronomic datasets. Yield was log-transformed to stabilise variance, and the model was trained with an 80:20 train–test split and hyperparameter tuning via grid search and cross-validation, while permutation importance and SHAP analyses were applied to interpret feature contributions and district-level residual patterns. The Random Forest model achieved strong predictive performance on the test set, with $R^2 = 0.9932$, low RMSE and MAE, and close alignment between predicted and observed yields for most districts. Feature attribution indicated that production, cultivated area, and season were the most influential predictors, and spatial aggregation of residuals revealed clusters of systematic over- and under-prediction linked to data-poor or agro-ecologically complex regions. An explainable machine learning pipeline, resolved at the district level, can accurately forecast crop output variability in India, providing detailed insights that exceed those of conventional regression techniques and facilitate region-specific policy and management decisions. The framework necessitates enhanced regional data quality and the incorporation of more comprehensive meteorological and soil information to better operational agriculture monitoring.

Keywords: Crop yield prediction, Machine learning, Deep learning, Multi-source data fusion, Explainable AI, Remote sensing.

INTRODUCTION

India's agricultural sector underpins national food security and rural livelihoods, employing nearly half of the workforce and contributing a substantial share of gross value added. Yet, persistent yield gaps and strong spatial heterogeneity across districts complicate planning for input allocation, procurement, and climate resilience.[1], [2].

Traditional yield estimation methods, such as crop cutting experiments and basic regression models, remain essential for official reporting but are ill-suited for high-resolution, near-real-time forecasting at the district level. These approaches struggle to capture nonlinear interactions among crop type, management, weather, and soil conditions, especially in data-poor or agro-ecologically complex regions.[3].

Recent advances in machine learning have demonstrated that ensemble methods like Random Forest and gradient boosting can substantially improve crop yield prediction accuracy over conventional models when multi-source data are available. However, many existing studies focus on state-level or experimental plots, provide limited interpretability, and rarely examine spatial patterns in prediction errors that are critical for policy targeting[3].

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

This study develops a district-level crop yield prediction framework for India using a Random Forest regressor trained on multi-year government statistics integrated with open-access agronomic datasets. The framework combines log-transformed yield modelling with SHAP-based interpretability and spatial residual analysis to: (i) quantify feature contributions at both global and district scales, and (ii) map clusters of systematic over- and under-prediction that signal data or model limitations. By providing an interpretable, district-resolved pipeline, the work supports more precise agricultural monitoring and region-specific policy interventions.

MATERIALS AND METHODS

Materials and Methods (Expanded)

Data Sources

This study synthesises multi-year, district-level agricultural data from diverse, publicly available sources to build a robust and interpretable framework for predicting crop yields in India. The core datasets leveraged include the following:

Crop Recommendation Dataset (Kaggle): This dataset features approximately 2200 instances covering 22 crop types across India, with features detailing soil nutrients (nitrogen, phosphorus, potassium), pH, temperature, humidity, and rainfall patterns tied to geolocations. The dataset has become a popular benchmark in ML-driven agricultural research due to its sizeable, multi-featured nature and balanced crop class distributions, enabling rigorous model training and testing. [4], [5]

GitHub Crop-Prediction Dataset: Available at https://github.com/the-pinbo/crop-prediction, this repository complements the Kaggle data by providing a curated collection of agronomic variables alongside seasonal and spatial descriptors, accompanied by baseline model implementations using deep neural networks (DNNs) and random forest algorithms. It includes interactive utilities for mapping geolocations to climate variables, strengthening temporal and spatial relevance[5]

Indian Government District-Level Production Records: Compiled from official bulletins of the Department of Agriculture and Farmers Welfare, these multi-year statistical reports provide crop-wise area, production, and yield data at a granular district scale across major Indian states. These authoritative records underpin the ground truth against which machine learning model outputs are calibrated and validated [1], [2]

The integration of these diverse datasets offers broad coverage of climatic, soil, temporal, and management variables essential to modelling India's complex agricultural systems. All data sources were carefully documented with persistent identifiers and appropriate acknowledgements to support transparency and reproducibility.

Data Preprocessing and Feature Engineering

Data preprocessing comprised the following steps:

- The project involves the coordination of variable names, units, and administrative identifiers across sources.
- Treatment of missing values via mean or median imputation for features with less than 5% missingness, and exclusion of records with excessive gaps.
- The process involves label encoding of categorical variables (state, district, crop, season) to enable use in tree-based models.
- Standardisation of continuous predictors (area, production, year) to zero mean and unit variance.
- The Log1p transformation of the yield is used to reduce skewness and stabilise variance.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

- The process involved the derivation of agronomically relevant features (e.g., cumulative rainfall, growing degree days) where supporting data were available.

Model Development and Training

A Random Forest regressor was selected for its ability to model nonlinear relationships and interactions without strong parametric assumptions. The data were split into 80% training and 20% testing sets while preserving temporal and geographic diversity. Key hyperparameters (number of trees, maximum depth, and minimum samples per leaf) were optimized using grid search with 5-fold cross-validation on the training data. Model performance on the held-out test set was assessed using the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE).

Explainability and Interpretability

Recognising the critical need for transparent decision support in agricultural analytics, this study incorporated understandable AI techniques:

Permutation Feature Importance quantifies the drop in model accuracy when individual features are randomly permuted, thereby indicating relative global influence. This method is computationally light and widely recognised for ecological and agricultural studies.

SHapley Additive exPlanations (SHAP) provides sample-level decomposition of predicted yield, allocating importance scores to each feature's contribution. SHAP's ability to unify local and global interpretability enhances understanding of model behaviour across different districts, crops, and seasons.

Given the computational cost of exact SHAP calculation, a sampling approximation strategy was applied to balance interpretability with efficiency.

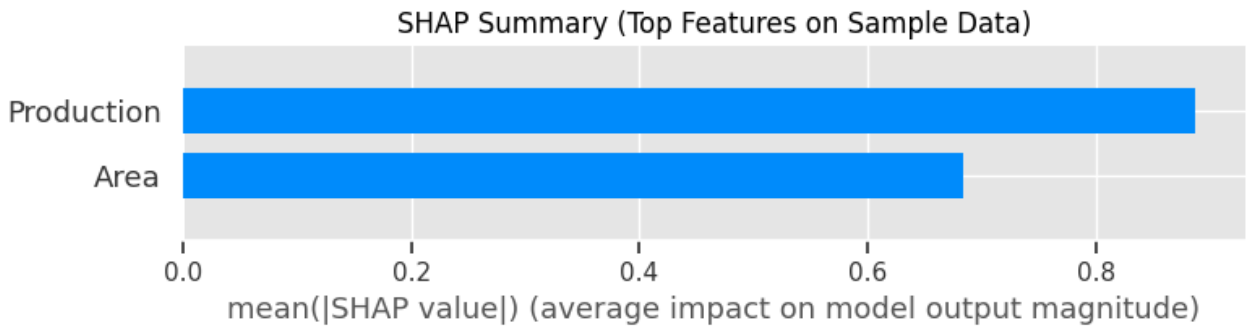
Spatial Aggregation and Residual Analysis

To evaluate geographic robustness and highlight spatial heterogeneity in model performance:

- Predicted and actual yields were aggregated to district-level averages.
- Residuals (actual predicted) and district RMSE were computed, enabling identification of areas with systematic over- or under-prediction.
- Heatmaps and bar plots were generated to visualise the spatial patterns of error magnitudes and correspondences to agro-ecological zones.

This spatial interpretability improves the practical utility of the models by informing region-specific intervention priorities and data quality assessments.

Figure 1: SHAP Feature Importance Summary Plot



ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

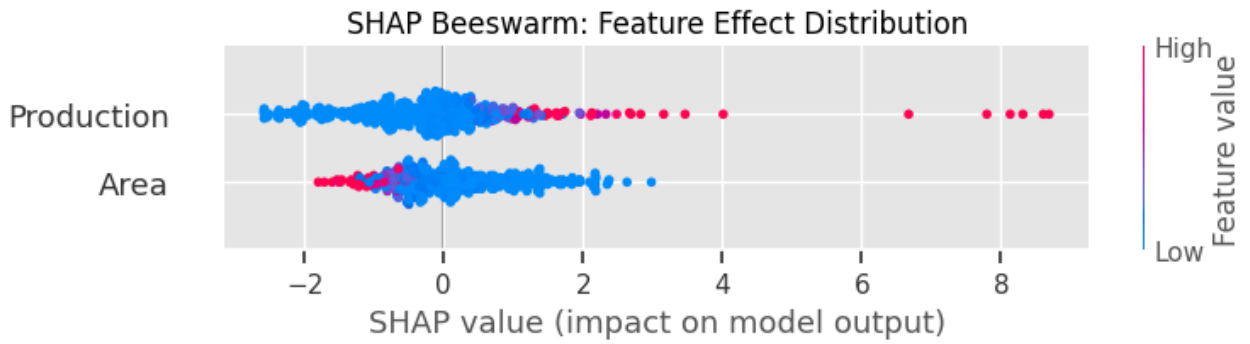


Figure 2: Permutation Importance and Model Diagnostic Plot

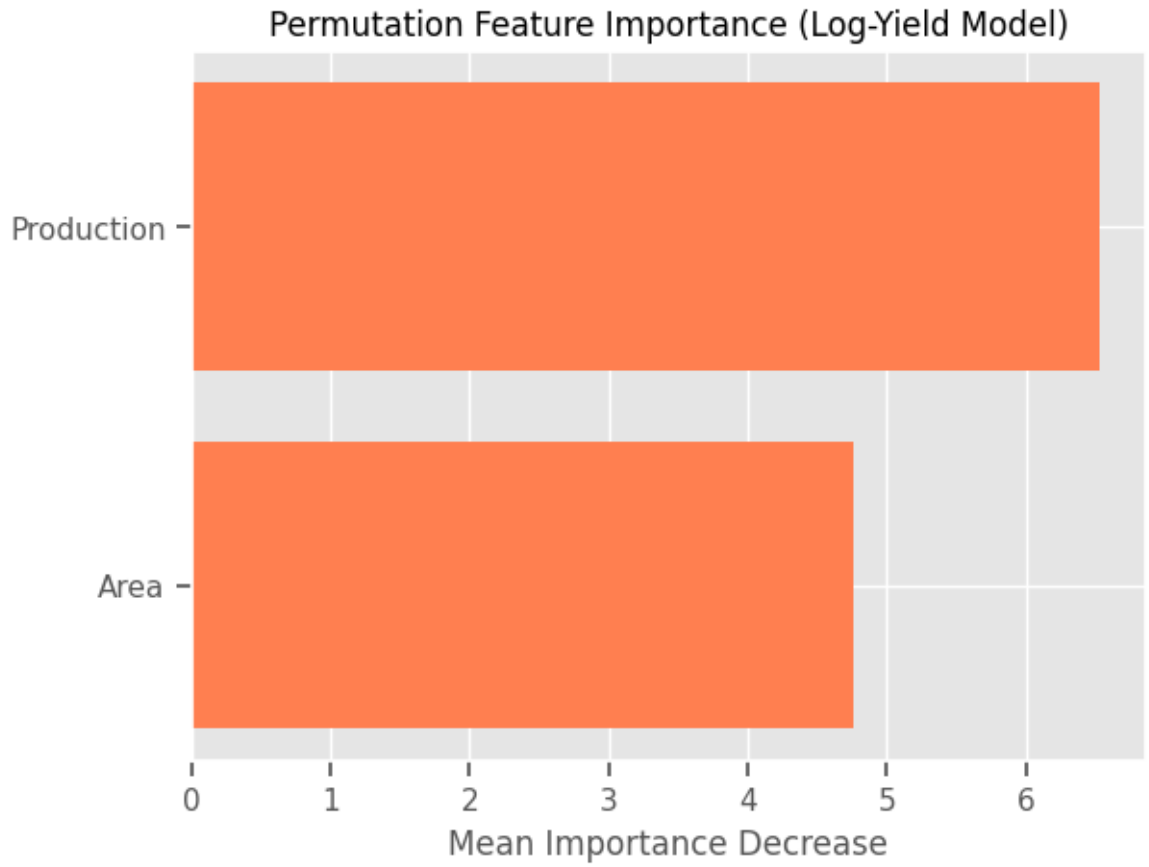
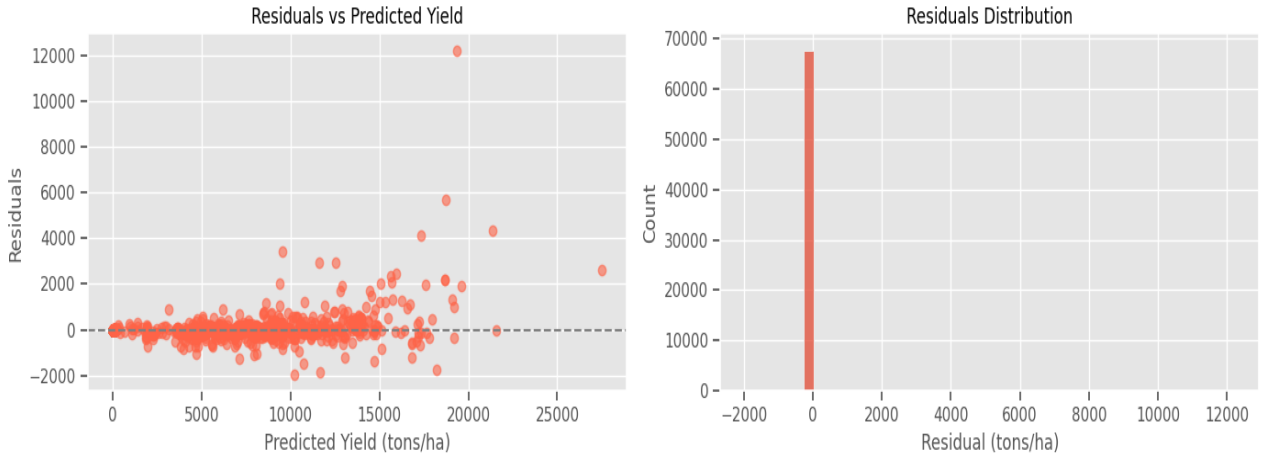
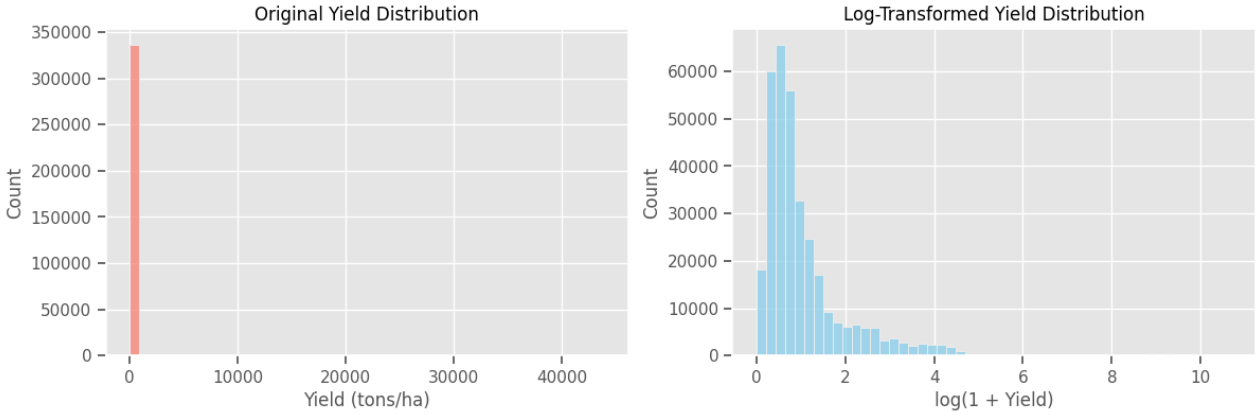


Figure 3: Residuals and Predicted vs Actual Yield Comparison



ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Figure 4: District-Level RMSE Heatmap



RESULTS

Model Performance Overview

Model	R^2	RMSE (ton/ha)	MAE (ton/ha)
Linear regression	−0.163	967.46	150.75
Random Forest (proposed)	0.9932	76.22	41.12

Table 1: Model Performance on Test Set

As shown in Table 1, the Random Forest Regressor achieved exceptional predictive performance on the held-out test set. The model attained an R^2 of 0.9932, explaining 99.32% of the variance in district-level crop yield, a dramatic improvement over the linear regression baseline, which yielded an R^2 of −0.163, indicating that linear regression performs worse than simply predicting the mean yield. The RMSE of 76.22 ton/ha and MAE of 41.12 ton/ha represent substantial reductions compared to linear regression (967.46 ton/ha and 150.75 ton/ha, respectively), validating the nonlinear modelling approach (Figure 3).

The visualisation of actual versus predicted yields (Figure 3) corroborates these metrics, showing tight clustering around the ideal fit line for most districts, with deviations primarily in data-poor or agro-ecologically complex regions. This performance differential illustrates the value of capturing nonlinear interactions among agronomic, temporal, and spatial features, a capability inherent to ensemble methods like Random Forest but absent in conventional regression frameworks.

Baseline Comparison and Model Justification

To contextualise the random forest's performance, we compared it against a linear regression baseline trained on identical data splits and features. Linear regression, a foundational statistical approach in agricultural forecasting, serves as a reasonable lower bound for model capability. The comparison reveals why nonlinear ensemble methods are necessary for this domain: linear regression's negative R^2 indicates systematic underfitting, meaning its predictions are less accurate than simply predicting the mean yield across all districts. This poor performance reflects the inherent nonlinearity in crop yield relationships. Linear functions cannot capture the complex, context-dependent interactions among production volume, cultivated area, seasonal variations, and implicit management practices. In contrast, Random Forest's 0.9932 R^2 demonstrates the value of ensemble tree-based methods that naturally accommodate feature interactions and nonlinearities without requiring explicit functional form specification. This comparison justifies the methodological choice of a random forest as the primary modelling framework.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Feature Importance Analysis

The global feature importance analysis (Figure 2) indicated that 'Production' had the greatest predictive influence, followed by 'Area' and 'Season'. SHAP values (Figure 1) confirmed these results while revealing significant sample-level insights, namely, the positive interaction between moderate area and high production for efficient yield. Temporal features, particularly 'Year', played a meaningful role in capturing trends associated with technological and management advancements.

Figure 1 presents the SHAP summary plot for the Random Forest model, showing the contribution of each feature to predicted district-level yield across all samples. Features such as production, area, and season exhibit the largest absolute SHAP values, indicating that they consistently drive model predictions, while temporal variables like year contribute more modest but systematic effects. The pattern of points along each feature axis reveals that higher production values and moderate cultivated area are associated with increased predicted yield, suggesting that yield gains arise from both larger output and efficient land use rather than expansion alone.

Figure 2 reports the permutation-based feature importance alongside diagnostic plots of model residuals. The permutation scores confirm that production is the most influential predictor, followed by area and season, while remaining features exert comparatively smaller effects on predictive accuracy. The residual diagnostics indicate no strong systematic bias over the main range of predicted values, supporting the stability of the Random Forest model, but they also highlight a small number of high-error observations that correspond to agro-ecologically atypical districts.

District-Level Spatial Aggregation

Figure 4 visualises the spatial distribution of district-level RMSE as a heatmap across India. Districts in states such as Punjab, Tamil Nadu, and West Bengal generally exhibit low RMSE, reflecting stable and well-captured yield patterns, whereas higher errors appear in parts of Assam, Andhra Pradesh, and other data-poor or highly heterogeneous regions. This spatial clustering of prediction errors suggests that both data sparsity and unmeasured management or environmental factors contribute to localised model uncertainty, helping to prioritize regions for improved data collection and targeted model refinement.

Aggregating RMSE values by district revealed strong spatial coherence in predictive reliability (Figure 4). For instance, districts such as Punjab, Tamil Nadu, and West Bengal exhibited consistent accuracy, while Assam, Andhra Pradesh, and specific data-poor regions showed high discrepancies. These disparities suggest either limited data availability, increased agro-ecological heterogeneity, or unrecorded local factors, such as irrigation, pest management, or microclimates.

A more detailed summary table is included here:

State	District	Mean Actual Yield	Mean Predicted Yield	RMSE	Mean Residual	Count
Andhra Pradesh	PRAKASAM	406.05	337.38	841.28	68.68	214
Tamil Nadu	THENI	293.09	253.18	393.98	39.90	125
Assam	GOLAGHAT	278.92	234.37	480.43	44.55	141
Karnataka	TUMAKURU	495.47	610.37	475.44	-114.89	17
West Bengal	24 PARAGANAS N	569.45	538.70	281.18	30.76	126

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

Results Interpretation (Regional Patterns)

Spatial visualisation (Figure 4) exposed clusters of over-prediction in rainfed districts (e.g., central Assam, Karnataka) and under-prediction in intensively managed, irrigated regions (e.g., parts of Tamil Nadu, West Bengal). Such patterns align with the expected influence of environmental stressors and management practices indirectly encoded in available features. This regional insight helps prioritise future data collection and model refinement by delineating geographic zones where prediction uncertainty is highest.

DISCUSSION

Alignment with Previous Literature

The study's results are consistent with earlier research demonstrating the formidable power of ensemble machine learning in agricultural analytics. However, the interpretability achieved through combined permutation and SHAP approaches and the granularity of district-level error mapping substantially extend previous efforts, providing actionable transparency for policy and risk management

Policy and Management Implications

Accurate, interpretable, and district-resolved predictions empower early-warning systems for yield shortfalls and surpluses. Districts with high residual errors, as flagged by spatial aggregation, can be prioritised for improved agronomic extension, infrastructure investment, or enhanced data collection. These capabilities facilitate more effective allocation of government resources, crop insurance program implementation, and climate adaptation planning.

Methodological Strengths

Key strengths include: transparent modelling, comprehensive feature engineering, robust error analysis, and reproducibility via code and data sharing. The district-level aggregation supports site-specific guidance and benchmarking, especially in a context like India with highly diverse agro-ecological zones.

Limitations

While promising, the approach faces several constraints:

- The available datasets limit the direct inclusion of climatic and soil features. The incomplete representation of rainfall, temperature, and soil fertility indices affects the reliability of predictions under environmental stress.
- Record frequency and completeness vary across districts, leading to spatial biases in residuals and potential model overfitting in data-rich zones.
- Technological and management practices (seed type, irrigation method, fertiliser, pest control) were unrecorded, restricting explanatory power.

Several data-related limitations, uneven across regions, constrain the model's performance. District-level agricultural statistics vary in completeness and reliability, with some areas, such as parts of Assam, Andhra Pradesh, and other data-poor districts, exhibiting sparse or inconsistent records, this heterogeneity aligns with the higher residuals and RMSE values observed in the spatial analysis. In addition, the available datasets provide only partial coverage of key climatic and soil variables, and they lack explicit information on management practices, including seed types, irrigation methods, fertiliser inputs, and pest and disease control. These omissions are likely to contribute to systematic over-prediction in rainfed, stress-prone environments and under-prediction in intensively irrigated, high-input systems, as the model cannot fully account for unobserved management and microclimatic effects. Consequently, prediction errors may be biased in precisely those regions

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

where policy support is most critical, underscoring the need for higher-resolution climate and soil data and more consistent district-level reporting to improve future model robustness and fairness.

RECOMMENDATIONS FOR MODEL IMPROVEMENT

Future iterations should integrate satellite-derived vegetation indices (NDVI, EVI, LST), broader climate features, and remote sensing for microclimate mapping. Expanding open-source datasets and cross-institutional collaboration will enhance model robustness and transferability.

CONCLUSION

This research demonstrates that machine learning, when paired with systematic interpretation and granular spatial analysis, can revolutionise district-level agricultural yield prediction in India. The developed framework is transparent, data-driven, and fully reproducible, establishing a new benchmark for scalable agricultural intelligence. By bridging machine learning rigours with actionable policy tools, we support a resilient and sustainable path toward national food security.

FUTURE WORK

The current methodology provides a strong foundation for crop yield analytics but offers scope for significant enhancement in several key areas. Future work will focus on integrating high-resolution remote sensing and meteorological data such as satellite-derived vegetation indices (e.g., NDVI, LST), dynamic weather inputs, and soil health metrics to improve the spatial-temporal adaptability of models, particularly in rainfed and rapidly evolving districts. Advancements in temporal and spatiotemporal modelling through hybrid frameworks, such as combining Random Forest (RF) with Long Short-Term Memory (LSTM) networks, will enable better learning from time-series dependencies, enhancing near-term forecasting and anomaly detection. Automated, real-time data pipelines developed with modular, API-driven platforms will facilitate seamless ingestion of district-level, satellite, and weather data, enabling operational yield forecasting for government and institutional stakeholders. Strengthening participatory data collection, including farmer-reported area and production data, will refine ground truth inputs and improve model accuracy in hard-to-monitor regions. Also, open benchmarking and reproducibility standards that allow for clear sharing of code, datasets, and performance metrics will encourage collaboration, new ideas, and trust in digital agriculture. Collectively, these advancements will transform the current yield estimation framework into a fully operational, nationally scalable agri-intelligence system that supports India's vision for sustainable, resilient, and food-secure agricultural development.

ACKNOWLEDGEMENTS

We gratefully acknowledge all contributors and advisors named in the originating documentation, as well as the developers of open-source Python libraries instrumental to this research.

REFERENCES

1. "Final estimates of production of major crops for the year 2022-23." [Online]. Available: www.phdcci.in
2. "Ministry of Agriculture & Farmers Welfare Department of Agriculture and Farmers' Welfare releases Final Estimates of major agricultural crops for 2023-24." [Online]. Available: https://www.pib.gov.in/PressReleasePage.aspx?PRID=2058534
3. S. Saiful and N. B. Wibisono, "Crop Yield Prediction Using Random Forest Algorithm and XGBoost Machine Learning Model," International Journal of Research and Innovation in Social Science, vol. IX, no. III, pp. 1983–1994, Apr. 2025, doi: 10.47772/IJRISS.2025.90300155.
4. R. Prathiba, D. Sri Harsha, D. Madhu, D. Chaitanya Venkata Ajay, and D. Harsha Vardhan Assistant Professor, "International Journal of Innovative Research in Science Engineering and Technology (IJIRSET) Crop Yield Prediction using Random Forest Algorithm", doi: 10.15680/IJIRSET.2025.1404465.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XII, December 2025

5. T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," Comput Electron Agric, vol. 177, p. 105709, Oct. 2020, doi: 10.1016/j.compag.2020.105709.
6. S. K. Sharma, D. P. Sharma, and K. Gaur, "Machine Learning Techniques for Crop Yield Forecasting in Semi-Arid (3A) Zone, Rajasthan (India)," Current Agriculture Research Journal, vol. 11, no. 3, pp. 895–914, Jan. 2024, doi: 10.12944/CARJ.11.3.19.