International Journal for Modern Trends in Science and Technology
Volume 12, Special Issue 01, pages 155-160.
ISSN: 2455-3778 online
Available online at: http://ijmtst.com/ICRCSET2026.html
DOI: https://doi.org/10.5281/zenodo.19536509

Zone-Aware Crop Yield Prediction System Using Agro-Climatic and Pesticide data

Unnava Lohitha, Pondugala Andy Mylin, Vucha Uday Kiran, Daggupati Balaji, Bonthagarla Manikanta

Department of CSE(Data Science), Bapatla Engineering College, Bapatla, Andhra Pradesh, India.

To Cite this Article
Unnava Lohitha, Pondugala Andy Mylin, Vucha Uday Kiran, Daggupati Balaji & Bonthagarla Manikanta (2026). Zone-Aware Crop Yield Prediction System Using Agro-Climatic and Pesticide data. International Journal for Modern Trends in Science and Technology, 12(SI01), 155-160. https://doi.org/10.5281/zenodo.19536509

Article Info
Received: 02 March 2026; Revised: 01 April 2026; Accepted: 04 April 2026.

KEYWORDS

Machine Learning, ANN, CNN, LSTM, LGBM, Ensemble Stacking, Agro-Climatic Zones.

ABSTRACT

Accurate crop yield prediction is essential for effective agricultural planning and food security. This study proposes a zone-aware hybrid machine learning framework for crop yield prediction using agro-climatic data from Andhra Pradesh. The model integrates Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Light Gradient Boosting Machine (LGBM) using an ensemble stacking approach to improve predictive performance. The dataset includes key features such as rainfall, temperature, humidity, pesticide usage, and agro-climatic zone information, capturing regional variability across different zones. Data preprocessing techniques, including normalization and feature engineering, are applied to enhance model efficiency. Each model is trained and evaluated using regression performance metrics such as R^2 score, and Mean Squared Error (MSE). The outputs of individual models are combined through a stacking strategy to generate the final prediction. Experimental results demonstrate that the proposed hybrid approach significantly improves accuracy and robustness compared to individual models. The developed system provides a reliable and efficient solution for crop yield prediction and supports data-driven decision-making in agriculture.

I. INTRODUCTION

Highlight Accurate crop yield prediction is essential for ensuring food security, optimizing agricultural resource use, and supporting effective policy decisions. Agriculture sustains millions of livelihoods worldwide, yet crop production is highly sensitive to environmental conditions. Variations in rainfall, temperature, and extreme weather events can significantly influence yields. Reliable forecasting tools allow farmers to make informed decisions about crop selection, sowing schedules, irrigation, and harvest timing, while also aiding policymakers in managing food distribution,

resource allocation, and subsidies [1], [2]. With global population projected to exceed 9 billion by 2050, precise and timely crop yield predictions are increasingly critical [1].

Conventional prediction methods, such as linear regression, time-series analysis, and crop simulation models, often struggle to capture the complex interactions among climate, soil, and management practices [3], [4]. These approaches generally assume linear relationships and may fail to adapt to heterogeneous datasets or diverse agro-climatic conditions, reducing prediction reliability across regions with varying soil types and seasonal patterns [5], [6].

Machine learning (ML) techniques offer a more flexible and powerful alternative for yield forecasting. Models like artificial neural networks (ANN), convolutional neural networks (CNN), and long short-term memory (LSTM) networks can learn nonlinear relationships and detect temporal patterns in large datasets [3], [6], [7]. By integrating information from multiple sources—including climate records, soil properties, and farm management data—ML models can provide more accurate and context-sensitive predictions. Hybrid approaches that combine multiple algorithms, or employ ensemble methods such as gradient boosting, can further improve performance by leveraging the strengths of individual models [12], [13].

II. RELATED WORK

A. Traditional Methods for Crop Yield Prediction

Early crop yield forecasting primarily relied on statistical approaches and crop simulation models. Techniques such as linear regression and time-series analysis have been widely used to estimate yields based on historical climate and soil data [1], [2]. While these methods are relatively simple to implement and interpret, they often assume linear relationships among input variables, limiting their ability to model complex interactions between climate, soil properties, and crop management practices. Their predictive accuracy tends to decline when applied across heterogeneous regions with varying agro-climatic conditions, reducing reliability for large-scale or multi-zone applications [1], [2].

B. Machine Learning Approaches

Machine learning (ML) methods have gained prominence as effective tools for crop yield prediction. Models like artificial neural networks (ANN), convolutional neural networks (CNN), and long short-term memory (LSTM) networks are capable of capturing nonlinear relationships and temporal dependencies in climate, soil, and agronomic datasets [3], [6], [10]. Numerous studies report that these approaches outperform traditional statistical methods in terms of accuracy and adaptability across different crops and seasons [7]. However, single ML models may not fully exploit the complementary strengths of other algorithms, and their performance can vary depending on input data types and regional characteristics [3], [6].

C. Hybrid and Ensemble ML Models

To address the limitations of individual models, hybrid and ensemble ML approaches have been introduced. These methods combine multiple algorithms or use ensemble techniques, such as gradient boosting (LightGBM) and stacked learning, to enhance robustness and accuracy [12], [13]. For example, integrating CNN with LSTM models allows capturing both spatial and temporal dependencies in crop-related data [6]. Hybrid frameworks are particularly effective for handling complex, large-scale datasets and learning intricate interactions between environmental and agronomic factors [3], [6], [12]. Nevertheless, increased computational requirements can pose challenges, especially when applying these models across multiple crops and regions.

D. Incorporation of Climatic and Agro-Climatic Data

Agro-climatic variables, including temperature, rainfall, soil characteristics, and humidity, significantly affect crop yields [5], [9]. Several studies incorporate these factors into ML models to improve prediction performance. However, many approaches focus on single climatic zones or specific crops, limiting their generalizability [5], [9]. Incorporating zone-specific and seasonal data enables ML models to generate more accurate, location-specific forecasts, making predictions more actionable for farmers and policymakers [8].

III. SYSTEM ARCHITECTURE

The proposed system architecture for crop yield prediction is designed as a multi-stage pipeline that integrates heterogeneous data sources, performs

preprocessing and feature engineering, and applies hybrid machine learning models followed by an ensemble approach for improved prediction accuracy.

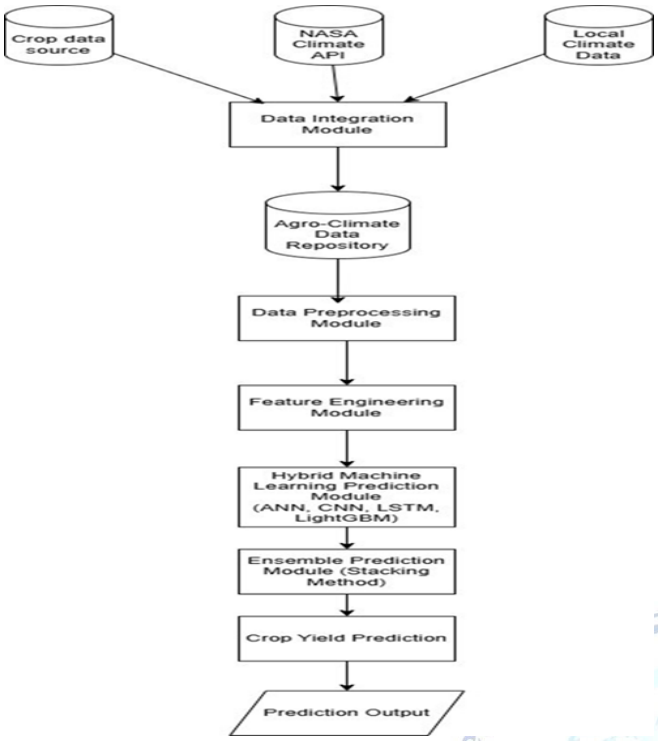

Figure-1. System Architecture

B. Data Sources

The system utilizes multiple data sources to capture diverse factors affecting crop yield:

- Crop Data Source: Historical agricultural data Including crop type, yield, soil characteristics, and farming practices.
- NASA Climate API: Provides satellite-based climate variables such as temperature, rainfall, solar radiation, and humidity.
- Local Climate Data: Region-specific meteorological data to enhance spatial accuracy.

These sources ensure comprehensive agro-climatic coverage for robust prediction. Data Sources

C. Data Integration Module

This module consolidates data from different sources into a unified format. Since the datasets may vary in structure, scale, and format, integration involves:
Data alignment (temporal and spatial)
Handling inconsistencies
Merging datasets into a common schema

D. Agro-Climatic Data Repository

The integrated data is stored in a centralized repository. This repository acts as:
A structured storage system, A reference for further processing, A scalable database for model training and evaluation

E. Data Preprocessing Module

Raw data is cleaned and prepared before model input. Key steps include:
Handling missing values, Removing outliers,
Data normalization and scaling
Encoding categorical variables (e.g., One-Hot Encoding)
This step ensures data quality and consistency.

F. Feature Engineering Module

In this stage, relevant features are extracted and transformed to improve model performance:
Creation of new variables (e.g., seasonal averages, rainfall indices) Feature selection to remove irrelevant attributes. Dimensionality reduction (if required) This module enhances the predictive power of the dataset.

G. Hybrid Machine Learning Prediction Module

Multiple machine learning models are trained in parallel to capture different data patterns:

- Artificial Neural Network (ANN): Captures non-linear relationships
- Convolutional Neural Network (CNN): Extracts spatial patterns
- Long Short-Term Memory (LSTM): Handles temporal dependencies
- LightGBM: Efficient gradient boosting for structured data

Each model contributes unique strengths to the prediction process.

H. Ensemble Prediction Module (Stacking Method)

To improve accuracy and generalization, predictions from individual models are combined using a stacking ensemble approach:
Base learners: ANN, CNN, LSTM, LightGBM
Meta-learner: Learns from base model outputs This reduces individual model bias and variance, resulting in more reliable predictions.

I. Crop Yield Prediction

The ensemble model generates the final predicted crop yield values based on processed agro-climatic features.

This output supports decision-making for crop planning, resource allocation, and risk management.

IV. Experimental Results and Discussion

The experimental setup utilizes an integrated agro-climatic dataset comprising historical crop yield data along with climate variables obtained from NASA and local meteorological sources for Andhra Pradesh. Multiple models were trained and evaluated. The performance of all models was assessed using Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2) metrics.

Table 1 is presents the performance evaluation of different machine learning models using RMSE and R^2 metrics. Among the individual models, LightGBM achieved the best performance with a low RMSE of 0.6432 and a high R^2 value of 0.8982, indicating its strong capability in handling structured agro-climatic data. The CNN and ANN models also demonstrated competitive performance, with RMSE values of 0.6754 and 0.6877 and R^2 values of 0.8878 and 0.8836, respectively, showing their effectiveness in capturing nonlinear relationships in the dataset.

TABLE 1. Performance Evaluation of Crop Yield Prediction Models

Model	RMSE (ton/ha)	R^2
ANN	0.6877	0.8836
CNN	0.6754	0.8878
LSTM	1.5273	0.4259
LightGBM	0.6432	0.8982
Stacked Ensemble	0.4450	0.9512

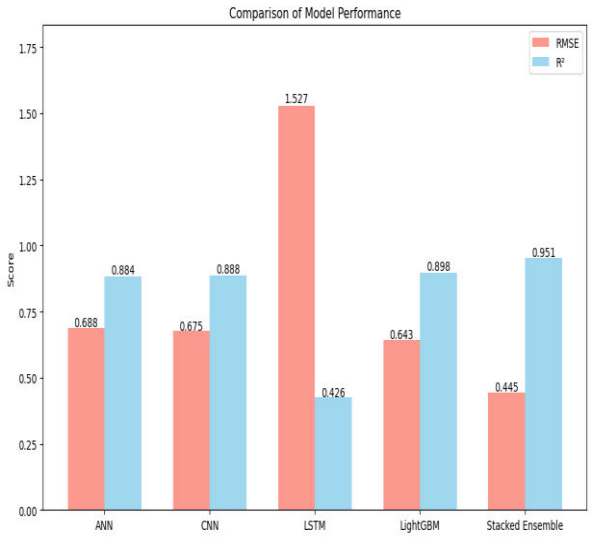

Figure 1. Performance Comparison of Crop Yield Prediction Models

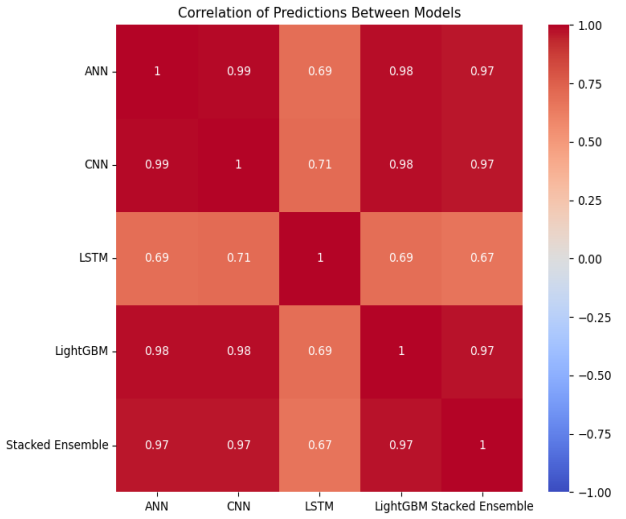

Figure 2. Correlation Matrix of Model Predictions

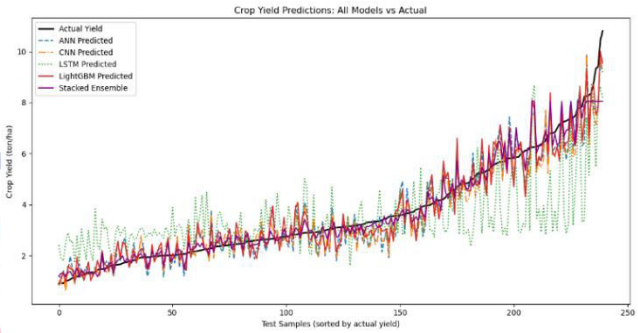

Figure 3. Actual vs Predicted Crop Yield Comparison Across Models

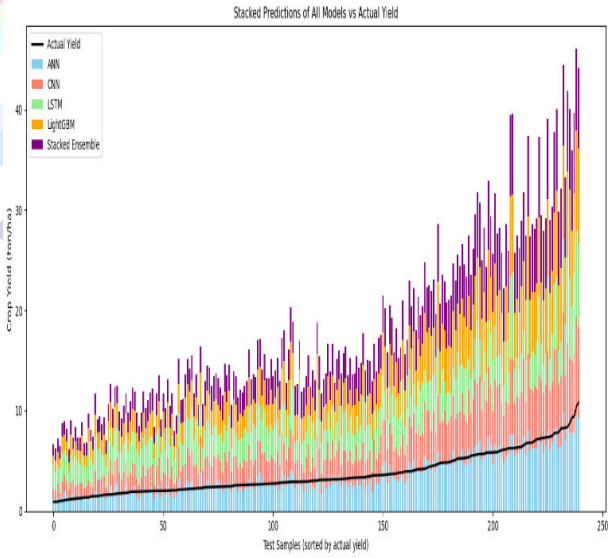

Figure 4. Stacked Model Predictions Compared with Actual Crop Yield

The proposed stacked ensemble model outperformed all individual models, achieving the lowest RMSE of 0.4450 and the highest R^2 of 0.9512. This demonstrates that combining multiple models through a stacking approach effectively leverages the strengths of each model, reduces prediction errors, and improves overall generalization.

Overall, the results indicate that ensemble learning provides superior performance compared to standalone models, making it a robust approach for accurate crop yield prediction.

V. Limitations and Future Work

Despite strong predictive results, this study has some limitations. The dataset is confined to historical weather and crop yield records from Andhra Pradesh, which may not fully reflect all environmental or socio-economic variables influencing crop production. Additionally, deep learning approaches like LSTM may require larger datasets and careful hyperparameter optimization to perform effectively. Future studies should aim to address these limitations to enhance model generalizability.

While the current study demonstrates that the Stacked Ensemble model provides accurate predictions for crop yields in Andhra Pradesh, there are several avenues to further enhance model performance and applicability:

1. Incorporation of Additional Features – Future research could integrate more environmental factors such as soil nutrient content, pest incidence, irrigation practices, and remote-sensing data, as well as socio-economic variables like farm size and labor availability.

2. Larger and Diverse Datasets – Expanding the dataset across multiple years and regions would improve model generalizability and robustness, allowing the models to better capture temporal and spatial variability in crop yields.

3. Advanced Model Architectures – Exploring more complex deep learning models, hybrid architectures, or attention-based mechanisms may further improve predictive accuracy, particularly for capturing non-linear and temporal patterns.

4. Real-Time and Operational Deployment – Developing real-time prediction systems or decision-support tools could assist farmers and policymakers in dynamic planning for crop management, irrigation scheduling, and market interventions.

5. Uncertainty Quantification – Incorporating methods to quantify prediction uncertainty could provide confidence intervals or risk estimates, making the forecasts more actionable for decision-making.

VI. Conclusion

In this research, several machine learning approaches were developed and assessed to forecast crop yields in Andhra Pradesh, India. The models included Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), LightGBM, and a Stacked Ensemble model that leveraged the strengths of all base models. Model performance was evaluated using Root Mean Squared Error (RMSE) and the Coefficient of Determination (R^2).

The findings reveal that the Stacked Ensemble model delivered the most accurate predictions, achieving the lowest RMSE of 0.4450 ton/ha and the highest R^2 of 0.9513, indicating its strong predictive ability. Among individual models, LightGBM performed best with an RMSE of 0.6432 ton/ha and R^2 of 0.8982, followed by CNN and ANN. The LSTM model showed relatively weaker performance, with an RMSE of 1.5273 ton/ha and R^2 of 0.4259, suggesting that deep learning models may need larger datasets or further optimization for precise yield forecasting.

Overall, the study demonstrates that ensemble methods can effectively integrate the advantages of different models, resulting in more reliable crop yield predictions. These insights can assist farmers and policymakers in making informed, data-driven decisions for improving agricultural efficiency and resource management. Future research could focus on incorporating additional environmental and socio-economic variables to enhance model accuracy further.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

References

[1] G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in A. K. Singh, R. K. Singh, and M. S. Yadav, "Machine learning techniques for crop yield prediction: A systematic review," Agritech Advances, vol. 2, no. 1, pp. 15–35, 2024.

[2] J. Li et al., "Predicting crop yields using remote sensing and machine learning approaches," Remote Sensing, vol. 12, no. 5, p. 845, 2020.

[3] F. Ahmed et al., "Hybrid deep learning model for crop yield prediction using climate data," Computers and Electronics in Agriculture, vol. 175, p. 105556, 2020.

[4] M. M. Hoque, M. A. Samad, M. S. Islam, et al., "Incorporating meteorological data and pesticide information to forecast crop yields using machine learning," IEEE Access, vol. 11, pp. 18045–18058, 2023.

[5] A. Kumar and S. Rani, "Remote sensing and crop modeling integration for yield prediction," MDPI Remote Sensing, vol. 14, no. 9, p. 1990, 2022.

[6] L. Zhang, X. Li, and Y. Wang, "CNN-LSTM-based models for smart farming yield forecasting," Computers and Electronics in Agriculture, vol. 177, p. 105707, 2020.

[7] R. T. Sousa et al., "Crop yield prediction using machine learning and climatic variables: A comparative study," Applied Sciences, vol. 13, no. 16, p. 9288, 2023.

[8] H. J. Kim et al., "Coupling crop modeling with ML techniques to improve yield prediction under climate variability," Scientific Reports, vol. 10, p. 12345, 2020.

[9] R. P. Singh and M. K. Sharma, "Influence of meteorological features on crop yield: A review," MDPI Remote Sensing, vol. 14, p. 1990, 2022.

[10] S. Patel and J. R. Mehta, "Climate-based crop yield prediction using ANN and RF models," Computers and Electronics in Agriculture, vol. 175, p. 105556, 2020.

[11] Y. Chen et al., "Integration of satellite-derived climate data in crop yield modeling," Remote Sensing, vol. 14, p. 1990, 2022.

[12] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," Advances in Neural Information Processing Systems, 2017.

[13] D. H. Wolpert, "Stacked generalization", Neural Networks, vol. 5, no. 2, pp. 241-259, 1992.

[14] Saha S., Kucher O.D., Utkina A.O., Rebouh N. Y. Precision agriculture for improving crop yield predictions: a literature review. Frontiers in Agronomy, p.1566201, 2025.