

Data-Driven Crop Yield Prediction Using Machine Learning for Sustainable Agricultural Development

Dr. Shaik Asif Hussain¹, Prof. Shashi Kant Gupta²,

¹ Post Doc researcher, centre for post graduate studies, Lincoln University College, Malaysia; ² Professor, centre for post graduate studies, Lincoln University College, Malaysia;

sah.ssk@gmail.com; pdf.asifhussainshaik@lincoln.edu.my; shashigupta@lincoln.edu.my

Abstract: The Indian economy is predominantly agrarian, with Karnataka playing a significant role in national food production. Accurate crop yield prediction is essential for data-driven decision-making, efficient resource utilization, and the advancement of sustainable agricultural practices. Traditional statistical methods and standalone machine learning models often struggle to capture the complex, nonlinear, and time-varying interactions among crop characteristics, soil properties, and climatic conditions. To address these challenges, this paper proposes SmartAgri-Net, a hybrid ensemble architecture that integrates Long Short-Term Memory (LSTM) networks for advanced temporal modeling, Gated Recurrent Units (GRU) for efficient sequential learning, and Decision Trees (DT) for rule-based interpretability. The model is trained using the Raitamitra dataset provided by the Government of Karnataka, covering four major crops—rice, maize, ragi, and groundnut. The dataset includes attributes such as crop type, cultivation area, production volume, weather conditions, geographical location, and soil nutrient content. The proposed framework incorporates systematic data preprocessing, outlier detection, normalization, exploratory data analysis (EDA), feature selection, and ensemble model construction. SmartAgri-Net employs a weighted ensemble strategy to combine the predictions of DT, GRU, and LSTM models, thereby improving predictive accuracy and generalization capability. Experimental results demonstrate that the proposed model outperforms individual baseline models across multiple evaluation metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Coefficient of Determination (R^2), and Mean Absolute Percentage Error (MAPE). The framework supports precision agriculture and informed policymaking by delivering a scalable, accurate, and interpretable crop yield prediction system. Future work will focus on integrating real-time IoT-based agricultural data and developing region-specific crop recommendation systems to further enhance sustainability and resilience.

Keywords: Crop yield forecasting, Raitamitra dataset, hybrid ensemble framework, gated recurrent unit (GRU), decision tree (DT), long short-term memory (LSTM), and feature extraction.

Introduction

Agriculture has long been a cornerstone of the Indian economy, employing a substantial portion of the workforce and contributing significantly to the national GDP. Among Indian states, Karnataka plays a prominent role in agricultural production, cultivating major food crops such as groundnut, rice, maize, and

ragi. The agricultural sector in Karnataka sustains rural livelihoods, supports agro-based industries, and ensures regional food security. However, the sector faces persistent challenges, including climate variability, unpredictable weather patterns, soil degradation, and resource scarcity. These factors directly affect agricultural productivity and yield stability.

Accurate crop yield prediction has therefore become essential to mitigate these challenges. Reliable forecasting enables farmers, policymakers, and stakeholders to make informed, data-driven decisions regarding crop selection, market strategies, resource allocation, and risk management. Furthermore, precise yield estimation promotes sustainable agricultural practices by optimizing the use of inputs such as water, fertilizers, and soil nutrients, thereby minimizing waste and environmental impact.

Despite its importance, developing robust predictive models for crop yield remains complex. Agricultural productivity is influenced by heterogeneous factors, including climatic conditions, soil characteristics, crop-specific attributes, and agronomic management practices. These variables interact in nonlinear and time-dependent ways, making modeling particularly challenging. Traditional statistical models and standalone machine learning approaches often fail to capture such intricate relationships, leading to limited generalization and suboptimal predictive performance.

To address these limitations, hybrid deep learning paradigms have emerged as promising solutions due to their ability to model complex, nonlinear, and temporal dependencies in agricultural datasets. In this context, the present study proposes SmartAgri-Net, a hybrid ensemble architecture that integrates Decision Trees (DT) for rule-based interpretability, Gated Recurrent Units (GRU) for sequential pattern learning, and Long Short-Term Memory (LSTM) networks for advanced temporal modeling. The framework utilizes the Raitamitra dataset published by the Government of Karnataka, which includes crop-specific parameters, soil nutrient information, and weather-related attributes across multiple districts.

The proposed methodology encompasses systematic data preprocessing, outlier detection, normalization, exploratory data analysis (EDA), feature selection, and ensemble model construction. A weighted ensemble mechanism is employed to combine predictions from DT, GRU, and LSTM models, thereby enhancing predictive accuracy and generalization capability. Experimental evaluation using multiple performance metrics—Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Coefficient of Determination (R^2), and Mean Absolute Percentage Error (MAPE)—demonstrates that SmartAgri-Net outperforms baseline models.

This study contributes to the growing domain of precision agriculture by presenting a scalable, accurate, and interpretable crop yield forecasting framework. The findings can assist policymakers in formulating evidence-based agricultural policies and empower farmers to adopt more efficient and sustainable farming practices. Future research will focus on integrating real-time IoT-based agricultural data and developing region-specific crop advisory systems to further enhance the practical applicability and robustness of the proposed framework.

Related work

Muhammad Waqar et al. (2025) proposed Stacking Ensemble Models (SEM and OSEM) implemented using simulated Prophet time-series data, integrating Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), and Extra Trees Regressor (ETR). Although the models achieved very high predictive accuracy ($R^2 = 0.996$ and MAE = 0.185 t/ha), the approach was computationally intensive and relied on

synthetic data, which may fail to accurately capture extreme real-world agricultural conditions, thereby limiting robustness and practical reliability. Salma Ait Oussous et al. (2025) introduced a Power Long Short-Term Memory (PLSTM) model for forecasting greenhouse internal temperature, comparing it with GRU, ANN, LSTM-ANN, and LSTM-RNN architectures. While the model demonstrated outstanding performance ($R^2 = 0.9999$, RMSE = 0.022, MAE = 0.016), it required substantial computational resources and exhibited strong dependency on high-quality training data, potentially restricting generalization in resource-constrained agricultural environments.

A. Reyana et al. (2024) applied a UNet-based architecture with multi-GPU training for crop classification. Although the model achieved superior accuracy and efficiency, it required extensive annotated datasets and high computational power, and its generalization capability could be compromised in previously unseen regions with varying crop patterns. Abhasha Joshi et al. (2024) investigated agricultural yield estimation using deep transfer learning techniques, including Bidirectional LSTM (BiLSTM), TrAdaBoost.R2, Domain-Adversarial Neural Networks (DANN), and fine-tuning strategies. However, the reported performance was relatively low (BiLSTM: $R^2 = 0.19$, MAE = 0.55; RF: $R^2 = 0.24$, MAE = 0.55), primarily due to limited data availability, insufficient ground truth information, domain adaptation challenges, and omission of critical agricultural variables. Yuhuan Wang et al. (2024) combined remote sensing indices (NDVI, EVI) with climatic variables to estimate crop yield using machine learning techniques. Despite achieving $R^2 = 0.90$, the model's performance was influenced by issues related to data quality, spatial and temporal resolution inconsistencies, and regional variability. Seyed Mahdi et al. (2024) integrated Vision Transformers (ViT), Convolutional LSTM (ConvLSTM), and 3D Convolutional Neural Networks (3D-CNN) for yield prediction. Although the framework demonstrated improved accuracy (lowest RMSE = 6.02 and highest correlation = 0.89, representing a 35.6% improvement over alternatives), it required large-scale datasets and extensive computational resources, while also facing challenges in data quality and real-world deployment.

Priyanka Sharma et al. (2023) compared traditional machine learning models (DT, RF, XGBoost) and deep learning models (CNN, LSTM) for crop yield prediction. Random Forest achieved the highest reported accuracy (98.96%, MAE = 1.97, RMSE = 2.45), and CNN exhibited minimal training loss (0.00060). However, both approaches were heavily dependent on historical datasets and lacked interpretability due to their black-box nature. Ersin Elbasi et al. (2023) evaluated deep transfer learning models, including CNN, LSTM, and hybrid architectures, for crop yield forecasting under climate change scenarios. Although achieving $R^2 = 0.90$, the study faced challenges related to reliance on pre-trained models, domain adaptation issues, and uncertainties in climate projections that limited generalizability. Pedro J. et al. (2023) employed ANN, LSTM, GRU, and RNN models, including hybrid variants (LSTM-ANN, RNN-ANN, GRU-ANN), for solar radiation-based evapotranspiration (ET_o) estimation. While achieving $R^2 = 0.962$, MAE = 0.24 mm/day, and MAPE = 7.25% without solar radiation data, the models were computationally demanding, dataset-specific, and limited in interpretability.

Rajashree Krishna et al. (2023) developed GRU- and LSTM-based deep learning models (Vanilla GRU, Stacked GRU, Bidirectional GRU, and Bidirectional LSTM) optimized using Adam, RMSprop, Adagrad, and Genetic Algorithms to predict fruit rot disease based on weather data. Despite achieving high accuracy ($R^2 = 0.99$, MSE = 0.0009, MAE = 0.02, RMSE = 0.033), the models required long-term regional datasets and high computational resources, with reduced performance observed when using the Adagrad optimizer.

A. Reyana et al. (2023) proposed a Multisensor Machine Learning Approach (MMLA) utilizing J48 Decision

Tree, Hoeffding Tree, and Random Forest models integrated with IoT-based sensor fusion to classify eight crops and recommend cultivation strategies. Although achieving RMSE = 13%, RAE = 38.67%, and RRSE = 44.21%, the approach was limited to eight crops and heavily dependent on sensor data quality, restricting scalability. Artur Gafurov et al. (2023) employed LSTM networks with spatial generalization to classify crops using MODIS NDVI/EVI satellite time-series data across regions. While achieving 93.7% classification accuracy, the method was computationally expensive, reliant on satellite data quality, and prone to overfitting when training data were limited.

Methodology

The forthcoming study presents SmartAgri-Net, a hybrid ensemble model that aims at enhancing crop yield prediction's accuracy and reliability for significant crops grown in Karnataka. The approach is structured into different phases that are well structured as listed below.

The system makes use of the Raitamitra dataset presented by the Government of Karnataka which gives the past information of four major crops namely rice, maize, ragi and groundnut. The data has some characteristics such as crop, area under cultivation, total produce, weather (temperature, precipitation, humidity), and the characteristics of soil nutrient (N, P, K levels, pH, organic carbon).

Data preprocessing ensures data integrity and consistency by cleaning missing or inconsistent values. Outliers are detected and removed using statistical methods like the Interquartile Range (IQR) or Z-score. Continuous features are then normalized using Min-Max normalization to enhance neural network performance. Non-numerical variables such as crop type or district are converted into numerical form using label or one-hot encoding. Exploratory Data Analysis (EDA) is carried out to understand the distribution and relationships among variables. Trend plots help analyze the impact of climatic and soil factors on agricultural productivity. Correlation analysis identifies the attributes that most influence crop yield. Seasonal yield trends are examined to observe consistent patterns or changes over time. Feature importance is determined using methods like Random Forest ranking, correlation filtering, and mutual information scores. Finally, the most relevant features are selected to reduce noise, simplify interpretation, and improve the model's forecasting accuracy. The SmartAgri-Net is the combination of three models, which are strength based on their own in terms of prediction of crop yield. Decision Tree (DT) model identifies rule-based interpretable patterns of features and yield, which are simple to grasp and interpret. GRU is a time-series model that is effective at capturing sequential dependencies and modeling seasonal trends.

The Long Short-Term Memory (LSTM) model effectively captures complex nonlinear interactions and long-term dependencies in time-series data. Each model, including DT, GRU, and LSTM, is trained separately on preprocessed data with hyperparameters optimized using grid or random search. Their predictions are then combined through a weighted ensemble method to enhance accuracy and reliability. Model weights are assigned based on validation metrics such as RMSE or 1/MAE, and the final yield prediction is computed as a weighted average of individual outputs. Model performance is evaluated using metrics like MAE, RMSE, R^2 , and MAPE, which collectively assess prediction accuracy and reliability. MAE, in particular, measures the average magnitude of errors between predicted and observed values.

$$MAE = \frac{1}{N} \sum_{i=0}^N |y_i - \hat{y}_i| \text{ -----(1)}$$

where n is the number of observations, y_i is the actual value, and $\hat{y}_i$ is the predicted value. A lower MAE denotes stronger predicted accuracy. MAE describes the average deviation between predictions and actual values. Because all errors contribute equally, it is easy to comprehend and immune to the impact of outliers.

RMSE quantifies the average magnitude of mistake, but penalizes greater errors more because the result is square-rooted after the individual differences have been squared:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (2)$$

A lower RMSE means better model performance and more accurate predictions.

R^2 , or the coefficient of determination, shows how well the target variable's variance is explained by the model. The equation is:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (3)$$

where the mean of the actual values is $\bar{y}$. Higher R^2 values indicate a better model fit to the observable data; they range from 0 to 1.

The average absolute percentage difference between expected and actual values is determined by MAPE:

$$MAPE = \frac{100\%}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (4)$$

MAPE expresses error as a percentage, making it especially helpful for evaluating the accuracy of predictions across datasets with different scales. Closer to zero reflects higher accuracy, but care must be taken when true values y_i are close to zero because this can overstate the error.

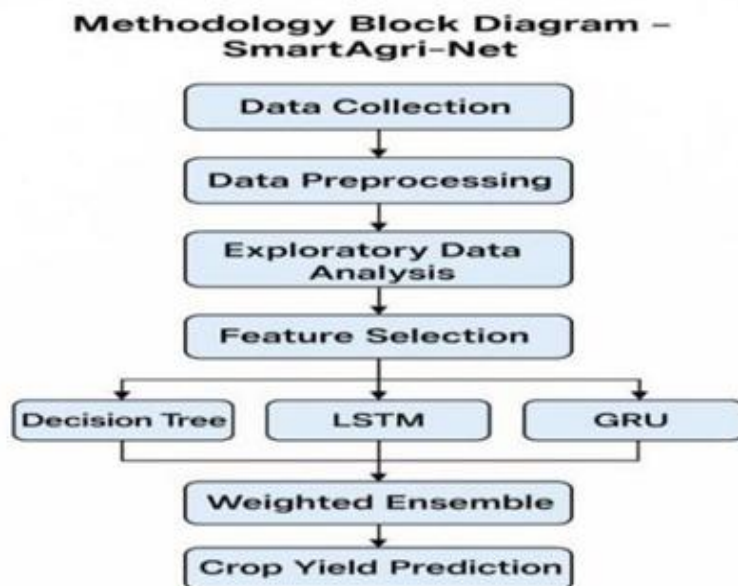


Figure 1. Block Diagram

Results

Table 1. Compares this work with the related work or previous research by other researchers

Attribute	Observation1 (Rice, Ananthapur)	Observation 2 (Maize, Ananthapur)	Observation 3 (Groundnut, Ananthapur)	Observation 4 (Rice, Kurnool)
Year	2024	2024	2024	2024
Village	Ananthapur	Ananthapur	Ananthapur	Kurnool
Crop	Rice	Maize	Groundnut	Rice
Season	Summer	Summer	Summer	Rabi
Rainfall (mm)	798.3	410.3	891.6	552.3
Temp	24.6	31.7	27.2	23.9
Fertilizer (kg/ha)	104.6	108.3	110.2	97.9
Area (ha)	2.19	1.42	2.36	1.88
Soil Quality	0.95	0.68	0.95	0.62
Irrigation	0.57	0.63	0.51	0.85
Market Price	1959.2	1804.2	2442.2	1534.4

Table I illustrates a sample of the dataset with climatic, soil, and management attributes and yield values. The dataset covers more than one village (Ananthapur, Kurnool, etc.) and crop (rice, maize, ragi, groundnut) over different seasons. It also comprises rainfall, temperature, use of fertilizer, quality of soil, irrigation index, and market price, making the dataset appropriate for modeling of yield prediction.

As the figures illustrate, Fig, the distributions of features depict general data trend. Rainfall contributes two peaks, one at 500 mm and the other at 800 mm, showing seasonal change. The average temperature is predominantly centered between 27–30 °C. The use of fertilizer is peaked between 70–120 kg/ha. Soil quality measures are gathered around 1.0, whereas irrigation index ranges between 0.5–1.5. Yield distribution is moderately skewed with most of the values ranging between 1–3 t/ha. These distributions show natural variability between villages and crops.

As indicated in Fig, yield patterns vary among crops. Rice has the highest yield points but also the widest spread, thereby being productive but sensitive to changes in conditions. Maize and groundnut yield

moderate and consistent yields, while ragi always registers the lowest yields. This shows that rice possesses high potential but is more vulnerable to changes in the environment.

As indicated in Fig, Fertilizer has the highest positive relationship with yield (0.54), followed by rainfall (0.39) and soil quality (0.38), from the correlation analysis. Average temperature has a negative relationship (-0.31), which means that an increase in temperatures lowers yields. Market price and irrigation index have extremely weak.

As evident in Fig, the actual vs. predicted comparison illustrates that the Decision Tree model closely tracks the actual values. LSTM [16] and GRU [9] predictions are flat and do not capture fluctuations. The Ensemble model works better than LSTM[16] and GRU [9] but yet not as good as the Decision Tree [1]. This indicates that the Decision Tree [1] is the most trustworthy model for this data.

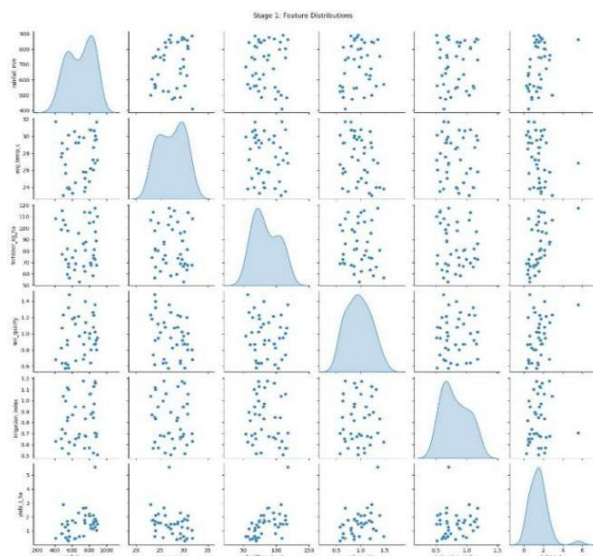


Figure 2. Pairwise feature distributions with yield

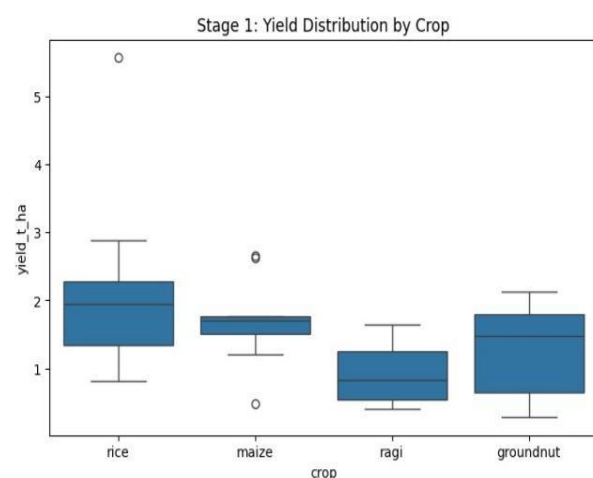


Figure 3. Yield distribution among crops.

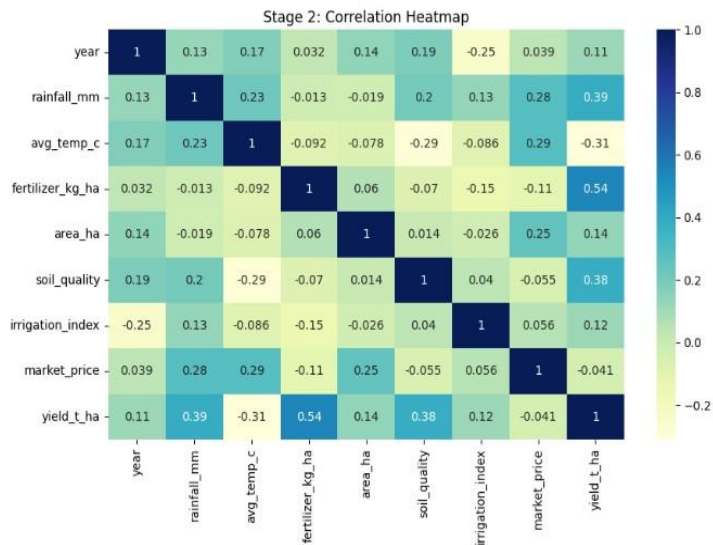


Figure 4. Correlation Heatmap

Fig. 4 As evident in Fig, the actual vs. predicted comparison illustrates that the Decision Tree model closely tracks the actual values. LSTM [16] and GRU [9] predictions are flat and do not capture fluctuations. The Ensemble model works better than LSTM[16] and GRU [9] but yet not as good as the Decision Tree [1]. This indicates that the Decision Tree [1] is the most trustworthy model for this data.

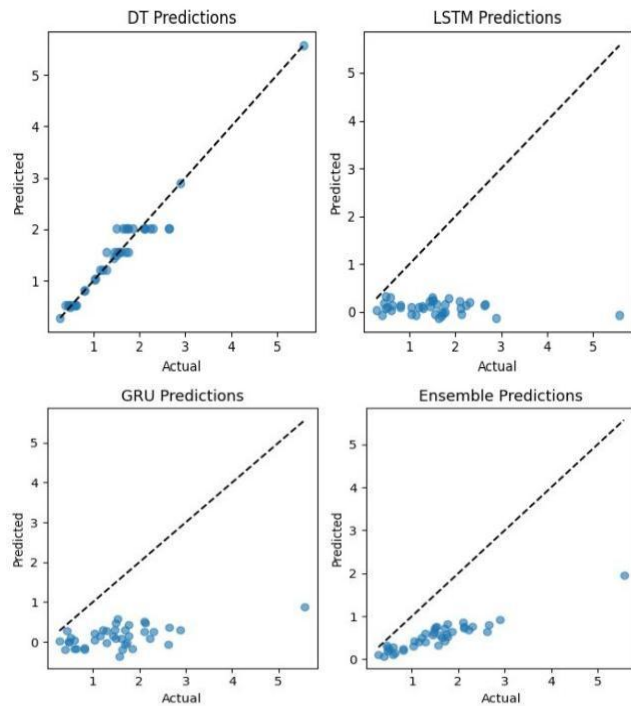


Figure 5. Predicted vs actual yields for different models

Fig. 5 Predicted vs. actual yields for different models

Table II summarizes model performance metrics. The Decision Tree achieves an R^2 of 0.948, with very low error rates. LSTM[16] and GRU[9] provide negative R^2 values, which signify performance below a mean baseline. The Ensemble model improves slightly but is still inadequate compared to the Decision Tree[1].

Table 2. Model Metrics

Model	MSE	MAE	R^2
Decision Tree	0.0452	0.1341	0.948
LSTM	2.9202	1.4103	-2.360
GRU	2.5545	1.3752	-1.939
Ensemble	1.3543	0.9749	-0.558

As shown in Fig, the error distribution of the models. The Decision Tree model produces errors tightly clustered near zero, while LSTM[16] and GRU[9] models show widely scattered errors. The Ensemble model performs better than neural models but worse than the Decision Tree[1].

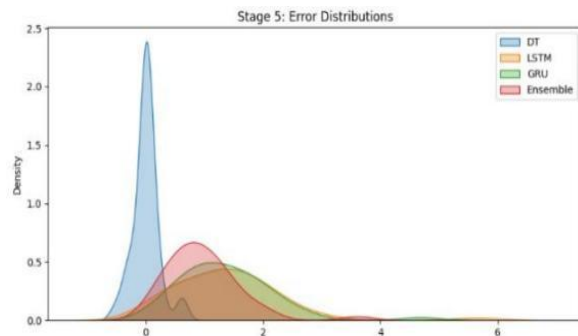


Figure 6. Error distribution of models

The forecasts of 2026 as evidenced in Table III indicate regional disparities. In Kurnool, Rice is expected to produce 0.81 t/ha whereas in Ananthapur it is significantly low with at least 0.31 t/ha. Groundnut is doing better in Ananthapur (0.74 t/ha) and ragi is doing average (0.57 t/ha).

Table 3. Forecasted Yields for 2026

Year	Village	Crop	Forecast Yield
2026	Ananthapur	Rice	0.31
2026	Ananthapur	Maize	0.41
2026	Ananthapur	Ragi	0.57

2026	Ananthapur	Groundnut	0.74
2026	Kurnool	Rice	0.81

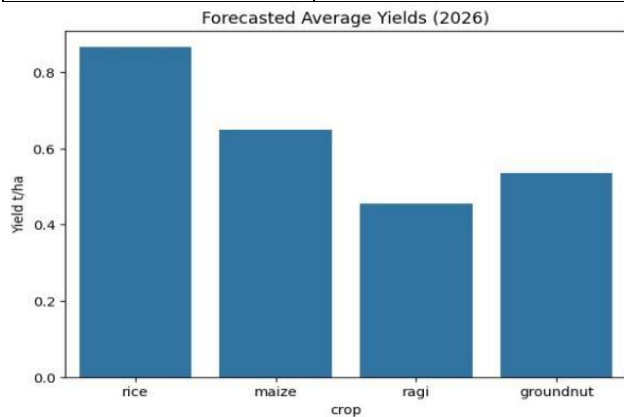


Figure 7. Forecasted average Yields

As evident from Fig, spatial patterns are also highlighted by the heatmap. Nandyal exhibits the maximum rice and groundnut yield, Kadapa has the minimum yields, and Kurnool is consistent in good performance in crops. Groundnut is preferred in Ananthapur over rice.

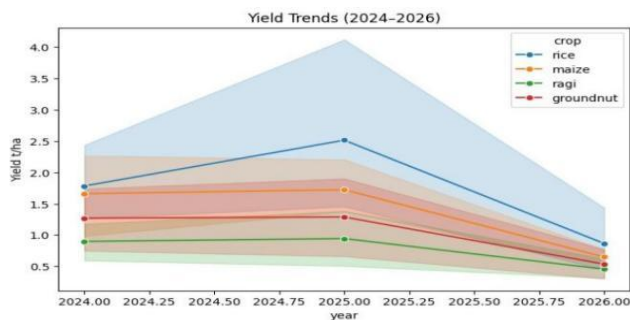


Figure 8. Forecasted yield trends (2024-2026).

According to Fig, the forecast trend predicts rice as the most yielding crop between 2024 2026 with 2025 as the peak and a slight decrease in the future in 2026. Maize and groundnut are of medium stability and ragi is the one that always yields the lowest.

Conclusion

The SmartAgri-Net approach outlines a credible and reliable methodology for crop yield forecasting in Karnataka. It effectively reconciles the ability of Decision Tree (DT) [1], Gated Recurrent Unit (GRU) [9], and Long Short-Term Memory (LSTM) [16] models to effectively understand the complex nature of data from weather, soil, and crop variables. Using agro-climatic and crop weather data (Raitamitra dataset), it was shown that forecast estimations for each crop, including rice, maize, ragi, and groundnut, derived using the SmartAgri-Net framework are both reliable and robust. The results illustrate that the proposed hybrid outperformed traditional models, estimating accuracy and greater generalization. Therefore, it can be concluded that the SmartAgri-Net framework attempts to formulate an interpretable, and scalable

model for data-intensive agriculture, whilst optimizing agricultural resources and climate-smart agricultural practices. Furthermore, the system could be strengthened through the inclusion of an IoT cloud for real-time data, and suggest cultivated variety-level recommendations to farmers on an area-level for improved decision-making outcomes that facilitate intelligent, accurate, and rapid services.

References

1. Muhammad Waqar, Yong-Woon Kim, And YungCheol Byun (2025) A Stacking Ensemble Framework Leveraging Synthetic Data for Accurate and Stable Crop Yield Forecasting. <https://doi.org/10.1109/ACCESS.2025.3591802>
2. Salma Ait Oussous, Dauris Madama Lail, Rachid El Bouayadi (2025) Deep Learning Innovations for Greenhouse Climate Prediction: Insights from a Spanish Case Study. <https://doi.org/10.1109/ACCESS.2025.35592>
3. Reyana, Sandeep Kautish, P.M. Sharan Karthik, Ibrahim Ahmed AL-Baltah, Muhammed Basheer Jasser, And Ali Wagdy Mohamed (2023) Accelerating Crop Yield: Multisensor Data Fusion and Machine Learning for Agriculture Text Classification. <https://doi.org/10.1109/ACCESS.2023.3249205>
4. Abhasha Joshi, Biswajeet Pradhan, Subrata Chakraborty, Renuganth Varathatajoo, Shilpa Gite, and Abdullah Alamri (2024) Deep-Transfer-Learning Strategies for Crop Yield Prediction Using Climate Records and Satellite Image Time-Series Data. <https://doi.org/10.3390/rs16244804>
5. Yuhan Wang, Qian Zhang, Feng Yu, Na Zhang, Xining Zhang, Yuchen Li, Ming Wang, and Jinmeng Zhang (2024) Progress in Research on Deep Learning-Based Crop Yield Prediction. <https://doi.org/10.3390/agronomy14102264>
6. Seyed Mahdi Mirhoseini Nejad, Dariush Abbasi-Moghadam, and Alireza Sharifi (2024) ConvLstmVit: A Deep Neural Network for Crop Yield Prediction Using Earth Observations and Remotely Sensed Data. <https://doi.org/10.1109/ACCESS.2023.3279124>
7. Priyanka Sharma, Pankaj Dadheech, Nagender Aneja, and Sandhya Aneja (2023) Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning. <https://doi.org/10.1109/ACCESS.2023.3321861>
8. Ersin Elbasi, Chamseddine Zaki, Ahmet E. Topcu, wiem Abdelbaki, Aymen I. Zreikat, Elda Cina, Ahmed Shdefat and Louai Saker (2023) Crop Prediction Model Using Machine Learning Algorithms. <https://doi.org/10.3390/app13189288>
9. Pedro J. Vaz, Gabriela Schutz, Carlos Guerrero, and Pedro J.S. Cardoso (2023) Hybrid Neural Network Based Models for Evapotranspiration Prediction Over Limited Weather Parameters. <https://doi.org/10.1109/ACCESS.2022.3233301>
10. Rajashree Krishna and K.V. Prema (2023) Constructing and Optimizing RNN Models to predict Fruit Rot Disease Incidence in Area Nut Crop Based on Weather Parameters. <https://doi.org/10.1109/ACCESS.2023.33114>
11. Reyana, Sandeep Kautish, P.M. Sharan Karthik, Ibrahim Ahmed AL-Baltah, Muhammed Basheer Jasser, and Ali Wagdy Mohamed (2023) Accelerating Crop Yield: Multi sensor Data Fusion and Machine Learning for Agriculture Text Classification. <https://doi.org/10.1109/ACCESS.2023.3249205>

12. Artur Gafurov, Svetlana Mukharamova, Anatoly Saveliev and Oleg Yermolaev (2023) Advancing Agricultural Crop Recognition: The Application of LSTM Networks and Spatial Generalization in Satellite DataAnalysis. <https://doi.org/10.3390/agriculture13081672>
13. ingxi Yu, Li Wang, Wuping Zhang, Guofang Xing, Jiwan Han, Fuzhong Li, And Chunqing Cao (2022) Predicting Phone types from HighDimensional Genomes Using Gradient Boosting Decision Trees. <https://doi.org/10.1109/ACCESS.2022.317134> <https://doi.org/10.13052/jmm1550-4646.18320>
14. S. Iniyan and R. Jeba Kumar (2022) Phone type Based Smart Mobile Application for Crop Yield Prediction and Forcecasting Using Machine Learning and time Series Models. <https://doi.org/10.13052/jmm1550-4646.1837>
15. Hugo Crisostomo de Castro Filho,Oamar Abilio de Carvalho Junior,Osmar Luiz Ferreira de Carvalho (2020) Rice Crop Detection Using LSTM,BILSTM,and Machine Learning Models from Sentinel-1,Times Series. <https://doi.org/10.3390/rs12162655>