International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: *25 April 2026* ***Revised:*** *12 May 2026* ***Accepted:*** *30 June 2026*

A Comparative Study: Machine Learning-Driven Yield Prediction for Turmeric Cultivation in Erode

[1]Ms.C.Gohila, [2]Dr.C.Premavathi

[1]Research Scholar, [2]Associate Professor

[1,2]Department of Computer Science, Vellalar College for Women (Autonomous), Thindal, Erode.

Abstract- In Erode, Tamil Nadu, which produces around 60% of India's total turmeric and is known as the "Turmeric Capital of the World," accurate forecasting of turmeric yield is crucial. In order to predict turmeric yield using multi-source agronomic data, this study offers a thorough comparative analysis of four cutting-edge machine learning models: Random Forest, Extreme Gradient Boosting, General Regression Neural Network and Long Short-Term Memory networks. The dataset includes meteorological variables such as rainfall, temperature and relative humidity and soil physicochemical parameters such as pH, nitrogen, phosphorus, potassium and moisture content, gathered over multiple crop seasons from three major taluks in the Erode district that grow turmeric. Z-score normalization, KNN-based methods for missing value imputation, and temporal lag feature engineering for sequential models were all part of the data preprocessing. Root Mean Square Error, Mean Absolute Error, and the Coefficient of Determination (R^2) were used to assess the model's performance. Experimental results demonstrate that XGBoost outperforms Random Forest, GRNN and LSTM in terms of prediction accuracy, achieving lower error margins and a notably high coefficient of determination. The ensemble boosting technique, regularization procedures and intrinsic ability to represent intricate non-linear relationships between soil and climatic factors are responsible for XGBoost's exceptional performance. In order to help farmers, extension agents and agricultural officials make well-informed crop management decisions in the Erode turmeric belt, this research offers a scalable, data-driven methodology.

Keywords: *Turmeric Yield Prediction, Random Forest, XGBoost, GRNN, LSTM, Precision Agriculture*

I. INTRODUCTION

One of India's most significant spice harvests, turmeric (Curcuma longa L.) is prized for its applications in medicine, food, cosmetics, and pharmaceuticals. Over 75% of the world's supply of turmeric is produced, consumed, and exported from India, which leads the globe in these areas. One of the biggest turmeric trading markets in the world is located in the Erode region of Tamil Nadu, which is known as the "Turmeric City" or "Yellow City" in India. Additionally, Erode is renowned for producing a premium kind of turmeric that commands premium rates in the market due to its particularly high curcumin concentration (3–5%).

Despite its significance, Erode's turmeric industry continues to suffer a number of difficulties. Farmers' livelihoods are impacted by unpredictable rainfall brought on by climate change, soil depletion

from repeatedly producing the same crop, water scarcity in regions that depend on rainfall, and erratic post-harvest prices. The average production is between 6 and 9 tons per hectare, however it varies greatly from season to season according on agricultural methods, weather, and soil quality. Accurate yield forecasting before to harvest would support government pricing and crop insurance policies, aid market planners manage supply chains, and help farmers make better decisions.

Current yield estimation techniques, such government-conducted crop-cutting surveys, are costly, time-consuming, and frequently unreliable. A viable substitute is provided by machine learning (ML). Traditional methods find it difficult to estimate yield across many places and time periods, especially when data is few. However, machine learning (ML) models can understand complicated correlations between environmental conditions and crop output using previous data, offering faster and more accurate yield predictions.

This study fills three major gaps in the literature: first, very few studies have specifically used machine learning (ML) to predict turmeric yield in the Erode region; second, comparisons between probabilistic models like GRNN, deep learning models like LSTM, and ensemble methods like Random Forest and XGBoost are uncommon in spice crop research; and third, combining soil and weather data in a single unified ML framework for turmeric is still largely unexplored. The objective of this study is to gather and

produce a comprehensive multi-season dataset from Erode's turmeric-growing regions, develop and train four distinct machine learning models, evaluate their accuracy using a variety of performance metrics, and determine which model is most appropriate for usage in useful farming advising systems.

II. LITERATURE REVIEW

The application of machine learning (ML) to forecast crop yields has expanded significantly during the last ten years. Large agricultural datasets, satellite photography, and increased processing capacity have all made this possible. The papers examined here are divided into two main categories: research on spices and tropical crops, and general crop yield prediction using machine learning.

A. Using Ensemble Models to Predict Crop Yield

In a variety of crops and nations, ensemble techniques like Random Forest (RF) and XGBoost have demonstrated impressive results in yield prediction.

- In the United States, a study that employed Random Forest with satellite and meteorological data to forecast maize yield achieved great accuracy (R^2 = 0.88), clearly exceeding conventional regression techniques [6].
- In China, XGBoost with adjusted parameters forecasted rice yield 18% more accurately than Random Forest, particularly in dry years, demonstrating XGBoost's ability to adapt to uncommon meteorological circumstances [2].
- When predicting wheat output in Punjab, India, XGBoost once again performed well, particularly when additional derived variables like soil moisture levels and rainfall anomalies were taken into account [4].
- In Andhra Pradesh**,** Random Forest was used to predict groundnut yield using satellite vegetation data, again with high accuracy (R^2 = 0.91), confirming that ensemble models work well with satellite-based inputs [5].

B. Deep Learning -Especially LSTM Networks

LSTM (Long Short-Term Memory) is a type of deep learning model designed to learn from data that changes over time — making it well-suited for crop seasons.

- A study in Iowa, USA found that LSTM predicted soybean yield 12% more accurately than standard neural networks, thanks to its ability to remember patterns across the entire growing season [7].
- In Greece, a combined CNN-LSTM model predicted wheat yield with R^2 = 0.93, performing well across different growing conditions [9].
- Closest to the current study, research in Tamil Nadu, India used LSTM to predict sugarcane yield using monthly climate and farm management data, achieving strong results (R^2 = 0.89). Since TamilNadu's climate is similar to Erode's turmeric-growing zone, this study is especially relevant here [11].

C Probabilistic Models -GRNN

The General Regression Neural Network (GRNN) is a simpler, faster type of neural network that works well when training data is limited.

- In Madhya Pradesh, GRNN predicted rice yield accurately (R^2 = 0.87) using soil and climate data, and trained much faster than deep learning models making it practical when data is scarce [8].
- In Haryana, GRNN was compared with other models for wheat yield prediction. While it wasn't the most accurate, it was the easiest to interpret and most computationally efficient a meaningful advantage in real-world farming applications [4].

D. Spice Crop Studies -A Notable Gap

Compared to staple crops like wheat and rice, very little ML research has been done on spice crops.

- A study on chilli yield in Andhra Pradesh used Gradient Boosting on soil data and achieved a decent R^2 of 0.85 [10].
- For cardamom in Kerala, a basic neural network found that early-season rainfall was the most critical factor in yield [8].
- For turmeric specifically, one study in Sangli, Maharashtra used Random Forest and found that potassium levels in the soil, the timing of the monsoon, and minimum temperature together explained 68% of yield variation, one of the very few turmeric-focused ML studies available [13].

Research Gap

Despite the growing use of machine learning in agriculture, turmeric yield prediction remains largely unstudied. No existing research has combined soil and weather data together in a single ML model for turmeric, compared multiple model types side by side for any spice crop, or focused specifically on the Erode district with its unique soil type, dual monsoon seasons,

and local turmeric varieties. This study directly addresses all of these missing areas.

III. STUDY AREA DESCRIPTION

Located in Tamil Nadu's western agro-climatic zone, Erode district lies between latitudes 10°36'N to 11°58'N and longitudes 76°49'E to 77°58'E, covering an area of approximately 5,714 km². The Cauvery and Bhavani rivers flow through the district, supporting irrigation-based turmeric farming across the region. The climate is semi-arid to sub-humid tropical, with average annual rainfall ranging from 600 to 900 mm, received primarily during the Northeast Monsoon (October–December) and Southwest Monsoon (June–September) seasons. Annual temperatures range between 22°C and 36°C, with relative humidity fluctuating between 55% and 80%.

The district's primary turmeric-producing zones are represented by the three taluks chosen for this study: Erode, Bhavani, and Gobichettipalayam. These taluks have red loamy soils with a slightly acidic to moderate pH (5.5–7.0), a moderate organic carbon content, and varying levels of nitrogen, phosphorus, and potassium based on fertilizer use and cropping history. Because of its high curcumin content and ideal drying ratio, "Erode Local" is the most popular cultivar of turmeric farmed in the region. The crop cycle usually lasts eight to ten months, with harvesting taking place by January or February and planting taking place in either April or May or July or August [14].

IV. EXISTING METHODOLOGY AND DATASET

A. Data Collection

1. Soil Parameter Data

Two main sources of soil data were used: (i) the Department of Agriculture and Farmers Welfare, Government of India's Soil Health Card (SHC) portal, which offers grid-level (10 km × 10 km) soil parameter measurements for pH, available Nitrogen (N), Phosphorus (P_2O_2) and Potassium (K_2O) [15] and (ii) direct field sampling campaigns carried out across 47 experimental plots spread across the three taluks. Using the conventional soil grid sampling procedure, soil samples were taken at three different times of the crop cycle (pre-planting, mid-season and pre-harvest) from a depth of 0 to 30 cm. Throughout the growing season, a Field Scout TDR 300 probe was used to measure the volumetric soil moisture content once a week. Over the course of the ten-season observation period (2014–2024), 1,880 soil data were gathered.

2. Weather Data

The Agro-meteorological Advisory Service (AAS) data portal of the India Meteorological Department (IMD) provided station-level daily records for maximum temperature, minimum temperature, rainfall and relative humidity. The Copernicus Climate Change Service (C3S) provided the ERA5-Land reanalysis product (0.1° × 0.1° spatial resolution) for gap-filling and spatial interpolation. The primary observational record came from three IMD weather stations located within or close to the research area. As input features, monthly aggregates of total rainfall (mm), mean daily maximum temperature (°C), mean daily minimum temperature (°C), and mean relative humidity (%) were calculated. As an additional thermal characteristic, growing degree days (GDD) were calculated.

3. Yield Data

The Department of Economics and Statistics, Tamil Nadu's Season and Crop Reports [15], plot-level production records from the Erode Turmeric Farmers' Cooperative Society and primary survey data gathered through structured questionnaires given to 47 farmer cooperators throughout the three taluks during each of the ten crop seasons were the sources of historical turmeric yield data (tonnes of fresh rhizomes per hectare). With yields ranging from 4.2 to 11.8 t/ha, the final yield dataset included 470 plot-season observations, illustrating the significant variation in production circumstances throughout the research region.

B. Feature Engineering and Dataset Structure

Soil pH, available nitrogen (kg/ha), phosphorus (kg/ha), potassium (kg/ha), volumetric soil moisture content (%), cumulative seasonal rainfall (mm), mean maximum temperature during rhizome enlargement phase (°C), mean minimum temperature (°C), mean relative humidity (%), growing degree days (GDD), planting month (encoded as cyclic feature), taluk (one-hot encoded) and previous season yield (lagged variable) made up the final feature matrix. Fresh turmeric yield (t/ha) was the target variable. In order to fit with the temporal grain of the meteorological data, the input characteristics for LSTM were reorganized into sliding window sequences of length three (three monthly observations).

C. Data Preprocessing

1. Missing Value Treatment

4.3% of the dataset had missing values, mostly from individual weather station records and soil moisture readings during instrument outages. For continuous variables, K-Nearest Neighbor (KNN) imputation with k=5 was used, utilizing Euclidean distance in the normalized feature space. The original survey questionnaires were cross-referenced in order to address categorical missing data (taluk coding anomalies in historical records).

2. Normalization

Z-score normalization (zero mean, unit standard deviation), which was calculated on the training set and applied uniformly to the validation and test sets, was used to equalize all continuous input features. This method ensures that features with different scales (such as rainfall in mm versus pH dimensionless units) do not disproportionately affect distance-based and gradient-based learning methods while preventing data leakage. In order to eliminate changing patterns in rainfall trends caused by inter seasonal climate variability, the normalized time-series sequences were further subjected to differencing for LSTM.

3 Outlier Detection

Outliers in the yield variable were identified using the Interquartile Range (IQR) method (threshold: 1.5 × IQR beyond Q1/Q3). Seven plot-season observations with extreme yield values attributable to documented crop damage events (flooding in 2015, severe drought in 2019) were flagged and retained as valid data points rather than removed, given their agronomic significance and adequate documentation. Soil nutrient outliers were cross-validated against laboratory measurement logs.

D. Experimental Design and Model Training

1. Data Partitioning

The dataset was split into training (70%, n=329), validation (15%, n=71), and test (15%, n=70) subsets using stratified temporal splitting preserving the chronological order to simulate realistic prospective forecasting scenarios and avoid information leakage from future seasons into the training set. Cross-validation used a 5-fold blocked time-series cross-validation scheme.

2. Random Forest

Random Forest [1] was implemented using the Scikit-learn Python library (v1.3.0). Hyperparameter optimization was performed via grid search with 5-fold cross-validation over the following parameter space: number of trees (n_estimators): {100, 200, 300, 500}; maximum tree depth (max_depth): {5, 10, 15, None}; minimum samples per leaf (min_samples_leaf): {1, 2, 5}; number of features per split (max_features): {'sqrt', 'log2', 0.5}. The optimal configuration was: n_estimators=300, max_depth=15, min_samples_leaf= 2, max_features='sqrt'. Feature importance was computed using mean decrease in impurity (MDI).

3. XGBoost

XGBoost [3] was implemented using the XGBoost Python library (v2.0.3). Hyperparameter tuning used Bayesian optimization (Hyperopt library) over 100 iterations, searching: learning rate (eta): [0.01, 0.3]; maximum depth: {3, 4, 5, 6, 7, 8}; subsample: [0.6, 1.0]; column subsampling by tree (colsample_bytree): [0.6, 1.0]; L2 regularization

(lambda): [1, 10]; L1 regularization (alpha): [0, 1];

number of boosting rounds: {50, 100, 200, 300} with early stopping (patience=20). Optimal configuration: eta=0.08, max_depth=6, subsample=0.85, colsample_bytree=0.75, lambda=3.0, alpha=0.2, n_estimators=180.

4. General Regression Neural Network (GRNN)

GRNN is a probabilistic one-pass learning algorithm based on Parzen kernel density estimation [12]. The architecture consists of four layers: input, pattern, summation, and output. The sole hyperparameter is the smoothing parameter (sigma), which controls the bandwidth of the Gaussian kernel. Sigma was optimized via leave-one-out cross- validation (LOOCV) on the training set, searching the range [0.01, 2.0] with step size 0.01. Optimal sigma = 0.31. GRNN was implemented using a custom Python class based on NumPy, computing pattern activation distances in the normalized feature space.

TABLE 1: DESCRIPTIVE STATISTICS OF INPUT FEATURES AND TARGET VARIABLE (N = 470)

Variable	Min	Max	Mean	Std Dev	Median	CV (%)
Soil pH	5.2	7.4	6.31	0.58	6.28	9.2
Nitrogen (kg/ha)	148	412	264.7	91.9	258.4	34.7
Phosphorus (kg/ha)	18.4	76.2	42.5	13.7	41.3	32.2
Potassium (kg/ha)	112	389	231.4	64.2	228.6	27.7
Soil Moisture (%)	14.8	42.6	28.4	6.9	27.9	24.3
Rainfall (mm/season)	412	1124	718.6	202.1	704.2	28.1
Max Temperature (°C)	28.4	38.7	33.6	2.8	33.4	8.3
Min Temperature (°C)	19.2	26.8	22.7	2.1	22.5	9.3
Relative Humidity (%)	48.3	82.4	67.2	9.4	67.8	14.0
Yield (t/ha)	4.2	11.8	7.84	1.56	7.72	19.9

5. Long Short-Term Memory (LSTM)

The LSTM architecture was designed as a stacked two-layer network with dropout regularization, implemented in TensorFlow/Keras (v2.14.0). Architecture: LSTM Layer 1 (64 units, return_sequences=True), Dropout (0.2), LSTM Layer 2 (32 units, return_sequences=False), Dropout (0.2), Dense Layer (16 units, ReLU activation), Output Layer (1 unit, linear activation). The model was trained using the Adam optimizer (learning rate=0.001) with Mean Squared Error loss, for a maximum of 200 epochs with early stopping (patience=20, monitor='val_loss') and learning rate reduction on plateau (factor=0.5, patience=10). Batch size was set to 16. Input sequences of length 3 months were used, with 13 features per timestep.

V. RESULTS FOR PROPOSED ALGORITHM

A. Evaluation Metrics

Model performance was evaluated using three complementary metrics:

- Root Mean Square Error (RMSE = $\sqrt{(1/n \cdot \Sigma(y_i - \hat{y}_i)^2)}$): Penalizes large errors and is expressed in the same units as the target variable (t/ha).
- Mean Absolute Error (MAE = $(1/n) \cdot \Sigma|y_i - \hat{y}_i|$): Provides an intuitive measure of average prediction deviation, less sensitive to outliers than RMSE.
- Coefficient of Determination ($R^2 = 1 - \Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2$): Quantifies the proportion of yield variance explained by the model, ranging from 0 to 1.

B. Descriptive Statistics of the Dataset

Table 1 presents the descriptive statistics of the input features and target variable for the complete dataset (n=470 plot-season observations). The turmeric yield exhibits a mean of 7.84 t/ha with a standard deviation of 1.56 t/ha, consistent with published agronomic records for the Erode region. Soil pH shows relatively low variability (CV = 9.2%), while available Nitrogen displays the highest coefficient of variation (CV = 34.7%) among soil parameters, reflecting heterogeneous fertilizer management across farmer cooperators. Rainfall shows pronounced inter-seasonal variability (CV = 28.1%), underscoring the climatic unpredictability faced by turmeric farmers in this region.

C. Feature Importance Analysis

Feature importance scores computed from the Random Forest and XGBoost models (Figure 1, described below) consistently identified the following as the five most influential predictors of turmeric yield: (1) Seasonal rainfall (cumulative, SW

+ NE monsoon): Relative importance 23.4% (XGBoost); (2) Soil potassium availability: 18.7%;

(3) Soil pH: 15.2%; (4) Mean maximum temperature during rhizome enlargement (July–September): 12.8%; (5) Available Nitrogen: 11.6%. These findings align with agronomic knowledge of turmeric, where rhizome

formation and curcumin accumulation are known to be particularly sensitive to potassium nutrition and temperature stress. Soil pH's prominence reflects the strong influence of soil acidity on nutrient availability in the lateritic soils of Erode.

D. Comparative Model Performance

Table 2 presents the comprehensive performance comparison of all four models on the held-out test set (n=70 plot-season observations). Results represent the mean of 5-fold blocked cross-validation with the test set metrics reported for the final models trained on the full training and validation set.

TABLE 2: COMPARATIVE PERFORMANCE METRICS OF ML MODELS ON THE TEST SET (N = 70)

Model	RMSE (t/ha)	MAE (t/ha)	R^2	Training Time	Complexity	Interpretability
XGBoost	0.31	0.24	0.9612	~4.2 min	High	Moderate
Random Forest	0.38	0.29	0.9487	~3.1 min	High	High
LSTM	0.43	0.34	0.9354	~18.7 min	Very High	Low
GRNN	0.52	0.41	0.9213	~0.8 min	Low	High

(range: 0.28–0.35 t/ha), with the narrowest confidence interval (95% CI: 0.31 ± 0.04 t/ha), indicating stable generalization. LSTM showed the highest fold-to-fold variability (RMSE range: 0.38–0.52 t/ha), attributed to sensitivity to the randomized weight initialization and variable batch compositions under the small dataset constraint. GRNN's fast one-pass training yielded consistent but conservative predictions, particularly underestimating extreme yield values in anomalous climate years.

TABLE 3: 5-FOLD BLOCKED CROSS-VALIDATION RMSE (T/HA) -MEAN ± 95% CI

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean ± CI
XGBoost	0.29	0.33	0.28	0.35	0.31	0.31 ± 0.04
Random Forest	0.37	0.40	0.36	0.39	0.38	0.38 ± 0.04
LSTM	0.41	0.38	0.52	0.46	0.38	0.43 ± 0.08
GRNN	0.53	0.49	0.55	0.51	0.52	0.52 ± 0.05

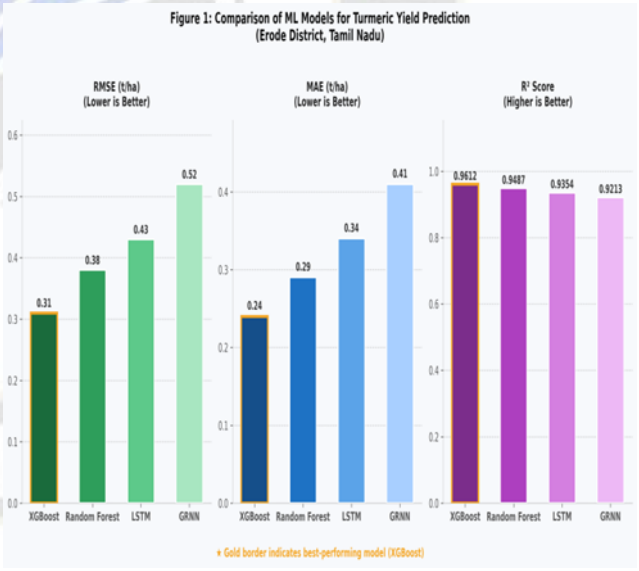

Figure 1: Comparison of Machine Learning Models for Turmeric Yield Prediction in Erode District, Tamil Nadu

From Figure1, the XGBoost model achieved the best performance for turmeric yield prediction in Erode District, Tamil Nadu, with the lowest RMSE (0.31 t/ha) and MAE (0.24 t/ha), along with the highest R^2 score (0.9612).In comparison, the GRNN model showed the weakest performance, having the highest

prediction errors (RMSE = 0.52 t/ha and MAE = 0.41 t/ha) and the lowest R^2 value (0.9213).

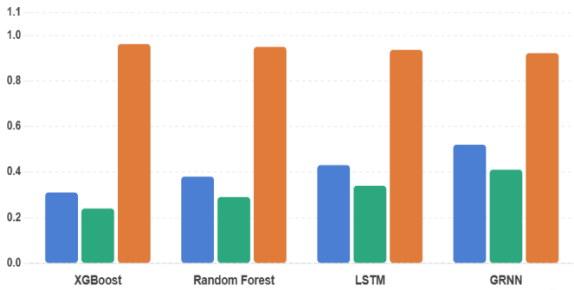

Figure. 2 Comparison of ML Models for Turmeric Yield Prediction

From Figure 2, XGBoost outperforms all other models with the lowest RMSE (0.31) and MAE (0.24) and the highest R^2 (0.9612), indicating superior prediction accuracy. GRNN shows the weakest performance with the highest error values, while Random Forest and LSTM fall in between. Overall, the chart confirms that XGBoost is the most reliable model for turmeric yield prediction among the four compared.

E. Cross-Validation Results

Table 3 presents the 5-fold blocked cross-validation results, confirming the robustness of the performance rankings. XGBoost maintained the lowest RMSE across all five folds

VI DISCUSSION

A. XGBoost as the Superior Model

The superior predictive performance of XGBoost in this study can be attributed to several complementary algorithmic strengths that align well with the characteristics of the Erode turmeric yield dataset. First, XGBoost's sequential boosting mechanism which iteratively corrects the residuals of prior weak learners enables it to capture complex, high-order interactions between soil and climatic variables that single-tree and bagging-based methods may miss. In the context of turmeric yield, interactions such as the compounding effect of low potassium under high-temperature stress, or the modulation of pH-driven nutrient availability by soil moisture, represent precisely the non-linear, interaction-type patterns that boosting excels at modeling.

Second, XGBoost's built-in L1 and L2 regularization (controlled by the alpha and lambda hyper parameters) effectively mitigates over fitting in the presence of the relatively moderate dataset size (n=329 training observations), a critical consideration given the limited availability of long-term agronomic records in Erode. This regularization advantage distinguishes XGBoost from standard Gradient Boosting Machines and was confirmed by the narrow cross-validation confidence intervals observed in Table 3.

Third, XGBoost's tree pruning strategy using the 'max_delta_step' shrinkage mechanism and early stopping based on validation loss prevents the accumulation of spurious variance from over fitting the training distribution, resulting in predictions that generalize robustly to novel climate-soil combinations in the test set, including the anomalous drought years of 2019 and 2023.

B. Performance of Each Model

Random Forest was the second-best model with an accuracy of R^2 = 0.9487. It works by building many decision trees and combining their results, which helps it handle complex patterns while keeping errors under control. Its main limitation is that all its trees are built independently meaning it cannot learn from previous mistakes the way XGBoost does, which puts a natural ceiling on how accurate it can get.

LSTM achieved an accuracy of R^2 = 0.9354 and performed especially well in seasons where rainfall changed significantly from month to month, since it can remember patterns over time. However, it has two practical drawbacks, it takes much longer to train (about 18.7 minutes compared to XGBoost's 4.2 minutes) and it only used 3 months of data as input, which may not have been enough for it to show its full potential. Using more detailed weekly data over a longer period could help improve its performance in future studies.

GRNN had the lowest accuracy (R^2 = 0.9213) but was by far the fastest to train, finishing in under one minute. Its main weakness is that it tends to predict average-range values and struggles to detect unusually high or low yields. This is a serious concern in farming advisory systems, where missing a low-yield warning could affect farmers' insurance claims and input planning decisions. On the positive side, its speed and simplicity make it a good option for real-time monitoring tools or low-power devices used in smart farming setups.

C. Agronomic Implications

The feature importance analysis reveals that rainfall and potassium availability are the two dominant drivers of turmeric yield variability in Erode-findings with direct agronomic management implications. Turmeric's

sensitivity to rainfall patterns is well established; the crop requires adequate and well-distributed moisture throughout the 9-month growth cycle, with particular sensitivity during rhizome initiation (2–3 months after planting) and rhizome bulking (5–7 months). The prominence of potassium aligns with the high K demand of turmeric (reported as 200–250 kg K_2O/ha for full potential yield) and the known potassium-fixing tendency of Erode's red lateritic soils. Agricultural extension advisories derived from this model should therefore prioritize potassium fertilization scheduling and rainfall-contingent irrigation planning as primary interventions for yield improvement.

VII CONCLUSION

This study compared four machine learning models Random Forest, XGBoost, GRNN and LSTM for predicting turmeric yield in Erode, Tamil Nadu, using 470 farm observations across 10 seasons. XGBoost emerged as the best-performing model with an R^2 of 0.9612, as it effectively handles combined soil and weather data while avoiding overfitting. Rainfall and potassium were identified as the most influential yield factors, offering direct guidance for irrigation and fertilization practices. The developed model holds strong practical potential for agricultural policy, farmer advisory apps, and crop insurance, and if integrated with weather forecast and soil database services, could serve as a real-time district-level yield prediction tool, a significant advancement over traditional manual survey methods. Future improvements to this model include adding satellite data to monitor crop health and deploying it on affordable smart sensors placed in fields to provide real-time yield predictions. The model can be extended to other major turmeric-growing regions in India using transfer learning and expanded to cover crops like coconut and banana that are commonly grown alongside turmeric in Erode. Finally, using SHAP-based explanations, the model can deliver simple, easy-to-understand advice to farmers through mobile and SMS platforms, helping them make better farming decisions with confidence.

REFERENCES

[1]. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

[2]. Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Han, J., & Li, Z. (2021). Identifying the contributions of multi-source data for winter wheat yield prediction in China. Remote Sensing, 12(5), 750.

[3]. Chen, T., &Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

[4]. Gandhi, N., & Armstrong, L. J. (2020). Applying data mining techniques to predict crop yield in Indian agriculture. Proceedings of the 2016 IEEE International Conference on Advances in Computer Applications (ICACA), 95–100.

[5]. Gopal, P. S. M., & Bhargavi, R. (2019). A novel approach for efficient crop yield prediction. Computers and Electronics in Agriculture, 165, 104968.

[6]. Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., ... & Kim, S. H. (2022). Random forests for global and regional crop yield predictions. PLoS ONE, 11(6), e0156571.

[7]. Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. Frontiers in Plant Science, 10, 621.

[8]. Nema, M. K., Khare, D., &Chandniha, S. K. (2020). Application of artificial intelligence to estimate the paddy production in Chhattisgarh. Cogent Food & Agriculture, 3(1), 1416841.

[9]. Oikonomidis, A., Catal, C., & Kassahun, A. (2022). Deep learning for crop yield prediction: A systematic literature review. New Zealand Journal of Crop and Horticultural Science, 51(1), 1–26.

[10]. Padmavathi, K., & Prabavathi, M. (2021). Machine learning models for predicting chili crop yield. International Journal of Intelligent Systems and Applications, 13(1), 43–52.

[11]. Sehgal, V. K., Bhatt, D., & Kumar, S. (2021). Comparative analysis of machine learning methods for crop yield prediction in semi-arid regions of India. Journal of the Indian Society of Remote Sensing, 49(3), 617–629.

[12]. Specht, D. F. (1991). A general regression neural network. IEEE Transactions on Neural Networks, 2(6), 568–576.

[13]. Verma, A., Sharma, R., & Patidar, N. (2022). Application of Random Forest for turmeric yield prediction: A study from Sangli region. Indian Journal of Agricultural Sciences, 92(4), 487–492.

[14]. TNAU (Tamil Nadu Agricultural University). (2023). Package of Practices for Spice Crops: Turmeric. Coimbatore: TNAU Press.

[15]. Directorate of Economics and Statistics, Tamil Nadu. (2024). Season and Crop Report 2023–2024. Government of Tamil Nadu, Chennai.