Journal of Agricultural Science and Technology (JAST), 29(5)
In Press, Pre-Proof Version

Forecasting Sugarcane Yield at the District Level in India using a Hybrid Hidden Markov-Random Forest Model

D. Gowri Shankar[1]*, and R. Jayaparvathy[1]

Abstract

Accurate district-level forecasting of sugarcane yield is critical for effective planning and management of this important crop. This study develops a novel hybrid Hidden Markov Model–Random Forest (HMM–RF) framework for forecasting sugarcane yield across India. The framework integrates hidden agro-climatic production regimes identified using a Gaussian Hidden Markov Model with the nonlinear predictive capability of Random Forest regression. The model was developed using 16,172 district-year observations covering 1966–2017 across 311 districts and 20 states. Historical productivity, climatic, agronomic, irrigation, fertilizer, infrastructure, and economic variables were used as predictors, together with temporal features derived exclusively from preceding observations. Model evaluation employed leakage-free chronological validation, with 1966–2010 used for training and 2011–2017 reserved for testing. **The proposed HMM–RF achieved the best testing performance, with R^2 = 0.9754, RMSE = 5.56 tonnes per hectare MAE = 1.55 tonnes per hectare and MAPE = 4.11%. Compared with the benchmark Random Forest model, HMM–RF reduced RMSE by 22.1% and MAE by 41.6%.** Statistical validation further confirmed significant performance differences among the evaluated models. SHAP analysis showed that historical productivity variables were the dominant predictors, while HMM-derived production regimes, climatic variables, and management-related factors provided complementary predictive information.

Keywords: District-level forecasting, Hidden Markov Model, India, Machine learning, SHAP, Random Forest, Sugarcane yield prediction, Temporal regime identification.

1. Introduction

Sugarcane (*Saccharum officinarum* L.) is a strategic crop, mainly cultivated for sugar, ethanol, and fodder with an annual global production of 190 million tonnes from 26 million hectares (Food and Agriculture Organization of the United Nations, 2025). Apart from Brazil, India is the second-largest producer and consumer of sugarcane, where it constitutes a significant agro-industry sector (Solomon, 2014). Being a climate-sensitive crop, it has a considerable implication on livelihood security, food and energy security. Hence, enhancing

[1] Department of Electrical and Electronics Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India.
*Corresponding author; e-mail: gshankar.rdg@gmail.com

knowledge about yield forecasting mechanisms for informed agricultural policy and production planning assumes paramount importance.

Forecasting crop yield has been a matter of great interest to researchers since conventional statistical models like multiple linear regression, and ARIMA suffer from performance deterioration due to nonlinearity of the process and stationarity requirement. Moreover, the interaction between various socioeconomic, climatic, and agricultural variables requires a more intelligent approach to comprehend yield-generation mechanisms (Wilson, 2016; Hyndman & Athanasopoulos, 2021). The emergence of Machine Learning (ML) algorithms has enabled researchers to develop a sound understanding of the crop-yield generation process by capturing the intricate relationship between predictor and response variables. Various studies have demonstrated the efficacy of ML algorithms over conventional statistical models in agricultural yield forecasting (Jeong et al., 2016; Chlingaryan et al., 2018).

Among the diverse ML algorithms, Random Forest (RF) has been widely used for its robustness against overfitting and high dimensional data analysis capabilities along with variable importance ranking. Still, most of these ML algorithms adopt an independence assumption among observations, ignoring temporal characteristics in long time series studies.

Deep Learning methods can help forecast agricultural yields pretty efficiently, because they capture nonlinear patterns through artificial neural networks. Recent studies show that Autoencoders, Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and Transformers really work well for agricultural yield forecasting (Khaki & Wang, 2019; Sun et al., 2019). In particular, LSTMs do well at capturing longer term temporal connections, and that makes them especially useful.However, hybridization of different algorithms, for instance, LSTM-ARIMA and CNN-LSTM, has been proposed to capture both short and long-term temporal associations. Still, most of the studies overlook the district-level yield forecasting and agro-climatic transition analysis.

Furthermore, the LSTM-based studies usually require a large amount of data for training the model and demand high-end graphic processing units (GPUs) for acceleration, which can be cost-prohibitive for agricultural researchers. A recent study reveals that machine learning algorithms help forecast sugarcane yields with good accuracy. However, district-level yield forecasting and agro-climatic transition analysis studies are still unexplored, which limits the application of explainable AI algorithms and statistical tests for validating the performance of

the forecasting model. These research gaps motivate the current study to propose a novel forecasting framework with several interesting features.

This work presents a novel hybrid methodology consisting of a Hidden Markov Model and Random Forest for forecasting sugarcane yield at the district level. The proposed Hidden Markov Model–Random Forest (HMM–RF) framework uses Random Forest algorithm for variable selection and employs a Hidden Markov Model for extracting state information. The state-specific posterior probabilities along with the chosen variables were used to develop a Random Forest regression model, which captures the nonlinear association between predictors and the district yield. The model's variable importance and statistical validity were checked using SHapley Additive exPlanations (SHAP) and statistical tests, respectively. The practical applicability of the proposed framework was demonstrated by forecasting sugarcane yield for Tamil Nadu districts for the period 2011–2017. Table 1 shows the previous work on crop and sugarcane yield forecasting.

This work makes the following contributions:

i. Development of a novel hybrid Hidden Markov Model–Random Forest (HMM–RF) framework for district-level sugarcane yield forecasting

ii. Identification of hidden agro-climatic regimes using Hidden Markov Model

iii. Comparative analysis of the performance of the proposed framework with statistical, machine learning, deep learning, and time series models using leakage-free chronological cross-validation

iv. Interpretation of the variable importance using SHAP values

v. Evaluation of yield prediction performance using statistical significance tests and spatial yield prediction analysis.

Table 1. Representative literature on crop and sugarcane yield forecasting.

Reference	Study Area	Crop	Forecasting Model	Validation	Major Limitation
Jeong et al. (2016)	USA	Multiple crops	Random Forest	Train–Test	No temporal dependency modelling; no hidden-state representation
Khaki & Wang (2019)	USA	Maize	Deep Neural Network	Train–Test	Limited model interpretability
Sun et al. (2019)	China	Crop forecasting	CNN–LSTM	Train–Test	High computational complexity; no hidden-state modelling
Kouame et al. (2025)	Ghana	Maize	Random Forest	Independent validation	No temporal dependency or hidden production-state modelling
Khosravani et al. (2025)	Iran	Winter wheat	XGBoost	Independent validation	Single global prediction model; no temporal state identification
Jabed & Murad (2024)	Review	Multiple crops	ML/DL Review	Review	Identified limited interpretability and temporal modelling in existing studies
Shawon et al. (2025)	Review	Multiple crops	ML Review	Review	Highlighted lack of probabilistic temporal modelling
Proposed Work	**India (20 states; 331 districts)**	**Sugarcane**	**HMM–RF**	**Hidden states + SHAP + statistical validation**	**Proposed Work**

2. Materials and Methods

2.1 Study Area and Dataset

The study used district-level sugarcane data from major producing regions across 20 states of India, obtained from the ICRISAT District Level Database (DLD) (ICRISAT, 2024). The database contains annual observations for 1966–2017, covering sugarcane yield and production, cultivated area, fertilizer use, irrigation, climate, land use, infrastructure, labour, market accessibility, and socioeconomic variables. **After data screening, variable selection, temporal alignment, and application of predefined eligibility criteria, the final dataset comprised 16,172 district–year observations from 311 districts.** Sugarcane yield (tonnes per hectare) was the target variable, while the remaining selected variables served as candidate predictors. Original variables are summarized in Table 2.

Table 2. List of variables extracted from the ICRISAT District Level Database.

Variable category	No. of variables	Variables
Identification and time	5	District code, Year, State code, State name, District name
Sugarcane production	4	Sugarcane production, Sugarcane yield, Sugarcane gur harvest price, Sugarcane irrigated area
Land use and cropping	11	Total area, Forest area, Barren and uncultivable land, Land put to non-agricultural use, Cultivable waste land, Permanent pastures, Other fallow area, Current fallow area, Net cropped area, Gross cropped area, Cropping intensity
Fertilizer use	15	Nitrogen consumption, Nitrogen share in NPK, Nitrogen per ha of NCA, Nitrogen per ha of GCA, Phosphate consumption, Phosphate share in NPK, Phosphate per ha of NCA, Phosphate per ha of GCA, Potash consumption, Potash share in NPK, Potash per ha of NCA, Potash per ha of GCA, Total consumption, Total per ha of NCA, Total per ha of GCA
Labour	4	District male field labour, District female field labour, State male average field labour, State female average field labour
Market and infrastructure	4	Principal markets, Sub-markets, Total markets, Road length
Climate	13	January–December rainfall, Annual rainfall
Total	**56**	—

2.2 Exploratory Data Analysis

Exploratory data analysis (EDA) was performed to understand the distribution, variability, spatial heterogeneity, and temporal behaviour of the sugarcane yield and production. A descriptive and graphical analysis of the data at the national and state level is presented in Figure 1. Uttar Pradesh had the highest production of sugarcane, mainly because of the largest area under cultivation, while Tamil Nadu had the highest yield. The production and yield were considered two separate but highly connected variables. Since production could be a target variable, care was taken to not include any production-related features that might cause target leakage. Features that have a contemporaneous target relationship were either removed or transformed using temporal transformations to ensure that they followed the same temporal hierarchy as the target. The analysis of temporal persistence suggested building features that followed the lag structure of the target. Features such as lagged values and rolling statistics at a district level were made to capture historical production trends.

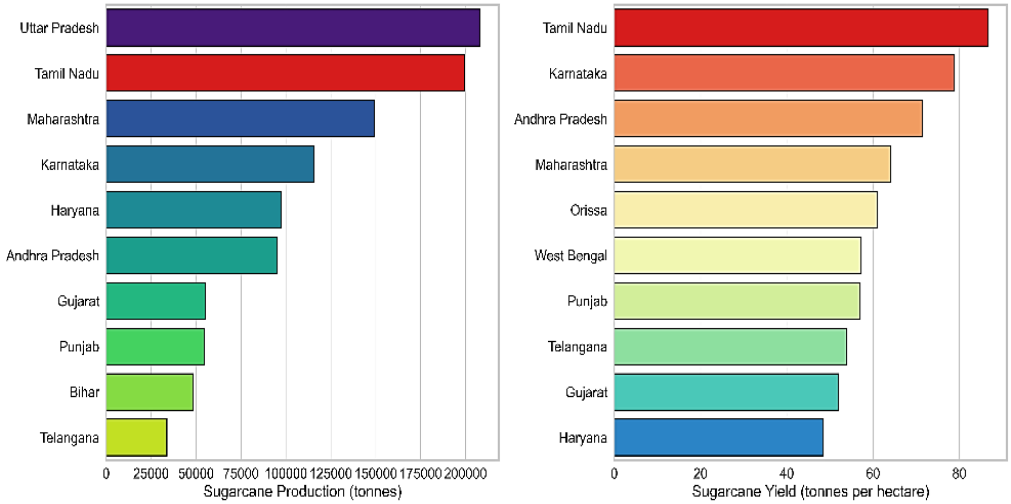

Figure 1. State-wise comparison of sugarcane production and yield.

2.3 Data Preprocessing and Feature Engineering

A systematic preprocessing procedure included data screening, missing-value treatment, temporal feature construction, multicollinearity assessment, and predictor selection. Invalid and infinite numerical values were treated as missing, while variables with excessive missingness were excluded. Missing values in retained variables were imputed using training-derived parameters only, which were subsequently applied to the testing data. After chronological sorting within each district, historical yield features were generated as (YieldLag_k(t)=Yield(t-k)), producing 1-, 2-, 3-, and 5-year lags without using current or future yield. Rolling statistics, production change, yield trend, area growth, and fertilizer-use efficiency were similarly calculated using historically available observations only. Multicollinearity was assessed using correlation analysis and VIF, with (VIF>10) indicating high redundancy. Identifiers and target variables were excluded from VIF analysis. Final predictors were selected using Random Forest importance based exclusively on training data.

2.4 Chronological Dataset Partitioning

A chronological hold-out strategy was adopted to evaluate forecasting performance under an independent future-period scenario. The dataset was divided into a training period (1966–2010) and testing period (2011–2017). All model-development procedures, including preprocessing, feature selection, HMM estimation, Random Forest construction, and hyperparameter optimization, were restricted to the training period. The testing data were withheld until final evaluation.

Temporal features were formed independently of the prediction year and subsequent years. In other words, the 2011 prediction was made using historical information as of 2010, and all

subsequent test years' predictors were formed using observations preceding a given year. In this way, the chronological nature of observations was respected, and the test results reflected the accuracy of models using observations that were not employed during the estimation or calibration period.

2.5 Hidden Markov Model

A Gaussian Hidden Markov Model was used to identify hidden production regimes associated with temporal variation in district level sugarcane yield. The HMM assumes that observations are conditionally dependent on an underlying Markov chain of unobserved states (Rabiner, 1989). Let s_t be the state of the Markov process at a given time (t) with transition probabilities defined by equation (1). A state is associated with a particular multivariate agricultural and agro-climatic observation vector.

$$a_{ij} = P(s_{t+1} = j \mid s_t = i), \quad (1)$$

subject to $$\sum_{j=1}^{N} a_{ij} = 1.$$

The HMM was fit globally using the training observations from all the 20 states and 311 districts, not separately for each district, to capture general production regimes characteristic of the study region. The HMM was estimated for the period 1966–2010; fitting it to the independent test data (2011–2017) was not performed.

The HMM observation vector consisted of the predictors chosen during the previous steps, including preprocessing, temporal feature engineering, multicollinearity assessment, and predictor selection. Specifically, the observation vector included historical agroclimatic, socioeconomic, and agricultural production attributes. The predictors related to land use and irrigation, fertilizer use, rainfall, and other climate variables, as well as agricultural infrastructure, were also included. On the other hand, the contemporaneous target variable, sugar cane yield, was excluded from the observation vector to avoid target leakage during state identification.

Because the observation variables were continuous, each state was represented using a multivariate Gaussian emission distribution is given by equation (2):

$$b_j(o_t) = \frac{1}{\sqrt{(2\pi)^d \mid \Sigma_j \mid}} \exp\left[-\frac{1}{2}(o_t - \mu_j)^T \Sigma_j^{-1} (o_t - \mu_j)\right] \quad (2)$$

where μ_j and Σ_j are the state-specific mean vector and covariance matrix. A diagonal covariance structure was adopted for the states of the system, which assumes covariance

between the states to be zero; however, state-specific variances are estimated. This reduced the number of variance-covariance parameters in the model and improved the numerical stability of the multivariate feature space.

The HMM was implemented using the GaussianHMM class from the hmmlearn Python library. The maximum-likelihood estimation was performed with a maximum of 1000 iterations, a tolerance of (10-4), and a random_state of 42. For all candidate models, the same settings were used.

The number of hidden states was determined empirically based on a set of models with 2–14 states trained with the training data. The models were compared using a Bayesian information criterion (BIC) (and other criteria, including AIC), posterior state probabilities, state occupancy, and transition stability as per equations (3) and (4):

$$AIC = 2k - 2\ln(L) \quad (3)$$

$$BIC = k\ln(n) - 2\ln(L) \quad (4)$$

where (k) is the number of estimated parameters, (L) the maximized likelihood, and (n) the number of training observations. The parameter count reflected the diagonal covariance specification. The final five-state model was selected based on the combined statistical and interpretability criteria.

The most probable state was estimated using the Viterbi algorithm. For each district – year, the hidden regime was identified according to the most probable state. For each of the 5 states, specific statistics for yield, production, rainfall, and other variables were estimated. Then, based on these variables' values, the state was assigned a label: Very Low, Low, Moderate, Moderate–High, and Very High Yield regimes. It is important to note that the labels were assigned during the estimation procedure not a priori.

The resulting HMM state was included as an additional categorical predictor in the Random Forest (RF) model. Consequently, the RF predictions used both explicit agricultural and historical predictors plus information about the broader hidden production regime. For the testing period, the states were generated using the HMM estimated on training data (without using future testing-period yield observations), thus maintaining the leakage-free chronological procedure.

2.6 Random Forest Regression

Random Forest regression model was selected as the primary nonlinear prediction approach because it can capture complicated nonlinear patterns and interactions between climatic factors,

agricultural, socioeconomic conditions, infrastructure, and historical variables, without us without the need of having any predefined functional form.

The RF model was constructed using bootstrap samples and random subsets of predictors at each tree split, with the final prediction obtained by averaging the predictions of the individual regression trees. The principal hyperparameters, including the number of trees, maximum tree depth, minimum samples required for a split, and number of predictors considered at each split, were optimized using grid-search cross-validation on the training dataset only. The testing period was not used during hyperparameter optimization. The selected configuration was subsequently refitted using the complete training dataset and evaluated on the 2011–2017 testing period.

2.7 Proposed HMM–RF Framework

The proposed HMM–RF framework integrates hidden-state modelling with nonlinear Random Forest regression. First, the district-level data were chronologically organized and subjected to preprocessing, temporal feature engineering, multicollinearity assessment, and predictor selection using the training period.

The selected agricultural and agro-climatic observations were then supplied to the Gaussian HMM to identify hidden production regimes. The resulting HMM-derived state information was combined with the selected observed predictors and supplied to the Random Forest regression model.

The hybrid framework therefore combines two complementary sources of information: (i) explicit historical, climatic, agronomic, production, infrastructure, and socioeconomic predictors and (ii) a hidden production-regime indicator derived from the multivariate temporal structure of the data. The inclusion of the HMM state was intended to provide contextual information about the prevailing production regime beyond individual lagged variables.

All preprocessing, feature selection, HMM estimation, state determination, and RF hyperparameter optimization were conducted using the chronological training period. The 2011–2017 observations were retained exclusively for independent testing. This design enabled the predictive contribution of the hidden production regimes to be evaluated under a realistic temporal forecasting setting.

The overall workflow of the proposed HMM–RF framework is presented in Figure 3.

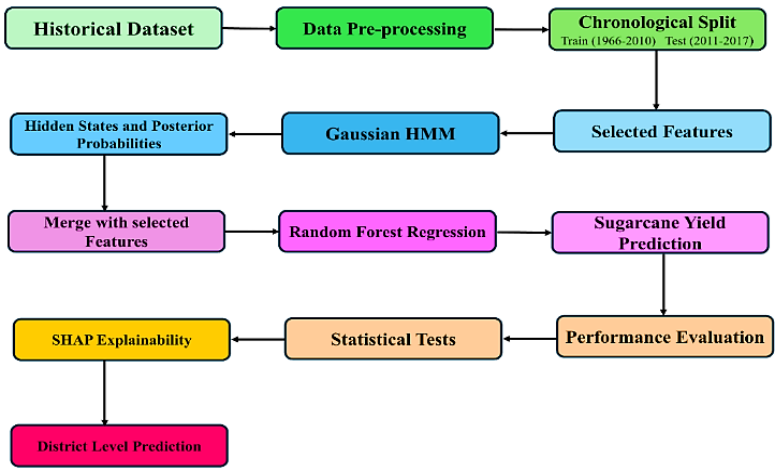

Figure 3. Overall workflow of the proposed Hidden Markov Model–Random Forest (HMM–RF) framework.

2.8 Benchmark Models

The proposed HMM–Random Forest (HMM–RF) framework was benchmarked against statistical, machine-learning, and deep-learning models. Feature-based models included Decision Tree, Random Forest, Gradient Boosting, AdaBoost, XGBoost, and CatBoost, using the same predictor matrix and chronological data partition. LSTM was evaluated as a deep-learning benchmark after training-only standardization and sequence construction. ARIMA and Exponential Smoothing (ETS) were fitted separately for eligible districts using historical sugarcane yield alone. All models were trained using 1966–2010 observations and tested on the 2011–2017 period, which was excluded from model fitting, feature selection, and hyperparameter configuration.

2.9 Model Configuration and Hyperparameter Selection

All the models that were feature based got trained on 1966–2010 data, and for the tuning part we only used the training set, random_state=42 where it made sense. The suggested HMM–RF mixed a five-state Gaussian HMM with an optimized Random Forest, whereas the benchmark RF, Decision Tree, Gradient Boosting, AdaBoost, XGBoost, and CatBoost models each relied on their own tuned setups (Table 3). **As for the LSTM setup, it was a 32-unit LSTM layer, then dropout at 0.20, followed by a 16-neuron ReLU layer and a linear output. It was trained with Adam, MSE loss, batch size 64, plus early stopping.** District level ARIMA with (2,1,2) and additive trend ETS were estimated from yield histories that had at least 10 training observations. After that, every model projected 2011–2017, and no method had any exposure to test data. The final set of hyperparameters is shown in Table 3 below.

Table 3. Hyperparameter settings used for the proposed HMM–RF framework and benchmark models.

Model	Hyperparameters
HMM	5 states; Gaussian emissions; diagonal covariance; max iterations = 1000; tolerance = 1×10^{-4}; random state = 42
Decision Tree	max depth = 6; min samples split = 50; min samples leaf = 30; max features = 0.60; random state = 42
Random Forest	120 trees; max depth = 9; min samples split = 25; min samples leaf = 12; max features = 0.65; bootstrap = True; random state = 42
Gradient Boosting	75 estimators; learning rate = 0.04; max depth = 3; min samples split = 30; min samples leaf = 15; subsample = 0.70; max features = 0.70
AdaBoost	60 estimators; learning rate = 0.30; loss = linear; base tree max depth = 4; base tree min samples leaf = 15
XGBoost	120 estimators; learning rate = 0.05; max depth = 4; min child weight = 10; subsample = 0.70; colsample bytree = 0.70; gamma = 0.30; reg_alpha = 0.20; reg_lambda = 2.0
CatBoost	80 iterations; learning rate = 0.04; depth = 4; L2 leaf regularization = 15; random strength = 5; bagging temperature = 3; subsample = 0.70
LSTM	sequence length = 1; 1 LSTM layer; 32 units; dropout = 0.20; Dense = 16 ReLU; Adam; MSE; batch size = 64; maximum epochs = 80; validation split = 20%; early stopping patience = 10
ARIMA	district-wise ARIMA(2,1,2); minimum 10 training observations
ETS	district-wise additive trend; no seasonality; estimated initialization

2.10 Model Evaluation

Chronological validation is applied in order to maintain the time series characteristics of the database. The period from 1966 to 2010 was used for training the model, while the years 2011-2017 were used for testing purposes. The performance of the model was evaluated using the four standard criteria for regression tasks presented in equations (5)- (8).

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \quad (6)$$

$$MAE = \frac{1}{n}\sum|y_i - \hat{y}_i| \quad (7)$$

$$MAPE = \frac{100}{n}\sum\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (8)$$

2.11 Statistical Validation

Statistical validation was performed using the common testing observations, with model prediction errors treated as paired observations. Because the error distributions did not satisfy the normality assumption, non-parametric procedures were used. Absolute prediction errors were calculated for each model and testing observation.

The Friedman test was used to assess overall differences among the competing models. When significant ((P<0.05)), pairwise Wilcoxon signed-rank tests were conducted using HMM–RF as the reference model. Effect sizes were calculated for pairwise comparisons to assess the practical magnitude of differences rather than relying solely on (P)-values.

Where applicable, 95% confidence intervals for paired performance differences were estimated using bootstrap resampling of the common testing observations, preserving the paired structure across models. Intervals excluding zero were interpreted as evidence of a meaningful difference. Statistical significance was set at (P<0.05).

3. Results and Discussion

3.1 Dataset Characteristics and Exploratory Analysis

The processed dataset comprised 16,172 district–year observations from 311 districts across 20 major sugarcane-producing states of India, covering the period 1966–2017. Following data screening and preprocessing, the dataset contained 61 variables representing climatic, agronomic, land-use, socioeconomic, infrastructure, and production-related characteristics. A summary of the processed dataset is provided in Table 4.

Table 4. Summary of the processed dataset used in this study.

Metric	Value
Rows	16,172
Columns	61
States	20
Districts	311
Start Year	1966
End Year	2017

According to the exploratory analysis, there was a significant variation in the cane yield (in tonnes per hectare) is shown in Figure 4 and Figure 5 below. In most districts, the yield was skewed towards the higher side with a value between 20-80 tonnes per hectare. However, a few had a yield exceeding 140 tonnes per hectare. It is evident that the average yield increased from roughly 36 tonnes per hectare in the late 1960s to over 50 tonnes per hectare after 2010.

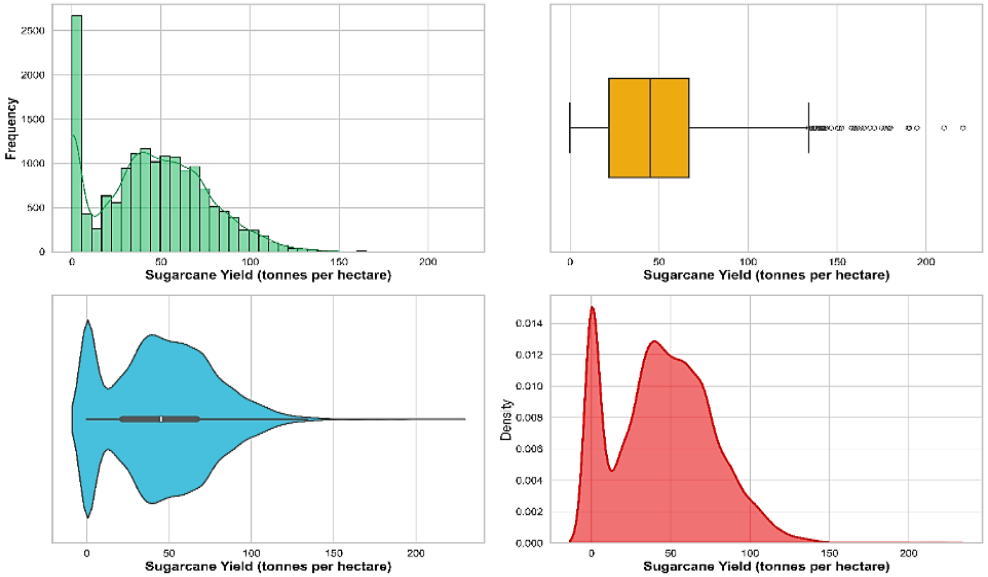

Figure 4. Distribution of district-level sugarcane yield across India (1966–2017).

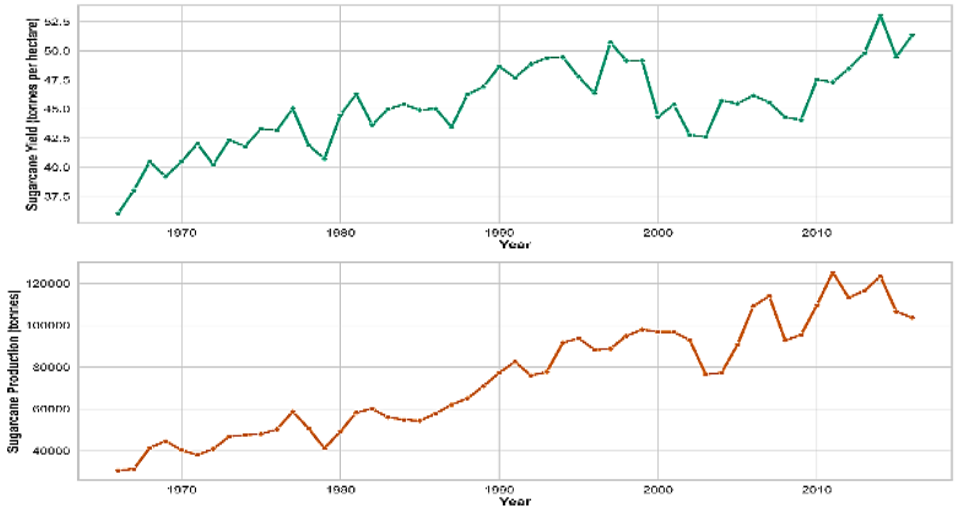

Figure 5. Temporal trends in average sugarcane yield and production.

3.2 Feature Engineering and Predictor Selection

Feature engineering was conducted to capture temporal dependence and production dynamics using only information available before each prediction year. Engineered variables included lagged yield, rolling yield statistics, production change, yield trend, area growth, and fertilizer-use efficiency is shown in Figure 6. Historical yield predictors showed the strongest association with subsequent yield, with mutual-information scores of 0.8862 (Yield Lag1), 0.7952 (Lag2), 0.7543 (Lag3), and 0.6986 (Lag5), indicating strong temporal persistence. Predictor redundancy was assessed using correlation and VIF analysis, followed by Random Forest importance, recursive feature elimination, and mutual information for predictor

selection. The final feature set is presented in Table 5, dominated by historical yield variables, while production (0.4017), road length (0.3507), production change (0.1209), rolling production variability, potash application, and climatic variables provided complementary information. All procedures followed the chronological modelling framework to prevent temporal information leakage.

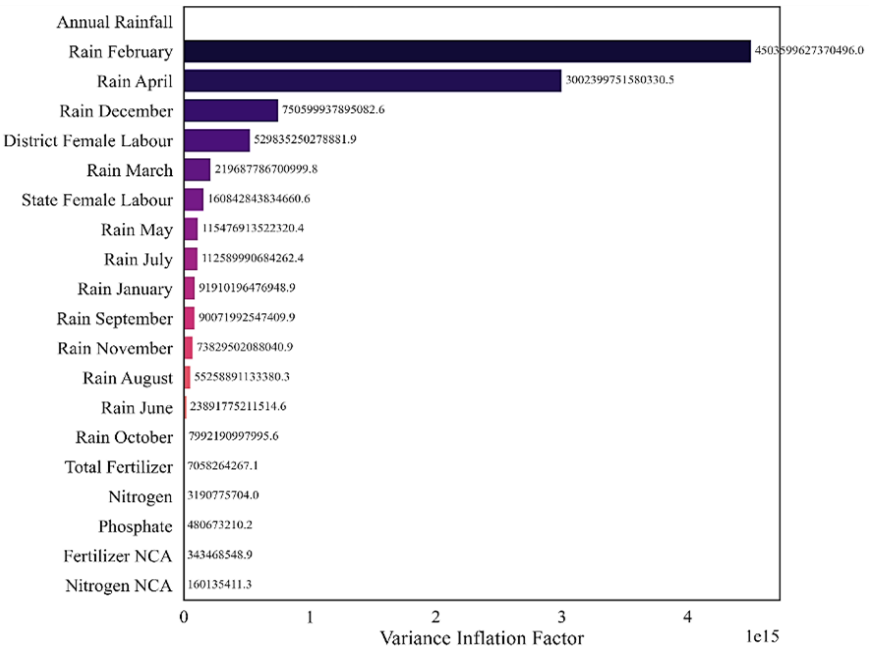

Figure 6. Variance Inflation Factor (VIF) analysis of highly correlated predictors.

Table 5. Final Selected Predictor Variables Used in the Proposed HMM–RF Framework.

Variable	Importance	Mutual Information	Variable	Importance	Mutual Information
Yield Lag1	0.054901	0.886168	Road Length	0.008599	0.350719
Yield Lag3	0.025607	0.754316	Production Change	0.0084	0.120908
Yield Lag5	0.015972	0.698575	Yield Rolling Std3	0.007907	0.157267
Yield Lag2	0.014652	0.795246	Production Rolling Std3	0.007462	0.279987
Production	0.010397	0.401677	Potash	0.006469	0.209742

3.3 Hidden Agro-Climatic Yield Regime Identification

A Gaussian Hidden Markov Model (HMM) was used to identify hidden production regimes associated with temporal variation in district-level sugarcane productivity. Candidate models containing 5–14 hidden states were evaluated using AIC, BIC, posterior state probabilities, state occupancy, and transition stability (Figure 7). Although additional states increased model flexibility, they also produced more fragmented and sparsely occupied regimes with limited

interpretive benefit. Based on the combined statistical and structural criteria, the five-state HMM was selected.

The selected model achieved a mean posterior state-assignment probability of 0.9825, indicating high confidence in the classification of district–year observations. The identified states showed sufficient occupancy and distinct empirical characteristics, supporting their interpretation as meaningful production regimes rather than statistical fragmentation.

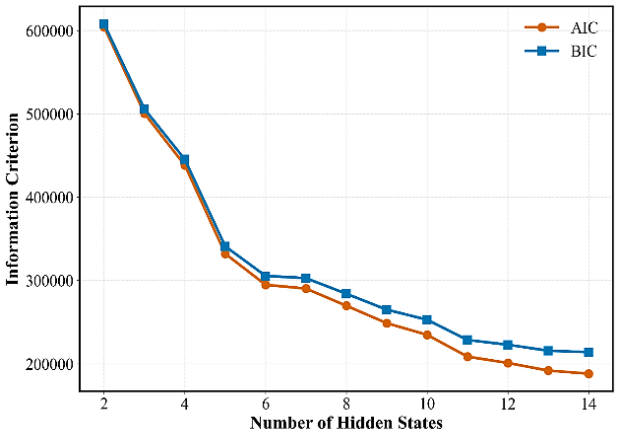

Figure 7. Hidden-state selection based on AIC and BIC.

The transition probability matrix of the final HMM is presented in Figure. 8. It can be seen that self-transitions in most of the regimes are higher than 0.83, which means a fairly high persistence of production regimes from one year to another. For the majority of districts, their production regime rarely changed over time, staying approximately the same at the next year. It can be noted that the Moderate Production regime has more transitions to other regimes than the rest, which can be explained by its intermediate position between other regimes. Therefore, the Moderate Production can be regarded as the most variable production regime.

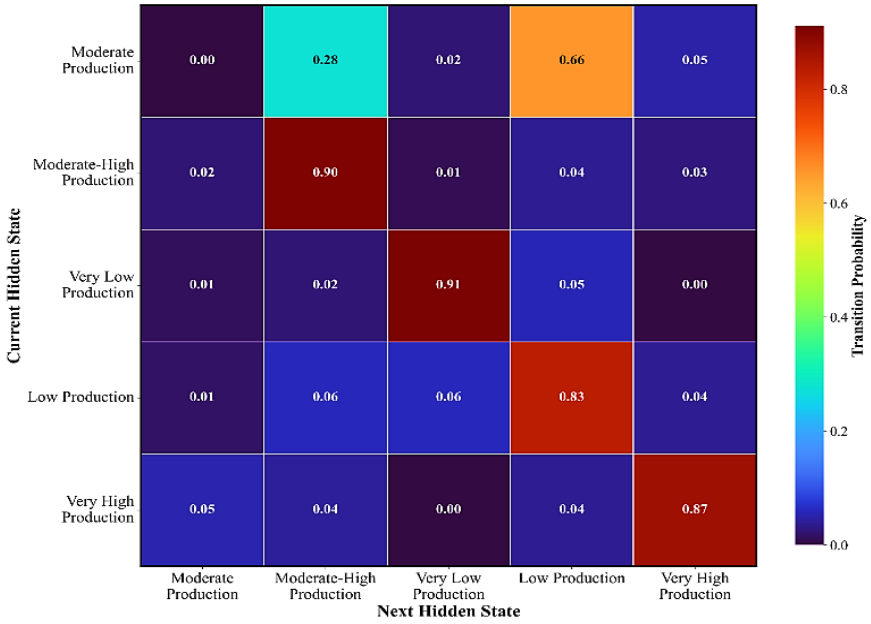

Figure 8. Transition probability matrix of the final five-state Hidden Markov Model.

The five states were characterized using empirical distributions of yield, production, rainfall, and other relevant agricultural variables are presented in Table 6 . Based on the characteristics, the regimes were interpreted as representing Very Low, Low, Moderate, Moderate–High, and Very High Production. The interpretations were made based on results from the HMM estimation and were not preset in the model fitting process.

The resulting hidden states therefore represent a compact summary of multivariate agricultural conditions and their persistence over time. In each case, the inferred state was added as an additional predictor to the Random Forest, enabling it to exploit the information contained in the observed agricultural, weather, production, and historical predictors, along with the hidden regime structure.

Importantly, HMM parameters were estimated exclusively from the 1966–2010 training period. Testing observations from 2011–2017 were not used for HMM estimation; their hidden-state information was generated using the fitted training model, thereby maintaining chronological separation and preventing testing-period target leakage.

Table 6. Quantitative characteristics and interpretation of the five latent production regimes identified by the HMM.

Hidden state	Regime interpretation	Yield (t ha⁻¹)	Production (t)	Crop intensity (%)	Potash	Road length (km)	July rainfall (mm)	n
State 1	Very Low Production	35.25	29,970.47	119.81	221.24	1,227.26	271.29	318
State 2	Low Production	42.24	63,738.27	143.25	2,983.19	2,611.47	315.32	4,758
State 3	Moderate Production	23.46	2,074.44	122.21	202.16	1,894.72	295.97	3,043
State 4	Moderate–High Production	49.08	15,829.76	125.39	2,062.39	5,008.26	296.72	3,725
State 5	Very High Production	73.38	275,557.40	131.65	12,895.70	7,584.33	218.08	2,151

3.4 Model Performance Comparison

The predictive performance of the proposed HMM–Random Forest (HMM–RF) framework was evaluated using the chronological partition described in Section 2, with 1966–2010 used for model development and 2011–2017 reserved for independent testing. Performance was assessed using R^2, RMSE, MAE, and MAPE, where higher R^2 and lower errors indicate better performance.

As shown in Table 7, HMM–RF achieved the highest training performance (R^2 = 0.9929; RMSE = 2.5525 tonnes per hectare; MAE = 0.6646 tonnes per hectare), followed by XGBoost (R^2 = 0.9799), Random Forest (0.9791), and LSTM (0.9766). Other models showed comparatively lower training performance. However, training accuracy was not considered sufficient evidence of forecasting superiority; independent testing was used to assess model generalization.

Table 7. Training performance of HMM–RF and benchmark models.

Model	Train R^2	Train RMSE	Train MAE
Proposed HMM-RF	0.9929	2.5525	0.6646
XGBoost	0.9799	4.3037	2.2167
Benchmark Random Forest	0.9791	4.3883	1.7196
LSTM	0.9766	4.6425	1.7942
Gradient Boosting	0.9446	7.1432	4.2907
CatBoost	0.9161	8.7909	5.368
AdaBoost	0.9016	9.5198	7.2753
Decision Tree	0.8616	11.2867	7.0831
ARIMA	0.8162	13.0193	7.6851
ETS	0.8109	13.1948	7.6851

HMM–RF the best testing performance, with R^2 = 0.9754, RMSE = 5.5625 tonnes per hectare, MAE = 1.5498 tonnes per hectare, and MAPE = 4.1128%. Compared to Random Forest, HMM–RF RMSE falls from 7.1385 to 5.5625 tonnes per hectare (22.1%) and also lowered MAE from 2.6517 to 1.5498 tonnes per hectare (41.6%). So, it shows extra predictive

power from HMM- based production regimes. XGBoost (R^2 = 0.9673; RMSE = 6.4128 tonnes per hectare), after which came LSTM and Random Forest. Gradient Boosting, AdaBoost, CatBoost, and Decision Tree ended up with larger errors though, while ARIMA and ETS were way worse (RMSE = 26.1282 and 26.7012 tonnes per hectare, respectively). Overall, it suggests that those multivariate, feature driven methods worked better than district-wise univariate forecasting, in a more robust kind of way.

The exceptionally large Decision Tree MAPE was not interpreted because percentage errors become unstable at very low observed yields; comparison therefore focused on R^2, RMSE, and MAE.

Table 8. Testing performance during 2011–2017.

Model	Test R^2	Test RMSE	Test MAE	Test MAPE	Improvement over Benchmark (%)
Proposed HMM-RF	0.9754	5.5625	1.5498	4.1128	0
XGBoost	0.9673	6.4128	2.9196	8.8054	0.84
LSTM	0.9626	6.8557	3.3094	7.3787	1.33
Random Forest	0.9594	7.1385	2.6517	6.7381	1.67
Gradient Boosting	0.9316	9.2699	5.4784	16.4327	4.7
AdaBoost	0.8988	11.2783	8.0844	29.1814	8.52
CatBoost	0.8744	12.5626	7.5292	19.7563	11.55
Decision Tree	0.8145	15.2681	8.7203	58862589	19.75
ARIMA	0.4567	26.1282	16.1428	27.5202	113.58
ETS	0.4326	26.7012	16.2147	29.3284	125.47

Observed-versus-predicted plots (Figure 9) showed HMM–RF predictions closely distributed around the one-to-one line. Temporal plots (Figure 10) demonstrated good tracking of major yield patterns, while the Taylor diagram (Figure 11) indicated strong correlation and comparatively low centred RMSE difference.

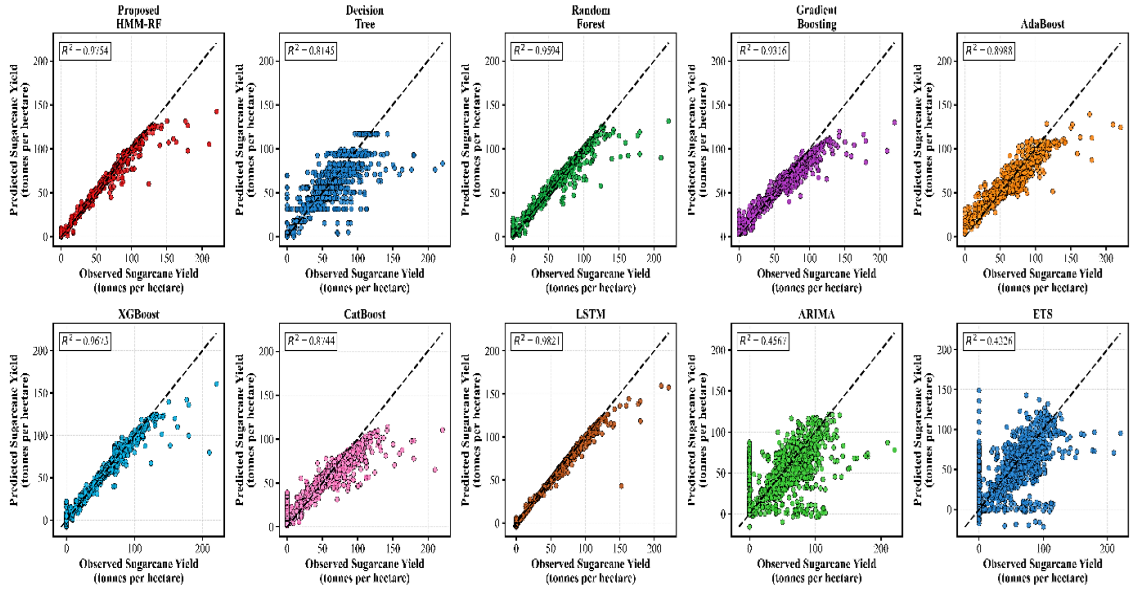

Figure 9. Observed versus predicted sugarcane yield.

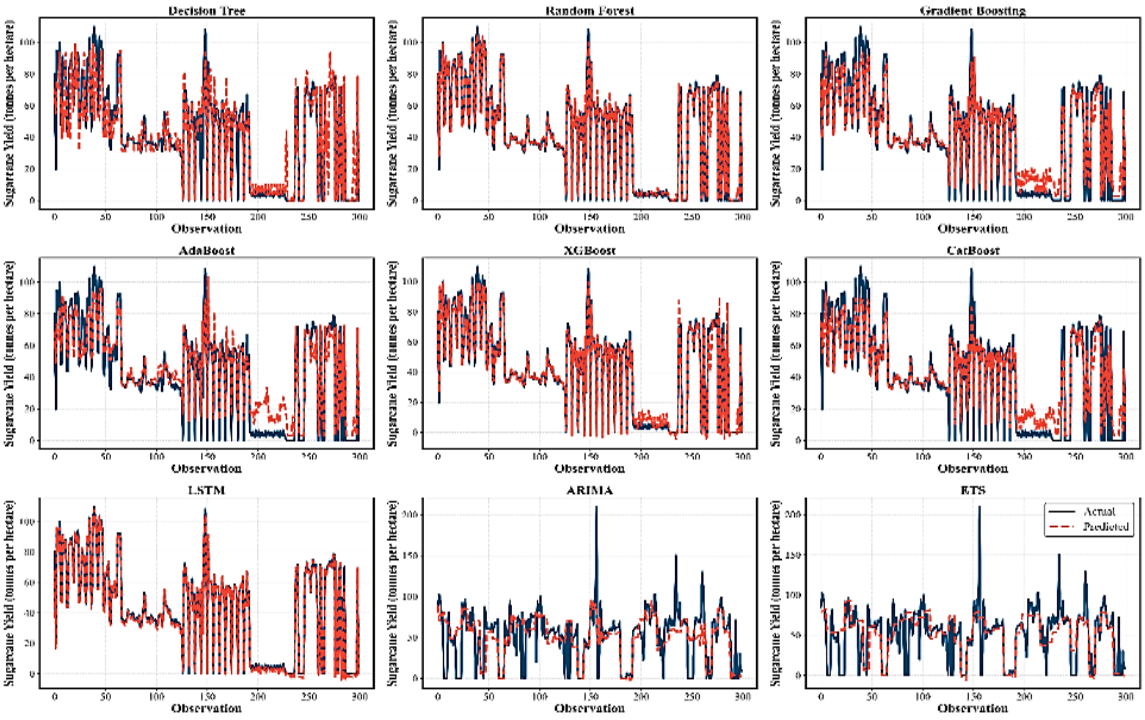

Figure 10 a.

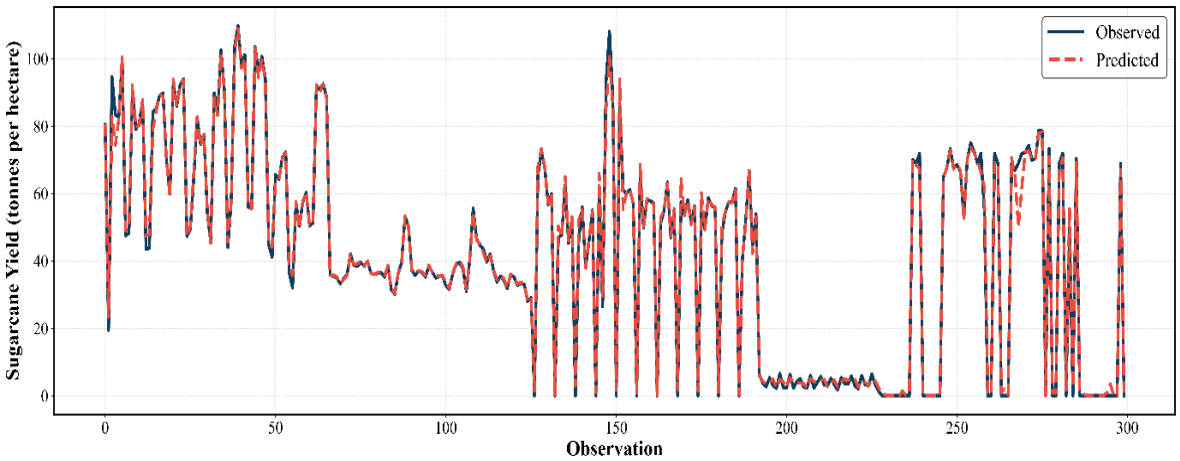

Figure 10 b.

Figure 10. Temporal comparison of observed and predicted yield.

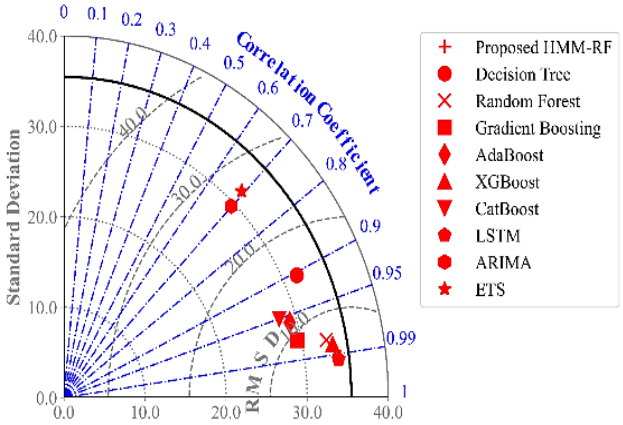

Figure 11. Taylor diagram of model performance.

3.5 Statistical Validation

Statistical validation was conducted using paired prediction errors from the independent testing dataset. As the error distributions were non-normal, non-parametric tests were applied. The Friedman test revealed significant differences among the evaluated models (χ^2=500.57, p=4.33×10^{-102}) rejecting the hypothesis of equivalent predictive performance. Post-hoc Wilcoxon signed-rank comparisons showed that HMM–RF differed significantly from Decision Tree, Random Forest, Gradient Boosting, XGBoost, CatBoost, LSTM, ARIMA, and ETS, whereas no significant difference was observed with AdaBoost (p=0.1735p=0.1735p=0.1735). Effect sizes and bootstrap-derived 95% confidence intervals for RMSE and MAE differences were also calculated to assess practical significance.

Statistical checks were carried out with paired prediction errors coming from the separate testing dataset. Since the error patterns looked non-normal, instead non-parametric tests were chosen. The Friedman test used to clear differences across the checked models (χ^2=500.57,p=4.33×10^{-102}) and so the idea of similar predictive performance got rejected. After that, post hoc Wilcoxon signed-rank comparisons indicated that HMM–RF came out significantly different versus Decision Tree, Random Forest, Gradient Boosting, XGBoost, CatBoost, LSTM, ARIMA, and ETS, but with AdaBoost there was no obvious gap (p=0.1735p=0.1735p=0.1735). To judge whether the effect was also meaningful in practice, effect sizes plus bootstrap based 95% confidence intervals were computed for the RMSE and MAE differences.

A supplementary Wilcoxon test showed no significant difference between observed and HMM–RF predicted yields (p=0.982p=0.982p=0.982), supporting the consistency of model predictions.

Table 9. Wilcoxon signed-rank test comparing observed and predicted sugarcane yields.

Model	Wilcoxon Statistic	p-value	Significant Difference ($\alpha = 0.05$)
Proposed HMM–RF	856,490	0.982	No
Decision Tree	794,647	0.176	No
Random Forest	867,900	0.896	No
Gradient Boosting	858,805	0.602	No
AdaBoost	869,171	0.939	No
XGBoost	863,441	0.747	No
CatBoost	849,932	0.366	No
LSTM	827,045	0.059	No
ARIMA	577,726	<0.001	**Yes**
ETS	612,372	<0.001	**Yes**

Table 10. Friedman Test and Pairwise Wilcoxon Comparisons.

Test	χ^2 Statistic	P-value	Conclusion
Friedman Test	500.57	4.33×10^{-102}	Significant difference among models

Table 11. Pairwise Wilcoxon Comparison (HMM–RF vs Benchmark Models).

Benchmark Model	P-value	Significant
Decision Tree	4.31×10^{-8}	Yes
Random Forest	6.97×10^{-5}	Yes
Gradient Boosting	5.51×10^{-5}	Yes
AdaBoost	0.1735	No
XGBoost	5.00×10^{-16}	Yes
CatBoost	1.05×10^{-4}	Yes
LSTM	7.05×10^{-131}	Yes
ARIMA	6.62×10^{-4}	Yes
ETS	0.0217	Yes

3.6 Model Explainability

SHAP analysis shown in Figure 12 is used to examine the historical yield variables as the most influential predictors within the HMM–RF framework, with Yield Lag1 coming up as the top contributor, then Yield Trend, and other lagged variables after that, it also reinforces the strong temporal persistence part. Production, Production Change, rolling statistics, potash application, road length, plus a few chosen climatic variables, they all gave extra complementary predictive signal. The HMM-derived hidden production regime made things even better, basically by summarizing multivariate agricultural conditions and their time persistence. In the end, the HMM–RF leaned mostly on past productivity, and it was supported by production related, agronomic, infrastructure, climatic, and this latent-regime information too. SHAP contributions point to predictive relevance, and they shouldn't be treated as causal effects, even if it feels tempting.

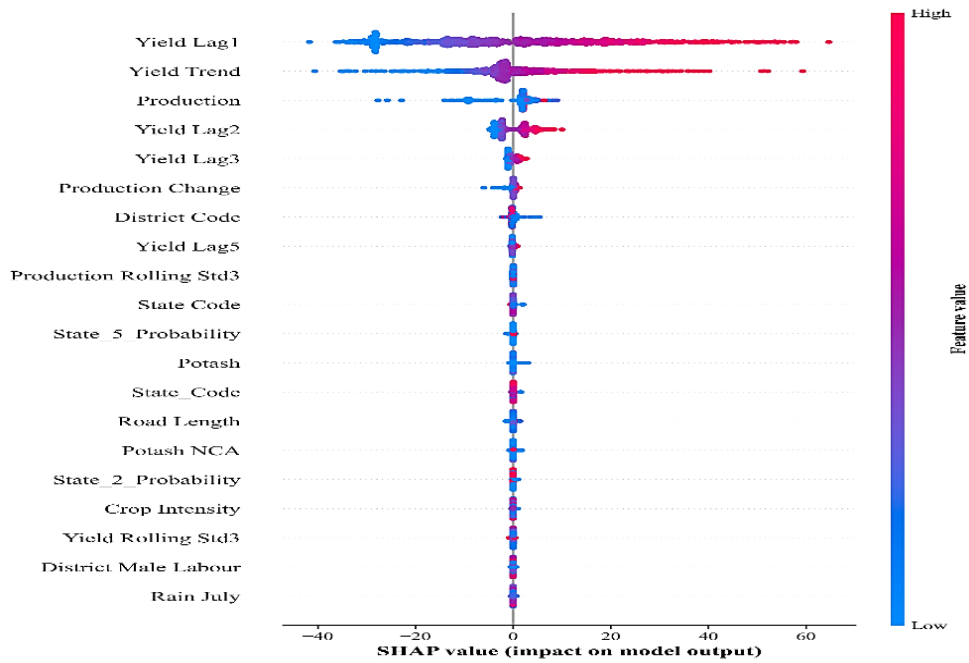

Figure 12. SHAP-based feature importance and predictor contributions for the proposed HMM–RF framework.

3.7 District-Level Validation

District-level validation was conducted to assess whether the proposed HMM–Random Forest (HMM–RF) framework maintained predictive performance across districts with different sugarcane production conditions. Tamil Nadu was selected because its districts represent diverse productivity and agro-environmental conditions. Validation used the independent 2011–2017 testing period, which was excluded from model fitting and predictor selection.

Overall, HMM–RF reproduced observed yield patterns well across the evaluated districts (Figure 13; Table 12). Close agreement was observed for Cuddalore, Tiruchirapalli, Kanchipuram, and Thanjavur, while Vellore and Ramanathapuram also showed good performance. The largest deviation occurred in Nilgiris, where observed and predicted mean yields were 31.33 and 24.75 tonnes per hectare, respectively, with an absolute error of 6.59 tonnes per hectare and RMSE of 12.84 tonnes per hectare.

Across the evaluated districts, the framework achieved an average R^2 of 0.9751, RMSE of 2.92 tonnes per hectare, MAE of 1.98 tonnes per hectare, and MAPE of 3.59%, with district-level R^2 ranging from 0.9167 to 0.9958 (Table 13). These results indicate generally strong

district-level predictive performance, although accuracy varied across local production environments.

This analysis represents spatial subgroup validation rather than external geographical validation, as the Tamil Nadu districts originated from the same national dataset used for model development. Therefore, the results demonstrate applicability under the study conditions but do not establish transferability to completely unseen regions. Further evaluation using independent datasets or additional states would provide stronger evidence of geographical generalization.

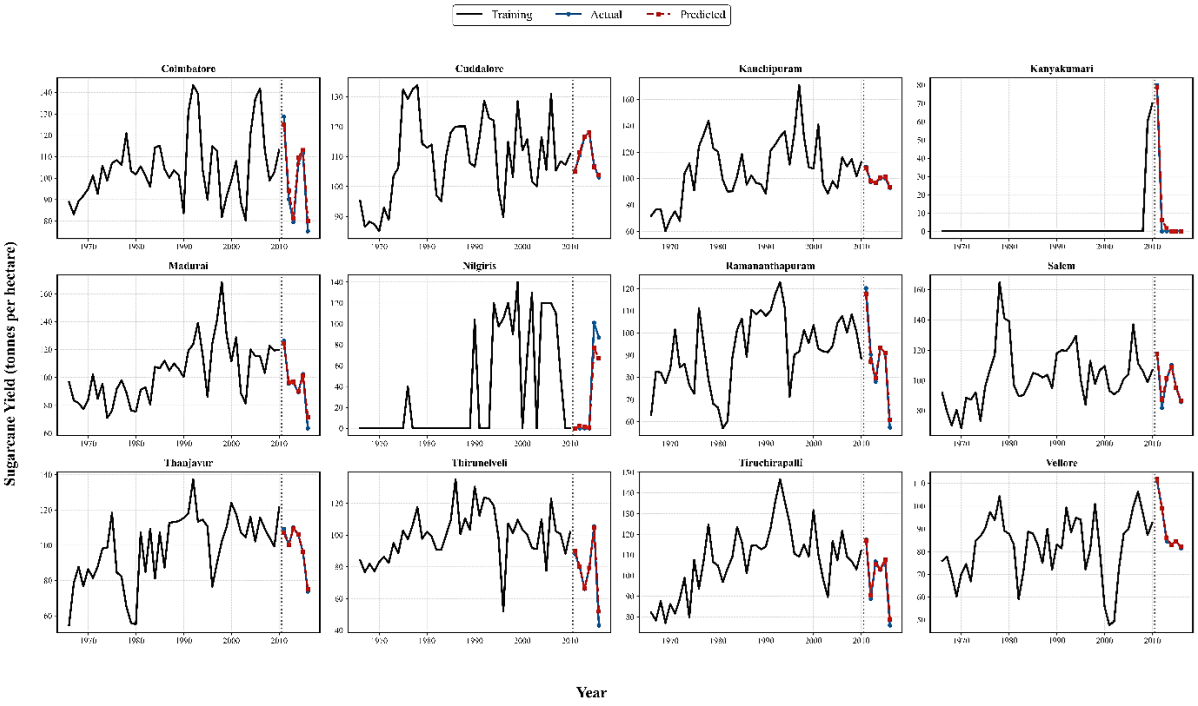

Figure 13. District-wise prediction performance of HMM–RF during 2011–2017.

Table 12. District-level prediction performance of HMM–RF during 2011–2017.

District	Actual Yield (tonnes per hectare)	Predicted Yield (tonnes per hectare)	RMSE	MAE	Absolute Error
Cuddalore	110.09	110.22	0.62	0.50	0.13
Tiruchirapalli	99.76	100.49	1.57	1.32	0.73
Kanchipuram	99.40	99.71	0.75	0.69	0.31
Thanjavur	99.22	98.98	1.05	0.73	0.23
Coimbatore	98.71	100.46	3.21	2.96	1.75
Salem	98.51	99.49	2.36	1.49	0.98
Madurai	95.60	96.87	3.46	2.18	1.27
Vellore	90.69	91.08	0.70	0.55	0.39
Ramanathapuram	88.42	88.25	2.26	1.83	0.17
Tirunelveli	77.00	78.71	3.65	1.95	1.70
Nilgiris	31.33	24.75	12.84	8.11	6.59

Table 13. Summary of district-level prediction performance of HMM–RF.

Metric	Value
Average R^2	**0.9751**
Average RMSE (tonnes per hectare)	**2.92**
Average MAE (tonnes per hectare)	**1.98**
Average MAPE (%)	**3.59**
Best-performing district	**Cuddalore**
Highest district R^2	**0.9958**
Lowest district R^2	**0.9167**

5. Conclusions

This work develops a hybrid Hidden Markov Model–Random Forest (HMM–RF) framework for predicting sugarcane yield in Tamil Nadu that uses the time series production data and incorporates the machine learning algorithm to improve prediction accuracy. The proposed framework has a high prediction accuracy ($R^2 = 0.9754$) and outperforms the conventional models. In addition, the framework helps understand the production regime and contributes to model interpretation using SHAP analysis. The main limitation of the study is the use of data at the district level, which does not account for the within-season variability. The study can be extended by including other variables, such as remote sensing data and socio-economic factors, to make it more effective in agricultural production monitoring and management.

References

1. Breiman, L. 2001. *Random forests*. Mach. Learn. 45(1):5–32. https://doi.org/10.1023/A:1010933404324
2. Chlingaryan, A., Sukkarieh, S. and Whelan, B. 2018. *Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review*. Comput. Electron. Agric. 151:61–69. https://doi.org/10.1016/j.compag.2018.05.012
3. Food and Agriculture Organization of the United Nations. 2025. *Sugar cane production (1961–2024).* Our World in Data. https://ourworldindata.org/grapher/sugar-cane-production
4. Friedman, J.H. 2001. *Greedy function approximation: A gradient boosting machine*. Ann. Stat. 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451
5. Hyndman, R.J. and Athanasopoulos, G. 2021. *Forecasting: Principles and practice.* 3rd ed. OTexts, Melbourne, Australia. https://otexts.com/fpp3/
6. International Crops Research Institute for the Semi-Arid Tropics. 2024. *District Level Database (DLD).* ICRISAT Development Center, Hyderabad, India. https://data.icrisat.org/district-level-data/
7. Jabed, M.A. and Murad, M.A.A. 2024. *Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for*

future research and sustainability. Heliyon 10(24):e40836. https://doi.org/10.1016/j.heliyon.2024.e40836

8. James, G., Witten, D., Hastie, T. and Tibshirani, R. 2021. *An introduction to statistical learning: With applications in R.* 2nd ed. Springer, Cham, Switzerland. https://doi.org/10.1007/978-1-0716-1418-1
9. Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K.-M., Gerber, J.S., Reddy, V.R. and Kim, S.-H. 2016. *Random forests for global and regional crop yield predictions.* PLoS One 11(6):e0156571. https://doi.org/10.1371/journal.pone.0156571
10. Khaki, S. and Wang, L. 2019. *Crop yield prediction using deep neural networks.* Front. Plant Sci. 10:621. https://doi.org/10.3389/fpls.2019.00621
11. Khosravani Shariati, S.A. and Abbasi, A. 2025. *Machine learning-based winter wheat yield prediction using multisource data.* Agric. Water Manage. 322:109951. https://doi.org/10.1016/j.agwat.2025.109951
12. Kouame, A.K.K., Heuvelink, G.B.M. and Bindraban, P.S. 2025. *Unraveling drivers of maize (Zea mays L.) yield variability in Ghana: A machine learning approach.* Comput. Electron. Agric. 237:110647. https://doi.org/10.1016/j.compag.2025.110647
13. Kuhn, M. and Johnson, K. 2019. *Feature engineering and selection: A practical approach for predictive models.* 1st ed. Chapman and Hall/CRC, Boca Raton, FL. https://doi.org/10.1201/9781315108230
14. Little, R.J.A. and Rubin, D.B. 2019. *Statistical analysis with missing data.* 3rd ed. Wiley, Hoboken, NJ. https://doi.org/10.1002/9781119482260
15. Rabiner, L.R. 1989. *A tutorial on hidden Markov models and selected applications in speech recognition.* Proc. IEEE 77(2):257–286. https://doi.org/10.1109/5.18626
16. Shawon, S.M., Ema, F.B., Mahi, A.K., Niha, F.L. and Zubair, H.T. 2025. *Crop yield prediction using machine learning: An extensive and systematic literature review.* Smart Agric. Technol. 10:100718. https://doi.org/10.1016/j.atech.2024.100718
17. Solomon, S. 2014. *Sugarcane agriculture and sugar industry in India: At a glance.* Sugar Tech 16(2):113–124. https://doi.org/10.1007/s12355-014-0303-8
18. Sun, J., Di, L., Sun, Z., Shen, Y. and Lai, Z. 2019. *County-level soybean yield prediction using deep CNN-LSTM model.* Sensors 19(20):4363. https://doi.org/10.3390/s19204363

19. Wilson, G.T. 2016. *Time series analysis: Forecasting and control, 5th ed., by G.E.P. Box, G.M. Jenkins, G.C. Reinsel and G.M. Ljung.* J. Time Ser. Anal. 37(5):709–711. https://doi.org/10.1111/jtsa.12194