MAUSAM, **77**, 3 (July 2026),791-806

MAUSAM

DOI: https://doi.org/10.54302/mausam.v77i3.6995
Homepage: https://mausamjournal.imd.gov.in/index.php/MAUSAM

UDC No.633.61:551.5:004.8(540.51)

Integrating meteorological insights and machine learning for sugarcane yield prediction in South Gujarat

V. B. VIRANI[1]*, D. R. VAGHASIYA[2], VIBHA TAK[1], N. D. BARIA[3] and N. M. CHAUDHARI[4]

[1]*Agricultural Meteorological Cell, Navsari Agricultural University, Navsari, Gujarat, India*

[2]*Agricultural Meteorological Cell, Junagadh Agricultural University, Junagadh, Gujarat, India*

[3]*Dept. of Agronomy, Navsari Agricultural University, Navsari, Gujarat, India*

[4]*Dept. of Soil Science and Agril. Chemistry, Navsari Agricultural University, Navsari, Gujarat, India*

(*Received 16 March 2025, Accepted 7 August 2025*)

***Corresponding author's email: vivekvirani.vv@gmail.com**

सार – कृषि प्रबंधन और खाद्य सुरक्षा को बनाए रखने के लिए फसल की पैदावार का सटीक पूर्वानुमान आवश्यक है। यह अध्ययन गुजरात के प्रमुख गन्ना उत्पादक जिले में गन्ने की पैदावार का एक मजबूत पूर्वानुमान मॉडल विकसित करने के लिए मौसम संबंधी मापदंडों और मशीन लर्निंग तकनीकों का लाभ उठाता है। यह अध्ययन गुजरात, भारत के चार जिलों (नवसारी, भरूच, सूरत और तापी) में गन्ने की पैदावार की भविष्यवाणी के लिए स्टेपवाइज मल्टीपल लीनियर रिग्रेशन (SMLR) और तीन मशीन लर्निंग (ML) मॉडलों: "आर्टिफिशियल न्यूरल नेटवर्क (ANN)," "रैंडम फॉरेस्ट रिग्रेशन (RFR)," और "सपोर्ट वेक्टर रिग्रेशन (SVR)" के प्रदर्शन का मूल्यांकन करता है। मॉडलों को प्रशिक्षित और परीक्षित करने के लिए ऐतिहासिक पैदावार डेटा (2001-2019) और मौसम चर का उपयोग किया गया था, जिसमें एक होल्डआउट डेटासेट (2020-2022) पर सत्यापन किया गया था। परिणाम इंगित करते हैं कि ANN ने अधिकांश जिलों में अन्य मॉडलों से बेहतर प्रदर्शन किया, जिससे न्यूनतम त्रुटियां और उच्चतम पूर्वानुमान सटीकता प्राप्त हुई। विशेष रूप से, भरूच में, ANN ने 2491.28 टन/हेक्टेयर का RMSE और 3.57% का MAPE प्राप्त किया; सूरत में, RMSE 8139.02 टन/हेक्टेयर और MAPE 9.92% था; तापी में, RMSE 3630.44 टन/हेक्टेयर और MAPE 4.43% था। नवसारी में भी मॉडल ने 5388.97 टन/हेक्टेयर के RMSE और 8.55% के MAPE के साथ अच्छा प्रदर्शन किया। SMLR ने नवसारी में मजबूत प्रदर्शन किया लेकिन अन्य क्षेत्रों में इसे और अधिक अनुकूलन की आवश्यकता थी। RFR और SVR ने सूरत और तापी में महत्वपूर्ण त्रुटियों के साथ मिश्रित परिणाम दिखाए, जो क्षेत्रीय परिवर्तनशीलता को दर्ज करने में चुनौतियों पर प्रकाश डालते हैं। विशेषता महत्व विश्लेषण (Feature importance analysis) से पता चला कि मौसम संबंधी चर (Weather Variables), जैसे सापेक्षिक आर्द्रता (Relative Humidity) और वर्षा (Rainfall), सभी जिलों में उपज पूर्वानुमान के लिए सबसे महत्वपूर्ण पूर्वानुमानक (Predictors) थे। अध्ययन इस बात पर बल देता है कि रिमोट सेंसिंग (Remote Sensing) डेटा को मौसम संबंधी चरों (Meteorological Variables) के साथ एकीकृत (Integrate) करने से मॉडल की सटीकता में उल्लेखनीय सुधार किया जा सकता है, विशेष रूप से SMLR (स्टेपवाइज़ मल्टीपल लीनियर रिग्रेशन) मॉडल के लिए। अध्ययन के अनुसार, भारूच (Bharuch), सूरत (Surat) और तापी (Tapi) जिलों में ANN (कृत्रिम तंत्रिका नेटवर्क) गन्ने की उपज का पूर्वानुमान लगाने के लिए अधिक उपयुक्त है, जबकि नवसारी (Navsari) जिले के लिए SMLR अधिक उपयुक्त पाया गया। ये निष्कर्ष गन्ने की पैदावार के पूर्वानुमान मॉडल में सुधार करने, टिकाऊ कृषि प्रथाओं का समर्थन करने और संसाधन आवंटन तथा जोखिम प्रबंधन में नीति निर्माताओं की सहायता करने के लिए मूल्यवान अंतर्दृष्टि प्रदान करते हैं।

ABSTRACT. Accurate crop yield forecasting is essential for sustainable agricultural management and food security. This study leverages meteorological parameters and machine learning techniques to develop a robust yield prediction model for sugarcane in the major sugarcane growing district of Gujarat. This study evaluates the performance of Stepwise Multiple Linear Regression (SMLR) and three machine learning (ML) models: "Artificial Neural Networks (ANN)," "Random Forest Regression (RFR)," and "Support Vector Regression (SVR)" for predicting sugarcane yield in four districts of Gujarat, India (Navsari, Bharuch, Surat, and Tapi). Historical yield data (2001–2019) and weather variables were used to train and test the models, with validation performed on a holdout dataset (2020–2022). Results indicate that ANN outperformed other models in most districts, achieving the lowest errors and highest predictive accuracy. Specifically, in Bharuch, ANN achieved an RMSE of 2491.28 t/ha and MAPE of 3.57%; in Surat, the RMSE was 8139.02 t/ha and MAPE 9.92%; in Tapi, the RMSE was 3630.44 t/ha and MAPE 4.43%. In Navsari, the model also performed well with an RMSE of 5388.97 t/ha and MAPE of 8.55%. SMLR demonstrated strong performance in Navsari

but required further optimization in other regions. RFR and SVR showed mixed results, with significant errors in Surat and Tapi, highlighting challenges in capturing regional variability. Feature importance analysis revealed that weather variables, such as relative humidity and rainfall, were critical predictors across all districts. The study underscores the importance of integrating remote sensing data with meteorological variables to enhance model accuracy, particularly for SMLR. ANN is recommended for yield forecasting in Bharuch, Surat, and Tapi, while SMLR is suitable for Navsari. These findings provide valuable insights for improving sugarcane yield prediction models, supporting sustainable agricultural practices, and aiding policymakers in resource allocation and risk management.

Key words – Machine learning, Meteorology, Yield forecasting, Sugarcane, ANN and SMLR.

1. Introduction

Sugarcane (*Saccharum officinarum*) is a vital crop globally due to its widespread use in daily life and its economic importance in both dietary and industrial sectors (Zulu *et al.,* 2019). India ranks as the world's second-largest producer of sugar, following Brazil, and its sugar industry is also the country's second-biggest agro-processing sector (Tyagi *et al.,* 2023). With the population steadily growing, the demand for transportation has surged, leading to a rise in greenhouse gas emissions. Moreover, limitations in renewable energy production could pose challenges in meeting the energy needs of electric vehicles (Tarei *et al.,* 2021). In response to these challenges, biofuels have gained attention as a sustainable alternative in the transport sector. The Indian government launched the National Policy on Biofuels in 2018, which permitted the utilization of sugarcane juice for ethanol production (Azad *et al.,* 2024). As a result, sugarcane has become a key crop for both sugar and ethanol manufacturing.

Sugarcane is a vital cash crop in South Gujarat, but its cultivation faces intertwined climatic, infrastructural, and socio-economic challenges. Despite high rainfall, uneven distribution, and poor drainage cause waterlogging in low-lying areas, while upland regions suffer from falling groundwater and unreliable canal irrigation. Small landholdings further hinder timely adoption of modern practices. Socio-economic issues also impact productivity-farmers report low procurement rates by co-operative societies, high labour costs and shortages, dishonest weighing, and high transport costs. Delays in harvesting, late payments, lack of mechanization, and poor advisory services worsen the situation (Chaudhari and Trivedi, 2023). In conclusion, farmers primarily face challenges related to low procurement prices, high labour costs, labour scarcity, and dishonest practices at the weighbridge.

Crop yield forecasting is a crucial yet challenging task, especially in countries like India that experience unpredictable weather patterns (Paudel *et al.,* 2021). Accurate mid-season yield prediction is essential for supporting farmers, policymakers, agronomists, and other stakeholders in the agricultural supply chain (Chipanshi *et al.,* 2015). Several methods are employed for yield forecasting, including field surveys, crop simulation models (CSM), and empirical approaches such as regression models (Basso and Liu, 2019). Field surveys, though effective, are labor-intensive, time-consuming, and prone to errors due to inadequate sampling and reliance on farmers' responses. Empirical models establish statistical relationships between predictors and crop yield, while mechanistic models use equations to simulate crop growth based on environmental conditions (Dimov *et al.,* 2022). However, the application of CSM on a regional scale is limited due to insufficient data and challenges in validating model outputs (Dhakar *et al.,* 2022). Regression models, on the other hand, are widely used in agricultural monitoring due to their simplicity and lack of the complex calibration process required for mechanistic models (Lobell and Asseng, 2017; Mathieu and Aires, 2018). Additionally, advancements in remote sensing and machine learning are enhancing yield prediction accuracy by integrating satellite-derived data and climatic variables (Virani *et al.,* 2024).

Various empirical techniques have been implemented to enhance the accuracy of sugarcane yield estimation, showcasing significant progress in predictive modeling. SMLR has been effectively used to establish relationships between variables (Kumar *et al.,* 2014; Dubey *et al.,* 2018; Verma *et al.,* 2020; Kumar *et al.,* 2022). SVR has gained recognition for managing nonlinear data patterns (Nihar *et al.,* 2022; Abebe *et al.,* 2022). ANN have demonstrated the ability to capture intricate data patterns (Abebe *et al.,* 2022; Krupavathi *et al.,* 2022). RFR is widely acknowledged for its robustness and efficiency in handling large datasets (Singla *et al.,* 2020; Shendryk *et al.,* 2021; Nihar *et al.,* 2022). Moreover, Extreme Gradient Boosting (XGBoost) has emerged as a powerful technique for improving prediction accuracy (Nihar *et al.,* 2022; Zhu *et al.,* 2023). These diverse methodologies reflect ongoing advancements in improving the precision and reliability of sugarcane yield forecasting.

2. Data and methodology

2.1. *Study area*

The key sugarcane-growing districts of South Gujarat were selected for this study due to their significant

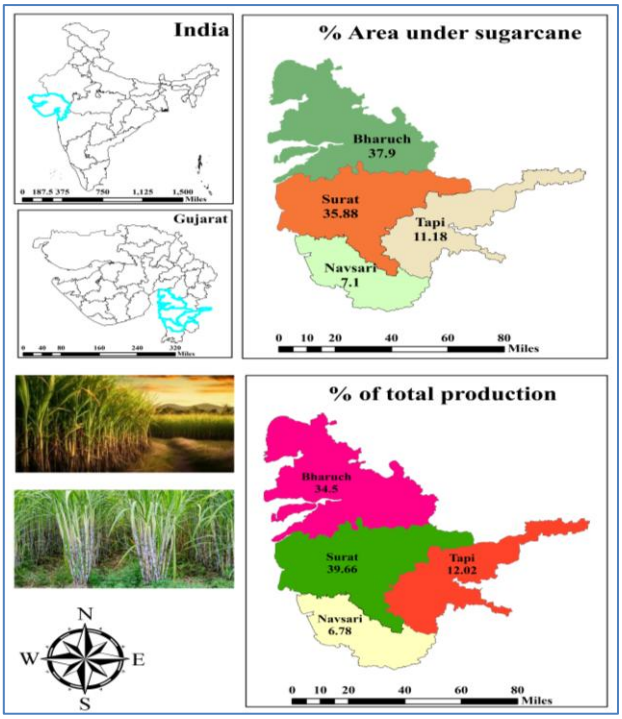

Fig. 1. Map showing the forecasting districts (Navsari, Surat, Bharuch, and Tapi) of South Gujarat

contributions to the state's sugarcane cultivation (Fig. 1). These four districts-Bharuch, Surat, Tapi, and Navsari-collectively account for approximately 83% of Gujarat's total sugarcane area and production. Bharuch district emerged as the leader in sugarcane cultivation, with the largest planting area of 68.7 thousand hectares. However, Surat district recorded the highest production, reaching 6,577 thousand metric tonnes. Despite having a comparatively smaller cultivation area and lower production, Navsari district made a modest contribution to the state's sugarcane output. Notably, Tapi district achieved the highest productivity at 83,175 t/ha, closely followed by Surat district with 81,827 t/ha, reflecting superior crop management and favorable agro-climatic conditions in these regions.

2.2. *Datasets*

The study employed meteorological data collected over a 21-year period (2001-2021), which included variables such as maximum temperature (T_{max}), minimum temperature (T_{min}), rainfall (RF), morning and afternoon relative humidity (RH I and RH II), and bright sunshine hours (BSSH). Historical sugarcane yield data for the same period was obtained from the Directorate of Agriculture, Gujarat. For the analysis, it was assumed that biophysical conditions, soil characteristics, and agricultural practices were consistent across the districts. The yield data served as

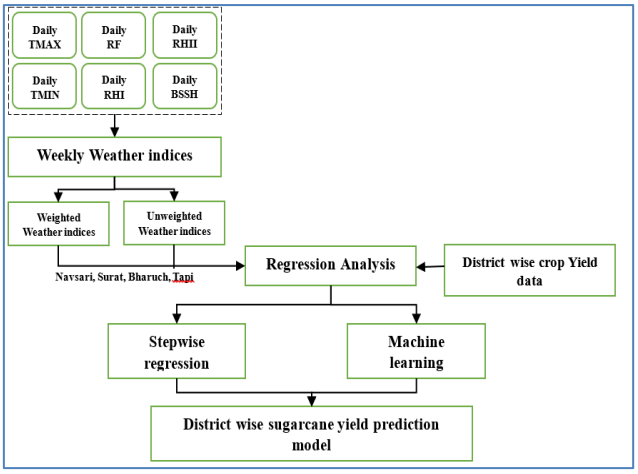

Fig. 2. Workflow employed for sugarcane yield estimation

the dependent variable, while the meteorological parameters were used as independent variables to develop the forecasting model. The dataset used in this study is complete with no missing values. Therefore, no imputation or data cleaning for missing data was required prior to model development and analysis.

2.3. *Methodology used for yield forecast*

2.3.1. *Generation of weather indices*

Daily meteorological observations were converted into weekly averages to develop weather indices. The unweighted indices were formed by simply summing these weekly means. For the weighted indices, correlation coefficients-reflecting the strength of association between weekly weather parameters and sugarcane yield-were used as weights to enhance the relevance of each variable. (Panwar *et al.*, 2018). A total of 42 indices were calculated, comprising 21 weighted and 21 unweighted indices. This set included 6 weighted weather indices, 15 weighted interaction indices, 6 unweighted weather indices, and 15 unweighted interaction indices (Table 1). The calculations were performed using the formulas outlined by Azfar *et al.*, 2015.

2.3.2. *Yield forecasting models development*

2.3.2.1. *SMLR*

"Stepwise regression is a technique for constructing regression models where the selection of predictive variables is automated." At every stage, variables were assessed for addition or removal from the pool of explanatory variables based on specific selection criteria (Dubey *et al.*, 2018). The complete SMLR analysis was performed using SPSS software.

TABLE 1

Unweighted and weighted weather indices

Parameter	Unweighted Weather Indices						Weighted Weather Indices					
	T_{max}	T_{min}	RF	RH-I	RH-II	BSSH	T_{max}	T_{min}	RF	RH-I	RH-II	BSSH
T_{max}	Z10						Z11					
T_{min}	Z120	Z20					Z121	Z21				
RF (rainfall)	Z130	Z230	Z30				Z131	Z231	Z31			
RH-$I_{morning}$	Z140	Z240	Z340	Z40			Z141	Z241	Z341	Z41		
RH-$II_{afternoon}$	Z150	Z250	Z350	Z450	Z50		Z151	Z251	Z351	Z451	Z51	
BSSH	Z160	Z260	Z360	Z460	Z560	Z60	Z161	Z261	Z361	Z461	Z561	Z61

TABLE 2

Statistics of sugarcane yield variability in the forecasting districts

District	Mean	Maximum	Minimum	Std	CV (%)	Shapiro–Wilk Test	
						Statistic	p Value
Navsari	65907	73380	57731	4565.70	6.92	0.954	0.315
Bharuch	68352	74640	61420	4482.89	6.56	0.911	0.051
Surat	72883	78820	65550	3420.48	4.69	0.971	0.728
Tapi	65731	85260	44350	10030.39	15.26	0. 950	0.263

TABLE 3

Sugarcane yield prediction equations using SMLR for different districts of Gujarat during 2023

District	Regression Equation	Weather variable in the Equation	R^2	Adjusted R^2	F
Navsari	Y = 8536.126+ (Z121*12.53) + (Z351*0.134)	T_{max}*T_{min}, Rf*RH II	0.712	0.659	13.57
Bharuch	Y = 69558.945 + (Z451*2.196) + (Z130*0.194) + (Z40*-29.79) + (Z120*1.712)	RH I*RH II, T_{max}*Rf, RH I, T_{max}*T_{min}	0.898	0.857	21.93
Tapi	Y = 151028.691 + (Z151*3.568) + (Z41*157.232) + (Z20*-151.015)	T_{max}*RH II, RH I, T_{min}	0.852	0.824	30.60
Surat	Y = 60000.63 + (Z131*0.625) + (Z451*2.406) + (Z450*0.091)	T_{max}*Rf, RH I, RH II	0.883	0.851	27.70

TABLE 4

District-wise error percentage of sugarcane yield (kg/ha) validated for the years 2020, 2021, and 2022 using a SMLR model

District	2020			2021			2022		
	Predicted Yield (kg/ha)	Observed Yield (kg/ha)	Error (%)	Predicted Yield (kg/ha)	Observed Yield (kg/ha)	Error (%)	Predicted Yield (kg/ha)	Observed Yield (kg/ha)	Error (%)
Navsari	66009	63773	**3.3**	64771	74690	**-15.0**	67351	68100	**-1.1**
Bharuch	65159	71000	**-8.9**	64763	70360	**-8.6**	74147	67800	**8.5**
Surat	72588	85000	**-17.0**	67784	86560	**-27.0**	74783	74420	**0.4**
Tapi	61837	76000	**-22.0**	84450	84200	**0.30**	94455	85230	**9.7**

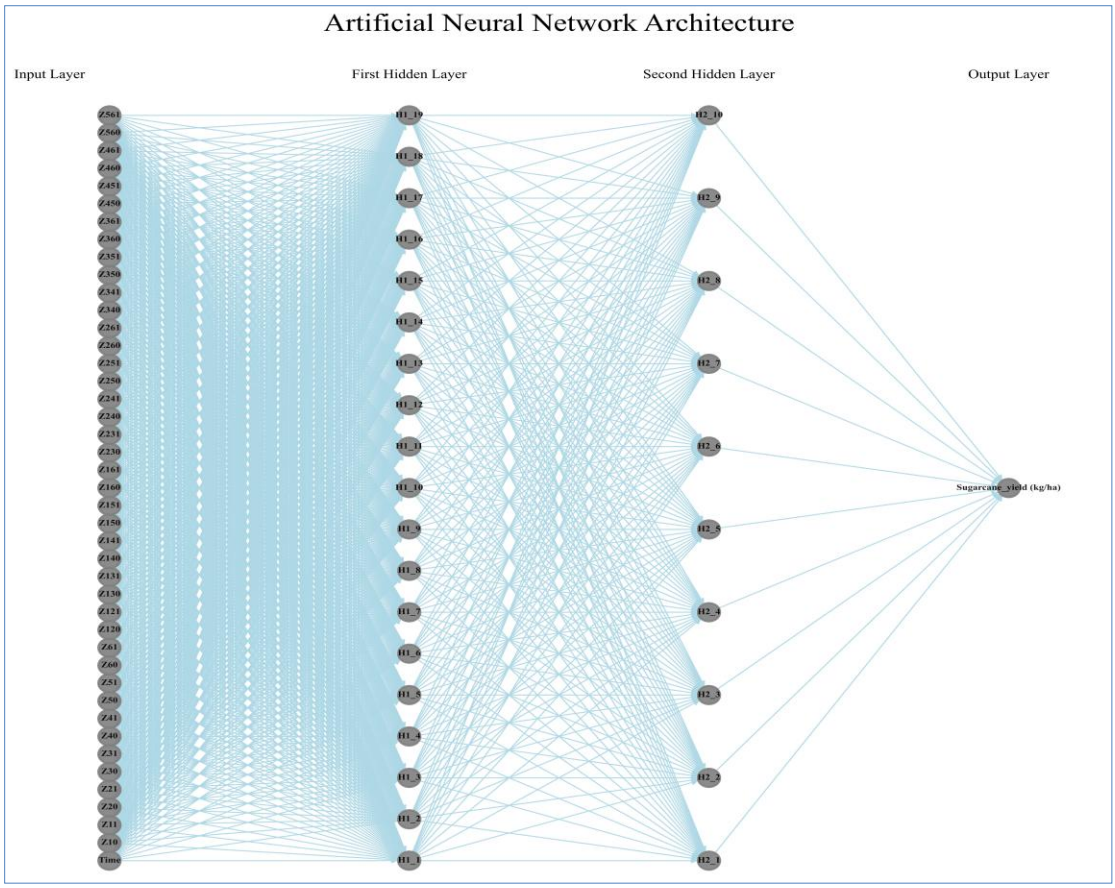

Fig. 3. Architecture of ANN model

2.3.2.2. *ML model*

The research utilized three machine learning algorithms-RFR, SVR, and ANN-each known for their unique strengths. The selection of these models was guided by the specific features and complexity of the dataset. Since no single algorithm is optimal for every scenario due to varying performance across different data types, it is essential to compare multiple approaches. Therefore, all three models were implemented to determine which was best suited for predicting sugarcane yield under the given conditions. Fig. 4 and Fig. 5 display the pseudocode used for building the RFR, SVR, and ANN models (Fig. 3), respectively. Additionally, Fig. 2 presents the procedural flowchart for sugarcane yield forecasting.

The models were developed using Python 3.11 within the Jupyter Notebook environment. The RFR and SVR were implemented through the "scikit-learn 1.2.2" library, while the ANN models were constructed and trained using "Keras 2.11.0" For handling numerical computations efficiently, the "NumPy 1.24.2" library was employed. To optimize the performance of RFR and SVR, hyperparameter tuning was carried out using "GridSearchCV," which explores a predefined set of parameters to determine the most effective configuration. For ANN, hyperparameter tuning was carried out using the "Keras Tuner 1.3.5" ensuring optimal model configuration and improved predictive accuracy. The description of the hyperparameters used for the ANN, SVR, and RFR models is provided in Table 5.

Data visualization, including the generation of feature importance plots, was carried out using the "Matplotlib 3.7.1" library. The "Pandas 1.5.3" library was essential for data preprocessing, manipulation, and cleaning throughout the workflow. To evaluate model performance, the dataset was split into training and testing subsets in an 80:20 ratio. The models were trained by feeding 80% dataset and after model fitting, the accuracy assessment was conducted based on 20% testing data and further assessed using an additional three years of holdout data (validation dataset) (2020-2022), which was completely unseen data for models. The validation dataset consists of data from three recent years (e.g., 2020–2022), selected to capture a representative range of interannual climate variability characteristic of South Gujarat. This

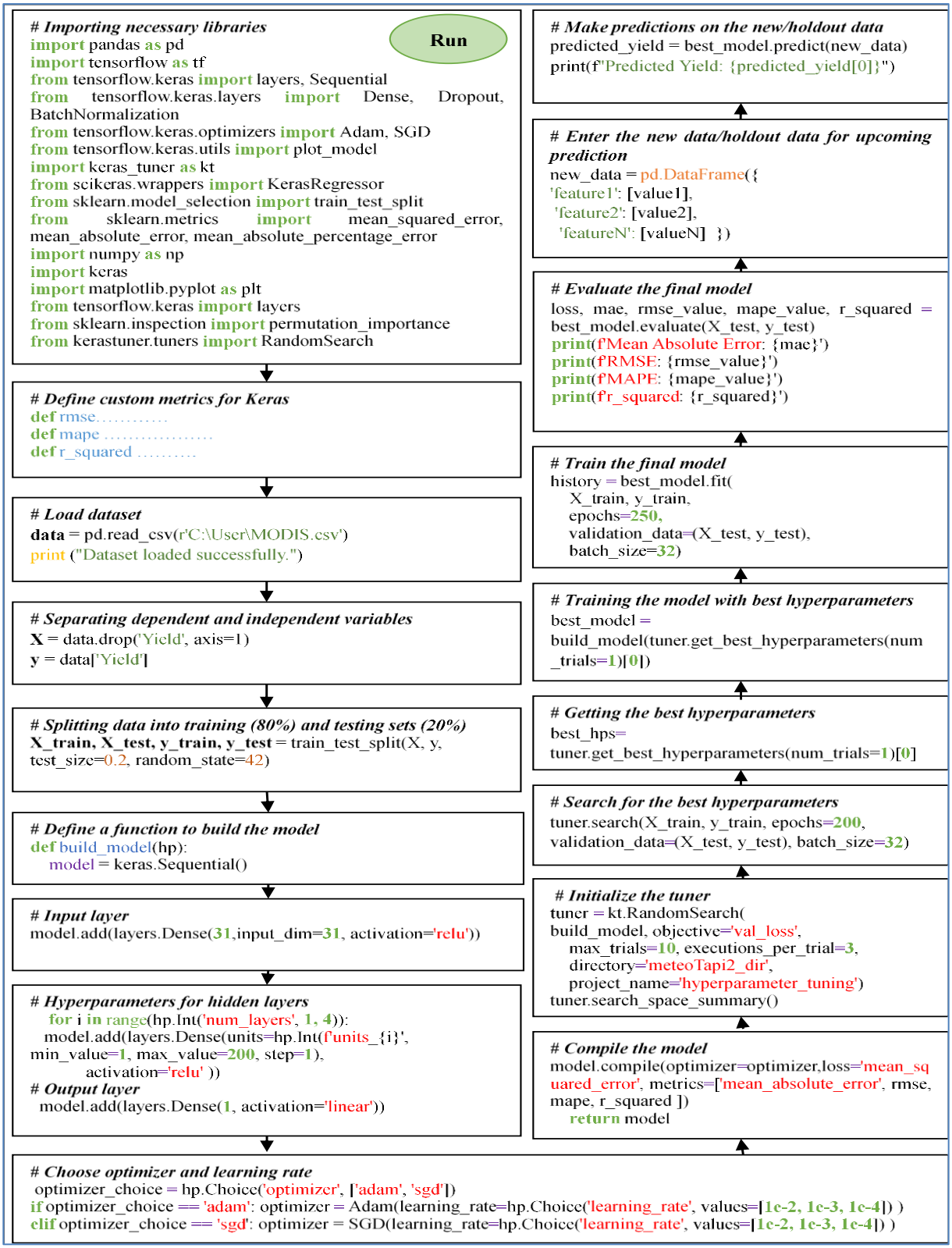

Fig. 4. Pseudocode for building RFR and SVR model

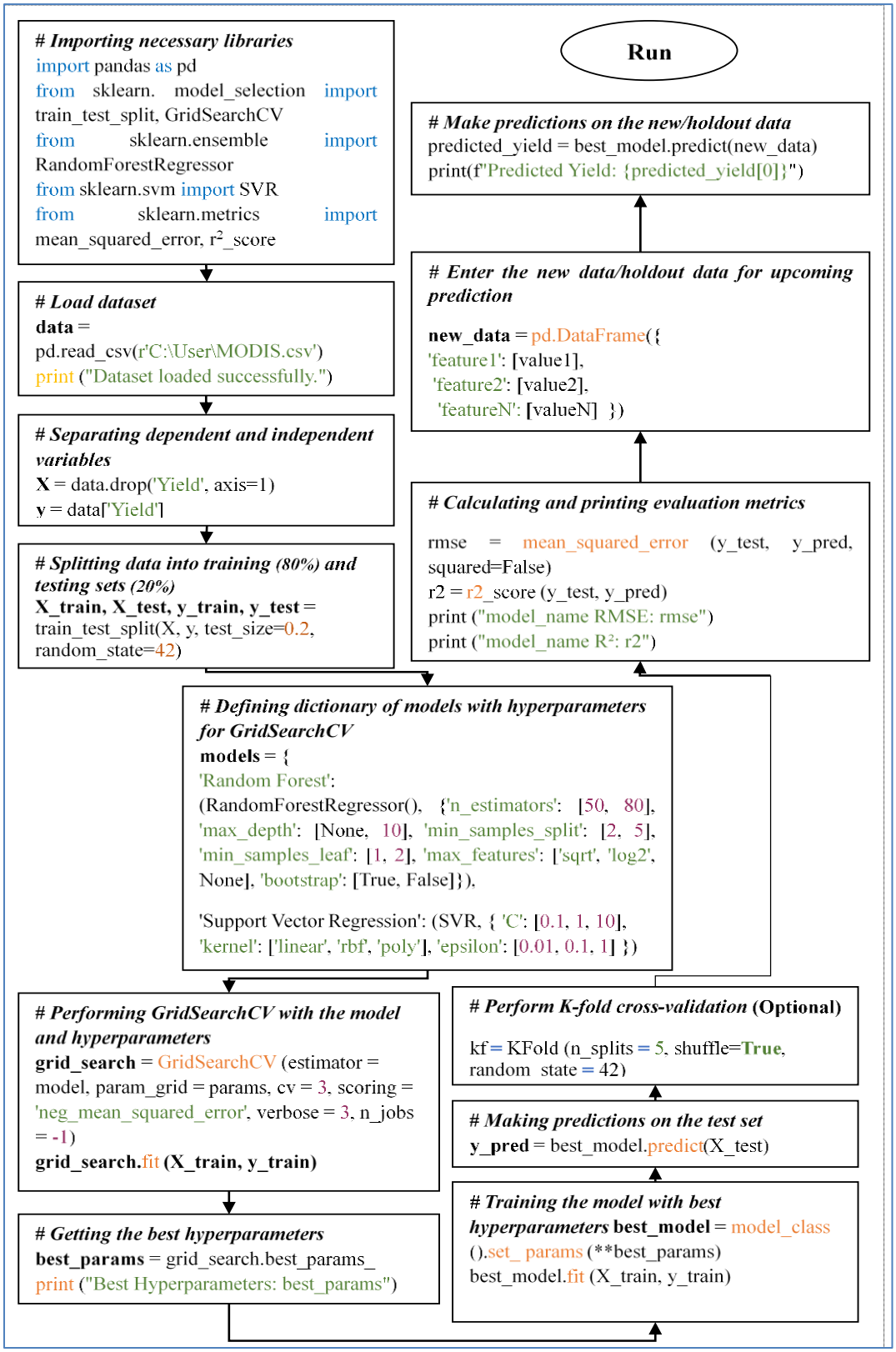

Fig. 5. Pseudocode for building ANN model

TABLE 5

Description of hyperparameters of ANN, SVR and RFR model

Sr. No.	Name of hyperparameter	Description
1	Number of Layers & Neurons	Determines the network's capacity to learn complex patterns, where more layers and neurons can increase learning capability but also risk overfitting
2.	Learning Rate & Optimizer	The learning rate controls how fast the model learns, while the optimizer helps adjust the model's settings to improve its performance based on errors made during training.
3.	epochs	The number of epochs determines how many times the model will go through the entire training dataset. More epochs can help the model learn better, but too many can lead to overfitting.
4.	batch size	The batch size is the number of training examples used to update the model weights in one iteration. A larger batch size can speed up training but requires more memory, while a smaller batch size provides more frequent updates and can help the model generalize better.
5.	C	Controls the trade-off between achieving a low error on the training data and minimizing the model complexity. A higher value of C emphasizes fitting the training data more accurately, while a lower value encourages a simpler model.
6.	Gamma	Determines the influence of a single training example. A low gamma means that the influence is far, leading to smoother decision boundaries, while a high gamma means that the influence is close, leading to more complex decision boundaries.
7.	Epsilon	Specifies the margin of tolerance where no penalty is given for errors. In regression tasks, it defines a tube around the predicted values within which errors are considered acceptable.
8.	Kernel	The kernel function transforms the input data into a higher-dimensional space where it can be more easily separated.
9.	n_estimators	The number of trees in the forest.
10.	max_depth	The maximum depth of each tree. It controls how deep the trees in the forest can grow, affecting the model's complexity and risk of overfitting.
11.	min_samples_split	The minimum number of samples required to split an internal node. It controls whether a node can be further split and can affect overfitting.
12.	min_samples_leaf	The minimum number of samples required to be at a leaf node. It ensures that leaf nodes have a minimum number of samples, which can help prevent overfitting.
13.	max_features	The number of features to consider when looking for the best split. It controls the randomness in the model, with fewer features per split increasing randomness and potentially reducing overfitting.

timeframe was chosen to ensure the model's robustness across diverse weather patterns affecting sugarcane growth, while maintaining an adequate dataset for training. Historical climate records confirm that these years reflect typical fluctuations in rainfall and temperature relevant to crop yield.

Models training and testing were performed on a workstation equipped with an NVIDIA GeForce RTX 3080 GPU (10 GB VRAM), Intel Core i7-12700K CPU (12 cores), and 16 GB RAM, running on Windows 11 Pro (64-bit). The average training time for ANN models was approximately 40 minutes, depending on the number of epochs and batch size selected during tuning, while RFR training completed in around 22 minutes, and SVR training took approximately 13 minutes, including hyperparameter tuning processes.

2.3.3. *Accuracy assessment metrics*

To evaluate the model's accuracy, this study used metrics including "root mean square error (RMSE)," "mean absolute error (MAE)," "mean bias error (MBE)," and "mean absolute percentage error (MAPE)" to assess the differences between predicted and observed values.

TABLE 6

Hyperparameter configuration for the ANN

District	Hyperparameters						Extra information	
	Number of layers	Number of neurons in layer	Learning rate	Optimizer	epochs	Batch Size	Executions per trial	Maximum trials
Navsari	3	1:40 2:34 3:30	0.01	adam	200	32	5	10
Bharuch	1	1:99	0.001	adam	150	32	3	10
Surat	3	1:12 2:73 3:29	0.001	adam	250	16	3	50
Tapi	1	1:147	0.001	adam	250	32	3	10

TABLE 7

District-wise error percentage of sugarcane yield (kg/ha) validated for the years 2020, 2021, and 2022 using ANN model

District	2020			2021			2022		
	Predicted Yield (kg/ha)	Observed Yield (kg/ha)	Error (%)	Predicted Yield (kg/ha)	Observed Yield (kg/ha)	Error (%)	Predicted Yield (kg/ha)	Observed Yield (kg/ha)	Error (%)
Navsari	69578	63773	8.34	70013	74690	-6.68	62483	68100	-8.99
Bharuch	67555	71000	-5.10	70193	70360	-0.24	70393	67800	3.68
Surat	74417	85000	-14.22	78295	86560	-10.56	78712	74420	5.45
Tapi	74416	76000	-2.13	81511	84200	-3.30	78771	85230	-8.20

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} |y(i) - \hat{y}(i)|^2}{n}} \quad (1)$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y(i) - \hat{y}(i)|}{n} \quad (2)$$

$$\mathrm{MBE} = \sum_{i=1}^{n} \left| \frac{y(i) - \hat{y}(i)}{n} \right| \quad (3)$$

$$\mathrm{MAPE} = \frac{1}{n} x \sum_{i=1}^{n} \left| \frac{y(i) - \hat{y}(i)}{y(i)} \right| \times 100 \quad (4)$$

where n is the "total number of data points," $y(i)$ is the "actual value for the i^{th} data point," and $\hat{y}(i)$ is the "predicted value for the i^{th} data point."

3. Results and discussion

3.1. *Overview of sugarcane yield variability across the forecasting districts*

The summary statistics for the yield data from the four districts of Gujarat (Navsari, Bharuch, Surat, and Tapi) are presented in the Table 2. The maximum yield observed was in Bharuch district (74640 kg ha^{-1}), while the minimum yield was recorded in Tapi district (44350 kg ha^{-1}). The yield variability, measured by the standard deviation, ranged from 3420.48 to 10030.39 kg ha^{-1}, with the highest variation seen in Tapi district. The coefficient of variation (CV) showed the highest yield variability in Tapi (15.26%), followed by Navsari (6.92%), Bharuch (6.56%), and Surat (4.69%).

To assess the normality of the yield data, the Shapiro-Wilk test was conducted for each district. The findings showed that the yield data for all districts exhibited a normal distribution, as validated by the Shapiro-Wilk test, with p-values exceeding 0.05. The normal Q-Q plots also showed that the data points aligned along a diagonal line, supporting the assumption of normality. Furthermore, homoscedasticity in the data was assessed using the Breusch-Pagan (BP) test. The BP test results indicated p-values greater than 0.05 for all four districts, confirming that the residuals exhibit constant variance. These findings validate the suitability of the data for fitting into parametric model like SMLR.

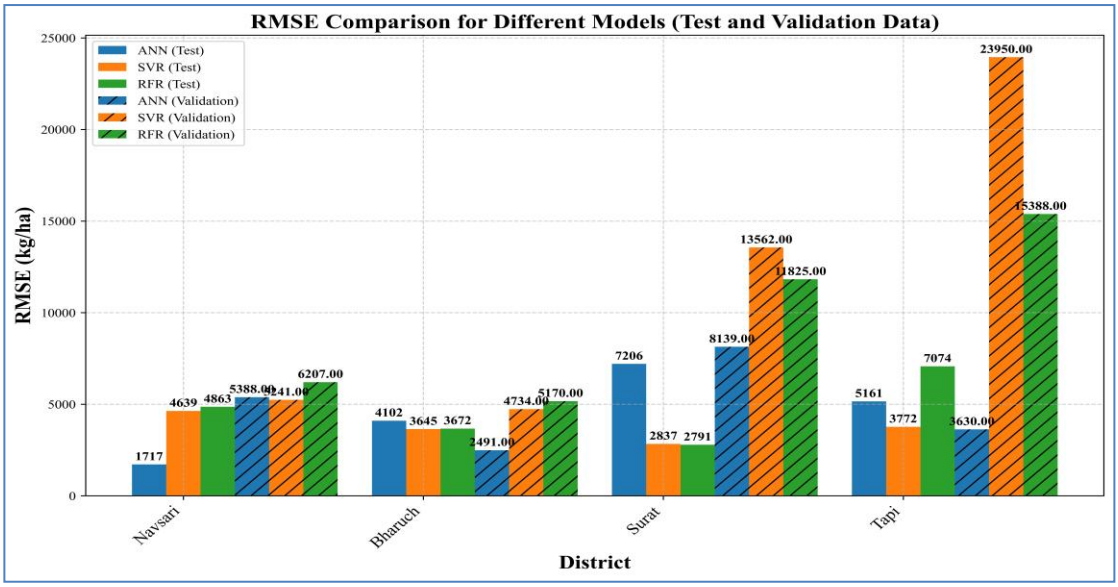

Fig. 6. RMSE comparison for ML models during testing and validation stage

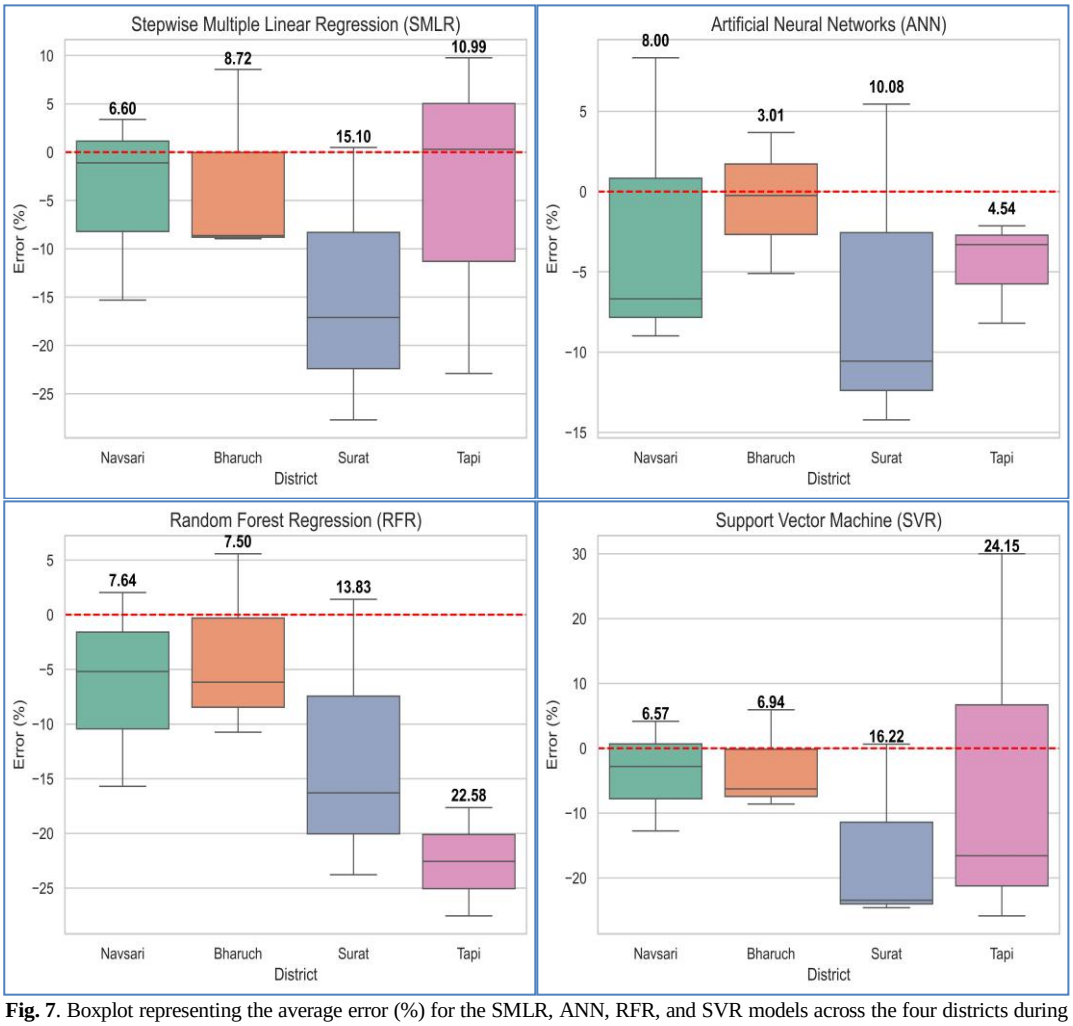

Fig. 7. Boxplot representing the average error (%) for the SMLR, ANN, RFR, and SVR models across the four districts during the validation stage

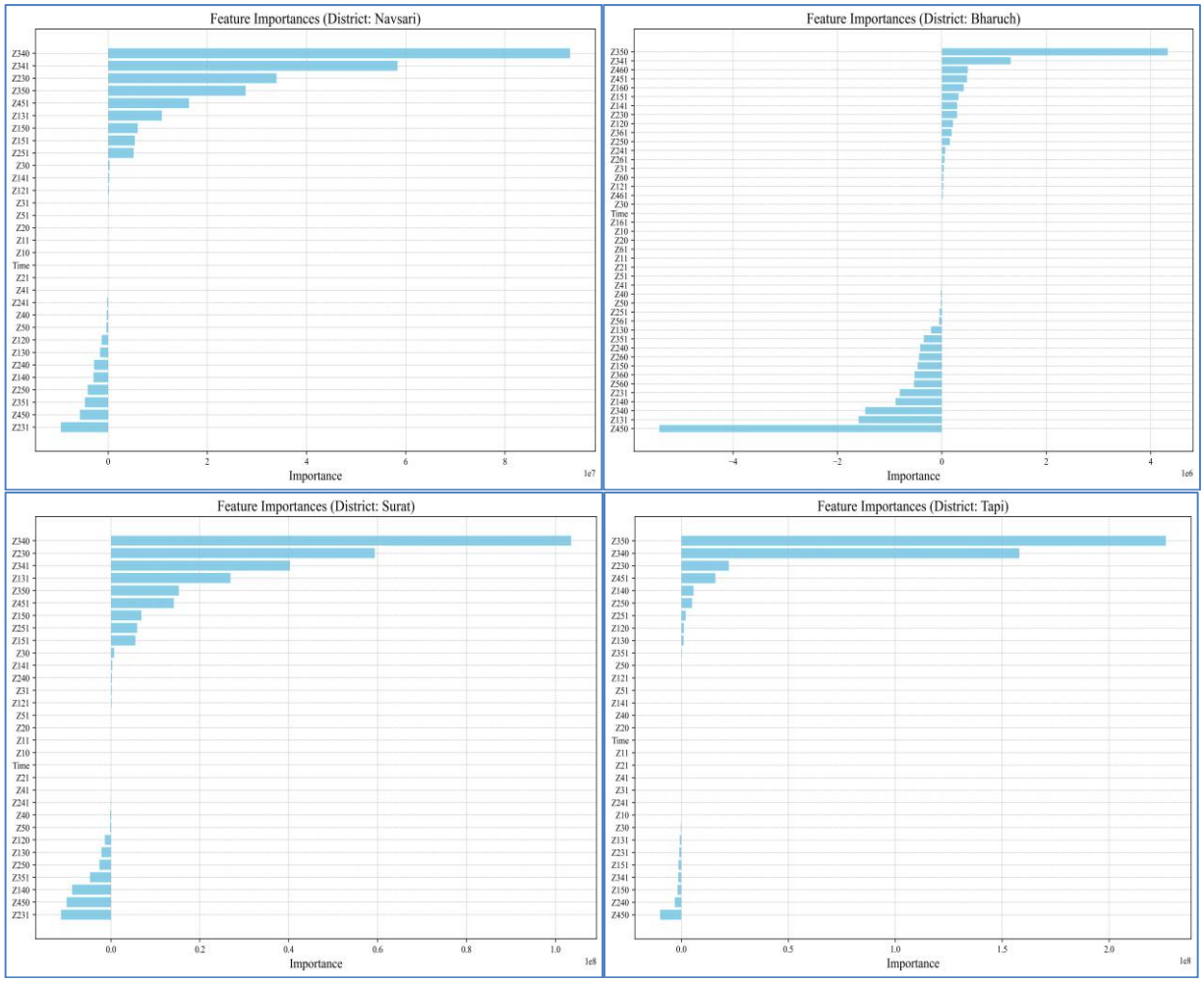

Fig. 8. Feature importance of weather variables in the ANN model for sugarcane yield prediction in Navsari, Bharuch, Surat & Tapi districts

3.2. *Sugarcane yield forecasting models*

3.2.1. *SMLR model*

In 2023, sugarcane yields for Navsari, Bharuch, Surat, and Tapi districts were forecasted using SMLR in SPSS. The regression equation, influenced by weather variables, along with R^2, adjusted R^2, and F-statistic values, are presented in Table 3. The highest adjusted R^2 (0.86) was observed in Bharuch, while Navsari had the lowest (0.66), indicating limited predictive power. The models for Bharuch, Tapi, and Surat showed strong predictive capabilities, highlighting the significant impact of weather variables on sugarcane yield. However, the Navsari model required further optimization.

The model was validated for the period 2020-2022, and the district-specific yield deviations from the observed values are presented in Table 4. The yields were generally underestimated, with error percentages for Navsari, Bharuch, Surat, and Tapi at 3.3%, -15.0%, and -1.1% in 2020, 2021, and 2022, respectively. Bharuch had errors of -8.9%, -8.6%, and 8.5%, Surat showed -17.0%, -27.0%, and 0.4%, and Tapi exhibited -22.0%, 0.3%, and 9.7%. While some districts had minor errors, others, like Surat and Tapi, showed more significant discrepancies, often exceeding the acceptable ±10% range. Further adjustments and the inclusion of additional variables are necessary to enhance model accuracy, especially for Navsari and Surat.

3.2.2. *ANN model*

In this study, a feed-forward ANN was developed in Python for predicting sugarcane yield. The ANN architecture, including the number of hidden layer

TABLE 8

Hyperparameter configuration for the RFR

District	Hyperparameters				
	n_estimators	max_depth	min_samples_split	min_sample_leaf	Max_features
Navsari	100	100	2	3	log2
Bharuch	900	None	7	1	0.5
Surat	100	100	2	1	0.6
Tapi	600	20	6	1	sqrt

TABLE 9

District-wise error percentage of sugarcane yield (kg/ha) validated for the years 2020, 2021, and 2022 using RFR model

District	2020			2021			2022		
	Predicted Yield (kg/ha)	Observed Yield (kg/ha)	Error (%)	Predicted Yield (kg/ha)	Observed Yield (kg/ha)	Error (%)	Predicted Yield (kg/ha)	Observed Yield (kg/ha)	Error (%)
Navsari	65093	63773	2.03	64562	74690	-15.69	64742	68100	-5.19
Bharuch	64112	71000	-10.74	66273	70360	-6.17	71808	67800	5.58
Surat	73090	85000	-16.29	69931	86560	-23.78	75484	74420	1.41
Tapi	64440	76000	-17.94	66016	84200	-27.54	69543	85230	-22.56

neurons per layer, and other hyperparameters, was fine-tuned using the "Keras Tuner." The optimized hyperparameters for each district, including learning rate, batch size, epochs, and number of neurons, are displayed in Table 6. The model was trained and tested on data from 2001 to 2019 (80% for training and 20% for testing) and validated on a holdout dataset from 2020 to 2022 to assess prediction accuracy.

The ANN model showed varying accuracy across districts (Table 7). In Navsari, it slightly overestimated yield in 2020 (8.34%) but underestimated in 2021 (-6.68%) and 2022 (-8.99%). In Bharuch, predictions were most accurate in 2021 (-0.24%) but had minor errors in 2020 (-5.10%) and 2022 (3.68%). Surat showed underpredictions in 2020 (-14.22%) and 2021 (-10.56%) and a slight overestimation in 2022 (5.45%). Tapi had small errors in 2020 (-2.13%) and 2021 (-3.30%), with a larger underestimation in 2022 (-8.20%). As the error was within the acceptable ±10% range for all districts, the ANN model was considered superior to the SMLR model.

Fig. 8 highlights feature importance derived from the ANN model, identifying key predictors for each district. For Navsari, Z340 and Z341 were the most influential features. In Bharuch, Z350 and Z341 played a critical role, while in Surat, Z340 and Z230 were significant. In Tapi, Z350 and Z340 were the primary predictors. These results emphasize the importance of weather variables, such as relative humidity and rainfall, in enhancing yield forecasting accuracy across all districts.

3.2.3. *RFR model*

The configuration of the RFR model, including the number of trees, maximum depth, and other hyperparameters (Table 8), was optimized using "GridSearchCV." The model was trained and tested on data from 2001 to 2019 (80% for training and 20% for testing) and validated on a holdout dataset from 2020 to 2022 to assess its prediction accuracy.

The RFR model's performance varied across districts. In Navsari, the model achieved minimal error in 2020 (2.03%) but showed higher deviations in 2021 (-15.69%) and 2022 (-5.19%). Bharuch consistently underpredicted yields in 2020 (-10.74%) and 2021 (-6.17%) but slightly overpredicted in 2022 (5.58%). Surat experienced significant underestimations in 2020 (-16.29%) and 2021 (-23.78%), with improved accuracy in 2022 (1.41%). Tapi showed the highest errors, with deviations of -17.94%, -27.54%, and -22.56% in 2020, 2021, and 2022, respectively (Table 9).

The results indicate that the RFR model struggled to account for regional factors, especially in Surat and Tapi, where errors exceeded the acceptable ±10% range. Incorporating additional data, such as soil moisture and crop health indices, could enhance model accuracy.

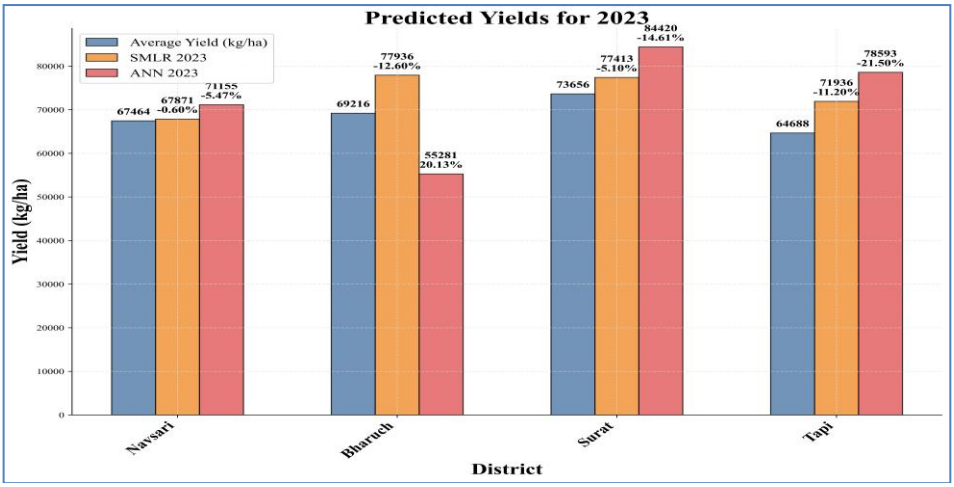

Fig. 9. District-wise average sugarcane yield and predicted yields using a SMLR and ANN during 2023 year

TABLE 10

Hyperparameter configuration for the SVR

District	Hyperparameters				
	C	gamma	epsilon	kernel	degree
Navsari	10	1e-05	0.0001	linear	2
Bharuch	100	1e-05	5	linear	3
Surat	100	1e-05	5	linear	2
Tapi	1000	0.0001	0.001	linear	2

3.2.4. *SVR model*

The SVR model was fine-tuned through hyperparameter optimization, including kernel type, regularization parameter (C), and kernel coefficient (gamma), as detailed in Table 10. The model's performance varied across districts, as shown in Table 11. In Navsari, the SVR model achieved an error of 4.16% in 2020, which increased to -12.75% in 2021 before improving to -2.81% in 2022. Bharuch showed consistent under-prediction in 2020 (-8.61%) and 2021 (-6.27%), followed by a slight over-prediction in 2022 (5.93%). Surat experienced significant errors in 2020 (-24.59%) and 2021 (-23.44%) but showed improved accuracy in 2022 (0.62%). Tapi exhibited the highest variation, with errors of -25.85% in 2020, -16.59% in 2021, and a notable over-prediction of 30.00% in 2022.

The results indicate that the SVR model performed reasonably well in Navsari and Bharuch, with errors within the acceptable range. However, it struggled to capture complex regional factors in Surat and Tapi, where errors were significantly higher. The model's performance highlights the need for further refinement and the integration of additional features to enhance prediction accuracy in these regions.

3.3. *Comparison of models for the predictability of regional sugarcane yield*

The comparison of models for predicting regional sugarcane yield involves evaluating the performance of different ML algorithms, including SMLR, ANN, RFR, and SVR. The ML models were trained and tested using historical data from various districts, with the goal of assessing their ability to accurately predict sugarcane yield across diverse agricultural and climatic conditions. The accuracy achieved during the testing and validation stages are presented in Fig. 6, with RMSE used as the accuracy metric. Fig. 6 shows that, in most cases, the accuracy achieved during the testing stage exceeds that of the validation stage. However, for the ANN model, higher accuracy was observed during the validation stage for the Bharuch, Surat, and Tapi districts.

Fig. 7 represents the average error (%) boxplot for the SMLR, ANN, RFR, and SVR models across the four districts during the validation stage (2020 to 2022). This boxplot visually compares the distribution of errors for each model, showcasing the spread, median, and potential outliers in prediction performance. The ANN model performs well in Surat, Bharuch, and Tapi districts, with the error (%) being relatively smaller compared to the other models. This suggests that the ANN model is more accurate in predicting sugarcane yield in these districts during the validation stage. Its lower error indicates better alignment with observed yield values in these areas, making it a more reliable model for those regions, while SMLR model is more suitable for sugarcane yield forecasting in the Navsari district.

TABLE 11

District-wise error percentage of sugarcane yield (kg/ha) validated for the years 2020, 2021, and 2022 using SVR model

District	2020			2021			2022		
	Predicted Yield (kg/ha)	Observed Yield (kg/ha)	Error (%)	Predicted Yield (kg/ha)	Observed Yield (kg/ha)	Error (%)	Predicted Yield (kg/ha)	Observed Yield (kg/ha)	Error (%)
Navsari	66538	63773	4.16	66245	74690	-12.75	66240	68100	-2.81
Bharuch	65369	71000	-8.61	66207	70360	-6.27	72076	67800	5.93
Surat	68226	85000	-24.59	70122	86560	-23.44	74886	74420	0.62
Tapi	60391	76000	-25.85	72217	84200	-16.59	121750	85230	30.00

TABLE 12

Statistical evaluation of sugarcane yield (kg/ha) prediction models across districts using RMSE, MAPE, MAE, and MBE metrics

Model	District	RMSE	MAPE	MAPE*	MAE	MBE
ANN	Navsari	5388.97	8.55	Excellent	5366.33	-1496.33
	Bharuch	2491.28	3.57	Excellent	1806.09	-339.66
	Surat	8139.02	9.92	Excellent	7713.33	4852.0
	Tapi	3630.44	4.43	Excellent	3244.0	-3244.0
SMLR	Navsari	5886.34	7.82	Excellent	4301.33	-2810.66
	Bharuch	5936.55	8.51	Excellent	5928.33	-1697.0
	Surat	12996.50	15.85	Fair	10517.0	-10275.0
	Tapi	9759.67	11.92	Good	7879.33	-1562.66
SVR	Navsari	5241.59	7.61	Excellent	4356.66	-2513.33
	Bharuch	4734.26	6.79	Excellent	4686.66	-1836.0
	Surat	13562.10	16.54	Fair	11226.0	-10915.33
	Tapi	23950.94	29.27	Poor	21370.66	2976.0
RFR	Navsari	6207.38	9.01	Excellent	4935.33	-4055.33
	Bharuch	5170.82	7.42	Excellent	4994.33	-2322.33
	Surat	11825.16	14.42	Good	9867.66	-9158.33
	Tapi	15388.01	18.80	Fair	15143.66	-15143.66

** MAPE classes: <10% = Excellent, 10–15% = Good, 15–25% = Fair, and >25% = Poor*

Table 12 represented the statistical evaluation of sugarcane yield prediction models across different districts (Navsari, Bharuch, Surat, and Tapi) reveals varying performance levels for the ANN, SMLR, SVR, and RFR models. The ANN model achieved excellent performance in most districts, particularly in Bharuch with a low RMSE of 2491.28 and a MAPE of 3.57%, while Surat showed higher errors with a RMSE of 8139.02 and MAPE of 9.92%. SMLR demonstrated good performance in Navsari and Bharuch but had fair or poor accuracy in Surat and Tapi. The SVR model had excellent accuracy in Navsari and Bharuch, though it performed poorly in Tapi with a very high RMSE and MAPE. RFR had mixed results, performing well in Bharuch but showing higher errors in Surat and Tapi. Overall, the ANN model outperformed others in terms of accuracy in most districts, while RFR and SVR faced challenges, especially in Surat and Tapi. Thimmegowda *et al.,* (2023) and Khaki and Wang (2019) reported that ANN outperformed other ML models included in their studies for crop yield prediction.

The ANN outperformed SMLR, SVR, and RFR due to its superior ability to capture complex, nonlinear relationships among the diverse meteorological and remote sensing variables used in the study. Unlike SMLR, which is limited to linear patterns, or SVR and RFR, which can struggle with high-dimensional or multicollinear data, ANN's adaptive architecture allows it to model intricate interactions more effectively. This makes it particularly suitable for agricultural yield prediction, where data patterns are often non-linear and influenced by multiple interacting factors. Van Klompenburg *et al.,* (2020) observed that ANN are the most used algorithm for crop yield prediction due to their robustness in making accurate predictions.

The ANN model is recommended for the weather data-based sugarcane yield forecasting in Bharuch, Surat, and Tapi districts due to its superior performance and accuracy across these regions. For Navsari district, the SMLR model is preferred due to its strong performance and reliability in sugarcane yield prediction. SMLR outperformed ANN likely due to the relatively linear relationship between climatic variables and yield in specific districts like Navsari. When data patterns are simpler, with less noise or fewer nonlinear interactions, linear models like SMLR can be more stable and less prone to overfitting than ANN. To further improve the model accuracy and increase the adjusted R^2 value from 0.66 onward, it is recommended to incorporate additional remote sensing variables alongside weather data. This integration is expected to capture more complex patterns in yield dynamics, improving prediction accuracy and providing more reliable forecasts for the agricultural community.

Islam *et al.* (2023) found that using only an NDVI product derived from remote sensing (RS) was inadequate for accurate crop yield prediction. However, incorporating supplementary meteorological variables into machine learning models led to a marked improvement in estimation accuracy. Similarly, Dubey *et al.,* (2018) noted that relying solely on VCI did not provide satisfactory accuracy and highlighted the importance of including other datasets, such as additional vegetation indices (VIs) and meteorological information, to improve predictive performance. In line with these findings, this study adopts an integrated approach, combining remote sensing data with meteorological data to improve the adjusted R^2 and accuracy of the SMLR model for Navsari district.

3.4. *District wise sugarcane yield prediction using ANN and SMLR during 2023*

The predicted sugarcane yields for 2023 across Navsari, Bharuch, Surat, and Tapi were compared with historical averages (Fig. 9) because the Directorate of Agriculture, Government of Gujarat (GoG), has not yet published the 2023 sugarcane productivity data.

4. Conclusions

This study presents an affordable method for estimating sugarcane yield by examining the correlation between weather indices and annual yield. The accuracy and performance of different models vary according to the district. The ANN model demonstrates superior forecasting accuracy among the selected models due to its ability to capture complex, non-linear relationships between input features, such as weather indices. Based on the study, when only meteorological data is available, the ANN model is the most effective choice for yield forecasting due to its capacity to model these complex relationships.

Weather data-based yield forecasting faces several limitations. The variability and uncertainty in weather patterns can cause inaccuracies, particularly in regions with extreme conditions. Weather data alone does not account for other factors influencing crop growth, such as soil quality and pest dynamics. Additionally, the daily temporal resolution of weather data may not align with crop development stages, and the uneven distribution of weather stations can lead to data gaps. These limitations emphasize the need to integrate other data sources, like satellite imagery, to improve forecasting accuracy.

Acknowledgments

The authors would like to acknowledge their own efforts in contributing to the development and execution of this study. Special thanks are extended to the development teams of QGIS, SPSS, and Google Earth Pro, as well as to the creators of the scikit-learn library. Additionally, the authors acknowledge the invaluable contributions from all those who supported the development and execution of this research.

Authors' Contributions

V. B. Virani: Manuscript writing, Data collection, Data Analysis, Machine Learning model development.
D. R. Vaghasiya: Manucript up gradation, Data arrangement. (*email*: *dhimantagmet@gmail.com*).
Vibha Tak: Data filtering, Data analysis. (*email*: *tak_vibha@yahoo.com*).
N. D. Baria: Facilitator, Manuscript up gradation. (*email*: *nayan17398@gmail.com*).
N. M. Chaudhari: Facilitator, Data analysis. (*email*: *nishantc909@gmail.com*).

Disclaimer: The opinions and conclusions presented in this research paper/article are those of the authors and do not necessarily represent the views of the organizations with which they are affiliated.

References

Abebe, G., Tadesse, T. and Gessesse, B., 2022, "Combined use of Landsat 8 and Sentinel 2A imagery for improved sugarcane yield estimation in Wonji-Shoa, Ethiopia", *Journal of the Indian Society of Remote Sensing*, **50**, 1, 143-157. https://doi.org/10.1007/s12524-021-01466-8.

Azad, V. K., De, K. and Majumder, S., 2024, "Ethanol blending and its environmental impacts: A case study of India", *Energy for Sustainable Development*, **79**, 101385. https://doi.org/10.1016/j.esd.2024.101385 .

Azfar, M., Sisodia, B. V. S., Rai, V. N. and Devi, M., 2015, "Pre-harvest forecast models for rapeseed & mustard yield using principal component analysis of weather variables", *Mausam*, **66**, 4, 761-766. https://doi.org/10.54302/mausam.v66i4.583.

Basso, B. and Liu, L., 2019, "Seasonal crop yield forecast: Methods, applications, and accuracies", *Advances in Agronomy*, **154**, 201-255. https://doi.org/10.1016/bs.agron.2018.11.002.

Chipanshi, A., Zhang, Y., Kouadio, L., Newlands, N., Davidson, A., Hill, H., Warren, R., Qian, B., Daneshfar, B., Bedard, F. and Reichert, G., 2015, "Evaluation of the Integrated Canadian Crop Yield Forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape", *Agricultural and Forest Meteorology*, **206**, 137-150. https://doi.org/10.1016/j.agrformet.2015.03.007.

Chaudhari, V. P. and Trivedi, S. M., 2023, "Problems Faced by Sugarcane Growers in Gujarat State", *International Journal of Agricultural Sciences*, **15**, 2, 12209-12210.

Dhakar, R., Sehgal, V. K., Chakraborty, D., Sahoo, R. N., Mukherjee, J., Ines, A. V., Kumar, S. N., Shirsath, P. B. and Roy, S. B., 2022, "Field scale spatial wheat yield forecasting system under limited field data availability by integrating crop simulation model with weather forecast and satellite remote sensing", *Agricultural Systems*, **195**, 103299. https://doi.org/10.1016/j.agsy.2021.103299

Dimov, D., Uhl, J. H., Löw, F. and Seboka, G. N., 2022, "Sugarcane yield estimation through remote sensing time series and phenology metrics", *Smart Agricultural Technology*, **2**, 100046. https://doi.org/10.1016/j.atech.2022.100046

Dubey, S. K., Gavli, A. S., Yadav, S. K., Sehgal, S. and Ray, S. S., 2018, "Remote sensing-based yield forecasting for sugarcane (Saccharum officinarum L.) crop in India", *Journal of the Indian Society of Remote Sensing*, **46**, 1823-1833. https://doi.org/10.1007/s12524-018-0839-2.

Islam, M. D., Di, L., Qamer, F. M., Shrestha, S., Guo, L., Lin, L., Mayer, T. J. and Phalke, A. R., 2023, "Rapid rice yield estimation using integrated remote sensing and meteorological data and machine learning", *Remote Sensing.*, **15**, 9, 2374. https://doi.org/10.3390/rs15092374.

Khaki, S. and Wang, L., 2019, "Crop yield prediction using deep neural networks", *Frontiers in Plant Science*, **10**, 621. https://doi.org/10.3389/fpls.2019.00621.

Krupavathi, K., Raghubabu, M., Mani, A., Prasad, P. R. K. and Edukondalu, L., 2022, "Field-scale estimation and comparison of the sugarcane yield from remote sensing data: A machine learning approach", *Journal of the Indian Society of Remote Sensing*, **50**, 2, 299-312. https://doi.org/10.1007/s12524-021-01448-w.

Kumar, N., Pisal, R., Shukla, S. and Pandey, K., 2014, "Crop yield forecasting of paddy, sugarcane and wheat through linear regression technique for South Gujarat", *Mausam*, **65**, 3, 361–364. https://doi.org/10.54302/mausam.v65i3.1041.

Lobell, D. B. and Asseng, S., 2017, "Comparing estimates of climate change impacts from process-based and statistical crop models", *Environmental Research Letters*, **12**, 1, 015001. https://doi.org/10.1088/1748-9326/aa518a.

Mathieu, J. A. and Aires, F., 2018, "Assessment of the agro-climatic indices to improve crop yield forecasting", *Agricultural and Forest Meteorology*, **253**, 15-30. https://doi.org/10.1016/j.agrformet.2018.01.031.

Nihar, A., Patel, N. R. and Danodia, A., 2022, "Machine-Learning-Based Regional Yield Forecasting for Sugarcane Crop in Uttar Pradesh, India", *Journal of the Indian Society of Remote Sensing*, **50**, 8, 1519-1530. https://doi.org/10.1007/s12524-022-01549-0.

Panwar, S., Kumar, A., Singh, K. N., Paul, R. K., Gurung, B., Ranjan, R., Alam, N. M. and Rathore, A., 2018, "Forecasting of crop yield using weather parameters–two step nonlinear regression model approach", *The Indian Journal of Agricultural Sciences*, **88**, 10, 1597-1599. https://doi.org/10.56093/ijas.v88i10.84230.

Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C. and Athanasiadis, I. N., 2021, "Machine learning for large-scale crop yield forecasting", *Agricultural Systems*, **187**, 103016. https://doi.org/10.1016/j.agsy.2020.103016.

Shendryk, Y., Davy, R. and Thorburn, P., 2021, "Integrating satellite imagery and environmental data to predict field-level cane and sugar yields in Australia using machine learning", *Field Crops Research*, **260**, 107984. https://doi.org/10.1016/j.fcr.2020.107984.

Singla, S. K., Garg, R. D. and Dubey, O. P., 2020, "Ensemble machine learning methods to estimate the sugarcane yield based on remote sensing information", *Revue d'Intelligence Artificielle*, **34**, 6. https://doi.org/10.18280/ria.340607.

Tarei, P. K., Chand, P. and Gupta, H., 2021, "Barriers to the adoption of electric vehicles: Evidence from India", *Journal of Cleaner Production*, **291**, 125847. https://doi.org/10.1016/j.jclepro.2021.125847.

Thimmegowda, M. N., Manjunatha, M. H., Huggi, L., Shivaramu, H. S., Soumya, D. V., Nagesha, L. and Padmashri, H. S., 2023, "Weather-based statistical and neural network tools for forecasting rice yields in major growing districts of Karnataka", *Agronomy*, **13**, 3, 704. https://doi.org/10.3390/agronomy13030704.

Tyagi, S., Chandra, S. and Tyagi, G., 2023, "Statistical modelling and forecasting annual sugarcane production in India: Using various time series models", *Annals of Applied Biology*, **182**, 3, 371-380. https://doi.org/10.1111/aab.12825.

Van Klompenburg, T., Kassahun, A. and Catal, C., 2020, "Crop yield prediction using machine learning: A systematic literature review", *Computers and Electronics in Agriculture*, **177**, 105709. https://doi.org/10.1016/j.compag.2020.105709.

Virani, V. B., Kumar, N. and Mote, B. M., 2024, "Integration of Remote Sensing and Meteorological Data for Rapid Sugarcane Yield Estimation Using Machine Learning", *Journal of the Indian Society of Remote Sensing*, 1-16. https://doi.org/10.1007/s12524-024-02066-y.

Zhu, L., Liu, X., Wang, Z. and Tian, L., 2023, "High-precision sugarcane yield prediction by integrating 10-m Sentinel-1 VOD and Sentinel-2 GRVI indexes", *European Journal of Agronomy*, **149**, 126889. https://doi.org/10.1016/j.eja.2023.126889.

Zulu, N. S., Sibanda, M. and Tlali, B. S., 2019, "Factors affecting sugarcane production by small-scale growers in Ndwedwe Local Municipality, South Africa", *Agriculture*, **9**, 8, 170. https://doi.org/10.3390/agriculture9080170.