Forecasting the sugarcane yields based on meteorological data through ensemble learning

Sumit Kumar[1], Millie Pant[1,2], and Atulya Nagar[3]
[1]Department of Applied Mathematics and Scientific Computing, Indian Institute of Technology Roorkee, Uttarakhand, India
[2]Mehta Family School for Data Science and Artificial Intelligence, Indian Institute of Technology Roorkee, Uttarakhand, India
[3]School of Mathematics, Computer Science, and Engineering, Liverpool Hope University, England, United Kingdom
skumar1@as.iitr.ac.in, pant.milli@as.iitr.ac.in, atulya. nagar@hope.ac.uk

ABSTRACT Accurate prediction of sugarcane yields is crucial, particularly for developing countries like India, due to its economic significance and impact on farmers' livelihood. Unexpected fluctuations in production can affect farmers' income and the stability of the market, emphasizing the necessity of accurate forecasting to avoid adverse economic consequences. This research aims to enhance the precision of sugarcane yield prediction in India by developing a stacking ensemble learning model. The developed model incorporates the least absolute shrink and selection operator (LASSO), artificial neural network (ANN), and random forest (RF) as base models alongside random forest regression (RFR) and Ridge regression (RR) as meta-models and utilizes principal component analysis (PCA) and SHAPLEY values to reduce dimensions and explore feature correlations within the dataset. The data used in the study is obtained from ICRISAT and NASA databases covering 40 years (1982 to 2021) of meteorological information and sugarcane yield data across 24 districts of Uttar Pradesh, India. The model's generalizability is further improved through 5-fold cross-validation. For comparison, the vector autoregression moving average (VARMA) statistical method was also applied and it was observed that the outcome was not desirable. The findings indicate competence of stacking ensemble model over individual models like LASSO, ANN, KNN, RF, and SVR.

INDEX TERMS Sugarcane Forecasting, Ensemble learning, Machine learning, Agriculture, Meteorological Data.

ABBREVIATIONS

VARMA, Vector Autoregression Moving Average; ML, Machine learning; EL, Ensemble learning; RF, Random Forest; ANNs, Artificial Neural Networks; KNN, k-nearest neighbour; RF, Random forest; LASSO, The least absolute shrink and selection operator regression; RR, Ridge regression, RFR, Random Forest Regressor; PCA, Principal Component Analysis; R^2, Coefficient of determination; RMSE, Root Mean Square Error; MAE, Mean Absolute Error; MSE, Mean Squared Error; MAPE, Mean Absolute Percentage Error; GDP, Gross Domestic Product; CCF, Comprehensive Climate Factor; UP, Uttar Pradesh; ICRISAT, International Crops Research Institute for the Semi-Arid Tropics; NASA, National Aeronautics and Space Administration; CV, Cross Validation; kPa, Kilo Pascal; m/s, Meter/Second; wfv, Water Fraction by Volume.

I. INTRODUCTION

India is one of the world's largest producers of sugarcane after Brazil [1]. Sugarcane is a crucial cash crop that plays a vital role in the country's agricultural economy. The cultivation of sugarcane serves as a dual powerhouse, boosting the national Gross Domestic Product and serving as the lifeblood for millions of farmers and industry laborers. This versatile plant produces various sugars and holds potential as a renewable energy source, capable of generating electricity, various bio-products, and biofuels. Due to its multifaceted nature, it is also known as "Wonder Cane" [2].

Accurate prediction of sugarcane yield holds significant importance in India's agricultural as well as economic sector [3]. The correlation between crop yield and market prices is deeply interconnected. Unexpected decline in production can lead to reduced surplus available for sale and diminished earnings for farmers, resulting in price hikes. Conversely, excessive production can trigger price drops, adversely affecting farmers' incomes [4]. The impact of price fluctuations of sugarcane plays a pivotal role in determining inflation rates, wages, salaries, and various economic strategies. The cost of raw materials of sugarcane for companies and their competitive positions in the market are directly impacted by production levels [5].

Various factors, from the unpredictable nature of weather patterns to developments in agricultural technology, impact the production of crops [6], [7]. Factors such as traits inherent in seeds, methods to combat pests, improved management techniques, and adjustments in the application of fertilizers play important roles in managing crop yield.

Other factors that influence the yield of sugarcane include diverse land types, limited resource availability, and fluctuating weather patterns [8], [9], [10].

Scientists across the globe are focusing on methods that can predict crop yields as accurately and as consistently as possible [11], [12]. Initially, sugarcane yield forecasting focused on statistical models such as ARIMA, SARIMA, and Exponential Smoothing [13], [14]. However, as research progressed, attention shifted towards exploring more advanced techniques, including regression analysis [15], machine learning algorithms[16], and crop simulation models. Nowadays, the integration of machine learning and artificial intelligence has encouraged their extensive utilization in agriculture, especially for forecasting crop yields [17] [18]. In conventional machine learning methods, predicting agricultural output from meteorological data typically involves utilizing a single model, like linear regression or decision trees [19]. While these models can provide reasonable predictions under certain conditions, they may not fully capture the complexities and nuances present in the relationships between meteorological factors and crop yields [20], [21], [22].

In contrast, Ensemble Learning (EL) offers a more efficient solution to enhance predictive accuracy by combining the strengths of multiple models [23]. EL combines several independent models, also referred to as weak learners or base models, to generate a prediction model that is more reliable and accurate [24]. The core principle behind EL lies in the belief that the collective wisdom of several models can outperform any individual model since the errors made by individual models often offset or diminish within the ensemble [25]. By obtaining reliable predictions, farmers can proactively prepare and implement suitable strategies to tackle future challenges in agriculture [26] [27]. Table I provides the prior works in the domain of yield forecasting.

The present research aims to exploit the potential of EL in developing a robust model for forecasting sugarcane yields in India.

The developed approach integrates five distinct machine learning models, from which the top three performers are chosen as base models for stacking, and then RR and RFR are used as meta models. Additionally, feature analysis is conducted to understand the relative importance of meteorological variables in predicting sugarcane yields. Furthermore, dimensionality reduction techniques such as PCA and Shapley values have been incorporated to improve computational efficiency and streamline model performance. For improving the generativity of the model 5-fold cross-validation technique has been used. Statistical technique VARMA has also been used for forecasting and comparing the results with EL techniques.

The rest of this paper is organized as follows: Section II describes the materials and methods. Section III discusses the experimental results and comparison. Section IV shows the analysis and discussion. Finally, the conclusion is presented in Section V.

II. MATERIALS AND METHODS

In this section, study area is described along with data collection process and the methodology developed for the proposed stacking EL model.

A. STUDY AREA

Study area taken is of Uttar Pradesh (UP), situated in the northern part of India with an area of 243286 km^2 with coordinates 26.85°N 80.91°E. UP covers 48% of the nation's sugarcane cultivation area and contributes 50% of the total production of India, making it the leading sugarcane-producing state in India [28] offering ideal growing conditions for sugarcane like sufficient sunshine, suitable temperature, simultaneous rainfall and heat, and moderate precipitation.

Sugarcane and climatic data were collected for twenty-four districts of UP, known for their significant sugarcane production. This data covers a period between 1982 to 2021. Figure 1 shows the geographical location of the area considered in this study [29].

B. DATA COLLECTION

1) YIELD DATA

Datasets with respect to yield were obtained from International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) databases from the prominent website https://www.icrisat.org//?s=sugarcane [30]. A comprehensive collection of 24 district-level datasets pertaining to sugarcane production and planting areas was compiled for around 40 years, from 1982 through 2021. Equation (1) given below is used for calculating sugarcane yield:

$$Y_i = \frac{P_i}{A_i} \tag{1}$$

where Y_i is the yield of sugarcane; P_i is the district-level sugarcane production and A_i is the planted area of sugarcane at the district level. Figure 2 displays a portion of data concerning the top five sugarcane-producing districts. It includes information on area, production, and yield for each district.

2) METEOROLOGICAL DATA

The meteorological data sets were collected from NASA https://power.larc.nasa.gov/data-access-viewer/. [31]. The geographic spread of these meteorological observation stations is illustrated in Figure 2. Table II provides the meteorological variables that are considered for the present analysis.

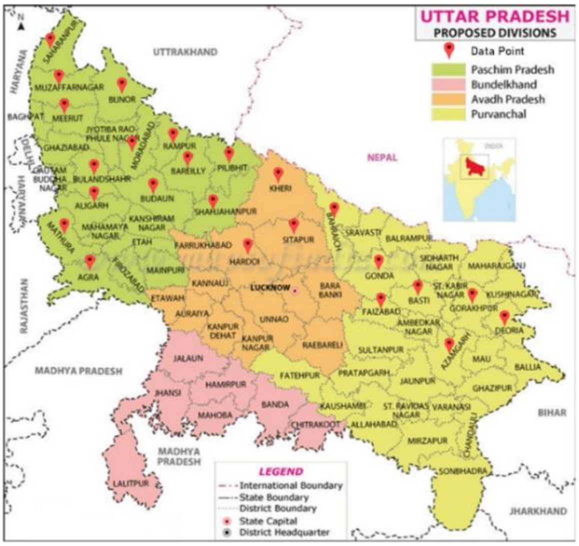


Figure 1. Geographical map of UP [29]

TABLE I
PREVIOUS WORKS DONE ON ESTIMATING THE YIELD FOR DIFFERENT CROPS

Year	Objective	Method(s) used	Data	Study Area
2011[32]	To forecast the sugarcane area, production, and yield	Univariate ARIMA models	Sugarcane production data from 1950 to 2007	Tamil Nadu, India
2011[33]	To forecast the sugarcane area, production, and productivity	Univariate ARIMA models	Sugarcane production data from 1950 to 2012	Uttar Pradesh and Tamil Nadu, India
2019[21]	EL for a cropping systems simulator (APSIM) to forecast the production of maize	LASSO Regression, RR, RF, Extreme Gradient Boosting, and their ensembles	Maize Management, cultivar, and environmental factors taken from 1983 to 2016	Kelley and Nashua, United States of America
2020[6]	To explore the influence of historical climate fluctuations on the yields of five major cereal crops	DSSAT crop model	Five main cereal crops: barley, maize, millet, sorghum, and wheat yield data from 1979 to 2014:	Ethiopia
2021[34]	To check the impact of climate variability on yield variation	CCF approach determines the sensitivity	Socio-economic and climate data for Rice, Maize, and Wheat from 1981 to 2015	North and South regions of China
2021[35]	To investigate the best ARIMA model for sugarcane yield forecasting	Different variants of ARIMA model	Sugarcane production data from 1950 to 2018	Uttar Pradesh, India
2021[36]	To forecast sugarcane yield using Various weighted and un-weighted weather indices	Statistical model using regression techniques	35 years of historical weather and sugarcane yield data from 1981 to 2015	Uttar Pradesh, India
2021[9]	To investigate varying trends in ten climate variable and their impact on the wheat yield in India	Regression modelling and correlation analysis	Wheat yield data from 1986 to 2015	India
2022[37]	To evaluate the effects of severe climatic indices and the mean climate on the actual and simulated yields of soybeans	DSSAT-CROPGRO soybean model	Weather and Soybeans yield data from 1981 to 2017	Northeast China
2023[38]	To develop climate-based soybean yield prediction models using machine learning	ML algorithms: MLR, MLP, SVM, RF, XG Boosting	Soybean yield data from 2002 to 2021	Mato Grosso do Sul, Brazil

2023[39]	To estimate the crop yield using ML techniques	ML algorithms: RF, SVR, LSTM, Gradient Descent, LASSO	Yield data of Mustard Wheat, Barley, Bajra, Jowar, Onion, and Maize crops from 33 districts from 1997to 2019	Rajasthan, India
2023[24]	To develop a soybean forecasting model based on the stacking EL framework using weather parameters	KNN, RF and SVR used as the base-models and RR as meta model to establish stacking EL framework	Meteorological records along with Soybean yield data spanning a period of 34 years.	China

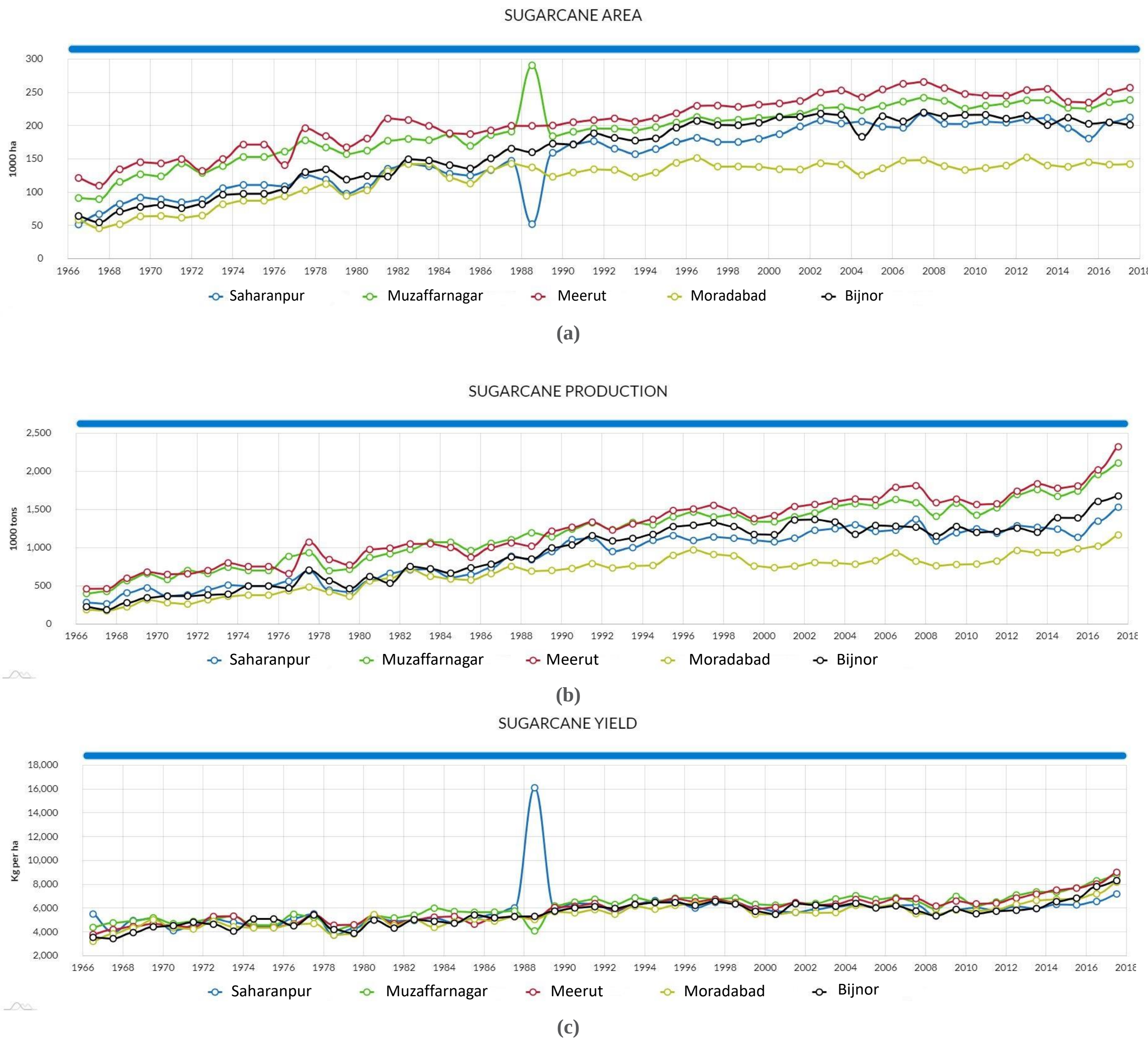


FIGURE 2. (a) **Area, (b) Production, and (c) Yield of five districts of UP**

TABLE II

PARAMETERS CONSIDERED IN THIS STUDY

Sugarcane Parameters		
Features	**Description**	**Units**
Area	Sugarcane Cultivation area	1000 Hectares
Production	Sugarcane production	1000 Ton

Yield (Productivity)	Production of sugarcane per unit area	Kg per Hectare
Meteorological Parameters		
Features	**Description**	**Units**
PS	Surface Pressure	kPa
T2M	Temperature at 2 Meters	°C
T2M_MAX	Temperature at 2 Meters Maximum	°C
T2M_MIN	Temperature at 2 Meters Minimum	°C
PRECTOTCORR	Precipitation Corrected	mm/day
QV2M	Specific Humidity at 2 Meters: It is the water vapor weight per unit of air.	grams of water vapor per kilogram of air (g/kg)
RH2M	Relative Humidity at 2 Meters measures the moisture in the air compared to its maximum capacity at a given temperature.	Percentage (%)
WS2M	Wind Speed at 2 Meters	m/s
WS2M_MAX	Wind Speed at 2 Meters Maximum	m/s
WS2M_MIN	Wind Speed at 2 Meters Minimum	m/s
GWETTOP	Surface Soil Wetness	wfv or m^3m^{-3}
GWETPROF	Profile Soil Moisture	wfv or m^3m^{-3}
GWETROOT	Root Zone Soil Wetness	wfv or m^3m^{-3}

C. METHODOLOGY

In order to achieve the set research objectives, several steps were undertaken. First, the meteorological data related to sugarcane yield at the district level were obtained. Next, the records of sugarcane meteorological factors were averaged annually. PCA and SHAPLEY were used to reduce dimensionality and analyse feature correlations. After applying VARMA Statistical method, five machine learning models were employed, and the best three were selected for further analysis using MAPE, RMSE, MAE, and R^2 values. Finally, a stacking model was developed, utilizing the prediction strengths of various machine learning algorithms to create an EL framework. We also incorporated a 5-fold cross-validation approach to enhance the model's robustness and generalizability. As shown in Figure 3, the research framework comprises four main components.

1) DATA PROCESSING

Effective data pre-processing is fundamental to ensure the accuracy and robustness of sugarcane yield forecasting models. This study involves the processing of yield and meteorological data from 24 sugarcane-producing districts in Uttar Pradesh over a 40-year period.

The dataset consists of 15 input variables and one output variable, organized into a time series format with 24 rows and 16 columns for each year. Each variable is represented numerically. To prepare the data for analysis, a thorough examination was conducted to identify and address missing values, outliers, and inconsistencies. The processed data for Meerut district is shown in the Figure 4.

The meteorological data obtained included a broad range of features, including three related to soil properties, eleven related to wind and pressure, four related to humidity and precipitation, twelve in solar fluxes, and eight related to temperature. However, after consulting the farmers and plant protection experts, the dataset was refined to include 13 key features: three related to soil properties, four related to wind and surface, three related to humidity and precipitation, and three related to temperature. The original meteorological data was recorded on a monthly basis. To facilitate analysis, the data was aggregated to an annual level by calculating the average of the monthly values for each year. Finally, a dataset containing 13 features for each of the 24 districts, spanning a 40-year period was obtained.

2) FEATURE REDUCTION AND FEATURE IMPORTANCE

PCA and SHAPLEY were incorporated in the study to optimize the dataset by reducing the dimensionality of the features. This analysis helped in identifying the correlation between meteorological factors and sugarcane yield.

3) DEVELOPING THE STACKING MODEL

Stacking EL model was developed by merging individual base models' unique strengths and traits to enhance the precision and robustness in forecasting sugarcane yields. Construction of the model is described in detail in the following subsections:

a. CONSTRUCTION OF THE STACKING ENSEMBLE LEARNING MODEL

Stacking, also known as Stacked Generalization, is a combination technique that merges multiple base models through meta-models. Unlike traditional methods such as bagging and boosting that aggregate outcome from base learners, stacking constructs a multi-layered learning system by integrating diverse base learners for model fusion. [40]. The different Base and Meta-models used in this study are described in Table III.

To enhance the model's stability and generalizability, we used the 5-fold cross-validation approach. This approach divides the dataset into five equal-sized folds, where the model is trained on four folds and validated on the remaining fold. This process is repeated five times, ensuring that each fold serves as the validation set once. By averaging the performance metrics across all folds, we obtain a more reliable and unbiased estimate of the model's predictive accuracy. This approach mitigates the risk of overfitting and ensures that the model's performance is not dependent on a particular data split, thereby improving its generalizability to unseen data.

For tuning the hyperparameters of the employed models, Grid Search CV is utilized to find the optimal hyperparameter settings. This technique systematically explores the hyperparameter space and selects the combination that results in the best model performance. Table IV displays the best-tuned parameters of all employed methods.

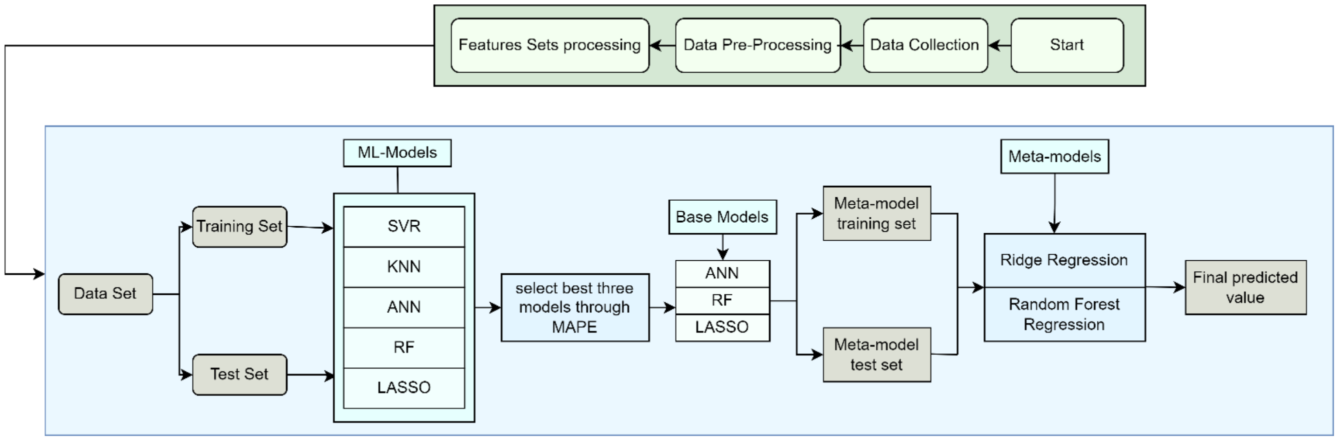


Figure 3. Sugarcane Yield Prediction Framework's Technological Roadmap

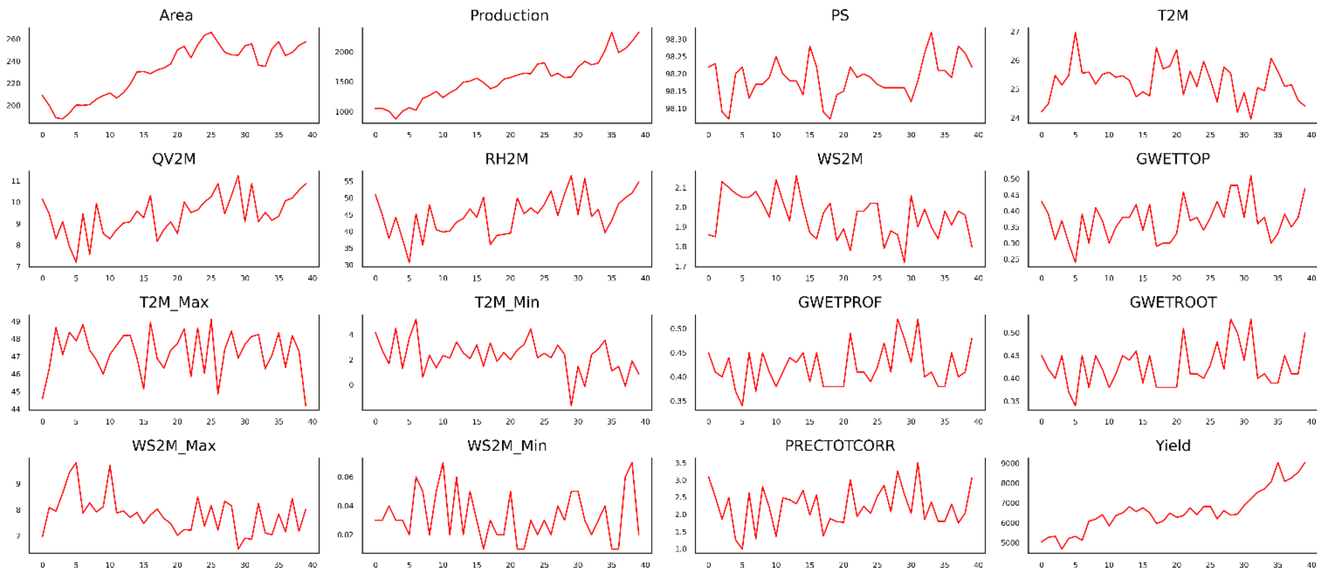


FIGURE 4. Dataset for Meerut District

TABLE III

METHOD & ALGORITHMS USED IN THIS STUDY

Algorithm	Definition	Applicability	Key Parameters	Advantages
VARMA [41]	It is a statistical method that combine vector autoregression and moving average	Dealing with interrelated time series data,	Lag order, error term, coefficient matrices	Captures complex dynamic relationship

		multivariate time series data		in multivariate time series data
SVR[42]	It a regression-based supervised ML algorithm that finds a hyperplane to best fit data while minimizing prediction errors.	When the data has non-linear patterns and complex relationships between variables.	Kernel type, kernel parameters (e.g., gamma, degree), regularization (C), epsilon (ε)	Effective in handling non-linear relationships. It can be adapted to various kernel functions.
KNN[43]	KNN, a versatile lazy learning algorithm, handles classification and regression tasks by predicting based on the majority class or the mean of the k nearest neighbors.	When data exhibits local patterns and neighbors' behavior is relevant.	Number of neighbors (k), distance metric, weight function (uniform or distance-based), etc.	Simple and easy to understand. Non-parametric and adaptable to data changes.
RF[44]	RF, an EL algorithm, combines multiple decision trees for predictions in classification and regression tasks.	When dealing with noisy data and complex relationships between variables.	Number of trees, tree depth, minimum samples per leaf, maximum features, bootstrap samples, etc	High accuracy and robustness. Handles feature importance and variable interactions well.
LASSO[45]	LASSO, L1 regularization [46], is a linear regression technique that adds a penalty term to the linear regression objective function. It can be used for feature selection and regularization.	Dealing with high-dimensional data and multi-collinearity issues.	Regularization strength (λ), alpha (α)	Feature selection by shrinking some coefficients to zero. Addresses multi-collinearity effectively
ANN[47]	It is a deep-learning algorithm inspired by the human brain that comprises interconnected nodes (neurons) arranged in layers.	For capturing complex temporal dependencies and patterns in data.	Number of layers, number of neurons per layer, activation functions, learning rate, batch size, etc.	Capable of handling large and complex datasets. Suitable for capturing non-linear relationships.
RR[48]	RR, identified as L2 regularization, integrates a penalty term into the linear regression objective function to mitigate overfitting in linear regression techniques.	Dealing with multi-collinearity and overfitting issues.	Regularization strength (λ), alpha (α)	Addresses multi-collinearity by shrinking coefficients. Stabilizes parameter estimates.

D. Assessment metric

The predictive method's accuracy was evaluated using four evaluation metrics: R^2, RMSE, MAE, MSE, and MAPE. R^2 assesses the adequacy of the prediction model with a value closer to 1, indicating a superior fit of the regression equation. RMSE quantifies the variance between predicted and observed values, and a value closer to 0 signifies a more accurate prediction model. MAE provides insight into the actual error of the estimated values. A value closer to 0 implies a more precise model.

TABLE IV

VALUES OF PARAMETERS USED AFTER FINE TUNING

Methods	Parameters used
VARMA	Maxlags = 1, order = (1,1)
SVR	Kernel = Linear, C= 10, Degree = 2, Epsilon = 0.01
RF	Min-samples-leaf= 2, n-estimators = 200
ANN	Hidden layer sizes= (100, 50), Learning rate init=0.1, Max iteration=1500, Optimizer= "lbfgs" stands for Limited-memory Broyden–Fletcher–Goldfarb–Shanno.
LASSO	Alpha = 1 (constant, controlling regularization strength)
KNN	n neighbors = 5, power parameter (p) =2, Weights = uniform

A lower MSE value indicates that the predicted values are closer to the actual values, signifying a more accurate model. MAPE calculates the average relative error between predicted and measured values, directly revealing the variance between predicted and actual results. A value closer to 0% indicates higher accuracy.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{6}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{9}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100\% \tag{10}$$

E. SYSTEM CONFIGURATION

The experiments were executed on Spyder 4.1.5 Integrated Development Environment (IDE) with Python 3.8 through an Anaconda distribution on an Intel Xeon W-1290P CPU @ 3.70GHz dual-processor system with 64 GB RAM, an Nvidia Quadro P1000 GPU, and a 64-bit Windows 11 Pro Operating System. All algorithms were implemented using Python, employing libraries such as Numpy, Pandas, Matplotlib, Sklearn, Seaborn, and Shap.

III. RESULTS

A. VARIABLE SCREENING AND DIMENSIONAL PROCESSING

This study chose a feature set comprising 15 variables and one target variable, encompassing 13 annual meteorological indicators. The results for Meerut district are being presented herein. The feature correlation matrix provides an extensive overview of the of the interrelationships among all variables considered in the analysis. It allows us to assess the strength and direction of correlations between each pair of variables, helping to identify patterns and associations within

Figure 5 displays the correlation matrix of 15 features, particularly emphasizing the variable area. From the figure, we observe that among the 15 features, 9 exhibit positive correlations, while 5 are negatively correlated with area. Production is highly correlated with area, while T2M_max is observed to have minimal impact.

In order to address potential issues of duplicated information and limited generalizability in the feature set, PCA [49] is employed. PCA allows the transformation of a set of correlated variables into a set of uncorrelated variables while retaining the essential information from the original feature set.

PCA aims to capture the maximum information in the first principal component, the maximum remaining information in the second component, and so on, as depicted in the Scree plot figure 6. The Elbow Point in the Scree Plot is where the cumulative explained variance levels off or plateaus. It indicates that additional principal components contribute less and less to the overall variance explained. The Elbow Point occurs at PC 8 in this figure, suggesting that the first eight principal components are crucial in capturing most of the dataset's variability.

In the context of feature importance, the Shapley value assesses the contribution of each feature to the prediction by considering all possible permutations of features [50]. Figure 7(a) indicates that PS is the most important feature, followed by Qv2m, among all 15 features that are correlated to the yield.

The bar chart in figure 7(b) represents the Shapley values in descending order of all features, where higher values indicate greater importance and lower values indicate lesser importance of the feature in predicting the output variable yield.

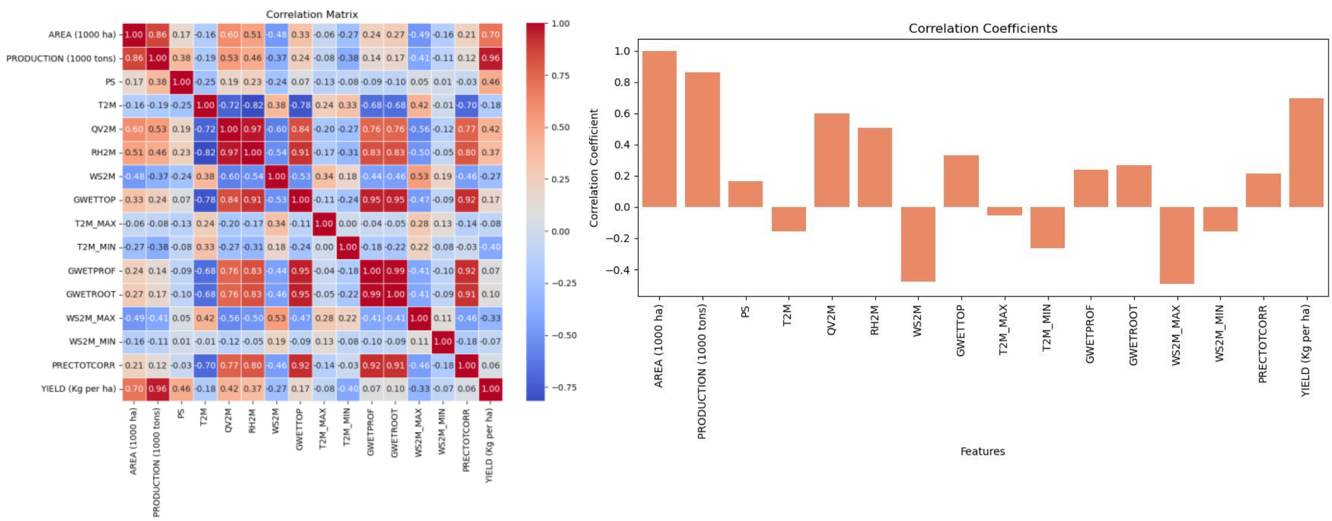


FIGURE 5. Feature Correlation matrix

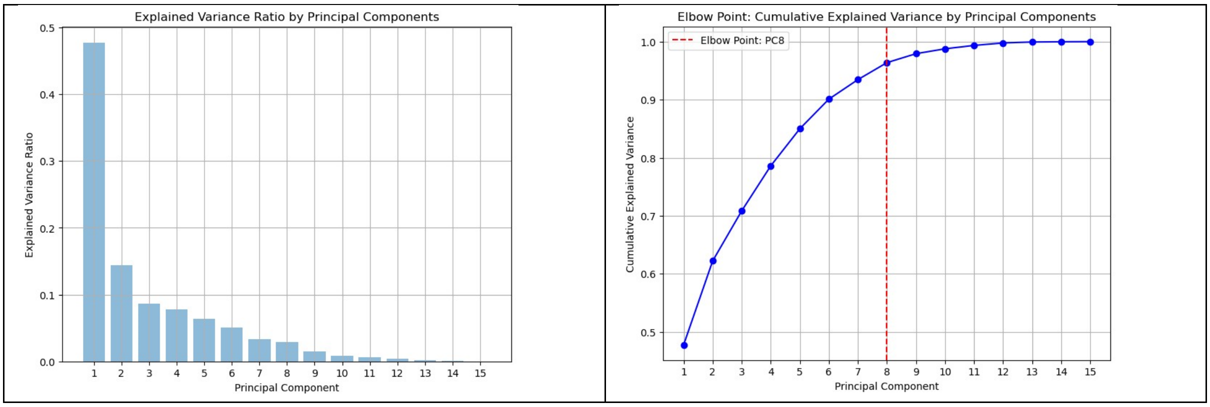
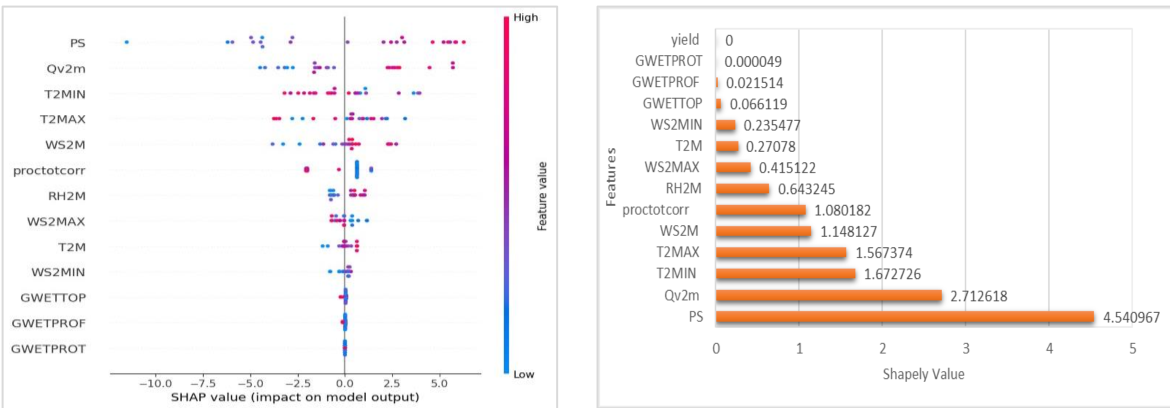


FIGURE 6. Explained variance ratio & Cumulative explained variance by PCA



(a) (b)

FIGURE 7. Feature importance obtained through Shapley value

B. Base Model Selection

In the initial phase, five machine learning models (RF, ANN, KNN, SVR, and LASSO) were utilized to forecast sugarcane yield. To ensure robustness, we partitioned our dataset with 80% for training and the remaining 20% for testing purposes. Here, the results for Meerut district are shown, offering valuable insights into model performance by comparing predicted yields with actual values on the test dataset. Among the models, SVR and KNN exhibited deviations from the actual values, indicating lower accuracy. In contrast, the RF, ANN, and LASSO models demonstrated closer alignment with actual values, suggesting superior performance. Moreover, we conducted an assessment of feature importance by generating column plots for each machine learning method, shown in figure 8. This facilitated the identification of the most influential features in sugarcane yield prediction for each respective model, contributing to a deeper understanding of the predictive dynamics involved.

After the analysis, we computed the evaluation metrics for all five ML methods on the test dataset, which were used to forecast the sugarcane yield. The results are summarized in Table V.

TABLE V
EVALUATION METRICS

Methods	MAE	MSE	RMSE	MAPE	R^2
VARMA	778.62	768545.25	876.66	8.87%	0.28
SVR	429.31	308126.45	555.09	6.48%	0.50
KNN	405.488	204871.46	483.47	6.42%	0.665
RF	**125.65**	**24864.71**	**157.68**	**1.92%**	**0.96**
ANN	**290.25**	**135923.74**	**368.68**	**4.38%**	**0.78**
LASSO	**93.91**	**9916.37**	**99.58**	**1.47%**	**0.98**

The superior performance of LASSO in comparison to other models is clearly visible through Table V for every evaluation metrics taken in this study and was it was therefore selected as the base model. Conversely, VARMA, SVR and KNN models exhibited notable underperformance, leading to their exclusion as base models for the stacking ensemble framework. RF and ANN did not perform as efficiently as LASSO but were around 50% better than SVR and KNN and were hence taken as base models along with LASSO.

FIGURE 8. Prediction of yield over time through ML Models

C. Yield forecast

Within the stacking EL framework, RR and RFR are utilized as meta-models. Figure 9 illustrates the stacking model's performance compared to individual single models. The ensemble models utilizing the stacking EL approach showed closer alignment with actual yield values, indicating superior performance and predictive accuracy in contrast to single models. A significant improvement in closeness to actual yield values was observed when RFR was utilized as the meta-model in the stacking framework, as opposed to RR.

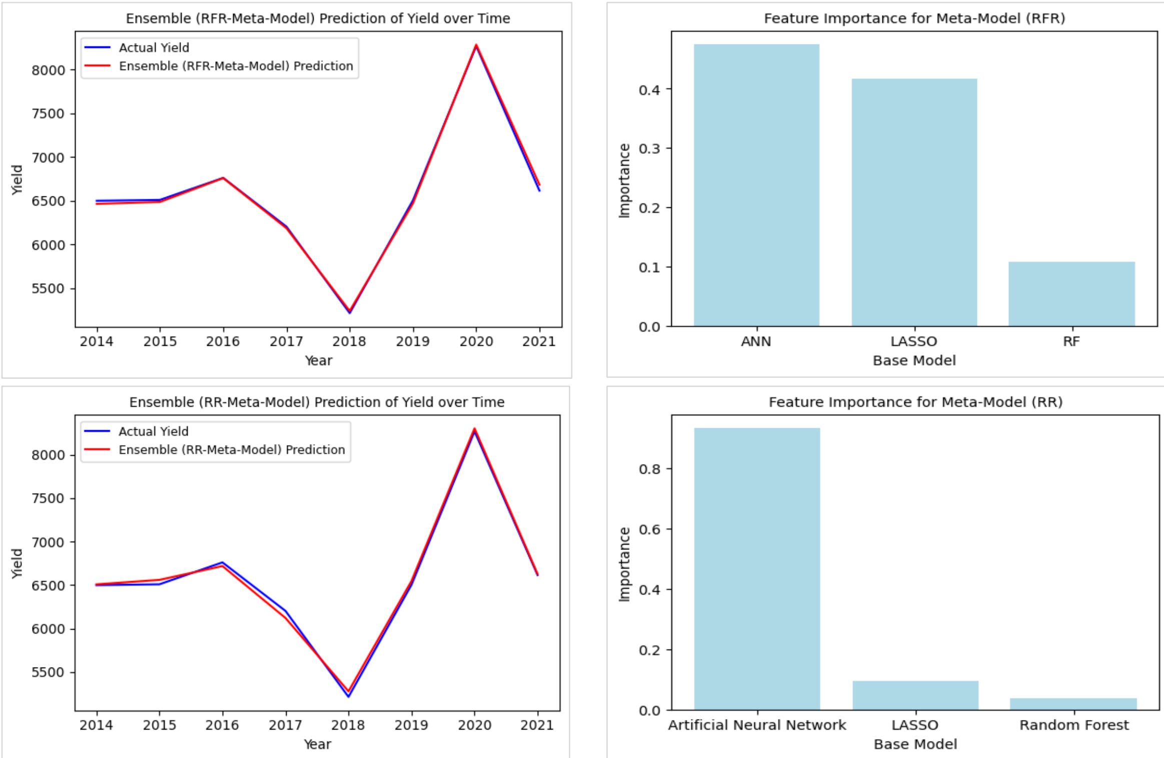


FIGURE 9. Prediction of yield over time through meta-models as RFR & RR

Table VI provides a comprehensive comparison of evaluation metrics between the stacking model and three base models. The stacking model always outperformed the individual single models. Stacking done through RFR meta-model performed better in comparison to stacking done through RR as meta model in terms of lower MAPE (0.43%), MAE (27.837), MSE (1106.698), RMSE (33.267) and R^2 (0.998).

TABLE VI
EVALUATION METRICS, INCLUDING STACKING

Index	ANN	RF	LASSO	Stacking (RR)	Stacking (RFR)
MAE	290.25	125.65	93.91	42.006	**27.837**
MSE	135923	24864.71	9916.37	2281.71	**1106.698**
RMSE	368.67	157.68	99.58	47.767	**33.267**
MAPE	4.38%	1.92%	1.47%	0.67%	**0.43%**
R^2	0.778	0.959	0.98	0.996	**0.998**

Figure 10 shows the results of yield forecasting using the stacking EL model, with RFR utilized as the meta-model, for the next ten years across five districts: Meerut, Moradabad, Muzaffarnagar, Bijnor, and Saharanpur. The dotted line represents the actual yield, while the asterisk line indicates the forecasted yield for future years. The graph illustrates an upward trajectory in yield over the forthcoming years.

IV. Discussion

This study addresses the need for improved sugarcane yield forecasting by utilizing meteorological data and introducing a stacking EL approach to overcome the limitations of traditional forecasting methods.

The model is designed to enhance predictive accuracy and flexibility, incorporating Principal Component Analysis (PCA) and SHAPLEY values for a better understanding of feature correlations. Accurate yield prediction is crucial for effective agricultural management, providing stakeholders with critical insights to support sustainable practices, optimize resource use, and stabilize the agricultural economy.

Previous studies on sugarcane yield forecasting have utilized simpler statistical models such as multiple linear regression and simple regression, and ARIMA, mainly considering the environmental variables such as Tmax, Tmin, Rainfall and humidity [51]. While these studies covered several years of data, they were unable to capture complex relationships between various meteorological factors. More recent work has used various ML algorithms such as RF, SVR, and ANN, while these methods improved the accuracy of prediction but did not employ ensemble approaches, limiting the full potential of combining model strengths.

The present study improves on previous work by using 13 meteorological variables related to soil, wind, humidity, and temperature, along with area, production and yield data of sugarcane. We also applied VARMA method for multivariate time series forecasting of sugarcane yield, but the results were not satisfactory. To improve the accuracy, we implemented a stacking EL approach, which provides stronger generalization and better captures complex relationships. This method significantly enhanced the accuracy and reliability of sugarcane yield prediction, outperforming traditional models in both precision and scalability. Table VII provides a comprehensive comparison of last five years of the research in the field of sugarcane yield forecasting in India, highlights the differences in methodologies and meteorological factors used in our study and the existing body of work.

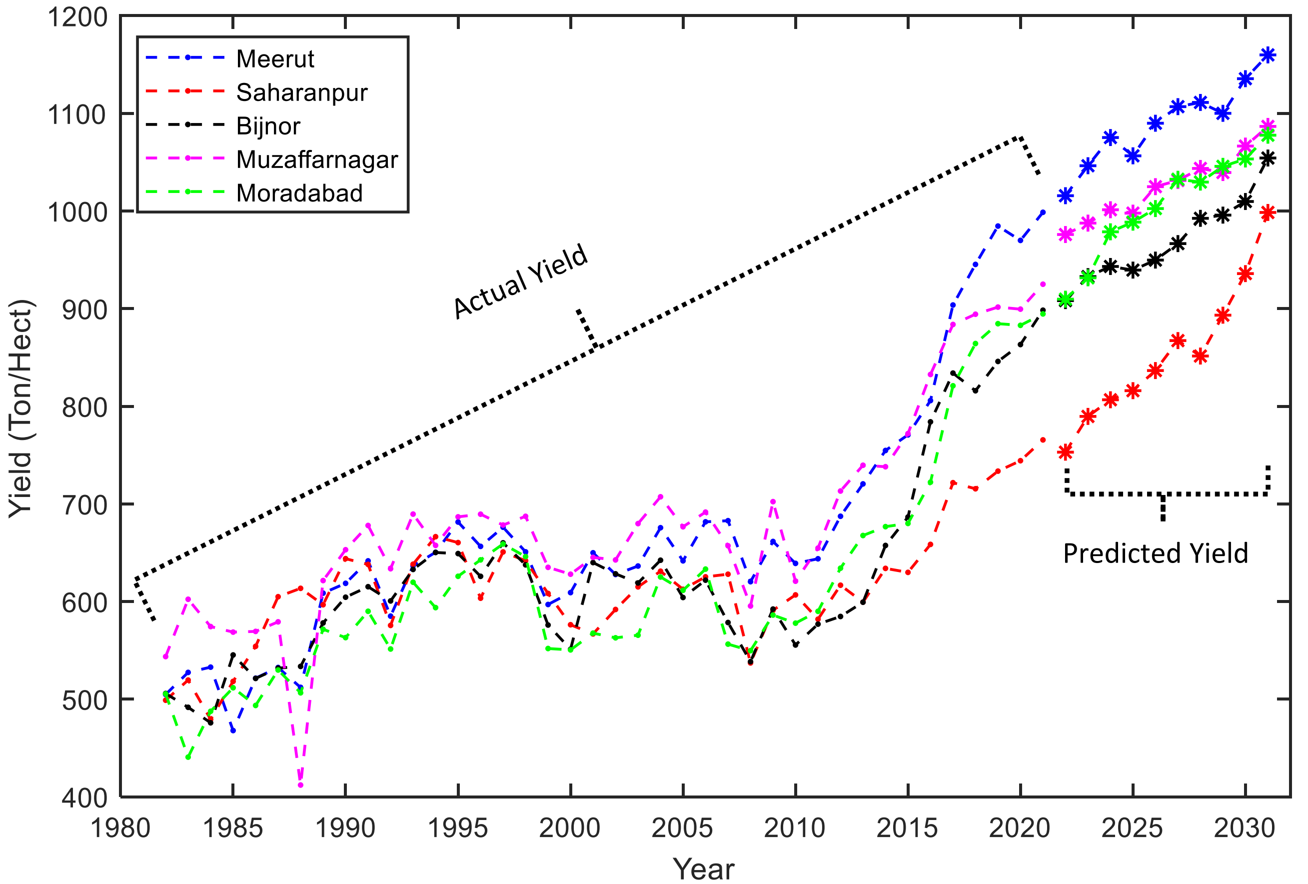


FIGURE 10. **Predicted yield when meta-model as RFR**

TABLE VII PREVIOUS WORK IN SUGARCANE FORECASTING

Year	No of Meteorological Parameters					Methods			Data	Study Area	
	Temperature	Humidity	Wind/ Pressure	Soil Properties	Total	Statistical	ML	Advanced		No. of Regions	State
2020 [52]	×	×	×	×	×	ARIMA	×	×	50 years (1967-2016)	2	Haryana
2021 [53]	×	×	×	×	×	ARIMA	×	×	69 years (1950-2018)	6	Andhra Pradesh, Uttar Pradesh, Karnataka, Tamil Nadu, Maharashtra
2021 [54]	2	3	×	×	4	Regression Techniques	×	×	35 years (1981-2015)	1	Uttar Pradesh
2022 [55]	2	1	×	×	3	ARIMA, ARIMAX	×	×	40 years (1979-2018)	3	Haryana
2023 [56]	2	4	×	×	6	Discriminant Function Analysis	×	×	57 years (1960-2016)	1	Tamil Nadu
2024 [57]	2	3	×	×	5	×	RF, SVM, SMLR, ANN	×	24 years (1997-2020)	10	Karnataka
Present Study	3	3	4	3	13	VARMA	SVR, ANN, KNN, RF, LASSO	Stacking Ensemble Learning	40 years (1982-20021)	24	Uttar Pradesh

The findings of our study are significant for precise sugarcane yield predictions, particularly with ongoing impacts of climate changes on farming. By using the developed model in real world agriculture, farmers can get valuable insights about future yields considering weather conditions. With this information, farmers and decision-makers may make better plans for maximizing their crops yield, including when to harvest, when to apply fertilizer and irrigation, and which seeds to use in response to changing weather patterns. Agribusinesses can also benefit from these predictions by optimizing their supply chains, emphasizing more efficient operations, cutting waste, and minimizing losses after harvest.

Given sugarcane's pivotal role in India's economy and the livelihoods of millions of farmers and laborers, accurate forecasting is essential for ensuring food security, stabilizing markets, and guiding policy decisions. Considering the significance of sugarcane to the Indian economy and the livelihoods of millions of people, reliable forecasting is essential for maintaining food security, stabilizing market prices, and assisting policymakers in formulating measures that better assist the agriculture sector.

V. Conclusion

Accurate prediction of sugarcane yield is crucial for effective agricultural management and early warning systems. We utilized VARMA method to forecast the sugarcane yield by incorporating this multivariate dataset but the results were not satisfactory. So, we developed a stacking EL model.

The model developed for this study comprised ANN, RF, and LASSO as base models and RFR as the meta-model and PCA aids in reducing dimensionality while preserving essential information, while SHAPLEY values offer insights into feature importance. These techniques not only enhance the predictive capability of the model but also provide interpretability. Numerical results and graphs demonstrated the efficiency of the model in terms of forecasting within primary sugarcane cultivation regions. Our approach provides a robust, data-driven framework that can be expanded to other crops and regions, contributing to more efficient and sustainable agricultural practices. Some concluding remarks that can be drawn from the current study are:

i. In comparison to utilizing VARMA and machine learning techniques LASSO, ANN, KNN, RF, and SVR in a standalone manner, the stacking model is better for all evaluation indicators, such as MAE, MSE, RMSE, MAPE, and R^2

ii. Utilizing the strengths of its base models, the stacking model notably enhances forecast accuracy, predicting the meteorological yield of sugarcane.

iii. The forecasting system developed in the study can help the farmers in planning cultivation strategies more effectively. It can also help the retailers n optimizing the supply chains, and policymakers to make informed decisions for resource allocation and agricultural sustainability.

Future research directions could explore refining the stacking EL approach by incorporating additional machine learning techniques or exploring alternative meta-models and uses of hybrid models that combine convolution layers to extract spatial features with recurrent layers. Also, one could explore the integration of real-time data sources, such as satellite imagery and IoT-based sensors, which could further enhance the model's responsiveness to changing environmental conditions. Additionally, expanding the model's application to different crops or geographic areas could provide valuable insights and contribute to a broader understanding of agricultural forecasting. Incorporating socio-economic factors—such as market trends, labor availability, and policy changes—into the forecasting model could provide a more holistic view of the factors influencing crop yields.

REFERENCES

[1] S. Gopikrishnan, G. Srivastava, and P. Priakanth, "Improving sugarcane production in saline soils with Machine Learning and the Internet of Things," *Sustain. Comput. Inform. Syst.*, vol. 35, p. 100743, Sep. 2022, doi: 10.1016/j.suscom.2022.100743.

[2] S. Kumar and M. Pant, "A Mathematical Model to Minimize the Total Cultivation Cost of Sugarcane," in *Soft Computing for Problem Solving*, M. Thakur, S. Agnihotri, B. S. Rajpurohit, M. Pant, K. Deep, and A. K. Nagar, Eds., Singapore: Springer Nature Singapore, 2023, pp. 529–542.

[3] S. R. K. Priya, R. K. Balambiga, P. Mishra, and S. S. Das, "Sugarcane yield forecast using weather based discriminant analysis," *Smart Agric. Technol.*, vol. 3, p. 100076, Feb. 2023, doi: 10.1016/j.atech.2022.100076.

[4] F. F. Bocca, L. H. A. Rodrigues, and N. A. M. Arraes, "When do I want to know and why? Different demands on sugarcane yield predictions," *Agric. Syst.*, vol. 135, pp. 48–56, May 2015, doi: 10.1016/j.agsy.2014.11.008.

[5] H. K. Sharma, K. Kumari, and S. Kar, "Forecasting Sugarcane Yield of India based on rough set combination approach," *Decis. Mak. Appl. Manag. Eng.*, vol. 4, no. 2, Art. no. 2, Jun. 2021, doi: 10.31181/dmame210402163s.

[6] M. Yang *et al.*, "The role of climate in the trend and variability of Ethiopia's cereal crop yields," *Sci. Total Environ.*, vol. 723, p. 137893, Jun. 2020, doi: 10.1016/j.scitotenv.2020.137893.

[7] A. K. Verma, P. K. Garg, K. S. H. Prasad, and V. K. Dadhwal, "Variety-specific sugarcane yield simulations and climate change impacts on sugarcane yield using DSSAT-CSM-CANEGRO model," Agric. Water *Manag.*, vol. 275, p. 108034, Jan. 2023, doi: 10.1016/j.agwat.2022.108034.

[8] J. Jägermeyr *et al.*, "Climate impacts on global agriculture emerge earlier in new generation of climate and crop models," *Nat. Food*, vol. 2, no. 11, Art. no. 11, Nov. 2021, doi: 10.1038/s43016-021-00400-y.

[9] A. Madhukar, K. Dashora, and V. Kumar, "Climate Trends in Temperature and Water Variables during Wheat Growing Season and Impact on Yield," *Environ. Process.*, vol. 8, no. 3, pp. 1047–1072, Sep. 2021, doi: 10.1007/s40710-021-00526-y.

[10] "Climate-Driven Crop Yield and Yield Variability and Climate Change Impacts on the U.S. Great Plains Agricultural Production | Scientific Reports." Accessed: Jul. 16, 2023. [Online]. Available: https://www.nature.com/articles/s41598-018-21848-2

[11] M. Christina *et al.*, "Impact of climate variability and extreme rainfall events on sugarcane yield gap in a tropical Island," *Field Crops Res.*, vol. 274, p. 108326, Dec. 2021, doi: 10.1016/j.fcr.2021.108326.

[12] A. A. Chandio, D. Ozdemir, K. K. Gokmenoglu, M. Usman, and Y. Jiang, "Digital agriculture for sustainable development in China: The promise of computerization," *Technol. Soc.*, vol. 76, p. 102479, Mar. 2024, doi: 10.1016/j.techsoc.2024.102479.

[13] J. Kumar, M. Devi, D. Verma, D. P. Malik, and A. Sharma, "Pre-harvest forecast of rice yield based on meteorological parameters using discriminant function analysis," *J. Agric. Food Res.*, vol. 5, p. 100194, Sep. 2021, doi: 10.1016/j.jafr.2021.100194.

[14] H. Jin, M. Li, G. Hopwood, Z. Hochman, and K. S. Bakar, "Improving early-season wheat yield forecasts driven by probabilistic seasonal climate forecasts," *Agric. For. Meteorol.*, vol. 315, p. 108832, Mar. 2022, doi: 10.1016/j.agrformet.2022.108832.

[15] S. McGree, S. Schreider, Y. Kuleshov, and B. Prakash, "On the use of mean and extreme climate indices to predict sugar yield in western Fiji," *Weather Clim. Extrem.*, vol. 29, p. 100271, Sep. 2020, doi: 10.1016/j.wace.2020.100271.

[16] L. Li *et al.*, "Crop yield forecasting and associated optimum lead time analysis based on multi-source environmental data across

China," *Agric. For. Meteorol.*, vol. 308–309, p. 108558, Oct. 2021, doi: 10.1016/j.agrformet.2021.108558.
[17] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Comput. Electron. Agric.*, vol. 177, p. 105709, Oct. 2020, doi: 10.1016/j.compag.2020.105709.
[18] "Sensors | Free Full-Text | Machine Learning in Agriculture: A Review." Accessed: Jul. 16, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/18/8/2674
[19] A. Nihar, N. R. Patel, and A. Danodia, "Machine-Learning-Based Regional Yield Forecasting for Sugarcane Crop in Uttar Pradesh, India," *J. Indian Soc. Remote Sens.*, vol. 50, no. 8, pp. 1519–1530, Aug. 2022, doi: 10.1007/s12524-022-01549-0.
[20] D. Elavarasan, D. R. Vincent, V. Sharma, A. Y. Zomaya, and K. Srinivasan, "Forecasting yield by integrating agrarian factors and machine learning models: A survey," *Comput. Electron. Agric.*, vol. 155, pp. 257–282, Dec. 2018, doi: 10.1016/j.compag.2018.10.024.
[21] M. Shahhosseini, R. A. Martinez-Feria, G. Hu, and S. V. Archontoulis, "Maize yield and nitrate loss prediction with machine learning algorithms," *Environ. Res. Lett.*, vol. 14, no. 12, p. 124026, Dec. 2019, doi: 10.1088/1748-9326/ab5268.
[22] L. Felipe Maldaner, L. de Paula Corrêdo, T. Fernanda Canata, and J. Paulo Molin, "Predicting the sugarcane yield in real-time by harvester engine parameters and machine learning approaches," *Comput. Electron. Agric.*, vol. 181, p. 105945, Feb. 2021, doi: 10.1016/j.compag.2020.105945.
[23] Y. L. Everingham, C. W. Smyth, and N. G. Inman-Bamber, "Ensemble data mining approaches to forecast regional sugarcane crop production," *Agric. For. Meteorol.*, vol. 149, no. 3, pp. 689–696, Mar. 2009, doi: 10.1016/j.agrformet.2008.10.018.
[24] Q. Li, S. Xu, J. Zhuang, J. Liu, Y. Zhou, and Z. Zhang, "Ensemble learning prediction of soybean yields in China based on meteorological data," *J. Integr. Agric.*, vol. 22, no. 6, pp. 1909–1927, Jun. 2023, doi: 10.1016/j.jia.2023.02.011.
[25] M. S. Divakar, M. S. Elayidom, and R. Rajesh, "Forecasting crop yield with deep learning based ensemble model," *Mater. Today Proc.*, vol. 58, pp. 256–259, Jan. 2022, doi: 10.1016/j.matpr.2022.02.109.
[26] S. Hao, D. Ryu, A. Western, E. Perry, H. Bogena, and H. J. H. Franssen, "Performance of a wheat yield prediction model and factors influencing the performance: A review and meta-analysis," *Agric. Syst.*, vol. 194, p. 103278, Dec. 2021, doi: 10.1016/j.agsy.2021.103278.
[27] A. C. dos S. Luciano, M. C. A. Picoli, D. G. Duft, J. V. Rocha, M. R. L. V. Leal, and G. le Maire, "Empirical model for forecasting sugarcane yield on a local scale in Brazil using Landsat imagery and random forest algorithm," *Comput. Electron. Agric.*, vol. 184, p. 106063, May 2021, doi: 10.1016/j.compag.2021.106063.
[28] Accessed: Aug. 01, 2023. [Online]. Available: https://farmer.gov.in/cropstaticssugarcane.aspx
[29] "About Us - Statistical, Government of Uttar Pradesh." Accessed: Aug. 01, 2023. [Online]. Available: https://upcane.gov.in/StaticPages/StatisticalAbout.aspx
[30] "ICRISAT – Science of discovery to science of delivery." Accessed: Aug. 01, 2023. [Online]. Available: https://www.icrisat.org/
[31] "POWER | Data Access Viewer." Accessed: Aug. 01, 2023. [Online]. Available: https://power.larc.nasa.gov/data-access-viewer/
[32] K. K. Suresh and S. R. Krishna Priya, "Forecasting Sugarcane Yield of Tamilnadu Using ARIMA Models," *Sugar Tech*, vol. 13, no. 1, pp. 23–26, Mar. 2011, doi: 10.1007/s12355-011-0071-7.
[33] K. P. Vishawajith, P. K. Sahu, B. S. Dhekale, and P. Mishra, "Modelling and Forecasting Sugarcane and Sugar Production in India," *Indian J. Econ. Dev.*, vol. 12, no. 1, p. 71, 2016, doi: 10.5958/2322-0430.2016.00009.3.
[34] Y. Xu, J. Chou, F. Yang, M. Sun, W. Zhao, and J. Li, "Assessing the Sensitivity of Main Crop Yields to Climate Change Impacts in China," *Atmosphere*, vol. 12, no. 2, Art. no. 2, Feb. 2021, doi: 10.3390/atmos12020172.
[35] P. Mishra *et al.*, "Modeling and Forecasting of Sugarcane Production in India," *Sugar Tech*, vol. 23, no. 6, pp. 1317–1324, Dec. 2021, doi: 10.1007/s12355-021-01004-3.
[36] A. K. Verma, P. K. Garg, K. S. Hari Prasad, V. K. Dadhwal, S. K. Dubey, and A. Kumar, "Sugarcane Yield Forecasting Model Based on Weather Parameters," *Sugar Tech*, vol. 23, no. 1, pp. 158–166, Feb. 2021, doi: 10.1007/s12355-020-00900-4.
[37] S. Guo *et al.*, "Impacts of mean climate and extreme climate indices on soybean yield and yield components in Northeast China," *Sci. Total Environ.*, vol. 838, p. 156284, Sep. 2022, doi: 10.1016/j.scitotenv.2022.156284.
[38] G. B. Torsoni, L. E. de Oliveira Aparecido, G. M. dos Santos, A. G. Chiquitto, J. R. da Silva Cabral Moraes, and G. de Souza Rolim, "Soybean yield prediction by machine learning and climate," *Theor. Appl. Climatol.*, vol. 151, no. 3, pp. 1709–1725, Feb. 2023, doi: 10.1007/s00704-022-04341-9.
[39] K. Jhajharia, P. Mathur, S. Jain, and S. Nijhawan, "Crop Yield Prediction using Machine Learning and Deep Learning Techniques," *Procedia Comput. Sci.*, vol. 218, pp. 406–417, Jan. 2023, doi: 10.1016/j.procs.2023.01.023.
[40] "Applied Sciences | Free Full-Text | Improvement of Wheat Grain Yield Prediction Model Performance Based on Stacking Technique." Accessed: Aug. 01, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/11/24/12164
[41] G. F. Dias and G. Kapetanios, "Estimation and forecasting in vector autoregressive moving average models for rich datasets," *J. Econom.*, vol. 202, no. 1, pp. 75–91, Jan. 2018, doi: 10.1016/j.jeconom.2017.06.022.
[42] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1023/A:1022627411411.
[43] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.
[44] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[45] R. Tibshirani, "A Comparison of Some Error Estimates for Neural Network Models," *Neural Comput.*, vol. 8, no. 1, pp. 152–163, Jan. 1996, doi: 10.1162/neco.1996.8.1.152.
[46] B. Das *et al.*, "Comparative evaluation of linear and nonlinear weather-based models for coconut yield prediction in the west coast of India," *Int. J. Biometeorol.*, vol. 64, no. 7, pp. 1111–1123, Jul. 2020, doi: 10.1007/s00484-020-01884-2.
[47] N. Chergui, "Durum wheat yield forecasting using machine learning," *Artif. Intell. Agric.*, vol. 6, pp. 156–166, Jan. 2022, doi: 10.1016/j.aiia.2022.09.003.
[48] M. Yuan, "Comment: From Ridge Regression to Methods of Regularization," *Technometrics*, vol. 62, no. 4, pp. 447–450, Oct. 2020, doi: 10.1080/00401706.2020.1805021.
[49] S. Singh, A. Prakash, N. R. Chakraborty, C. Wheeler, P. K. Agarwal, and A. Ghosh, "Trait selection by path and principal component analysis in Jatropha curcas for enhanced oil yield," *Ind. Crops Prod.*, vol. 86, pp. 173–179, Aug. 2016, doi: 10.1016/j.indcrop.2016.03.047.
[50] P. Csóka, F. Illés, and T. Solymosi, "On the Shapley value of liability games," *Eur. J. Oper. Res.*, vol. 300, no. 1, pp. 378–386, Jul. 2022, doi: 10.1016/j.ejor.2021.10.012.
[51] V. Pagani *et al.*, "Forecasting sugarcane yields using agro-climatic indicators and Canegro model: A case study in the main production region in Brazil," *Agric. Syst.*, vol. 154, pp. 45–52, Jun. 2017, doi: 10.1016/j.agsy.2017.03.002.
[52] E. Hooda, U. Verma, and B. K. Hooda, "ARIMA and State-Space models for sugarcane (Saccharum officinarum) yield forecasting in Northern agro-climatic zone of Haryana," *J. Appl. Nat. Sci.*, vol. 12, no. 1, Art. no. 1, Mar. 2020, doi: 10.31018/jans.v12i1.2229.
[53] P. Mishra *et al.*, "Modeling and Forecasting of Sugarcane Production in India," *Sugar Tech*, vol. 23, no. 6, pp. 1317–1324, Dec. 2021, doi: 10.1007/s12355-021-01004-3.
[54] A. K. Verma, P. K. Garg, K. S. Hari Prasad, V. K. Dadhwal, S. K. Dubey, and A. Kumar, "Sugarcane Yield Forecasting Model Based on Weather Parameters," *Sugar Tech*, vol. 23, no. 1, pp. 158–166, Feb. 2021, doi: 10.1007/s12355-020-00900-4.

[55] PUSHPA, CHETNA, ADITI, and U. Verma, “ARIMA and ARIMAX models for sugarcane yield forecasting in Northern Agro-climatic zone of Haryana,” *J. Agrometeorol.*, vol. 24, no. 2, Art. no. 2, Apr. 2022, doi: 10.54386/jam.v24i2.1086.

[56] S. R. K. Priya, R. K. Balambiga, P. Mishra, and S. S. Das, “Sugarcane yield forecast using weather based discriminant analysis,” *Smart Agric. Technol.*, vol. 3, p. 100076, Feb. 2023, doi: 10.1016/j.atech.2022.100076.

[57] S. Sridhara, S. B. R, and G. R. Kashyap, “Multistage sugarcane yield prediction using machine learning algorithms,” *J. Agrometeorol.*, vol. 26, no. 1, Art. no. 1, Mar. 2024, doi: 10.54386/jam.v26i1.2411.

Atulya Nagar received the M.Sc. and MPhil degree in mathematics from The MDS university of Ajmer, India, and DPhil in Applied Non-linear Mathematics from University of York. He is currently a Professor and PRO Vice Chancellor at Liverpool Hope University, Liverpool. He has authored or coauthored more than 550 papers in various journals conferences of national and international repute. His research interest include Mathematics Computational Optimization, Natural and Bio-Computing Systems Engineering.

Sumit Kumar received the M.Sc. degree in mathematics from CCS University Campus Meerut, India. He is currently a Ph.D. Scholar with the Department of Applied Mathematics and Scientific Computing, IIT Roorkee, India. His research interests include agri-food supply chain, MCDM methods, machine learning algorithms, and evolutionary algorithms.

.
MILLIE PANT received the M.Sc. degree in mathematics from CCS University, Meerut, and the Ph.D. degree from the Department of Mathematics, IIT Roorkee, India. She is currently a Professor and the Head of the Department of Applied Mathematics and Scientific Computing, IIT Roorkee, where she is also a Joint Faculty at the Mehta Family School of Data Science and Artificial Intelligence. She has authored or coauthored more than 350 papers in various journals and conferences of national and international repute. Her research interests include numerical optimization, operations research, evolutionary algorithms, supply chain management, swarm intelligence techniques, A.I. assisted decision making, data analysis, and image processing