Artificial Bee Colony Algorithm-based Feature Selection and Hybrid ML Framework for Efficient Rice Yield Prediction

Original Scientific Paper

Manasa Chitradurga Manjunath
Department of Computer Science and Engineering,
Presidency University, Bengaluru,
Karnataka, India
manasacm@presidencyuniversity.in

Blessed Prince Pallayan
Department of Computer Science and Engineering,
Presidency University, Bengaluru
Karnataka, India
blessedprince@presidencyuniversity.in

Abstract *– India's economy predominantly depends on monsoon and agricultural output. Agribusiness products contribute to nearly a quarter of its gross domestic product and 58% of its population depends on agriculture for their livelihood. Certain crops, like rice, are vital to its food security being the most widely grown crop and accounting for one-third production of foodgrains in India. Understanding and enhancing its production is critical in ensuring food availability and promoting sustainable agricultural practices. Rice yield prediction has been a most researched area in the agriculture domain. Machine Learning (ML) frameworks have been found to perform well in patches with large, complex datasets as insufficient feature engineering and temporal dependencies plague efficacy. In this paper, we propose a swam-based meta-heuristic artificial bee colony (ABC) algorithm for feature selection from the dataset sourced from the Agricultural Production and Statistical Division of the Department of Agriculture Cooperation and Farmers Welfare, Government of India. The feature engineering is further optimized by a hybrid model comprising a convolutional neural network (CNN) for learning hierarchical representations and identifying relevant attributes from the complex dataset and long short-term memory (LSTM) for temporal aspects. Finally, a random forest (RF) regressor provides the benefits of ensemble learning, which merges multiple decision trees to remove bias, and variance and improve prediction accuracy. From the results, it is observed that the proposed hybrid model outperforms existing state-of-the-art standalone and hybrid models with the highest coefficient of determination (R^2) and lowest mean square error (MSE) of 0.989 and 13613 respectively. The reliable and efficient hybrid model can aid farmers and policymakers in making informed decisions related to rice yield prediction leading to sustainable agricultural practices.*

Keywords: *Artificial bee colony algorithm, convolution neural network, feature selection, long short-term memory, machine learning, random forest, rice yield prediction.*

Received: November 5, 2023; Received in revised form: December 27, 2023; Accepted: January 25, 2024

1. INTRODUCTION

Rice is an important crop for nations worldwide, including India. Fig. 1 illustrates the annual yield of rice in India from the financial year (FY) 1991-2022 [1]. Its production holds significant importance for several reasons as it provides food security since rice is a staple food for most of the Indian population. Ensuring a high rice yield is crucial for the food security of the country's large and growing population. It is of great economic importance as high rice yields lead to increased agricultural income for the farmers, which, in turn, boosts rural livelihoods and helps alleviate poverty. It has a good export potential too as a robust rice yield not only fulfills domestic demand but also provides a surplus for export. India is one of the world's largest rice producers [2]. Rice exports contribute to foreign exchange and enhance its position in global agricultural trade. It provides employment generation as its cultivation employs a rural workforce, including farmers and laborers involved in planting, harvesting, and processing. A healthy rice yield supports these livelihoods.

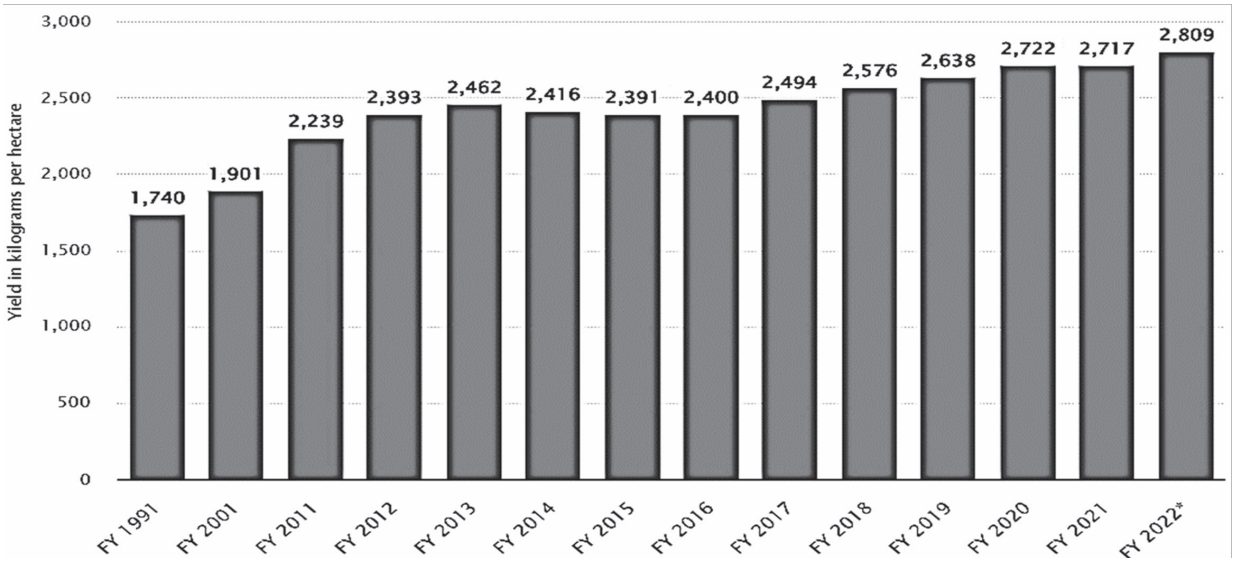


Fig. 1. Annual yield of rice in India FY 1991 to 2022 (in kilograms per hectare) [1]

It is also of big social and cultural significance as rice is deeply embedded in Indian culture, traditions, and cuisine. It is used in many religious rituals and daily meals across the country. It helps in rural development as a strong rice yield encourages investment in agricultural infrastructure, including irrigation systems, storage facilities, and research on improved farming practices. This contributes to rural development and modernization.

Given these reasons, ensuring a sustainable and high rice yield is of utmost importance for India's overall development, food security, and well-being of its citizens. Accurate and timely prediction of its yield can significantly benefit farmers, policymakers, and food distribution systems. By employing advanced computational techniques, such as machine learning, researchers have strived to develop models capable of accurately forecasting rice yields based on various influencing factors. However, this faces several challenges including data availability and quality. High-quality and comprehensive data on several factors that influence rice yield, such as weather conditions, soil characteristics, pest and disease occurrences, crop management practices, and historical yield data, are essential. In many cases, data may be sparse or unevenly distributed across different regions and years, leading to difficulties in building robust and representative predictive models. Rice yield is based on various interconnected factors, including weather patterns, soil health, irrigation practices, and crop management techniques [3]. Capturing the complex interactions and relationships between these variables requires sophisticated modeling techniques. Also, rice is a seasonal crop, and its yield prediction needs to account for seasonal variability, including changes in weather patterns, and pest, and disease outbreaks. The relationships between input variables (e.g., rainfall, temperature, fertilizer application) and rice yield may be non-linear. Thus, traditional linear models may not adequately capture these complex non-linearities. Building complex models to fit the training data too closely can also lead to overfitting, where the model fails to generalize accurately to new, unseen data and does well on training data. Accurate yield prediction often requires historical data over multiple years to identify trends and patterns. In some cases, limited historical data may be available, making it difficult to capture long-term effects accurately. And many advanced machine learning algorithms, such as deep learning models, can be challenging to interpret.

Addressing these challenges requires a combination of domain knowledge, careful feature engineering, model selection, and validation techniques. A proper feature selection technique can have a major influence on the accuracy of rice yield predictions in terms of improved model performance [4-7]. By eliminating noise and irrelevant information, the model can better capture the essential factors that directly influence rice yield, leading to improved prediction accuracy. Feature selection also reduces the complexity of the model, helping mitigate overfitting. A smaller set of relevant features reduces the computational complexity of the model. This results in faster training times and quicker predictions. When the model uses a reduced set of relevant features, it becomes easier to interpret the results. One can gain insights into which specific factors are driving the predictions, and to know the relationships between input variables and rice yield.

Different feature selection methods, such as correlation analysis, recursive feature elimination, feature importance from tree-based models, or advanced techniques like LASSO (L1 regularization), have been used [8-9] but with limited success. The key contributions of this paper are:

- A novel swam-based meta-heuristic artificial bee colony (ABC) algorithm for feature selection from the dataset.

- A reliable and accurate hybrid CNN-LSTM-RF model augmented by feature descriptors from the ABC algorithm for rice yield prediction outperforming existing state-of-the-art standalone and hybrid models.

The manuscript is divided into the following sections: A literature review is covered in section 2. It is followed by Section 3 which outlines the material and methods employed for implementing the proposed model. Section 4 presents the results and compares the performance with existing models. The conclusion is covered in section 5 and references at the end.

2. LITERATURE REVIEW

In the field of rice yield prediction, various techniques have been proposed in the literature. Authors recommend a crop-based prediction system based on site-specific parameters, achieving high accuracy and efficiency [10]. Their recommendation system employs an ML model with a majority voting technique, including RF, Naïve Bayes (NB), Support vector machine (SVM), Linear regression (LR), Decision tree (DT), and XGBoost which motivates to use these ML techniques. They also provide recommendations for suitable fertilizers. The authors in [3] suggested using crop yield projections to optimize fertilizer application. To increase production, they suggested practical fertilizer management practices and employed ML techniques. [2] focuses on predicting paddy yield in the Tamil Nādu Delta region using an MLR-LSTM (Multiple Linear Regression - LSTM) model. This approach aims to provide accurate predictions for paddy yield in this specific region, hence a motivation for using LSTM in a hybrid ML model. [8] employed various ML techniques to analyze and predict crop yields including regression models, DT, SVM, and ensemble methods. Performance evaluation of the best-suited feature subsets for yield prediction is carried out by [11] using ML algorithms. The authors assess different feature combinations and subsets to identify the most influential features for accurate crop yield prediction which motivates us to evaluate computational intelligence (CI) techniques. For capturing complex patterns and data relationships, a hybrid approach using RF and Deep Neural Network (DNN) was proposed by [12]. The results gave better prediction accuracy compared with traditional random forest and deep neural network algorithms, indicating hybrid models are a good fit. A Deep Reinforcement Learning (DRL) model was proposed by [13] for sustainable agricultural applications. DRL combines reinforcement learning techniques with deep learning architectures to optimize decision-making processes and predict crop yields based on environmental factors and other relevant variables. [14] focuses on the most cost-effective means of predicting yields and selecting crops. The study uses artificial neural networks, a reliable tool for modeling and prediction, with forty-six parameters and DNN for crop yield prediction. Both these studies indicate the efficacy of Deep Learning (DL) algorithms compared to ML techniques. IoT is used in [15] to remotely monitor crops with sensor data, and a Multisensor Machine-Learning Approach (MMLA) is proposed for classifying eight crops using the J48 Decision Tree, Hoeffding Tree, and RF algorithms. The RF algorithm proves effective for classifying agricultural text, demonstrating the lowest root mean squared error (RMSE) at 13%, and relative absolute error (RAE) at 38.67%. [16] employ the Normalized Difference Vegetation Index (NDVI) as a crop monitoring tool along with a correlation-based technique to estimate crop production, incorporating physical parameters like soil types and geographic data. They compare SVM, LR, and RF algorithms to improve accuracy and reduce error rates, with RF providing better results. For predicting losses caused by the insect grass grub, [17] uses NB, SVM, DT, RF, NN, K-nearest neighbor (KNN), and ensemble methods. Results from RF and Neural Networks (NN) outperform other classifiers, with ensemble models enhancing the results of weak classifiers. A hybrid model using evolutionary algorithms and data mining techniques is also proposed to enhance findings. All these studies indicate the superiority of RF and evolutionary algorithm efficacy for crop prediction tasks. [18] forecast various crops cultivated in India using the Kernel regression technique, Lasso, and Efficient neural network (ENet) algorithms for yield forecasting, and employ a stacking regression approach to improve algorithms and enhance forecast accuracy, motivating us to employ hybrid approaches.

The ensemble model is employed by [19] aiming to improve the prediction of traits that help overcome hunger-related issues. The study uses a Wild blueberry dataset, utilizing stacking regression (SR) and cascading regression (CR) with a novel combination of ML algorithms. The SR model, with an R^2 of 0.984 and RMSE of 179.898 performed the best. [20] use various data mining techniques to forecast crop production and summarize different crop prediction approaches with different machine learning algorithms. [21] highlight the relevance of remote sensing-based techniques for estimating crop output, comparing remotely derived datasets to in-field survey-based data. [22] introduce the eXtensible Crop Yield Prediction Framework (XCYPF) to predict agricultural yields in precision agriculture, combining relevant indices with information on rainfall and surface temperature for rice and sugarcane crop yield prediction. These studies use datasets with many complex parameters thus inducing the need for better feature selection techniques. [23] propose DL methods like CNN and LSTMs for strawberry yield prediction five weeks ahead, also utilizing yield and weather input data to predict strawberry prices. Their Attention (ATT) based CNN-LSTM model outperforms other ML and DL models for both yield and price prediction using weather data. Thus, CNN-LSTM is a good combination if used in a hybrid mode. [24] use an ensemble model with a majority voting technique includ-

ing NB, RF, Chi-square Automatic Interaction Detector (CHAID), and KNN as learners for a crop recommendation system that accurately and efficiently suggests crops for site-specific parameters. [3] aim to optimize fertilizer application based on crop yield predictions, using ML techniques, and proposing effective fertilizer management strategies to enhance productivity. [25] develops an accurate prediction model for rice yields by utilizing ML algorithms, specifically focusing on rice crop prediction, and presenting a framework that integrates various ML models for improved accuracy. [26] integrates ML techniques with Streamlit, a user-friendly interface, allowing users to analyze and predict crop production effectively, emphasizing the practical implementation of ML models. [27] proposes a comprehensive approach for predicting crop yield using hybrid ML algorithms, combining various techniques to efficiently predict rice crop yield by integrating factors such as weather information, soil properties, and historical yield data. [28] investigate the use of data mining (DM) techniques for crop yield prediction, employing ML algorithms to explore historical crop data and predict future yields, highlighting the importance of data mining in improving prediction accuracy. All these studies again provide the effectiveness of ML and hybrid techniques for crop yield prediction. [29] focus on rice crop yield prediction using Artificial neural networks (ANN), developing an ANN-based model to forecast rice yields based on various input parameters, highlighting the effectiveness of DL techniques in predicting rice crop yields and their potential for enhancing agricultural decision-making. [30] conduct a study on crop analysis and seed marketing, using regression and association rule mining techniques to analyze crop data and identify patterns. Filter, Wrapper, and embedded methods are some of the feature selection methods used by [5, 9], helping farmers make more informed decisions about crop management and improving yields. Authors in [31] used Feature shuffling and Feature performance feature selection methodology with a hybrid model comprising DT, XGBoost, and RF achieving a coefficient of determination (R^2) of 98.6. Filter methods evaluate features independently of the classification model and rank them based on their correlation or mutual information with the target variable [6]. Wrapper methods evaluate feature subsets by repeatedly training and evaluating a classification model on different subsets. They search for the optimal subset of features that gives the best model performance but can be computationally expensive [31]. Embedded methods integrate feature selection into the model training process itself [32]. These studies indicate the need for optimal feature selection methodologies to reduce computations but at the same time not affecting accuracy. However, the use of computational intelligence (CI) or genetic algorithms for feature selection in crop yield prediction has been limited in the literature.

3. MATERIAL AND METHODS

The overall block diagram of the proposed model is shown in Fig. 2. Feature selection plays a vital role in rice yield prediction as it enhances model performance, reduces dimensionality, improves interpretability, and optimizes resource allocation. By identifying the most relevant features, predictive models can provide accurate and actionable insights to aid farmers, agricultural researchers, and policymakers in making informed decisions and achieving higher rice yield sustainably. We propose a CI algorithm-based feature selection technique. It uses the Artificial Bee colony (ABC) algorithm for selecting optimum features.

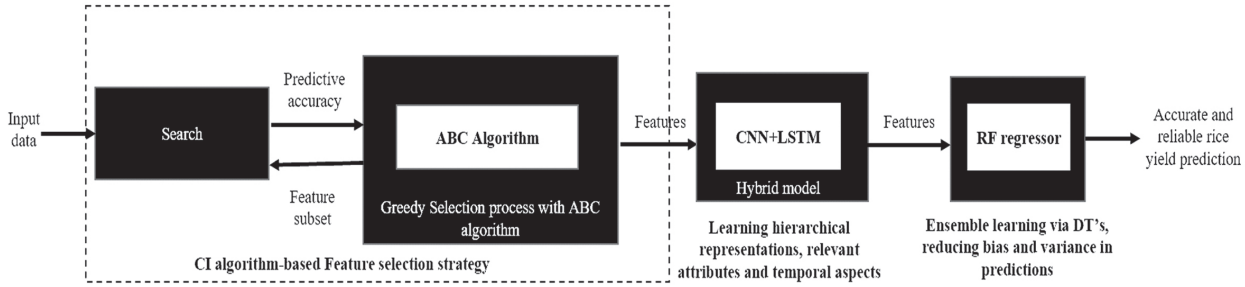


Fig. 2. Block diagram of the proposed framework

3.1. DATASET

The dataset used in this study was sourced from the Agricultural Production and Statistical Division of the Department of Agriculture Cooperation and Farmers Welfare, Government of India https://data.gov.in/sector/Agriculture. The dataset consists of seven features such as state, district, production, year, season, area, and the target variable, yield. Exploratory data analysis (EDA) is conducted to study the dataset characteristics.

3.2. ARTIFICIAL BEE COLONY (ABC) ALGORITHM

The ABC algorithm, introduced by Karaboga [33-35], is a stochastic optimization technique based on the foraging behavior of honeybee swarms. This method is versatile and can be applied to various tasks such as classification, feature selection, clustering, and optimization. The algorithm's flowchart is depicted in Fig. 3. We leverage the ABC algorithm as a tool for feature selection, drawing inspiration from the natural behavior of honeybees

in their colonies. Honeybees exhibit impressive communication, coordination, and self-organization skills in their foraging activities. Communication is facilitated through a behavior known as the 'waggle dance,' which effectively directs other bees to fruitful food sources. In this context, a swarm of bees, denoted as 'S' bees (forming the population), is established.

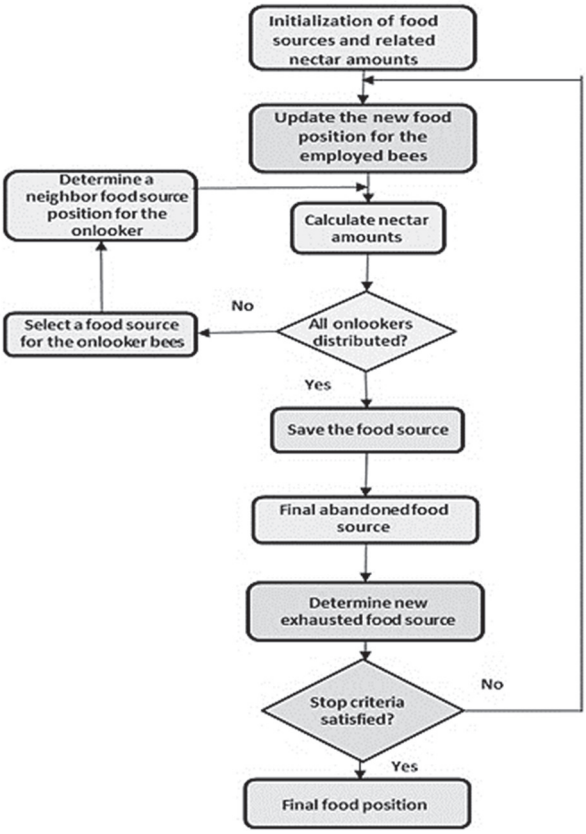


Fig. 3. Flow-chart of ABC algorithm

Potential solutions are represented as food sources, assigned to the bees in a d-dimensional space, aligning with the number of parameters in the optimization problem. The fitness (f_i) metric measures the quantity of nectar at a given food source. The honeybee colony consists of three groups: the employed bees (EB), the onlooker bees (OB), and the scout bees (SB). In each cycle, the algorithm proceeds in the following steps for all the categories of bees.

- Onlooker bees (OB) refer to bees waiting on the 'dance floor' inside the hive.
- In the first cycle, EBs move to random food sources, evaluate nectar amounts, and share this information with OBs.
- OB uses a greedy selection process based on the nectar information received from EB to update their positions.
- In the next cycle, EB selects new food sources in the vicinity of the sources found in the previous cycle, using their memory, and compares nectar amounts.
- OBs use this information to select food sources based on nectar value, and the probability of selecting a position increases with higher nectar amounts.

EB and OB both engage in the search for improved food positions during each cycle. If the nectar quantity at a food source does not improve within a limited number of cycles, the bees abandon those positions. New food sources are randomly generated and assigned to bees called SBs, effectively creating new sources of food. Initially, OBs and SBs are considered unemployed bees (UBs). The exploitation of nectar in food sources is primarily conducted by EB and OB. In every cycle, the food source with the highest present nectar quality is memorized. This food source's position represents the local optimal solution for that cycle. The entire process of food source search, involving EB, OB, and SB, is executed for a specified number of cycles (k_max). The best global solution among all the cycles yields the optimum solution.

The step-by-step procedure is as follows:

- Initialization step: Algorithm control parameters are set, including population, dimension, maximum cycles (k_max), and limits (x_min and x_max). Another limit (B) is established to determine when a food source should be abandoned if further improvement or exploitation is not possible.
- Employed Bee (EB) search: EB searches its neighborhoods, utilizing a greedy selection process to choose between the current food source and one within the neighborhood. If the new source is superior, it updates its position (x_{iD}) accordingly; otherwise, it retains the current position.
- Onlooker Step: Each OB is assigned a probability (Pi) proportional to the quality of the selected food source, as determined by Eq. (1).

$$P_i = \frac{f_i}{\sum_{i=1}^{s} f_n} \tag{1}$$

A new candidate position is generated based on existing memory, as represented in Eq. (2).

$$v_{ik} = x_{ik} + \in_{ik} (x_{ik} - x_{oi}) \tag{2}$$

values o= {1, 2, 3…N} and k= {1, 2, 3…D} are randomly assigned $\in_{ik}$ is a random number chosen from the range [-1, 1]. If the new solution's value exceeds x_min and x_max, it is adjusted to the acceptable limits, and its fitness is evaluated.

- Scout Step: To abandon a food source, a control parameter (B) is utilized. If a predetermined number of trials (T) surpasses B, the food source is abandoned, and SBs are generated. New food source positions are discovered by SBs, and existing food positions are updated randomly using Eq. (3):

$$x_i^j = x_{min}^j + rand(0,1)\left(x_{max}^j - x_{min}^j\right) \tag{3}$$

The ABC algorithm acts as a cluster to choose optimal features. At the end of the training, feature vectors named *food* about each class are obtained. These obtained features are fed into the hybrid CNN+LSTM model.

Table 1. ABC algorithm initialization parameters

Parameter	Number	Remarks
Population (S)	30	Max. population of bees
Food no.	15	No. of sources of food. ~50% of population
Limit (B)	100	Max. limit after which source of food is abandoned and cannot be further improved
No. of iterations	100	No. of foraging cycles
Runtime	25	No. of runs to see its robustness

3.3. CONSTITUENTS OF THE HYBRID MODEL

The CNN model is used to capture relevant informative patterns and relationships between regions and their impact on rice yield. CNN is effective in automatically learning hierarchical representations and identifying relevant attributes from complex datasets. This method is extremely valuable in dealing with diverse input features, which may have non-linear relationships with rice yield. By applying non-linear transformations and feature extraction, the CNN component captures intricate relationships and patterns that may not be evident through traditional linear models.

Next, the LSTM is tuned to manage the temporal aspect by capturing long-term dependencies in sequential data. The LSTM model is well-suited to manage variable-length sequences and effectively structure the temporal dynamics of rice growth. It can manage the input data with varying time steps and learn the patterns and dependencies present in the sequential data.

Finally, the RF regressor provides the benefits of ensemble learning, which merges multiple decision trees to improve prediction accuracy. Incorporating the RF regressor into the hybrid model provides the advantage of its ability to manage non-linear relationships, manage missing data, and provide robust predictions. Additionally, the RF provides interpretability by offering feature importance rankings, enabling us to identify the most influential variables contributing to the predictions. The RF ensemble further reduces bias and variance, leading to more reliable and accurate predictions for rice yield. The RF regressor complements the CNN-LSTM architecture by adding diversity and reducing prediction errors, leading to more reliable predictions for rice yield.

The idea behind the proposed hybrid model is to offer flexibility in incorporating several types of data sources, allowing us to include various features relevant to rice yield prediction. This flexibility enables the model to adapt to different datasets and capture domain-specific knowledge.

3.4. MODEL FLOW

The complete model flow is explained in Fig. 2. We initially use the ABC algorithm to capture optimal features suitable for the prediction and identifying the target variable. The selected features data, though optimal, are pre-processed to manage any anomalies like missing values and inconsistency. The processed data is fed as input to the hybrid CNN-LSTM model as shown in Fig. 2. The CNN model consists of three Conv1D layers which take the input of the pre-processed data comprising 512, 256, and 128 filters respectively. These layers are followed by two LSTM layers consisting of 100 units each. Three dense layers incorporate 512, 10, and 1 output units, respectively. These dense layers are responsible for structuring the output for making accurate predictions. The output from the combined model is fed to the RF regressor which as mentioned in section 3.3 offers the benefits of ensemble learning for reliable predictions. And finally, using the standard metrics the performance of the entire model is evaluated.

3.5. CONSTRUCTING THE MODEL

3.5.1. Import necessary package

The NumPy library in Python is used for managing the numeric data. Pandas package is used for data manipulation and restructuring. Matplotlib is used to visualize the data in the form of graphs. The scikit-Learn library consists of many key features and plays a critical role in pre-processing and getting the data ready to be fed into the model. TensorFlow Keras library manages the deep learning models.

3.5.2. Feature Selection and Pre-Processing

Feature selection is the most important part of constructing a model. For this, we have used the ABC algorithm as explained in section 3.2 and we capture the model performance with and without ABC feature selection. In the pre-processing stage, we employ label encoding to manage string values in four dataset columns, assigning unique numerical labels to each category. This enables effective processing and analysis of categorical data. Further, we utilize the min-max scaler to ensure feature equalization and compatibility. Scaling the numerical features within a specific range eliminates bias and discrepancies caused by measurement unit differences, preserving relative relationships between values. Incorporating label encoding and min-max scaling techniques transforms the dataset, enabling our hybrid model to manage categorical variables and achieve balanced feature representation.

3.5.3. Implementing and training the model

The CNN-LSTM model is implemented using the Sequential API. The model as presented in Fig. 4 consists of three Conv1D layers for feature extraction in CNN, two LSTM layers for temporal sequence modeling,

and three dense layers of output units for shaping and presenting the output. This nature of the hybrid model makes it completely capable of identifying and extracting the hidden features present in the dataset. The model is trained on the dataset using the *fit*() function. The inclusion of the early-stopping call-back method is for stopping the epochs to prevent overfitting. This trained model is used to make predictions on the train and the test data, and the extracted characteristics features feed the RF regressor. The model parameters of the CNN-LSTM model are displayed in Fig. 5. In the summary the first column contains the names of the layers present in the architecture. The output shape indicates the output tensors generated by the model, it comprises (*batch_size*, *timesteps*, and *filters/units*). The *Param #* column contains the number of trainable parameters that are present in each layer. The RF model is built with modified hyperparameters as shown in Table 2.

The *n_estimator* is used to define the number of trees that will constitute the RF model, *max_depth* controls the maximum depth of each decision tree, the *min_sample_split* is the minimum number of nodes that will be present before splitting the nodes and *min_sample_leaf* tells us about the minimum number of leaf's that will be present in each node. Finally, the Random Forest regressor is trained using the extracted attributes from the CNN-LSTM model to obtain the final output.

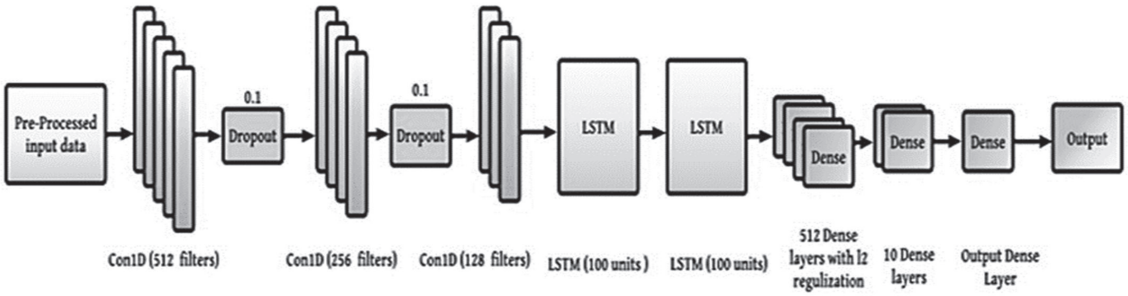


Fig. 4. Block diagram of CNN-LSTM model

3.5.4. Evaluation parameters

The evaluation parameters used in this study are Mean Square error (MSE), Mean absolute error (MAE), Mean absolute percentage error (MAPE), coefficient of determination (R^2), and Root mean square error, calculated using Eq. (4), (5), (6), (7), and (8) respectively.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}) \quad (4)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}| \quad (5)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - F_i|}{A_i} \quad (6)$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2\ (sum\ squared\ regression)}{\sum(y_i - \bar{y})^2 (total\ sum\ of\ the\ squares)} \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})} \quad (8)$$

where n=total number of values, y_i=actual value, $\hat{y}$=predicted value, A_i=actual value and F_i=predicted value.

Table 2. Random Forest parameters

Parameter	Value
Number of estimators	2000
Max. Depth	6
Min. samples split	5
Min. samples leaf	3

```
Model: "sequential_5"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_15 (Conv1D)           (None, 4, 512)            1536
dropout_10 (Dropout)         (None, 4, 512)            0
conv1d_16 (Conv1D)           (None, 4, 256)            262400
dropout_11 (Dropout)         (None, 4, 256)            0
conv1d_17 (Conv1D)           (None, 4, 128)            32896
lstm_10 (LSTM)               (None, 4, 100)            91600
lstm_11 (LSTM)               (None, 100)               80400
dense_15 (Dense)             (None, 512)               51712
dense_16 (Dense)             (None, 10)                5130
dense_17 (Dense)             (None, 1)                 11
=================================================================
Total params: 525,685
Trainable params: 525,685
Non-trainable params: 0
_________________________________________________________________
```

Fig. 5. CNN-LSTM model parameters

4. RESULTS AND DISCUSSION

The results of EDA performed on the dataset are first discussed. Table 3 illustrates a cross-section of the dataset used as mentioned in section 3.1.

Next, we examine the basic statistical properties of the data. We calculated the count of the data, unique values in columns, and descriptive statistics, such as mean, stan-

dard deviation, minimum, maximum, and quartile values for each numerical feature, namely, Area (A), Production (P), and Yield (Y). These statistics helped us understand the range and distribution of values within each variable. The statistical data is represented in Table 4 and Table 5.

Table 3. Cross section of the crop (rice) dataset

State	Dist.	Year	Season	Area	Prod	Yield
Bihar	Purnia	2001-02	Autumn	77801	63333	0.814
Bihar	Purnia	2001-02	Kharif	20748	43473	2.095
Bihar	Purnia	2001-02	Summer	21846	38924	1.781
Bihar	Purnia	2002-03	Autumn	20792	24332	1.170

Table 4. Statistics for categorical data

	State	District	Year	Season
Count	21611	21611	21611	21611
Unique	36	694	25	6
Top	Uttar Pradesh	Purulia	2019-20	Kharif
Frequency	2142	69	1153	9831

Table 5. Statistics of numerical values

	Area	Production	Yield
Count	21611	21611	21611
Mean	46079.02	103475.4	2.09
Std	65787.41	171058.4	1.01
Min	0.35	0	0
25%	2787	4211.5	1.34
50%	16810	29000	2.04
75%	66093.5	129257.5	2.7
Max	687000	1710000	22.37

Next, we explore the categorical variables, namely, State (St), District (D), Year (Y), and Season (S). We examined the unique values present in each category and assessed the frequency distribution of these values. The values are shown in Table 6.

Table 6. Statistics of categorical variables

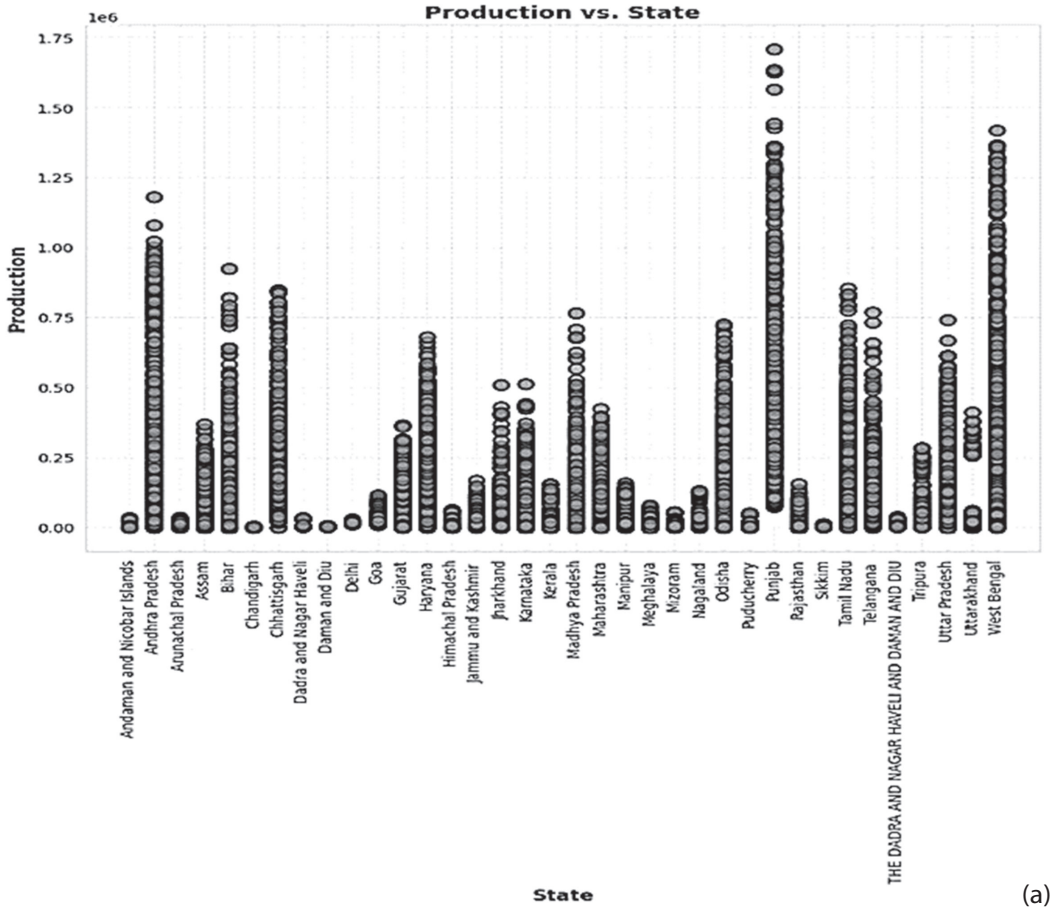
Feature	Unique values
State	36
District	694
Season	6
Year	25

To gain further insights, we visualized the data using various graphical techniques. Fig. 6 shows the distribution of rice cultivation across different states.

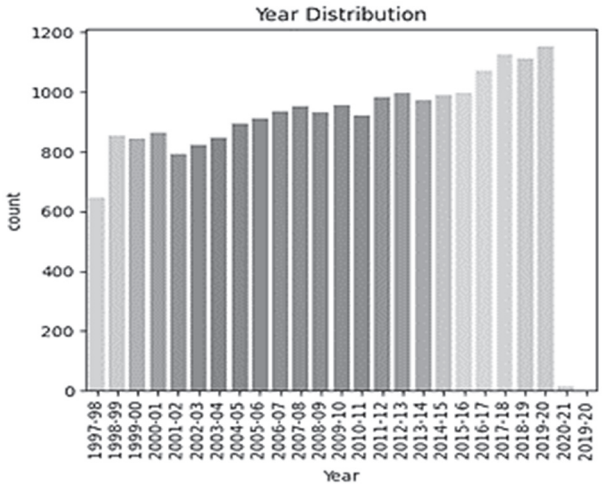
The result of the dataset description post using the ABC feature selection (FS) algorithm is shown in Table 7. It chooses five out of the initial seven features optimally.

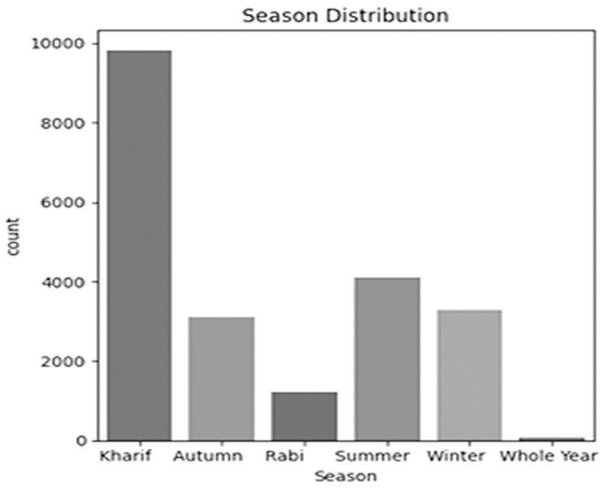
Table 7. Dataset description after ABC feature selection

FS method	St	D	Y	S	C	A	P
ABC algorithm			*	*	*	*	*



(a)





(b) (c)

Fig. 6. Results of exploratory data analysis; (a) State vs. Production for rice yields; (b) Year vs. Production for rice yields and (c) season vs. production for rice yields.

Also, the proposed model outperforms existing state-of-the-art models with an RMSE of 116.67, MAE of 7.43, RAE of 8.67, MAE of 7.43, MSE of 13613 and R^2 of 0.989. MAE scores provide information about the difference between actual and predicted values. A lower MAE value indicates a more efficient model in predicting yield. The MSE score indicates the squared difference between the true and predicted numbers. Computing the RMSE score helps measure the standard deviation of the residuals. The lower the MSE and RMSE scores, the better the model for calculating returns. The higher R^2 value of the proposed model indicates that it captures the spatiotemporal non-linear features well and uses them effectively in predicting the yield.

A comparison between the actual and anticipated values in the case of the proposed hybrid model is shown in Fig. 7. This graph provides a visual representation of the model's performance capturing the patterns and trends of the rice yield data. Since our model has a high R^2 value, the close alignment between the true & predicted values demonstrates a good model fit.

Table 8 presents the scaling characteristics of the LSTM model. Scaling characteristics are measured by increasing the number of epochs and *max_depth* by parameter hyper-tuning. From the results, it is observed that as both these values are increased, the R^2 value too increases, with the best value achieved at 50 epochs, but the time taken also increases. This time can be decreased by using graphical processing units (GPU) as part of future research.

The result of the hybrid model implemented with and without the ABC feature selection technique and comparing its performance with existing state-of-the-art models is captured in Table 9. ABC FS methodology along with the hybrid implementation of CNN+LSTM and RF regressor helps in improving the model performance compared to its performance without any feature selection.

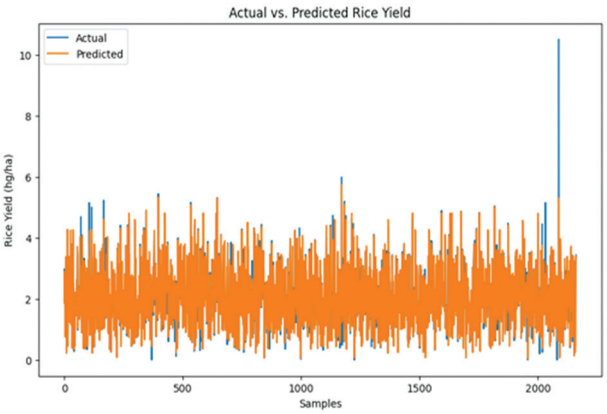


Fig. 7. Actual vs. predicted values of rice yield

Table 8. Scaling characteristics of the LSTM model

Epochs	max_depth	Total_time (mins)	R^2
5	6	15	0.9833
10	6	17	0.9830
20	6	25	0.9824
50	6	40	0.9844
5	8	18	0.9886
10	8	25	0.9878
20	8	42	0.9850
50	8	58	0.9899

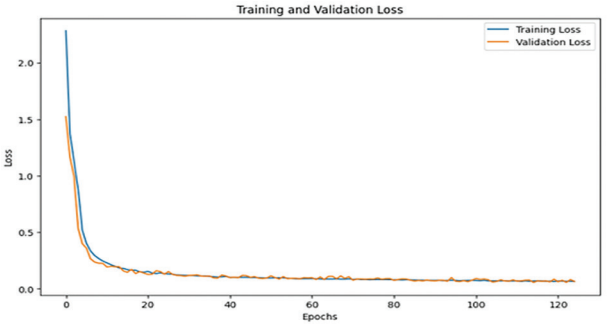


Fig. 8. Training and validation loss in the CNN-LSTM model

Table 9. Comparison of proposed hybrid model with existing models (best results in **bold**)

Reference	Method	R	RMSE	RAE	MAE	R^2	MSE
[25]	K-Star	0.95	365.22	26.65	223.43	0.910	133386
	LR	0.55	936.57	79.57	666.96	0.302	877163
	Gaussian process	0.72	790.54	65.64	548.63	0.525	300995
	MLP	0.76	760.77	68.29	572.48	0.588	327733
	RBF	0.09	1117.03	100.09	839.01	0.008	1247756
	Bagging Model	0.79	700.11	55.46	464.89	0.625	490154
	Additive Regression	0.53	949.21	81.21	680.73	0.285	900999
[26]	Rigid Regression	NA	NA	NA	NA	0.542	NA
	Gradient Boosting	NA	NA	NA	NA	0.667	NA
	Random Forest	NA	NA	NA	NA	0.967	NA
	XGBOOST_ Regression	NA	NA	NA	NA	0.867	NA
[28]	J48	NA	0.27	37.97	0.11	NA	NA
	LWL	NA	0.32	76.34	0.22	NA	NA
	LAD Tree	NA	0.41	68.88	0.19	NA	NA
	IBK	NA	0.30	35.86	0.10	NA	NA
[31]	Hybrid DT, XGBoost, RF with Feature shuffling and Feature performance	0.11	335.10	11.29	8.6023	0.879	112296
[36]	Random Forest	NA	NA	5.82	15	NA	NA
	K-star	NA	NA	12.8	34	NA	NA
	Bays-net	NA	NA	68.45	18.5	NA	NA
	J48	NA	NA	59.29	16	NA	NA
Proposed Model	CNN+LSTM+RF	0.98	122.7	13.1	5.57	0.98	15065
	CNN+LSTM+RF with features selected by the ABC algorithm	**0.99**	116.67	8.67	7.43	**0.989**	13613

The training versus validation loss for the CNN-LSTM model is shown in Fig. 8. which demonstrates the learning advancement of our model during the training process. The graph shows the convergence of the model with reducing training loss, and the low value of validation loss compared to training loss confirms the effectiveness of the training procedure.

Overall, with a high R^2 value of 0.989 and, a low MSE value of 13613 with optimal features selected by the ABC algorithm, and proposed hybrid CNN+LSTM model along with the RF regressor successfully captures the complex relationships between the input features and crop yield. It suggests that this hybrid model can be relied upon to provide accurate predictions of crop yields, which can be invaluable for agricultural planning, resource allocation, and the decision-making process of the agricultural community.

5. CONCLUSION

This research introduced a novel approach for predicting rice yield, leveraging the ABC algorithm for feature selection along with a hybrid CNN+ LSTM model. The RF regressor complements the hybrid model by adding diversity and reducing prediction errors, exhibiting superior performance compared to other existing models by achieving the highest scores for important evaluation metrics. The visualization presented in the research further confirmed the model's ability to capture underlying patterns in the dataset. The practical implications of this research go beyond the academic sphere, providing significant benefits to the agricultural community. Accurate crop yield predictions empower farmers to optimize their practices, allocate resources effectively, and minimize crop losses. Policymakers can utilize these predictions for devising strategies related to food security, distribution management, and sustainable agriculture. Additionally, the commercial sector can make informed decisions regarding crop procurement, storage, and pricing based on this information. Although our hybrid model has shown impressive performance, there is still room for improvement. Future studies could focus on incorporating additional features like soil characteristics, historical climate data, and socio-economic factors. Leveraging GPUs to accelerate the model's training can help in faster predictive results. Evaluating the model on larger and more diverse datasets would also validate its robustness and generalizability across various regions and timeframes.

6. REFERENCES

[1] S. Keelery, "Annual yield of rice India FY 1991-2022", https://www.statista.com/statistics/764299/india-yield-of-rice/ (accessed:2023)

[2] P. Sathya, P. Gnanasekaran, “Paddy yield prediction in Tamil Nadu delta region using MLR-LSTM model”, Applied Artificial Intelligence, Vol. 37, No. 1, 2023.

[3] S. M. Bharath, S. Manoj, P. Adhappa, P. L. Patagar, R. Bhaskar, “Crop yield prediction with efficient use of fertilizers”, Lecture Notes in Electrical Engineering, Vol. 783, 2021, pp. 937-943.

[4] F. B. Felipe, L. H. A. Rodrigues, "The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modeling”, Computers and electronics in agriculture, Vol. 128, 2016, pp. 67-76.

[5] C. Girish, F. Sahin, "A survey on feature selection methods", Computers & Electrical Engineering, Vol. 40, No. 1, 2014, pp. 16-28.

[6] Sánchez-Maroño, Noelia, A. Alonso-Betanzos, M. Tombilla-Sanromán, "Filter methods for feature selection--a comparative study", Lecture notes in Computer Science, Vol. 4881, 2007, pp. 178-187.

[7] N. El Aboudi, L. Benhlima, "Review on wrapper feature selection approaches", Proceedings of the IEEE International Conference on Engineering & MIS, Agadir, Morocco, 22-24 September 2016, pp. 1-5.

[8] P. S. M. Gopal, R. Bhargavi, “Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms”, Applied Artificial Intelligence, Vol. 33, No. 7, 2019, pp. 621-642.

[9] D. W. Christopher, J. M. Vance, H. K. Rasheed, A. Missaoui, K. M. Rasheed, F. W. Maier, "Using machine learning and feature selection for alfalfa yield prediction", AI, Vol. 2, No. 1, 2021, pp. 71-88.

[10] C. Gaurav, A. Chaudhary, "Crop recommendation system using machine learning algorithms", Proceedings of the IEEE 10th International Conference on System Modeling & Advancement in Research Trends, Greater Noida, India, 11-12 May 2021, pp. 109-112.

[11] S. V. Joshua et al. "Crop yield prediction using machine learning approaches on a wide spectrum", Computers, Materials & Continua, Vol. 72, No. 3, 2022, pp. 5663-5679.

[12] E. Banu, A. Geetha, “Rice crop yield prediction using random forest and deep neural network - an integrated approach”, SSRN Electron Journal, 2021.

[13] D. Elavarasan, P. M. Durairaj Vincent, “Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications”, IEEE Access, Vol. 8, 2020, pp. 86886-86901.

[14] T. Islam, T. A. Chisty, A. Chakrabarty, "A Deep Neural Network Approach for Crop Selection and Yield Prediction in Bangladesh", Proceedings of the IEEE Region 10 Humanitarian Technology Conference, Malambe, Sri Lanka, 6-8 December 2018, pp. 1-6.

[15] A. Reyana, S. Kautish, P. M. S. Karthik, I. Ahmed Al-Baltah, M. B. Jasser, A. W. Mohamed, "Accelerating Crop Yield: Multisensor Data Fusion and Machine Learning for Agriculture Text Classification", IEEE Access, Vol. 11, 2023, pp. 20795-20805.

[16] H. Jing, H. Wang, Q. Dai, D. Han, "Analysis of NDVI data for crop identification and yield estimation", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 7, No. 11, 2014, pp. 4374-4384.

[17] A. Umair, S. A. Moqurrab, "Predicting crop diseases using data mining approaches: classification", Proceedings of the IEEE 1st International Conference on Power, Energy and Smart Grid, Mirpur Azad Kashmir, Pakistan, 9-10 April 2018, pp. 1-6.

[18] N. P. Sai, P. S. Venkat, B. L. Avinash, B. Jabber, "Crop yield prediction based on Indian agriculture using machine learning", Proceedings of the IEEE International Conference for Emerging Technology, Belgaum, India, 5-7 June 2020, pp. 1-4.

[19] R. Seireg, Hayam, Y. M. K. Omar, F. E. Abd El-Samie, A. S. El-Fishawy, A. Elmahalawy, "Ensemble machine learning techniques using computer simulation data for wild blueberry yield prediction", IEEE Access, Vol. 10, 2022, pp. 64671-64687.

[20] G. Yogesh, "A study on various data mining techniques for crop yield prediction", Proceedings of the IEEE International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques, Mysuru, India, 15-16 December 2017, pp. 420-423.

[21] I. E. Mladenova, J. D. Bolten, W. T. Crow, M. C. Anderson, C. R. Hain, D. M. Johnson, R. Mueller, "Intercomparison of soil moisture, evaporative stress, and vegetation indices for estimating corn and soybean yields over the US", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 10, No. 4, 2017, pp. 1328-1343.

[22] A. Manjula, G. Narsimha, "XCYPF: A flexible and extensible framework for agricultural Crop Yield Prediction", Proceedings of the IEEE 9th International Conference on Intelligent Systems and Control, Coimbatore, India, 9-10 January 2015, pp. 1-5.

[23] N. Lobna, I. E. Okwuchi, M. Saad, F. Karray, K. Ponnambalam, P. Agrawal, "Prediction of strawberry yield and farm price utilizing deep learning", Proceedings of the IEEE International Joint Conference on Neural Networks, Glasgow, UK, 19-24 July 2020, pp. 1-7.

[24] S. Pudumalar, E. Ramanujam, R. H. Rajashree, C. Kavya, T. Kiruthika, J. Nisha, "Crop recommendation system for precision agriculture", Proceedings of the IEEE Eighth International Conference on Advanced Computing, Chennai, India, 19-21 January 2017, pp. 32-36.

[25] N. Gnanasankaran, E. Ramaraj, T. Manikumar, "An Intelligent Framework for Rice Yield Prediction using Machine Learning based Models", International Journal of Scientific Engineering and Research, Vol. 12, No. 1, 2021.

[26] A. Saxena, M. Dhadwal, M. Kowsigan, "Indian Crop Production: Prediction and Model Deployment Using ML and Streamlit", Turkish Journal of Physiotherapy and Rehabilitation, Vol. 32, No. 3, 2021, p. 3.

[27] K. S. Saravanan, V. Bhagavathiappan, "Relative temperature disparity and rice yield across seasons in Tamil Nadu", Journal of Agrometeorology, Vol. 6, No. 2, 2004, pp. 5-9.

[28] S. Mishra, P. Paygude, S. Chaudhary, S. Idate, "Use of data mining in crop yield prediction", Proceedings of the 2nd International Conference on Inventive Systems and Control, Coimbatore, India, 19-20 January 2018, pp. 796-802.

[29] N. Gandhi, O. Petkar, L. J. Armstrong, "Rice crop yield prediction using artificial neural networks", Proceedings of the IEEE Technological Innovations in ICT for Agriculture and Rural Development, Chennai, India, 15-16 July 2016, pp. 105-110.

[30] V. Sharma, M. Shukla, R. Mandal, "Crop Analysis and Seed Marketing using Regression and Association Rules of India", Proceedings of International Conference on Emerging Trends in Information Technology and Engineering, Vellore, India, 24-25 February 2020, pp. 1-5.

[31] C. M. Manasa, B. P. Palayyan, "An Efficient Crop Yield Prediction Framework Using Hybrid Machine Learning Model", Revue d'Intelligence Artificielle Journal, Vol. 37, No. 4, 2023, pp. 1057-1067.

[32] T. Jiliang, S. Alelyani, H. Liu, "Feature selection for classification: A review", Data classification: Algorithms and applications, CRC Press, 2014, pp. 37-64.

[33] D. Karaboga, C. Ozturk, "A novel clustering approach: artificial bee colony (ABC) algorithm". Applied Soft Computing Journal, Vol. 11, No. 1, 2011, pp. 652-657.

[34] B. Akay, D. Karaboga, "A modified artificial bee colony algorithm for real-parameter optimization", Information Sciences, Vol. 192, 2012, pp. 120-142.

[35] D. Karaboga, B. Akay, "A comparative study of artificial bee colony algorithm", Applied Mathematics and Computation, Vol. 214, No. 1, 2009, pp. 108-132.

[36] K. Lata, S. Khan, "Experimental analysis of machine learning algorithms based on agricultural dataset for improving crop yield prediction", International Journal of Engineering and Advanced Technology, Vol. 9, No. 1, 2020, pp. 3246-3251.