Neural Computing and Applications (2024) 36:20723–20750
https://doi.org/10.1007/s00521-024-10226-x

ORIGINAL ARTICLE

A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning

Mahmoud Abdel-salam[1] · Neeraj Kumar[2] · Shubham Mahajan[3]

Received: 26 August 2023 / Accepted: 12 July 2024 / Published online: 16 August 2024

Abstract

Accurately predicting crop yield is essential for optimizing agricultural practices and ensuring food security. However, existing approaches often struggle to capture the complex interactions between various environmental factors and crop growth, leading to suboptimal predictions. Consequently, identifying the most important feature is vital when leveraging Support Vector Regressor (SVR) for crop yield prediction. In addition, the manual tuning of SVR hyperparameters may not always offer high accuracy. In this paper, we introduce a novel framework for predicting crop yields that address these challenges. Our framework integrates a new hybrid feature selection approach with an optimized SVR model to enhance prediction accuracy efficiently. The proposed framework comprises three phases: preprocessing, hybrid feature selection, and prediction phases. In preprocessing phase, data normalization is conducted, followed by an application of K-means clustering in conjunction with the correlation-based filter (CFS) to generate a reduced dataset. Subsequently, in the hybrid feature selection phase, a novel hybrid FMIG-RFE feature selection approach is proposed. Finally, the prediction phase introduces an improved variant of Crayfish Optimization Algorithm (COA), named ICOA, which is utilized to optimize the hyperparameters of SVR model thereby achieving superior prediction accuracy along with the novel hybrid feature selection approach. Several experiments are conducted to assess and evaluate the performance of the proposed framework. The results demonstrated the superior performance of the proposed framework over state-of-art approaches. Furthermore, experimental findings regarding the ICOA optimization algorithm affirm its efficacy in optimizing the hyperparameters of SVR model, thereby enhancing both prediction accuracy and computational efficiency, surpassing existing algorithms.

Keywords Machine learning · Feature selection · CFS · Crop yield prediction · Information gain · Hyperparameter optimization · Crayfish optimization · Support vector machine · Support vector regression

Abbreviations

KM	K-means clustering algorithm
RFE	Recursive feature elimination
COA	Crayfish optimization algorithm
ICOA	Improved crayfish optimization algorithm
GA	Genetic algorithm
PLMDC	Proximity likelihood maximization data clustering
ANN	Artificial neural network
RMSE	Root mean square error
GBDT	Gradient boosting decision
LSTM	Long short-term memory
CNN	Convolutional neural network
RNN	Recurrent neural network
DL	Deep learning
DM	Data mining
X_i	The position of the current agent in the population
Max_iter	The maximum number of iterations
t	The current iteration number
CFS	Cluster feature selection
RMSE	Root mean square error
MAE	Mean absolute error

✉ Mahmoud Abdel-salam
mahmoud20@mans.edu.eg

Neeraj Kumar
neeraj.kumar@thapar.edu

Shubham Mahajan
mahajanshubham2232579@gmail.com

[1] Faculty of Computer and Information Science, Mansoura University, Mansoura, Egypt

[2] Department of CSE, Thapar Institute of Engineering and Technology, Deemed to Be University, Patiala 147004, India

[3] Department of Computer Science, CHRIST (Deemed to be University), Delhi-NCR, India

MedAE	Median absolute error
MAPE	Mean absolute percentage error
MIG	Mutual information gain

1 Introduction

Since farmers are responsible for producing a sizable proportion of the world's food supply, agriculture is one of the significant areas of social interest. Due to population growth and food shortages, hunger persists in many nations today. Increasing food production is an attractive strategy to eradicate hunger. The United Nations has set the year 2030 as a target date for accomplishing two of its most important goals: increasing food security and decreasing need. Policymakers in a country need accurate forecasts to make informed decisions about food exports and imports. Also, farmers and growers can use yield predictions to better plan their budgets and operations. However, due to many complicated factors, accurate predictions of agricultural yields are notoriously tricky. The success of a crop relies on several factors, including the weather, the soil, the terrain, the presence of pests, the water supply, the plant's genetic makeup, the crop's organization, and more [1, 2].

Researchers are making more precise predictions with the help of data-driven models [3]. To enhance the accuracy of data-driven models, Machine Learning (ML) methods are essential [4]. Machine learning enables computers to gain new abilities without the need for explicit programming. Agricultural frameworks, whether non- or linear-based, can be resolved by these procedures, which thus provide exceptional foresight [5]. Agronomic frameworks based on machine learning acquire their strategies through learning. The operations necessitate much practice before they can be executed successfully. Once the training phase is complete, the model will use its assumptions to validate the data.

While ML and its realization have made great strides, there are still some limits to what can be accomplished when relying solely on the data. ML predictions' accuracy and limitations are influenced by model parameters, data quality, and the relationship between target and input variables in the obtained datasets [6]. Incomplete or inaccurate data, biases, outliers, and noisy data can all severely weaken the prediction ability of models [7]. Multivariate regression, random forests, regression trees, neural networks and association rules are machine learning models engaged in numerous research to predict yields of agriculture. The output, crop yield, is viewed by machine learning models as a function of many input factors, such as soil conditions and weather components.

ML has developed as a powerful tool for obtaining insights and patterns from data, applied to various applications and domains including the agriculture environment. ML models can be generally split into unsupervised and supervised learning models. For supervised learning, where models are trained on labeled data to make predictions or decisions, offers a structured approach to solving predictive tasks. In contrast, unsupervised learning explores patterns and structures within unlabeled data, often uncovering hidden relationships or clusters. This paper focused on the supervised learning models such as SVR, kNN, RF, etc., for assessing the evaluation of the proposed framework. The supervised learning allows us to leverage historical datasets containing labeled information about environmental factors and corresponding crop yields. This labeled data provides a clear signal for the model to learn from, enabling it to capture complex relationships and make accurate predictions. By utilizing supervised learning, we aim to develop robust models that effectively forecast crop yields, thereby informing optimized agricultural practices and contributing to food security efforts.

More recently, Support Vector Machine has attracted the attention of researchers, practitioners, and statisticians due to its theoretical and practical superiority, which has been learnt to perform better in both classification and regression. SVM was originally applied to classification problems [22]. It has since been extended to handle nonlinear regression problems and named Support Vector Regression [22]. There are two distinct advantages to the application of SVR. Firstly, SVR guarantees convergence to optimal solutions using quadratic programming with linear constraints for learning data. Secondly, it is computationally efficient in modeling nonlinear relationships using kernel mapping. However, the computational efficiency of SVR depends on a couple of hyperparameters and factors that directly or indirectly affect finding the optimal solutions. Ordinarily, an exhaustive grid search is utilized to explore all the hyperparameter combinations. Cross-validation is conducted to evaluate the prediction capability of SVR. Despite its striking features, SVR has its limitations [26]. The important one is its inability to perform feature selection. In other words, it is incapable of feature selection [27].

On the other hand, feature selection plays an essential part in supervised learning to obtain more promising and efficient results. Feature selection filters unnecessary information from a dataset using statistical metrics to improve a learning algorithm. The primary goal of feature

selection is to collect a good set of characteristics that may be used to characterize and limit a dataset. The feature selection approach in machine learning reduces computation time [8], improves forecasting outcomes, and enhances data comprehension. Feature selection, then, is a typical preprocessing step for high-dimensional data. The goals are to make the data and the model easier to understand by reducing their dimensionality and improving forecasts' accuracy. In other words, feature selection involves identifying the most relevant input variables from a pool of potential predictors, such as weather conditions, soil attributes, and agricultural practices for improving and enhancing the crop prediction phase.

There are three distinct types of feature selection FS methodologies named as a wrapper, a filter, or a hybrid. Due to the increased complexity introduced by features of higher dimensions, this problem cannot be fixed by simply combining all possible solutions. The filter approaches can identify and eliminate irrelevant features; they cannot do the same for repeating features due to their failure to account for possible associations among features [9, 10]. For the filter method, it is the features of the data themselves that define which subset of features is most important, such as its correlation, Fisher score measure, information gain, mutual information, and entropy [11]. The wrapper feature selection approaches is wrapped within the induction process [12, 13]. It is helpful to use the wrapper method when problems arise. Several search methods can be used to find a subgroup of features by restricting the suitable objective function, including recursive feature elimination and backward and forward elimination passes [14]. Wrapper methods are easily recognized by the excellent quality of the features they select, although at the deprivation of a higher computational cost. Hybrid approaches are another approach that is occasionally studied to better special features. They employ methods that aim for an intermediate between computational complexity and speed [1, 15]. It strikes a good mix between accuracy and processing speed. In this study, we proposed a hybrid approach, utilizing elements of both the filter and wrapper approaches.

Despite the promising potential of supervised learning and feature selection techniques in agriculture, challenges persist in effectively integrating these methods to enhance crop yield prediction models. Agricultural datasets often exhibit high dimensionality and contain numerous variables, necessitating robust feature selection approaches to identify the most influential factors. Moreover, manual tuning of SVR hyperparameters can be labor-intensive and may not always yield optimal results.

To address these challenges, this paper proposes a new framework with three phases: Preprocessing, Feature Selection, and prediction. First, the k-means (KM) technique is used to cluster all the dataset's features. It strives to maintain the clusters as far apart as feasible while making their features consistent. Then, the CFS ranking method independently positions features in each cluster. These two techniques simplify the search space by addressing the high dimensionality and redundancy problems. After the top features from each cluster are chosen, the resulting reduced dataset is forwarded to the feature selection phase. Secondly, in the feature selection phase, a novel hybrid feature selection strategy is proposed to narrow down the pool of candidates to the top-performing features. Filter-type approaches, Fisher score and Mutual information gain, are applied. The intersection set of the resulting features from each process is fed into the wrapper approach. The wrapper-based approach, recursive feature elimination Random-Forest-based RFE, combined with the filer approaches to create a hybrid-based feature selection technique. Finally, for the prediction phase, a novel improved algorithm ICOA is proposed to optimize the hyperparameter of SVR model to enhance the prediction results of the final phase. The COA algorithm is enhanced with the chaotic map and the Levy distribution function to enhance exploration and exploitation phases of COA resulting in a novel ICOA algorithm. The paddy crop dataset is used with the proposed method to identify the best features for future crop production prediction.

This paper's main contributions are summed up as follows:

- A framework that integrates a novel hybrid feature selection approach with optimized SVR model to enhance the prediction results is proposed.
- This paper provides a hybrid approach to feature selection, combining heuristic techniques such as filter and wrapper methods.
- An improved variant of COA algorithm is proposed to enhance the exploration and exploitation phases of COA.
- The Levy flight and chaotic maps are integrated into the original COA resulting in a promising ICOA applied to optimize the SVR model.
- The dataset's redundancy and high dimensionality were mitigated through an unsupervised feature selection strategy in the preprocessing phase, such as combining KM clustering and the CFS ranking.
- Experimental results confirm that the proposed approach selects the most relevant features and enhances the prediction results.

The following sections make up this paper: Sect. 2 compiles previous research on predicting crop yields, and Sect. 3 explains the information collected for this study. Section 4 discusses the proposed framework and its components; Sect. 5 presents our results and discussions of the

experiments; Sect. 6 present the discussion of the obtained results and Sect. 7 offers the conclusion and future work in the related work.

2 Related work

Estimating crop yields is critical in today's world when an ever-growing population demands more and more food. It aids in the enhancement of management procedures vital to maximizing agricultural yield. ML methods, traditional regression methods, and crop models [16–18] have been used to estimate crop yields in the previous decade. Crop yield models are a type of crop growth model. According to these parameters, they are merely a simulacrum of actual scientific studies [19]. Providing reliable data on agricultural output, these models aid policymakers, farmers, and the government achieves maximum sustainability [20].

Vani and Rathi [21] described big data analysis as gathering, maintaining, and analyzing massive amounts of data to find connections and other insights. Big data was used to analyses harvest, soil, and climate data from internal and external sources for agricultural applications. Several machine learning algorithms grouped the data to estimate agricultural productivity. However, the grouping was inaccurate and deprived. On the other hand, Proximity Likelihood Maximization Data Clustering (PLMDC) uses fewer characteristics from vast and densely packed farming data to improve clustering and farmer crop output projections. An appropriate linear regression method was utilized to remove extraneous features from dense and sparse agricultural data. The Genetic Algorithm (GA) selected clustering data features for best fitness. The A-FP development methods evaluate the decision-support system's capability to predict agricultural yields using meteorological data and crop quality. The facts and observations showed that PLMDC was more effective than current methods.

Predictions of frost danger for Zhejiang tea plantations using ML methods have also been made [22]. Damage was calculated using meteorology, topography, and coordinate geometry (latitude and longitude). ANN and SVM were used for the estimation. The authors in [23] built a Spatio-temporal hybrid model using satellite-derived hydro-meteorological data from 20 sites for 20 years in Bangladesh. Dragonfly optimization and support vector regression (SVR) were employed in this research. This hybrid model reduced the relative error in predicting tea crops by 11%.

A. Reyana et al., [24] utilize data from IoT sensors to remotely monitor their crops. In contemporary agriculture, producers presently control the surrounding environment of their crops to optimize yields. The authors presented the MMLA, a novel method for recognizing multisensory information. The suggested recommendation system categorized eight crops. Different machine learning algorithms were utilized to classify different sorts of crops, specifically the Random Forest and J48 Decision Tree. The classifier's performance was evaluated using precision, F-measure and recall. These findings were then compared to the advanced classifications. The Random Forest method demonstrated superior efficiency in categorizing agricultural-related text, exhibiting the lowest error measures such as a 13% Root Mean Square Error (RMSE).

Paudel et al. [25] used Crop Yield Prediction model (MARS) data from the mutual Research Centre of the European Commission to evaluate NN crop yield forecasting models' accuracy and accessibility. The 1DCNN and LSTM could handle time-series data. A GBDT model with hand-crafted attributes was compared to effectiveness. Agriculture and crop yield forecasting experts used feature recognition algorithms to rate input parameters' significance. LSTM models outperformed GBDT models economically for wheat crop in Germany. LSTM models accurately predicted the impacts of yield pattern, static features including biomass and soil retention ability features on crop output, however high temperature and moisture circumstances were harder to measure. This study shows that DL can mechanically acquire characteristics and provide accurate crop output estimates the advantages and challenges of relating stakeholders' human in model understanding assessment.

Khaki et al. [26] used ML to successfully forecast corn production and yield difference among corn hybrids given either environmental or genotype data. Using remotely sensed data collected before the crop, You et al. [27] applied DL algorithms to estimate soybean production. An ANN model was also built to forecast environmental impacts and tea crops in Iran for black, green, and oolong [28]. After comparing the performance of deep fully connected neural networks, LASSO and RF found that combining a CNN and an RNN was superior for predicting soybean and corn yields [29]. Researchers created a decision-support system using information about the soil and the surrounding environment [30].

Swanth et al. [31] propose a new way of predicting crop yield using a hybrid classification model that incorporates an enhanced feature ranking fusion technique. The authors propose a new SMOTE algorithm for data enrichment to ensure the optimization of features that will be extracted. Their technique for feature extraction includes statistical features, improved correlation-based features, raw data, and entropy features. They also offer an enhanced way of combining feature rankings using the results of various feature selection techniques: Relief, RFE and Chi-square. Their hybrid model, which combines DBN and LSTM models, is used for prediction. The results of the authors

show that their approach improves upon traditional classifiers including LSTM, DBN, Bi-GRU, CNN, and SVM.

Fatma M. Talaat [32] introduces the Crop Yield Prediction Algorithm (CYPA), a new method that employs IoT techniques in precision farming. The authors integrate climate, meteorological, chemical data and agricultural yield into CYPA to enable policymakers to forecast annual crop yields. The authors developed a decision support tool to aid farmers and decision makers in predicting agricultural yields by analyzing meteorological circumstances specific to their regions. The researchers suggested an advanced machine learning approach for predicting agricultural yields. In addition, active learning was implemented in CYPA to optimize the model's performance by minimizing the amount of labeled data required for training. The CYPA can respond to adjusting field environments, such as pest outbreaks or weather by engaging in active learning. This involves actively choosing fresh samples for labelling that accurately reflect the current conditions.

The Levenberg–Marquardt technique was previously exploited to evaluate and forecast human gait [33], and can be utilized in surveys of forests and farms. Surveying with the old methods is difficult, time-consuming, and costly, especially in remote or rugged places or where a lot of vegetation is present, such as mountains, forests, or fields. In another paper, relevant operating laws and necessary weighted aggregation operators were devised [34]. Here, scalar multiplication and neutral addition operational rules define the properties of the neutral type in the group association degrees and the sum of probability. All facets of the proposed legislation are examined.

However, research on the use of DL for predicting tea yield is scant [35]. To estimate yield, ML and DL methods analyze data on climate, soil, crops, and satellite imagery [36]. The use of different microwave and spectral wavelengths, made possible by remote sensing data, enables crop status monitoring [37]. Predictions of wheat crops have been made using satellite and climate data [38]. A model for prediction for sorghum biomass was suggested with the sorghum crop model APSIM, the multi-layer perceptron, and SVM as input. After comparing other models, they decided that the MLP one was the most reliable [39].

Previous authors have used data mining (DM) techniques to identify and organize data corresponding to the relative importance of the critical features influencing sugarcane output and then to create mathematical models for predicting sugarcane yield [40]. Three different DM methods were used to analyze data from the databases of numerous sucrose mills in Brazil. Some DM strategies have been used to investigate relationships between weather conditions and plant care. An external dataset has been used to estimate the accuracy of the derived models. The RF algorithm was utilized for comparison.

In [41], CNN and LSTM are coupled to estimate county-level soybean yields using outdoor remote sensing data at both the end of the increasing season. There is a shortage of literature on using the deep learning approach to estimate agricultural yields in an indoor greenhouse setting, in contrast to outside application scenarios. The research in [42, 43] motivated us to apply a RNN with long-short temporal memory (LSTM) units to the problem of predicting crop yields for tomatoes and ficus. The evaluation results also demonstrate that the conventional machine learning algorithms are inferior to deep learning techniques regarding prediction accuracy and root mean square errors (RMSEs).

To define agricultural objectives for import and export, as well as to boost farmer incomes, crop yields must be predicted quickly and precisely in numerical and economic assessments. Crop production forecasts are one of the most difficult concerns in the agricultural industry, since they are used to estimate higher crop output utilizing machine learning techniques. According to previously reported related studies, it is found that the literature work that utilizes DT algorithm has overfitting concerns with the data, resulting in inaccurate predictions. The primary obstacles and issues in the associated work can be summarized as follows:

- More technique classifiers for agricultural yield prediction must be examined in the linked work.
- The associated work must consider the proposed technique in all variables of the agricultural sectors to enhance the forecasting process.
- The related work must add climatic data of the suggested method to boost the accuracy of prediction.
- The linked work must add more crop-related features into the suggested technique for accurate prediction.
- The linked work needs to analyze improved strategies for higher accuracy in crop forecast.
- The associated work does not use an optimized model, which can have a significant impact on prediction accuracy when compared to traditional models.

Therefore, this paper proposes a new framework that enhances the prediction performance by introducing a comprehensive framework that proposes a new hybrid feature selection approach and a novel algorithm for optimizing the different hyper parameters for the prediction process. The proposed framework helped to handle different issues by related works where the novel hybrid feature selection approach is more focused on the best features to reduce the dimensionality reduction. In addition, the new optimized model for the prediction enhanced the prediction results compared with the recent approaches.

Furthermore, a new set of climatic features are integrated to enhance the obtained results.

3 Prelilimary

This section provides a high-level overview of the tools and techniques used in this study.

3.1 Crayfish optimization algorithm (COA)

Jia et al. [44] developed the Crayfish Optimization Algorithm by mimicking the behaviors of crayfish. These include summer resorting, competition, and foraging. Foraging and competition behaviors imitate the exploitation and exploration processes of the optimization procedure, respectively, which can be controlled by temperature. In high temperatures, crayfish seek shelter in caves either for summer retreats or to compete for cave possession. In suitable temperatures, crayfish manifest foraging behavior as their means of exploration. They become more random in searching for the global solution through temperature adjustment. The following sections detail the process of COA.

3.1.1 Initialization

The COA randomly initialize the population X of N candidate solutions each of which with *dim* dimensions. The position of each solution $X_{i,j}$ is modeled as:

$$X_{i,j} = Lb + (Ub + Lb) \times rand \tag{1}$$

where Lb and Ub refer to the limit bounds of each of the dimension j.

As previously mentioned, temperature is a crucial factor in multiple phases of the crayfish and has been defined in Eq. (2). When the temperature exceeds 30 degrees, the crayfish relocates to a cooler area for its summer retreat. When the temperature is suitable, the crayfish initiates its foraging habit. The temperature range for foraging behavior is specified as 15 to 30 degrees. Therefore, the foraging behavior can be replicated using a normal distribution, which is influenced by the temperature. The mathematical representation of this relationship is presented in Eq. (3).

$$\text{temp} = \text{rand} \times 15 + 20 \tag{2}$$

$$p = C_1 \times \left(\frac{1}{\sqrt{2 \times \pi} \times \sigma} \times \exp\right)\left(\frac{(\text{temp} - \mu)^2}{2\sigma^2}\right) \tag{3}$$

where the temperature of the crayfish location is denoted by *temp* while μ refers to the temperature of the best crayfish. In addition, σ and C_1 parameters control the different temperatures of crayfish.

3.1.2 Summer resort phase

In this phase, if the temperature is larger than 30 degrees, then the crayfish enjoys the summer break at the cave X_{shade} which is formulated as:

$$X_{\text{shade}} = X_{\text{gbest}} + X_{\text{local}} \tag{4}$$

where the best position obtained is denoted by X_{gbest} and the current population position is defined as X_{local}. The crayfish usually compete and fight for the cave which is modeled as a random process. On the other hand, if $rand < 0.5$, then no competition among the crayfishes is hold and the crayfish directly hold the cave as follows:

$$X_{i,j}^{t+1} = X_{i,j}^{t} + C_2 \times \text{rand} \times (Xshade - X_i, j_t) \tag{5}$$

$$C_2 = 2 - \left(\frac{t}{\text{Max}_{\text{iter}}}\right) \tag{6}$$

where $(t + 1)$ in the next updated position of the crayfish, t is the present iteration defined and the maximum number of iterations is expressed by Max_{iter}.

3.1.3 Competition phase

In this phase, the crayfish fight for the possession of the cave. If the temperature is over 30 degrees and the random value is greater than 0.5, the other crayfish are attracted to the same cave. Consequently, they engage in conflict with one another in order to obtain possession of the cave, as indicated by Eq. (7).

$$X_{i,j}^{t+1} = X_{i,j}^{t} - X_{z,j}^{t} + X_{\text{shade}} \tag{7}$$

$$z = \text{round}(\text{rand} \times (N - 1)) + 1 \tag{8}$$

where the total number of population's agents is defined by N.

3.1.4 Foraging phase

In this phase, the crayfishes start the process of searching for food (optimal solution). Hence, when the temp is less than 30 degrees, the crayfish start searching for food at different locations. The location and size of food is formulated as:

This process simulates the process of searching for the optimal solution for a problem

$$X_{\text{food}} = X_G \tag{9}$$

$$Q = C_3 \times \left(\frac{\text{fit}_i}{\text{fit}_{\text{food}}}\right) \tag{10}$$

3.2 Chaotic maps

Chaos refers to the inherent unpredictability exhibited by a complex system. A chaotic map is a mathematical function that is used to associate or map chaotic behavior to a parameter in an algorithm. Chaotic maps are commonly employed in optimization issues because of their ergodic properties. It facilitates the dynamic exploration of the search space at a faster rate compared to stochastic searches that primarily depend on probability. Substituting the stochastic elements in meta-heuristic algorithms with chaotic maps rather than conventional probability distributions can provide benefits. The use of chaos techniques enhances the ability to search for a global optimum by overcoming getting stuck in local optimal values. This study considers 8 one-dimensional chaotic maps, defined in Fig. 1, in order to enhance the basic COA.

3.3 Levy flight

A Lévy Flight is a type of arbitrary walk where the steps taken follow a probability distribution known as the Lévy distribution, which has tails that are heavier than those of a normal distribution. The concept of Lévy-flight was initially developed by Paul Lévy in 1937 and later further elaborated by Benoit Mandelbrot [45]. Multiple studies indicate that as animals and insects look for food, their flight behavior often exhibits a characteristic pattern of random direction selection, which can be described as a Lévy-flight. In [46], Reynolds et al. investigated the movement patterns of fruit flies as they navigated their environment using a series of straight paths that were interrupted by a sudden 90° turn. This resulted in a search pattern known as a scale-free intermittent Lévy-flight. In [47], the authors have shown that Lévy-flight can be utilized to replicate particular light phenomena. The Lévy Flight can be defined as a stochastic process where the variance is infinite, and the mean is given by Eq. (11).

$$\mathrm{Levy}(\gamma) \sim \mathrm{u} = \mathrm{t}^{-1-\gamma}, (0 < \gamma \leq 2) \tag{11}$$

The Mantegna method is an effective approach for producing random step lengths that exhibit behavior similar to Lévy flights.

$$s = \frac{\mu}{|v|^{\frac{1}{\gamma}}} \tag{12}$$

where v and μ are normal distribution with $\mu \sim \mathrm{N}\left(0, {\sigma_\mu}^2\right)$ and $v \sim \mathrm{N}(0, {\sigma_v}^2)$ and $\gamma = 1.5$

$$\Gamma \text{ is a gamma function } \sigma_\mu = \left[\frac{\Gamma(1+\gamma) \times \sin\left(\pi \times \frac{\gamma}{2}\right)}{\left(\Gamma\left[\frac{(1+\gamma)}{2}\right] \times \gamma \times 2^{\frac{(\gamma-1)}{2}}\right)}\right]^{\frac{1}{\beta}} \tag{13}$$

where $\Gamma(\mathrm{z}) = \int_0^\infty t^{z-1} e^{-t} dt$.

4 Proposed framework

This section presents the proposed framework consisting of three main phases as shown in Fig. 2: preprocessing, hybrid feature selection, and prediction. The preprocessing phase includes the normalization task and clustering with the CFS task. The goal of this phase is to normalize the values of the dataset features, then to cluster the dataset into different clusters. The groups/ clusters help to extract the hidden information in the dataset. After that, the CFS is applied to reduce each cluster's features, presenting a new reduced dataset. In the second phase, a hybrid feature selection approach is proposed, which implements two stages of filter and wrapper stages. The most relevant

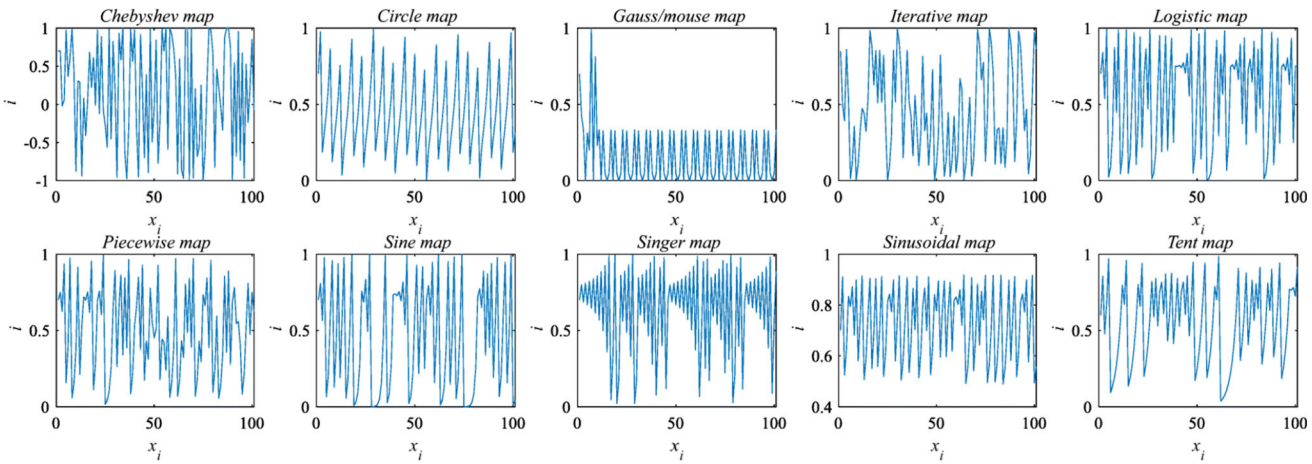

Fig. 1 The different types of chaotic maps theory

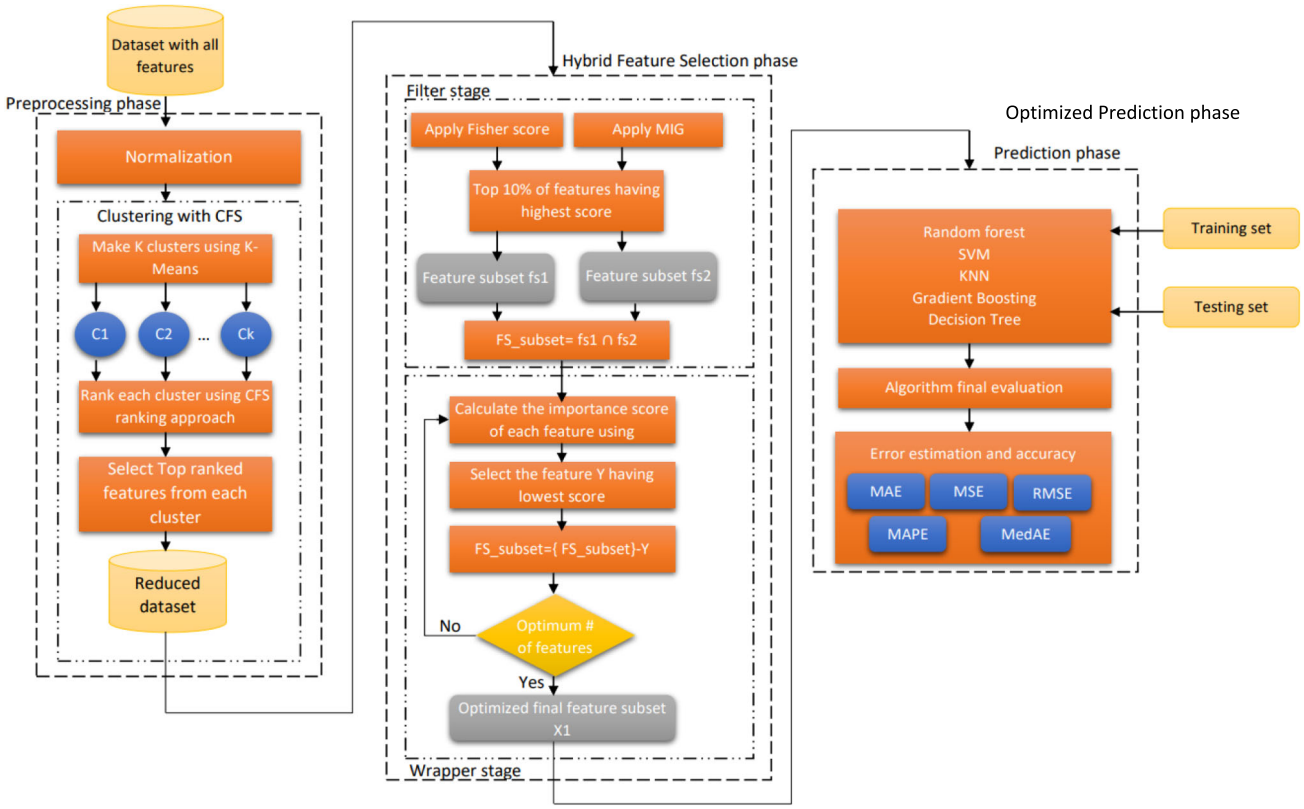

Fig. 2 The proposed framework phases

features are selected for the prediction phase in this phase. In the prediction phase, a novel variant of COA algorithm is proposed to optimize the different hyperparameters of different machine learning models particularly SVR. This is because the manual tuning of hyperparameters may not lead to a more promising solution.

4.1 Pre-processing phase

This section presents the preprocessing phase components. The preprocessing phase consists of two main stages, normalization, and clustering with CFS stages. Normalization is applied as a preprocessing stage to make the values of different features in a specified range. Feature values may be of various ranges, so normalization is used. Secondly, clustering is applied to cluster the dataset into groups of similar patterns and rank the features of each cluster using CFS ranking. Top-ranked features from each cluster are chosen to form a new minimized dataset for different phases of the proposed framework. The details of each stage of the preprocessing phase are discussed in the following subsections.

4.1.1 Normalization

Normalizing values is necessary for processing data. To acquire values associated with another variable, some normalization forms require merely a rescaling step. When we have data on the size of a crop's population, we can correct the mistakes. The population values can be regularly distributed instead of randomly distributed once the inaccuracies are corrected. Getting the z-score is the initial step in normalization. The z-score can be written as:

$$z = [(x - \mu)/\sigma] \tag{14}$$

where means of the crop population and standard deviations of the crop population are denoted by μ and, respectively.

4.1.2 Clustering with CFS

In this phase, we apply a preprocessing strategy that combines KM clustering and CFS ranking to deal with the high dimensionality of the input data. The KM approach is used on raw data to generate the initial '**k**' number of clusters. The suggested method is very similar to the traditional ones in that the 'k' number of clusters is predefined each time, with 'k' values of 8, 10, and 12 being considered. Each cluster's data is then ranked using CFS rating

and sorted in ascending order. A minimized dataset is obtained by choosing the top CFS-ranked features from each cluster. This reduced new dataset is sent to the next phase since it has less redundancy. KM clustering is used on training data to identify commonalities and create subsets. The theory behind this strategy is that clustering can help bring to light previously hidden information and highlight the underlying data structure that was not apparent before grouping.

The cluster analysis output can serve as a valuable guide when extracting and prioritizing critical features from many clusters. Using a KM approach, the training data is partitioned into k groups. The center of each cluster can be determined by taking the meaning of the data inside it. Choosing how many clusters should be created or what value k would have been a crucial challenge in the KM clustering process.

After the clusters have been generated, a CFS filter approach is applied to select the crucial features for each cluster's resulting minimal subset dataset. CFS uses statistical metrics to evaluate feature subsets as part of its filtering procedure. When one feature is highly correlated with another, it is considered redundant. In this case, the features are discrete assessments of the variable of interest's distinctive qualities. Given a set of features, if the relationship between each extrinsic variable and feature is known, and the relationships between all other pairs of features are also known, then Eq. 15 can be used to determine the relationship between the complex test, which includes all features, and the extrinsic variable,

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{15}$$

The formula mentioned above characterizes Pearson's correlation coefficient (PCC). Here, $\overline{x}$ and x_i define the mean and actual values related to the features under consideration. The average and actual values of the dataset class are denoted by $\overline{y}$ and y_i, respectively.

As the degree of similarity between the features and the classes grows, so does the importance of the resulting feature set. Furthermore, the overall feature-output correlation is denoted by $r_{ny} = p(X_n, Y)$, while the overall feature-feature correlation is denoted by $r_{nn} = p(X_n, X)(X_n, X_n)$. The relative importance of the chosen features is calculated using Eq. 16,

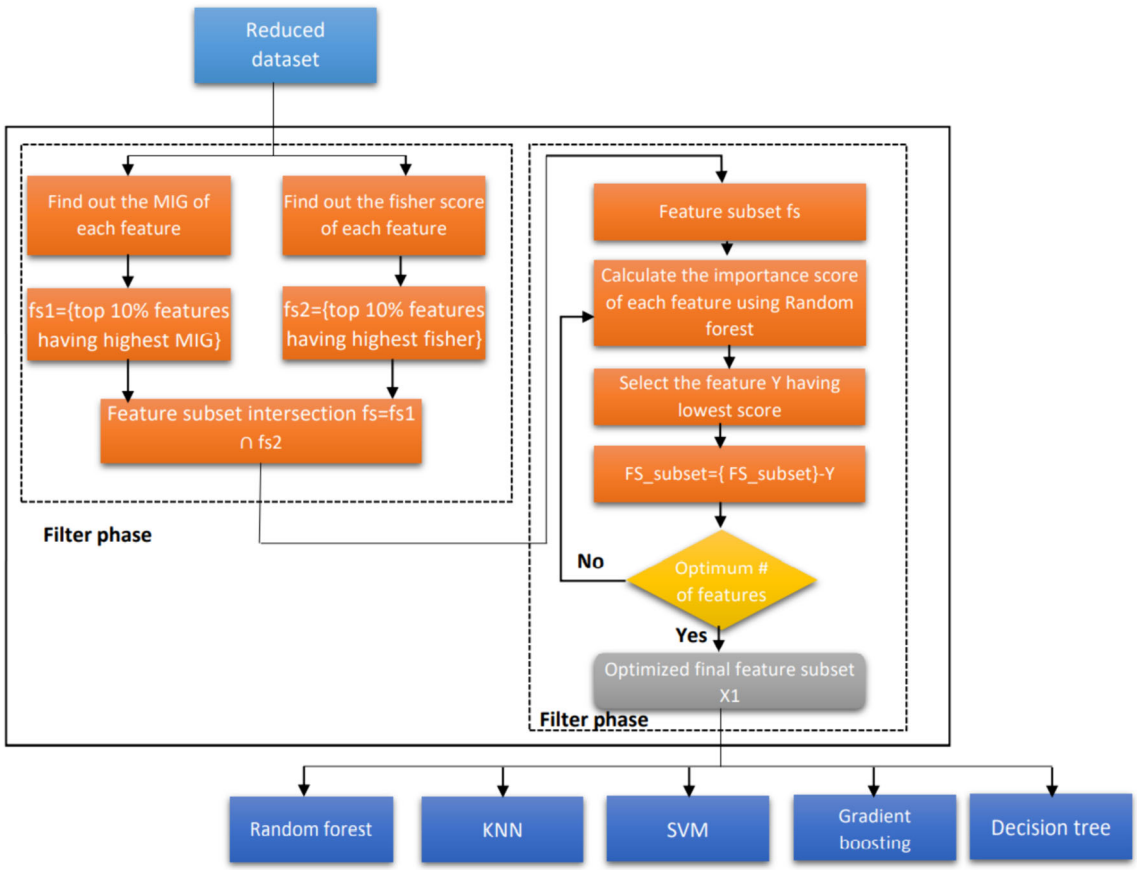

Fig. 3 The proposed hybrid feature selection approach

$$J(X_n, Y) = \frac{nr_{ny}}{\sqrt{n + (n-1)r_{nn}}} \quad (16)$$

This indicates that the relationship between an external feature and a group grows more vital as the size of the group increases. The steps for selecting from each cluster using the CFS filter are presented in pseudo-code form in Algorithm 1.

	Algorithm 1: Cluster-based-CFS
	Input: cluster dataset D_{train}, predictor P, number of features n
	Output: Selected feature set F_χ
1	def CFS (D_{train}, P, n):
2	$F_0 = \emptyset$
3	X = 1
4	**While** $\lvert F_x \rvert < n$ do
5	**if** $\lvert F_x \rvert < n - 1$ **then**
6	$F_x = CFS(D_{train}, P, F_{x\text{-}1})$
7	**else**
8	add the best ranked feature $\overline{f}$ to $F_{x\text{-}1}$
9	**end if**
10	x=x+l
11	**end while**
12	**return** F_x
13	**End def**

4.2 Hybrid feature selection approach

Because feature selection is optimized for the employed learning algorithm, wrappers typically produce better results than filter procedures [48]. This makes wrappers challenging to work with and prohibitively costly to execute, especially for large databases with many features.

In most cases, filters are faster to execute than the wrapper, making them a much more attractive option for scaling to extensive databases with many features. A filter can supply an intelligent initial feature subset for a wrapper when higher precision is desired for a specific learning algorithm. Wrapper and filter methods are combined in the proposed methodology to use the filter method's ranking data. In this approach, we take full advantage of filters and wrappers. Combining filter and wrapper approaches may boost the former's prediction accuracy while reducing the latter's runtime.

The proposed hybrid feature selection approach uses a Reverse Feature Elimination RFE with support vector machine predictor (RFE) as a wrapper approach and two filter approaches to form a hybrid approach to choose the most relevant subset of features.

Given the reduced dataset obtained from preprocessing phase, First, the top 10% of features are chosen based on their predictive quality, using a combination of the two feature ranking approaches, fisher and Mutual Information Gain (MIG). During this phase, features that are unlikely to classify samples in the dataset are eliminated. With filter-based procedures, the optimum feature set is found by obtaining the intersection of two feature sets generated by distinct feature rankers. The feature set obtained by filter methods is then wrapped using a wrapper technique, which is used to eliminate redundant features and boost accuracy.

Recursive Feature Elimination (RFE) is a cyclical procedure that removes features in reverse. Using a learning algorithm, this technique ranks feature subsets to produce a superior feature set. Figure 3 depicts the proposed hybrid feature selection approach's architecture. Each step of the proposed approach is outlined in greater detail below.

4.2.1 Feature ranking approaches—filter stage

It is crucial for classification that the best features be identified. On the other hand, classification algorithms benefit little from using high dimensions because they encompass all features and lengthen the processing time. Fisher Score and mutual information gain feature selection methods are popular supervised feature selection methodologies. By picking the most valuable features, the Fisher Score and MIG can be utilized to minimize the number of dimensions being considered drastically.

The features in the reduced dataset obtained from clustering with the CFS phase are ranked and sorted using the fisher score and MIG simultaneously. The output of each ranking approach is a feature subset with ranked features. The output of each ranking approach may be different. Therefore, the top-ranked 10% of features from each feature subset are chosen. The intersection between the top selected features is obtained to form a new final reduced dataset. Then, the new dataset is fed into the wrapper-based phase to select the optimal features for final prediction.

4.2.2 Recursive feature elimination (RFE)—wrapper stage

While filter approaches choose the most informative features to include in a feature subset, they do not consider the correlation between features in making their decision. Wrapper techniques can optimize the feature subset by employing a learning algorithm in the feature selection process. Compared to filter approaches, these are more costly because updating the feature subset necessitates re-creating the classification algorithm. Therefore, using a filter with a wrapper technique for feature selection can lead to a superior feature size and accuracy solution. Recursive feature elimination is a ranking method that iteratively removes features based on their importance. RFE ranks feature using a greedy approach.

RFE will always start by removing the minor essential features from the obtained feature set from the previous stage. The relative importance of particular defining features can change drastically when evaluating beyond a different subgroup of features throughout the step-by-step elimination phase, making the recursion necessary. One example is the ensemble-based classification or prediction method known as the random forest.

An absolute rule is applied to selecting the various predictors, and the resultant prediction is made for the supplied dataset. In addition, different trees are created using subsets of the training set to ensure that the insights they yield are independent and novel. Furthermore, the algorithm incorporates random chance in its quest for optimal splits [49], so there is a possible variation between the trees. Applications of random forests to this issue will dictate how far the wrapper stage of feature selection goes. Since every tree in a random forest is built from a bootstrap sample, some portions of the feature set are not used during training. This subset, known as out-of-bag, provides an objective way to quantify error projections.

The RFE-SVM feature selection is described in Algorithm 2. The RFE-SVM method relies on a recursive feature elimination procedure, the significance of which is proportional to the number of identifying features that are eliminated. Therefore, the idea behind an RFE-SVM wrapper model is that it's best to build a model consistently and then choose the best or worst feature. Once that step is complete, you can repeat it with the remaining steps for each feature. Until all the features in the reduced dataset have been used, this process will continue. After each feature is discarded, its place in the ranking is determined by its elimination order. On the other hand, it uses a greedy optimization search to find the optimal features. Algorithm 3 presents the overall steps of the proposed hybrid feature selection method components.

	Algorithm 2: RFE-SVM-Wrapper approach
	Input: reduced dataset D_{train}, set of n features F= [f1, f2, …, fn], Ranking method M(D_{train}, F), subset of features Fs= [1, 2, 3, …, m]
	Output: Final ordered optimal features X_1
1	def RFE-SVM-Wrapper (D_{train}, F, Fs):
2	Fs= [1, 2, …, m]
3	X_1 = []
4	**While** $Fs \neq []$ **do**
5	**for** x *in* [1:n]
6	Ranking features set using M(D_{train}, F)
7	$Fs(f^*) \leftarrow$ F's last ranked feature
8	X_1=(n-x+1) $\leftarrow Fs(f^*)$
9	$Fs(x_1) \leftarrow Fs(x_1) - Fs(f^*)$
10	**end for**
11	**end while**
12	**return** X_1
13	**End def**

	Algorithm 3: Proposed FMIG-RFE-SVM
	Input: $ReducedD_{train}$, feature subset F with size N
	Output: Final optimal features F_{final}, MAE, RMSE, MSE, MedAE, MAPE, R^2
1	def FMIG-RFE-SVM ($ReducedD_{train}$, F):
2	$F_{final} \leftarrow \emptyset$ //
3	`// Filter phase using both Fisher Score and Mutual Information Gain`
4	k=[0.1, 0.2] * N
5	**while** $i \leq N$ **do**
6	$FS_1 = FS$ — MutualInfoGain(i)
7	$FS_2 = FS$ — Fisher(i)
8	$i = i + 1$
9	**end while**
10	**foreach** $f \in FS_1, FS_2$ **do**
11	Feature set $Y1$ = *top* 10% with highest MIG $\in FS_1$.
12	Feature set $Y2$ = *top* 10% with highest Fisher $\in FS_2$.
13	**end foreach**
14	a $\leftarrow$ $\lvert Y1 \rvert$
15	b $\leftarrow$ $\lvert Y2 \rvert$
16	Y $\leftarrow Y1 \cap Y2$
17	`// Wrapper phase using RFE-SVM`
18	$F_{final} \leftarrow$ RFE-RE Wrapper ($ReducedD_{train}$, F, Y).
19	Apply classifiers C $\in$ *RF*, *Boosting*, *KNN*, *DT*, *SVM* using optimal subset F_{final}.
20	Apply 5 - *fold* cross validation for classifiers.
21	Apply evaluation metrics, MAE, RMSE, MSE, R^2, MAPE, MedAE
22	**return** F_{final},*MAE, RMSE, MSE,* R^2, *MAPE, MedAE*
23	**End def**

4.3 Prediction

In this phase, multiple ML predictors are used to predict the crop yield including SVR, kNN, DR, RF and Gradient boosting. In order to get the best combination of hyperparameters for the ML predictors, a new optimization

Table 1 Training models parameters and configurations

ML model	Parameter values
DT	Min. sampling split = 2, min. leaf = 1 max. depth = 5
kNN	K = 5, Euclidean distance
RF	N_estimators = 550, min. sample split = 2, min_leaf = 1, max_depth = 5
Gradient boosting	Learning rate = 0.02, N_estimators = 550, min. sample split = 2, min_leaf = 1, max_depth = 5

algorithm is proposed. ICOA is a new variant of COA algorithm which enhanced the searching process of COA to search for the best parameter combination to reduce the prediction error defined by MSE. The proposed ICOA is used to optimize the different parameters of ML models and the results obtained indicated that SVR is the best among all other algorithms in the crop yield prediction problem. The main strategy of enhancement defined to boost the performance of COA is using combination of chaotic mapping theory and Levy flight operators. Multiple studies indicate that involving Lévy flight trajectory enhances the equilibrium between exploration and exploitation in optimization algorithms. In this paper, the Lévy flight technique is employed to further adjust the positions of the gazelles. In addition, the chaotic maps are utilized to further explore the search space of optimizing the hyperparameters and obtain more promising parameter sets for different ML models. Therefore, the combination of chaotic and Levy flight operators helps the ICOA algorithm to avoid falling into local optima and enhance the search process of ICOA by balancing both exploration and exploitation phases. The mathematical model for the new formulated combination is defined as follows:

$$\mathrm{Gazelle_i}(\mathrm{t}+1) = \mathrm{Gazelle_i}(\mathrm{t}) + ch(i) \times \mathrm{sign}[\mathrm{r}_1 - 1/2] \times L\acute{e}vy(\gamma) \tag{17}$$

$Gazelle(t)$ indicates the position of the ith gazelle at the tth iteration, $r1$ is a stochastic number between 0 and 1, and $ch(i)$ is a chaotic value obtained by the chaos map. The stochastic random walk equation, denoted as Eq. (17), assists the COA in guaranteeing that the search agent will systematically explore the search area. This is achieved by increasing the step length over time, which helps to eliminate local minima. This study incorporates the Lévy flight trajectory with the chaos map applied into the COA. Therefore, the proposed ICOA can be used to optimize the parameters of SVR ML algorithm to enrich the high performance of the prediction results.

The regression performance of SVR is highly dependent on the values of the bandwidth of the Gaussian kernel σ in the Radial Basis Function (RBF) and the penalty factor C. The SVR's regression performance is enhanced by utilizing the ICOA to optimize the bandwidth of the Gaussian kernel σ and the penalty factor C. The mean square error (MSE) between the actual values and the predictive values of SVR is employed as the fitness function of ICOA. The ICOA-SVR model algorithm is depicted below.

1. ***Step 1.*** Initialize the population of ICOA with a set of candidate parameters, set the population size and the maximum number of required iterations. Also, the LB and UB of the optimized parameters (σ and C) are set. The initial population consists of a set of candidate solutions each of which is two-dimensional solution to represent the two parameters. The initial population is generated randomly.
2. ***Step 2.*** The fitness value of each solution (set of parameters) is determined using the MSE fitness function and X_G and X_L are obtained.
3. ***Step 3.*** The algorithm update formulas are executed according to the exploration and exploitation phased based on the *temp* variables which can be either less than or greater than thirty degrees.
4. ***Step 4.*** The Levy-chaotic update position (as shown in Eq. (17)) is utilized to modify the parameters with more enhanced parameter set.
5. ***Step 5.*** At the end of each iteration, the solution with the minimum MSE is recorded which indicates the best parameter set at this iteration.
6. ***Step 6.*** Save the optimal Crayfish overall the whole iterations to represent the optimal set of σ and C that minimized the MSE fitness value.
7. ***Step 7.*** The steps 2–6 are repeated until the maximum number of iterations is reached and output the optimal solution that represents the optimal parameter set. The flowchart of the proposed steps for optimizing the SVR parameters.

Table 2 Description of dataset parameters

SN	Parameter	Parameter description	Units
1	Net cropped area	The area that has had at least one planting of the crop in a given year	Integer (hectare)
2	Gross cropped area	Total area dedicated to growing crops across all growing seasons	Integer (hectare)
3	Net irrigated area	The sum of land that has been rinsed at some point during the year	Integer (hectare)
4	Gross irrigated area	How much land has been devoted to crops watered during the year's growth seasons	Integer (hectare)
5	Area rice	Cumulative acreage devoted to rice cultivation	Integer (hectare)
6	Quantity rice	Quantity of rice grown in the region	Integer (ton)
7	Yield rice	Acquired rice amount in total	Integer (ton)
8	Soil type	The soil type in the research location was considered—2—Red soil type, 1—Medium black soil type	Integer
9	Land slope	An increase or decrease in elevation	Integer
10	Soil PH	The soil pH scale measures both alkalinity and acidity	Integer
11	Topsoil depth	The top layer of soil is where most of the microbes and organic stuff are	Integer (meters)
12	N soil	The number of nitrogen molecules in the soil	Integer (kilogram/hectare)
13	P soil	The measure of soil phosphorus content	Integer (kilogram/hectare)
14	K soil	The potassium content of the ground	Integer (kilogram/hectare)
15	QNitro	The application rate of nitrogen fertilizers	Integer (kilogram)
16	QP2O5	The ratio of phosphorus-containing fertilizers used	Integer (kilogram)
17	QK2O	Use of potassium-based fertilizers in quantities	Integer (kilogram)
18	Precipitation	Condensation of atmospheric water vapor, or precipitation	Integer (millimeter)
19	Potential evapotranspiration	How much water evaporates from a given region given a sufficient supply	Integer (millimeter/day)
20	Reference crop evapotranspiration	The rate of evaporation and transpiration from an irrigated crop reference surface	Integer (millimeter/day)
21	Ground frost frequency	The total number of days that the soil temperature in the upper layer has been below the water freezing point	Integer (number of days)
22	Diurnal temperature range	Temperature variation between daily high and low	Integer (C)
23	Wet day frequency	The total number of days with rainfall of 0.2 mm or more	Integer (number of days)
24	Vapour pressure	Thermodynamic equilibrium is maintained thanks to the pressure exerted by water vapor in its condensed phase	Integer (hectopascal)
25	Maximum temperature	The highest measured ambient temperature	Integer (C)
26	Minimum temperature	The air temperature was the coldest ever measured	Integer (C)
27	Average temperature	The typical level of air temperature	Integer (C)
28	Humidity	Levels of atmospheric water vapor	Integer (percentage)
29	Wind speed	the velocity of the wind	Integer (miles/hour)
30	Aquifer area percentage	Groundwater transmission capacity is a fraction of an area bounded by a body of porous rock	Integer (percentage)
31	Aquifer well yield	The quantity of water extracted from an aquifer employing pumping	Integer (liters/minute)
32	Aquifer transmissivity	The amount of water that can be distributed horizontally if the aquifer were completely saturated throughout its whole thickness	Integer (meter2 / day)
33	Aquifer permeability	The rate at which fluids can move through a rock is a measure of this attribute	Integer (meter/day)
34	Post-electrical conductivity	We have a standardized electrical conductivity of groundwater after rain	Integer (siemens/meter)

Table 2 (continued)

SN	Parameter	Parameter description	Units
35	Pre–electrical conductivity	Standard groundwater electrical conductivity before the monsoons	Integer (siemens/meter)
36	Groundwater post-calcium	The typical calcium content of groundwater after rainfall	Integer (milligram/Liters)
37	Groundwater pre-calcium	Specific groundwater calcium content before the monsoons	Integer (milligram/Liters)
38	Groundwater post-magnesium	Groundwater magnesium levels are about average after a monsoon	Integer (milligram/Liters)
39	Groundwater pre-magnesium	Typical levels of magnesium in groundwater before the monsoon season	Integer (milligram/Liters)
40	Groundwater post-sodium	the specific concentration of sodium in groundwater after a monsoon	Integer (milligram/Liters)
41	Groundwater pre-sodium	Average sodium concentration in groundwater before the monsoons	Integer (milligram/Liters)
42	Groundwater post-potassium	Potassium concentration in groundwater, on average, following a monsoon	Integer (milligram/Liters)
43	Groundwater pre-potassium	Potassium concentration in the earth was about average before the monsoons hit	Integer (milligram/Liters)
44	Groundwater post-chloride	Chloride concentration in groundwater, on average, following a monsoon	Integer (milligram/Liters)
45	Groundwater pre-chloride	Level of chloride in groundwater, typically before the monsoons	Integer(milligram/Liters)

Table 3 Years of testing datasets with cross-validation

Fold	Test Data	Training Data	R^2 Score	Correlation value
1	2001 – 2003	1996 – 2000	0.814	0.766
2	2004 – 2006	1996 – 2003	0.854	0.839
3	2007 – 2009	1996 – 2006	0.70 1	0.608
4	2010 – 2012	1996 – 2009	0.833	0.758
5	2013 – 2016	1996 – 2012	0.882	0.858

5 Experimental results

In this section, a set of experiments are conducted to evaluate the performance of the proposed framework including the proposed hybrid feature selection and prediction phases.

5.1 Simulation environment

The crop yield forecast model was set up in Python. The Python version was "PYTHON 3.7", and the processor was "Intel(R) Core(TM) i7-10750G7 @ 2.40 GHz". Furthermore, the system contained 32.0 GB of RAM. A set of ML models are used for the evaluation of the ICOA approach including RF, kNN, SVM, Gradient Boosting and DT. The training models parameters and configurations are set according to Table 1.

5.2 Dataset description

With vast increases in both human intelligence and the availability of suitable tools, the field of machine learning has exploded in recent years, allowing for the development of novel circumstances to ascertain, evaluate, and value information-pervasive approaches in agricultural contexts.

The dataset for this study was collected from official Indian government websites belonging to several agricultural ministries. Three primary keys are used to piece together the data: states, years, and crops. There needs to be a lot of data for feature selection methods to work well. Data with flexible features simplify discovering patterns by filtering out details that aren't pertinent to the study's goals.

This section describes the data set utilized to make predictions about crop yield in the study. Factors such as rainfall, crop type, market price, and yield are collected to create a dataset that can forecast whether or not a crop will be profitable. Data from many sources is gathered, filtered, and combined using Python. Eands. dacnet.nic.in. [50], Agmarknet [51], and Mospi.nic.in. [52] are among the sources used.

Paddy crop production prediction in the Vellore district of southern India is the focus of the planned study. Ponnai, Sholinghur, Arcot, Thimiri, Ammur, and Kalavai are all part of the study's geographical focus. Since paddy is a significant cash crop in the area, it makes sense to look into the economy there. The information includes non-typical meteorological and soil features, such as the characteristics of the groundwater used by the crops and the amount of fertilizer applied to them. Parameters such as evapotranspiration, wet day frequency, groundwater nutrients, and aquifer features were examined in this study. Brief details regarding the study's many crop parameters can be found in Table 2.

A combination of paddy output (tonnes) cultivated area (hectares), and yield acquired (kg/hectare) is used to calculate the estimated paddy crop yield. Regular climatic factors were used, such as reference crop evapotranspiration, mean temperature, humidity, potential evapotranspiration, and precipitation. In contrast, unique climatic data such as diurnal temperature range, ground frost frequency, and wind speed were also considered. The climate information comes from the Indian Meteorological Department's online platform, metadata. Topsoil density, soil macronutrients and Soil pH are all examples of soil parameters. The analysis considers the many hydro chemical characteristics of groundwater, such as its permeability, aquifer type, transmissivity, electrical conductivity, and pre- and post-monsoon micro-nutrient (chloride, magnesium, potassium, calcium, and sodium) content. Table 3

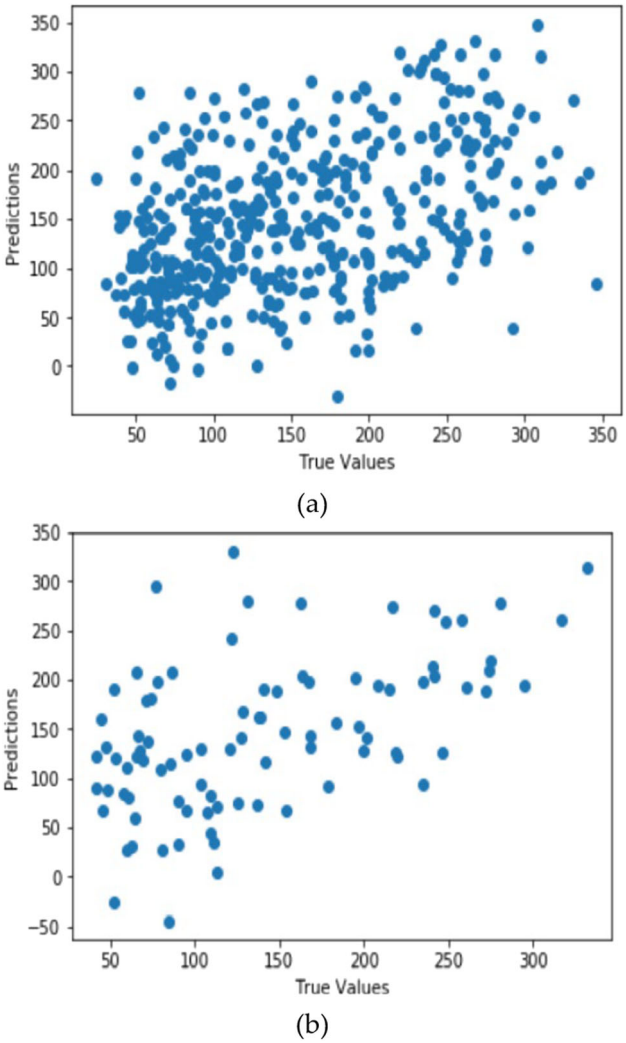

Fig. 4 The proposed forecasting model's predicted values are compared to actual values. **a** After cross-validation, **b** Before cross-validation

Table 4 Evaluation of ML models' performance using all features of the dataset

ML Model	The Efficiency Metric Using All Dataset Features					
	MAE	MSE	RMSE	R^2	MAPE (%)	MedAE
Support Vector Machine	0.291	0.093	0.304	0.452	20	0.330
Random Forest	0.301	0.097	0.311	0.415	33	0.340
Decision Tree	0.306	0.101	0.317	0.401	38	0.346
k-Nearest Neighbor	0.356	0.143	0.379	0.353	45	0.369
Gradient Boosting	0.304	0.103	0.321	0. 409	36	0.266

Table 5 Evaluation of ML models' performance using the inherent approach of feature importance

ML Model	The Efficiency Metric Using Inherent feature importance					
	MAE	MSE	RMSE	R^2	MAPE (%)	MedAE
Support Vector Machine	**0.276**	**0.089**	**0.298**	**0.472**	**21**	**0.316**
Random Forest	0.280	0.091	0.302	0. 441	29	0.320
Decision Tree	0.326	0.121	0.348	0.382	45	0.366
k-Nearest Neighbor	0.406	0.169	0.411	0.478	51	0.426
Gradient Boosting	0.286	0.095	0.309	0.427	33	0.326

Table 6 Evaluation of ML models' performance using the proposed FMIG-RFE-SVM approach

ML Model	The Efficiency Metric Using the FMIG-RFE-SVM approach					
	MAE	MSE	RMSE	R^2	MAPE (%)	MedAE
Support Vector Machine	**0.194**	**0.039**	**0.196**	**0.542**	**20**	**0.196**
Random Forest	0.238	0.060	0.245	0.415	35	0.230
Decision Tree	0.272	0.075	0.274	0.403	40	0.286
k-Nearest Neighbor	0.316	0.102	0.319	0.384	45	0.330
Gradient Boosting	0.252	0.064	0.253	0. 409	38	0.266

Table 7 Effectiveness of a hybrid FMIG-RFE-SVM approach for machine learning models

ML Model	Accuracy with All Features in the Dataset (%)	Accuracy with Inherent feature importance Method (%)	Accuracy with proposed FMIG-RFE-SVM approach (%)
Support Vector Machine	90.36	91.36	91.98
Random Forest	86.22	87.34	88.91
Decision Tree	79.25	81.35	83.46
k-Nearest Neighbor	78.21	79.61	80.69
Gradient Boosting	84.22	85.33	86.77

Table 8 Optimized Prediction models evaluation without using Hybrid FMIG-RFE-SVM as feature selection

ML Model	MAE	MSE	R^2	MedAE
ICOA-SVM	0.182	0.073	0.493	0.280
ICOA-RF	0.220	0.077	0.421	0.320
ICOA-KNN	0.212	0.091	0.412	0.318
ICOA-DT	0.233	0.103	0.381	0.351
ICOA-Gradient	0.211	0.083	0. 427	0.360

5.3 Evaluation metrics and machine learning models

Validating the proposed framework is essential for achieving the desired results. It is implemented alongside different machine learning algorithms: Decision Tree [53], Support Vector Machine [54], k-Nearest Neighbor [55], gradient boosting [56] and Random Forest [57]. The following subsections details the utilized evaluation metrics to assess the different components of the proposed model. The evaluation map consists of many folds including: the evaluation of the hybrid FS approach, the evaluation of the proposed optimized prediction phase and the evaluation of the full framework.

5.3.1 Metrics of evaluation

Model performance is defined by the evaluation measures used. Measures of evaluation and performance metrics are used to determine how effective a machine learning model is. Distinguishing between the outcomes of multiple learning models is a crucial part of the evaluation metric. Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Determination Coefficient (R2), Median Absolute Error (MedAE) and Root Mean Squared Error (RMSE) are some of the

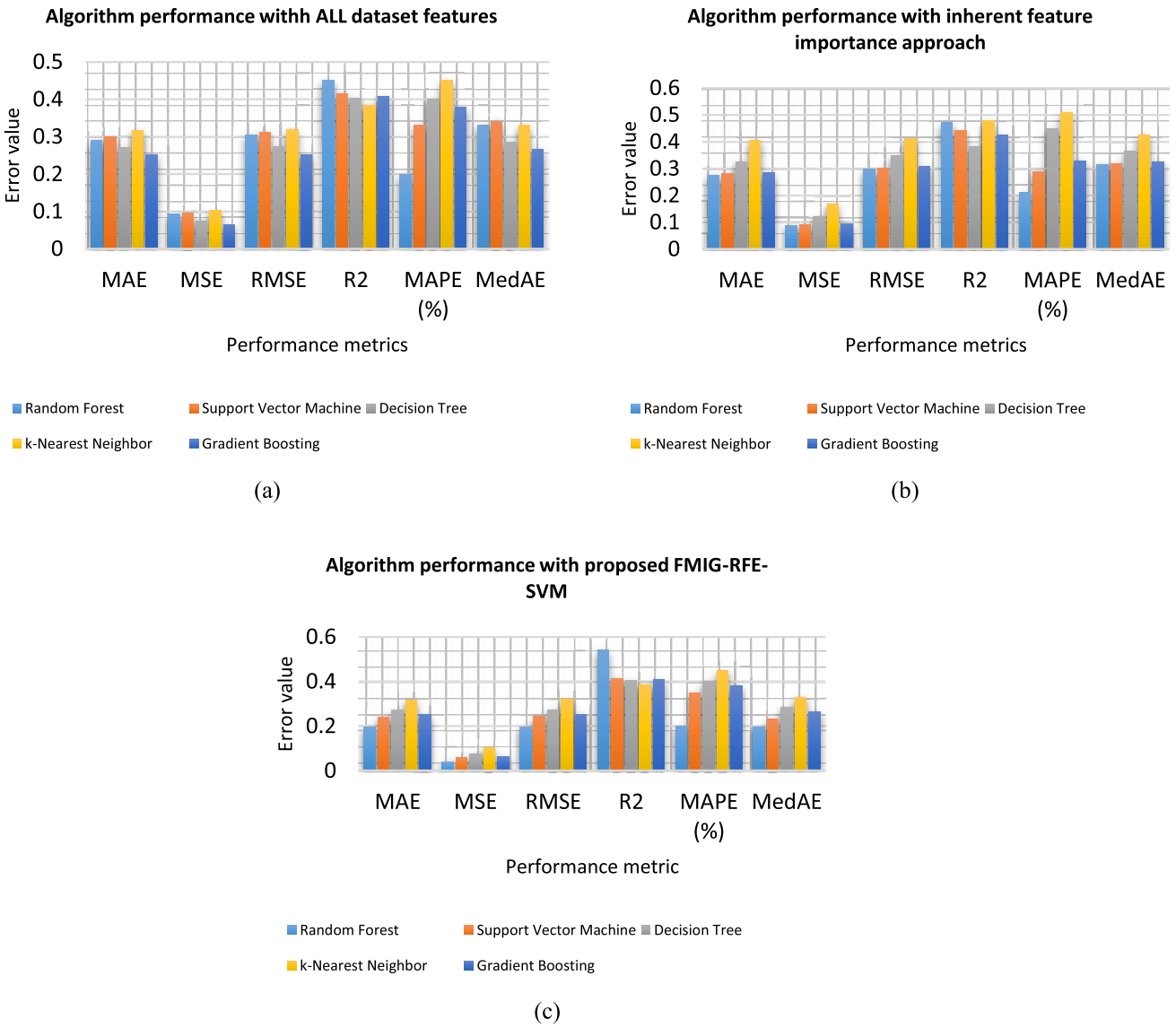

Fig. 5 Machine learning model performance metrics with **a** all features of the dataset, **b** selected features using the inherent feature importance approach, and **c** features obtained using the proposed FMIG-RFE-SVM approach

performance indicators considered in evaluating the work that has been developed.

- **Mean Absolute Error (MAE)** is a way to determine the typical significance of errors [58] by taking a set of predictions and averaging them over. According to the Equation, the mean absolute deviation from the expected value has been observed.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - y'_j\right| \tag{18}$$

where sample size, n, represents the population from which information is drawn; y_j represents the baseline measure of interest, and $y'j$ characterizes the predicted estimate of interest.

- **Mean Squared Error (MSE)** is an essential metric for evaluating an estimator's efficacy. Moreover, this characterizes how well a regressor line corresponds to the points in the dataset [59]. Mean squared error (MSE) can be calculated using the formula:

$$MSE = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - y'_j\right)^2 \tag{19}$$

- **Root mean square error (RMSE)** measures prediction uncertainty by squaring the difference between the observed and the predicted errors [60]. More specifically, it clarifies the degree to which the data is concentrated along the best fit line. Equation 6 represents the calculation of the (RMSE):

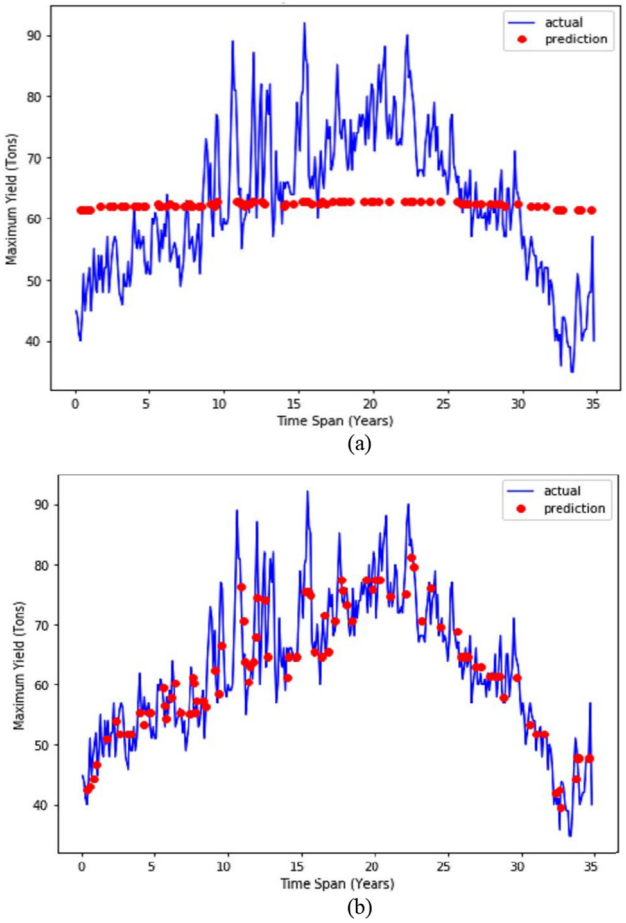

Fig. 6 The results of comparing the accuracy of the Gradient boosting with **a** all of the features in the dataset and **b** the proposed feature selection approach

$$\mathrm{RMSE} = \sqrt{\sum_{j=1}^{n} \frac{\left(y_j - y_j'\right)^2}{n}} \tag{20}$$

- **R-squared**, also known as the determination coefficient, is a statistical metric used to assess how well the regression model fits the data [61]. As such, the determination coefficient characterizes the extent to which the revised framework outperforms the original. Equation 18 provides the following definition:

$$R^2 = \left(\frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} \right)^2 \tag{21}$$

- **Mean Absolute Percentage Error (MAPE)** measures how far the model's forecast differs from the actual result. Simply put, it represents an average of the various percentage errors. It's calculated by dividing the total absolute error by each timeframe separately. As shown in Eq. 19, its definition is as follows:

$$\mathrm{MAPE} = \frac{1}{n}\sum_{j=1}^{n} \frac{y_j - y_j'}{y_j} \tag{22}$$

- **Median Absolute Error (MedAE)** is an attractive metric because it holds up well in the face of extremes. When determining the loss, the middle value of all absolute discrepancies between the prediction and the target is used. The median absolute error estimated over n samples is defined as follows, where $\hat{y}$ represents the predicted value for sample i, and the actual value is defined as yi.

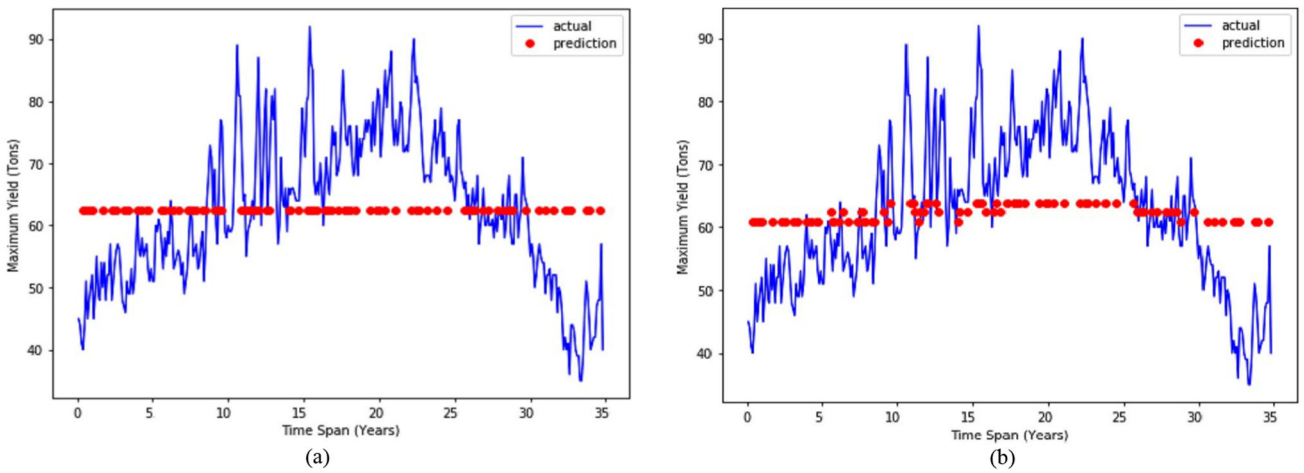

Fig. 7 The results of a comparison of the accuracy of the Random forest model with **a** all of the features in the dataset and **b** the proposed feature selection approach

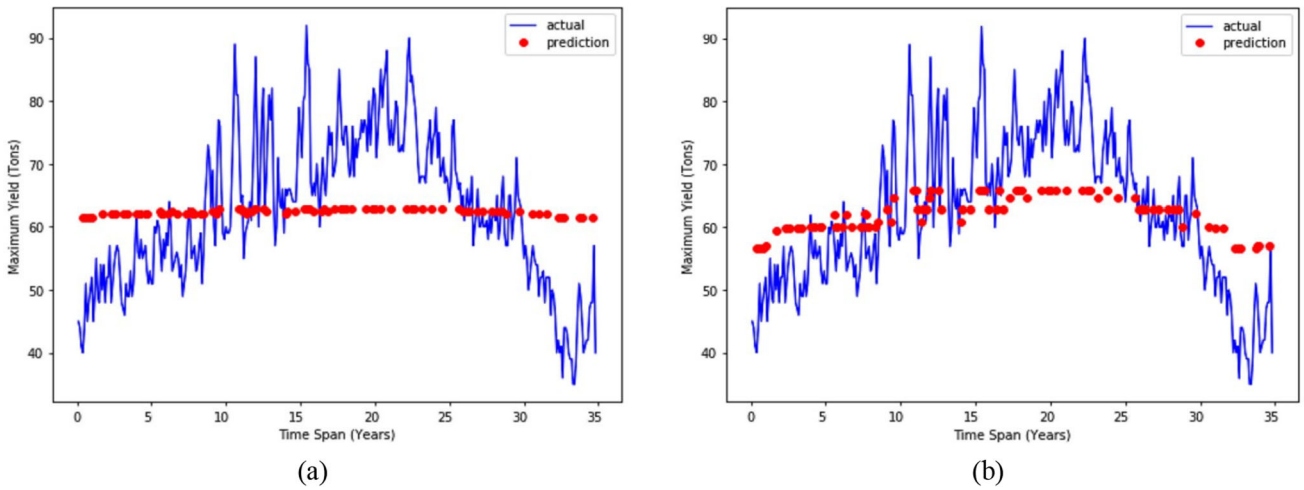

Fig. 8 The results of a comparison of the accuracy of the Decision-tree model with **a** all of the features in the dataset and **b** the proposed feature selection approach

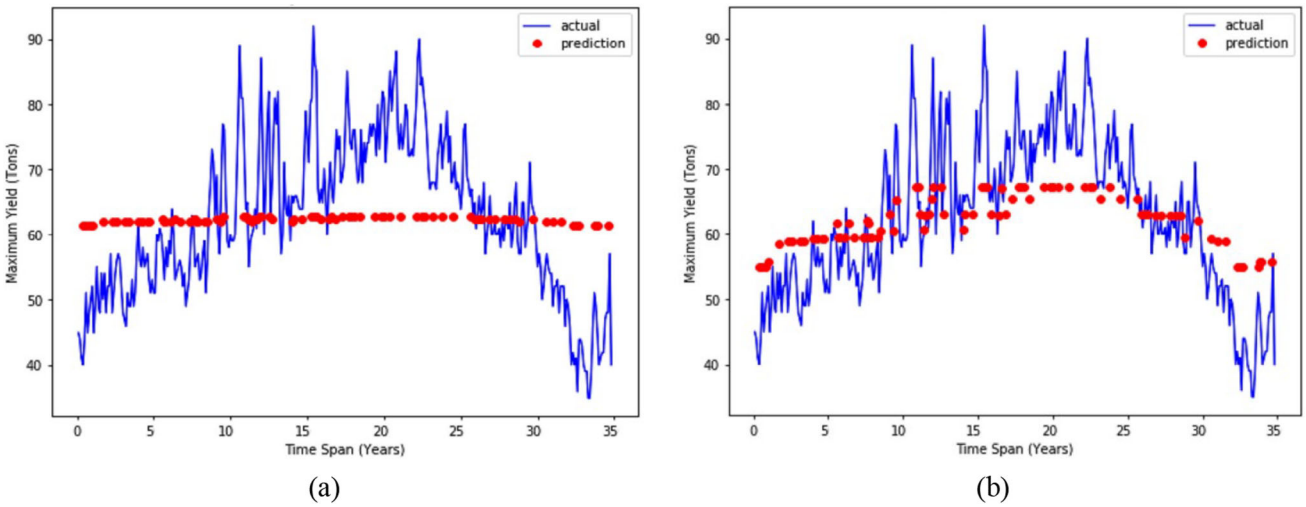

Fig. 9 The results of a comparison of the accuracy of the SVM model with **a** all of the features in the dataset and **b** the proposed feature selection approach

$$\text{MedAE}(y, \hat{y}) = \text{median}(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|) \quad (23)$$

5.3.2 Cross-validation

It is common to split the dataset into a test set and a training set when building a model of machine learning, with the more extensive set being given more weight as the model is refined. While the test dataset is minimal, there is always the chance that important data that could have helped the model was left out. High variance in the data is also a cause for concern. The method of K-fold cross-validation is used to deal with this issue. Attempting to model and predict time series data is a challenging and involved process. Cross-validation based on randomly splitting a time series does not work very well. A temporal dependency issue may arise because of the reliance on earlier observations and the possibility of leakage in lag variables due to the response variable. Because of this, the information space exhibits non-stationarity or a tendency toward fluctuating mean and variance values. In this case, a forward chaining method is more suitable for executing the cross-validation. The proposed approach builds its model on historical data and makes predictions using a five-fold cross-validation procedure. Table 4 summarizes these results.

It's similar to training on a small sample of data and then using that to make predictions about new data and assess how well those predictions hold up. The data points predicted are included in the next batch of training data, and predictions are made for the following data points. The

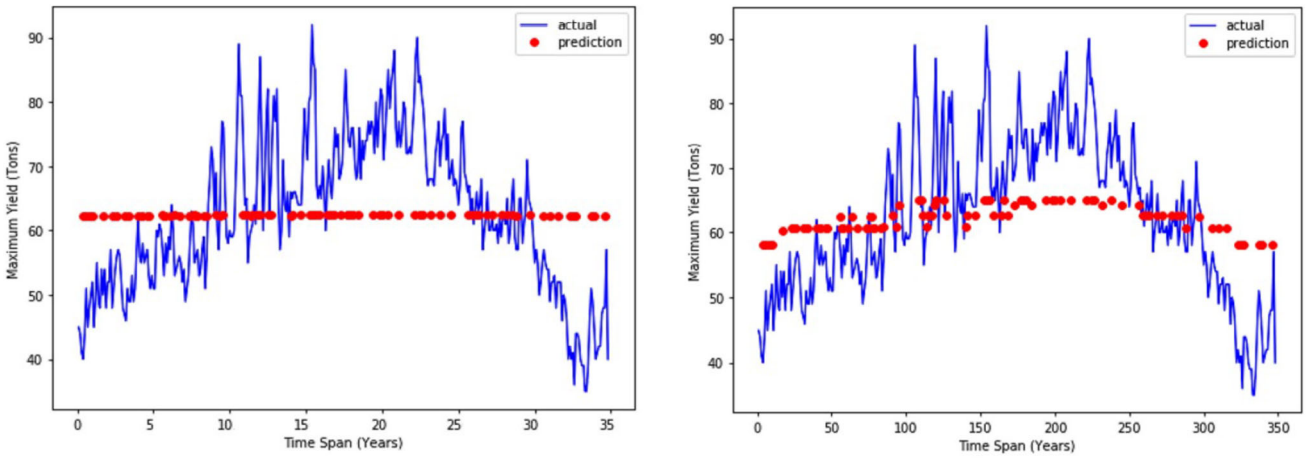

Fig. 10 The results of a comparison of the accuracy of the KNN model with **a** all of the features in the dataset and **b** the proposed feature selection approach

Table 9 Optimized Prediction model evaluation using Hybrid FMIG-RFE-SVM as feature selection

ML Model	MAE	MSE	R^2	MedAE
ICOA-SVM	0.151	0.062	0.572	0.216
ICOA-RF	0.205	0.070	0. 531	0.250
ICOA-KNN	0.201	0.088	0.472	0.306
ICOA-DT	0.206	0.098	0.518	0.316
ICOA-Gradient	0.196	0.071	0.537	0.231

Table 10 Prediction model evaluation without using Hybrid FMIG-RFE-SVM as feature selection

ML Model	MAE	MSE	R^2	MedAE
ICOA-SVM	0.151	0.062	0.572	0.216
COA-SVM	0.211	0.102	0.462	0.336
PSO-SVM	0.261	0.113	0.453	0.364
RUN-SVM	0.196	0.093	0.521	0.297
WOA-SVM	0.213	0.112	0. 489	0.268

Table 11 Prediction model evaluation without using Hybrid FMIG-RFE-SVM as feature selection

ML Model	MAE	MSE	R^2	MedAE
RF [62]	0.203	0.093	0.526	0.301
1DCNN [25]	0.193	0.085	0.513	0.296
LSTM-DBN [31]	0.194	0.101	0.561	0.243
CYPA [32]	0.166	0.099	0.566	0.231
Proposed	**0.151**	**0.062**	**0.572**	**0.216**

results of the cross-validation on the proposed method are shown in Fig. 4.

PyScikit-Learn, a machine learning library, is used to conduct cross-validation. The steps of preprocessing the dataset are performed. Sklearn's train test_split_function is brought in via the model_selection sub-library so that data can be split into test and training sets. To discover the best possible value for K, the cross_val_score function is used for fine-tuning the cross-validation hyperparameters. The data is divided into K-equal subgroups by specifying a value of 5 for the n_splits argument. In this work, we allocate 75% of the data to training and 25% to testing. K-fold cross-validation is where the error measure for the trained model is established. The R2 score is used to quantify the accuracy of the model and is refined at each iteration until an optimal value is reached.

Below is a detailed illustration of the experimental setup for the proposed hybrid feature selection approach for crop prediction, which may be used in conjunction with several different machine learning frameworks, including KNN, gradient boosting, SVM, decision trees, and random forest.

5.4 Experimental results

Here, we summarize the results of experiments that chose the proposed framework above the baseline machine learning models. The feature selection approach, which involves picking the most pertinent features from a dataset,

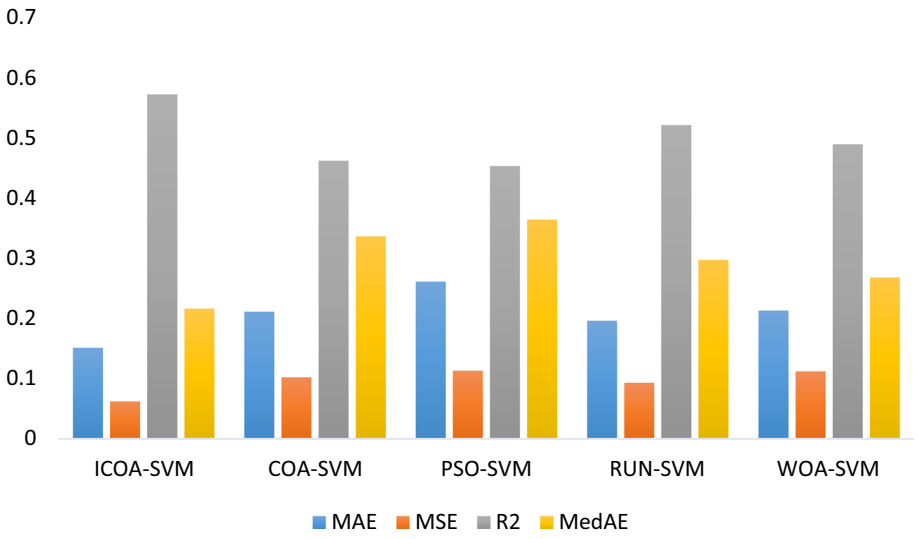

Fig. 11 Comparison between ICOA and several optimization algorithms to optimize SVM

Table 12 Prediction model evaluation without using Hybrid FMIG-RFE-SVM as feature selection

	RF [62]	1DCNN [25]	LSTM-DBN [31]	CYPA [32]	Proposed
Min	0.831	0.893	0.884	0.911	0.938
Max	0.853	0.924	0.914	0.939	0.949
Med	0.842	0.912	0.900	0.928	0.942
Mean	0.845	0.912	0.903	0.925	0.943
STD	0.025	0.036	0.022	0.029	0.011
p-value	1.63E-5	1.63E-5	2.45E-4	2.33E-3	–
Friedman rank	6.38	4.53	5.67	2.29	1.42

can boost the accuracy of a prediction model. In addition, the proposed prediction phase is optimized by using a novel ICOA algorithm to best obtain a set of parameters of ML models. To ensure that the proposed hybrid statistical feature selection technique work as intended, it is implemented with the following models of machine learning:

- Decision tree
- Random forest
- Gradient boosting.
- Support Vector machine.
- KNN

In some machine learning algorithms, a helpful built-in mechanism known as feature importance is included. These techniques are commonly used in forecasting, allowing close monitoring of the essential model variables. Depending on the situation, this data can be utilized to modify the current models by engineering new features or discarding the noisy feature data. The proposed hybrid feature selection framework is compared to this metric as one of its benchmarks. There are five stages to this model's analysis and evaluation.

First, Prediction results are verified by several statistical assessment measures once the models are built in the first step using all the features in the dataset. Second, the feature importance techniques included within the algorithm are used to create models, with only the most essential features being chosen. Third, the models are built using the proposed FMIG-RFE-SVM approach. The method proposes selecting the most critical features and assessing the predicted outcomes. Fourth, the prediction phase is evaluated using the novel ICOA compared with other optimization algorithms for optimizing the different ML parameters. Finally, the full framework is evaluated with some state-of-art approaches and models.

5.4.1 Estimating the effectiveness of hybrid proposed feature selection approach

To evaluate the efficacy of the proposed hybrid feature selection method, a set of experimental results are captured. The features obtained via the proposed feature selection methods, features gained via the built-in feature importance approach and all of the features are used to evaluate the accuracy of the various experimental models. The evaluation metrics define the effectiveness of the running model. The differences between the expected and actual values are measured by the residuals acquired in the

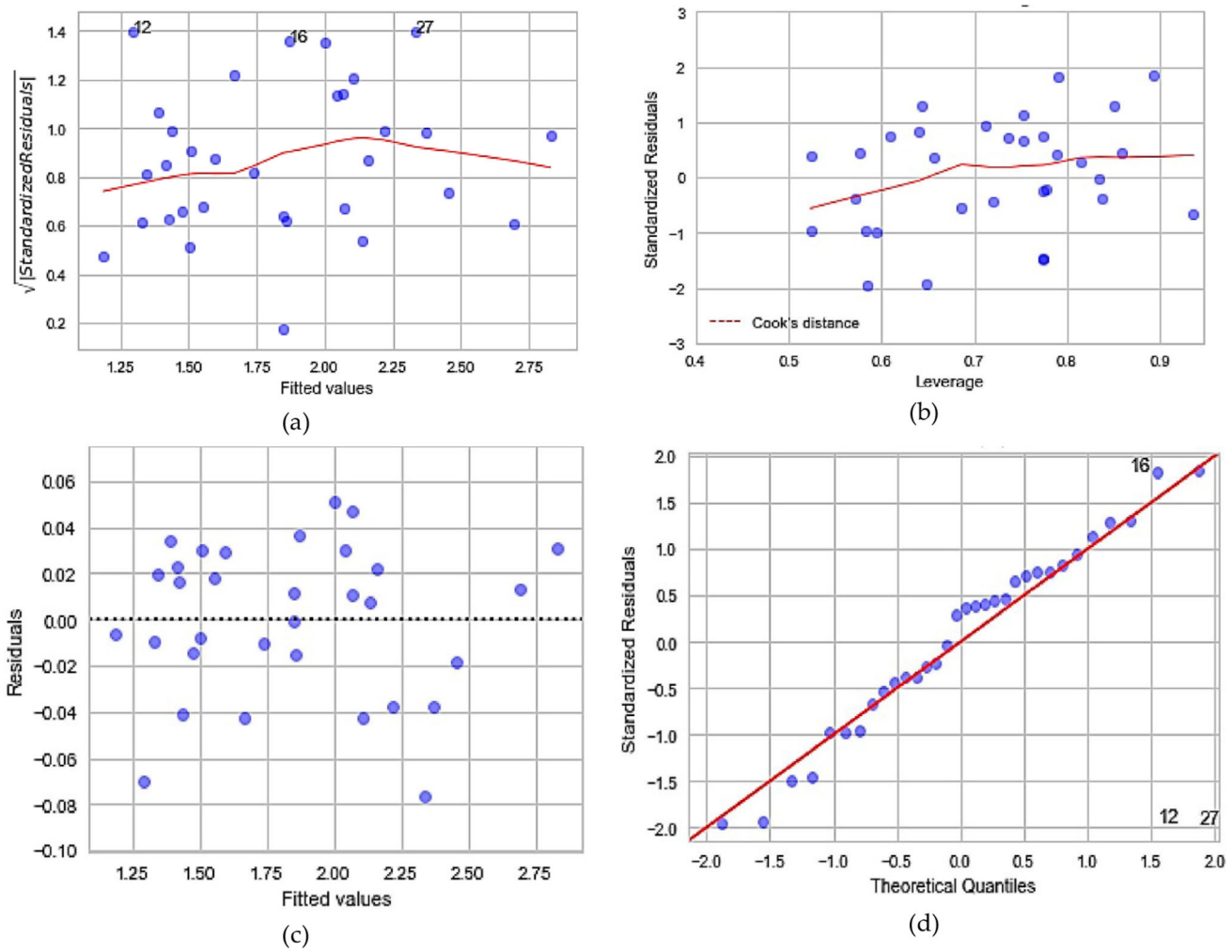

Fig. 12 Diagnostic residual plots for regression analysis. **a** the scale versus location plot, **b** the residuals against leverage plot, **c** The residuals vs fitted plot, and **d** the usual Q-Q plot

experiments. A model's efficacy and accuracy can be measured by examining the size of the residual spread. As can be shown in Tables 5, 6 and 7, the evaluation metrics attained via the proposed hybrid feature selection technique outperform those acquired via the other investigated methods.

Efficiency evaluation is a crucial part of developing a better model. For future iterations, it aids in determining the best possible framework for describing the data and putting it to use. A prediction's accuracy is evaluated by considering how close the prediction comes to the true value. It measures how often a model comes up with correct results. Accuracy measurements for the tested models are shown in Table 8 using the proposed FMIG-RFE-SVM hybrid approach for feature selection, the built-in feature importance approach, and the entire dataset.

When evaluated with the proposed feature selection approach, the results show higher accuracy. Figure 5 visually depicts the different types of machine learning models. Each was trained using either the complete set of features, the inherent feature importance approach or a subset of features generated using the proposed feature selection method.

Figures 6, 7, 8, 9, 10 show visual representations of the accuracy of machine learning models trained on the entire dataset and the proposed feature selection method. For instance, you can see in Fig. 6a how practical the proposed feature selection approach is by looking at the accuracy achieved by the Gradient boosting algorithm when fed with features produced in this manner. When using all of the features in the dataset, as shown in Fig. 6a, the Gradient boosting algorithm achieves an accuracy of 84.22%. Achieving an accuracy of 86.77% utilizing the proposed hybrid feature selection approach of the gradient boosting algorithm is depicted in Fig. 6b. In Fig. 7a, we can see that when the random forest algorithm is applied to the entire dataset, it achieves an accuracy measure of 90.84%. The proposed feature selection approach yielded an accuracy of

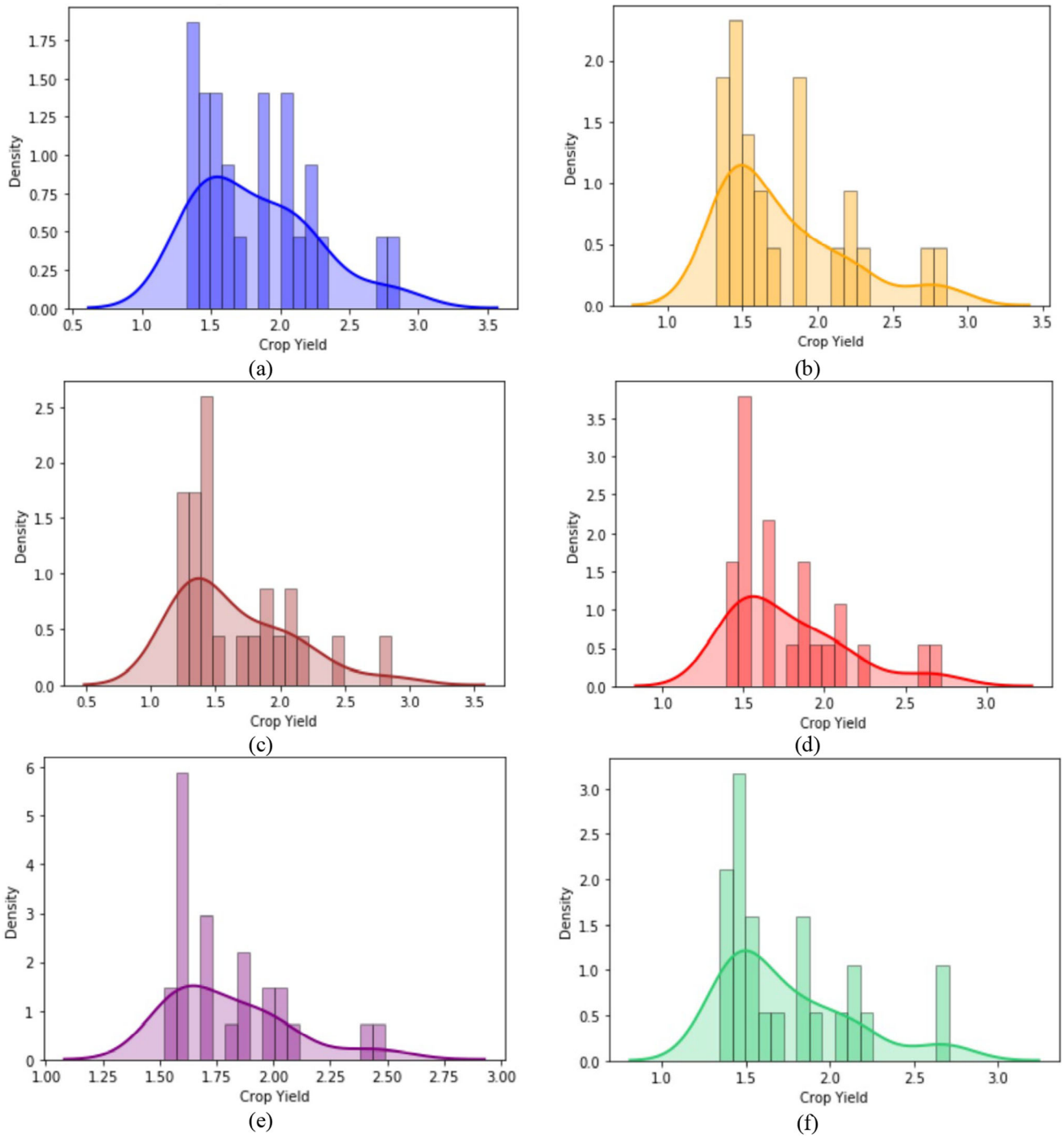

Fig. 13 presents the probability density curves of the following: actual data and data predicted by the proposed FMIG-RFE-SVM used with: **b** Random Forest, **c** Gradient Boosting, **d** Decision Tree, **e** Support Vector Machine, (**f**)K-Nearest Neighbor

91.98%when fed into the random forest algorithm (as shown in Fig. 7b). Figure 8a shows that an accuracy measure of 79.25% is obtained when the decision tree method is applied to all dataset features. As shown in Fig. 8b, the accurate measurement of the features was obtained after using the proposed feature selection method. Utilizing all features in the dataset with the SVM machine learning model is shown in Fig. 9a with an accuracy measure of 86.22%. Figure 10b shows an accuracy of 88.91% after applying the proposed feature selection approach with the SVM model. In Fig. 10a, we can see that when the KNN model is used for all of the features in the dataset, it achieves an accuracy measure of 78.21%. The

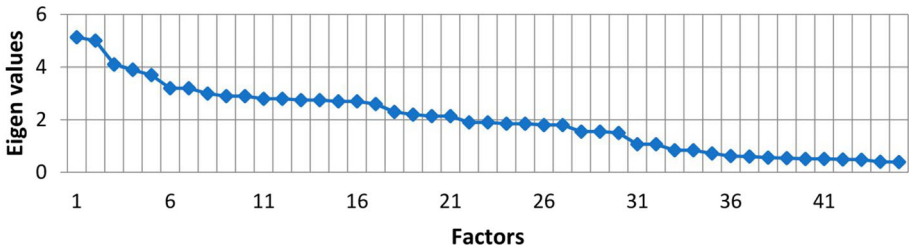

Fig. 14 Factor analysis scree plot defined with the number of factors

proposed feature selection approach yielded an accuracy of 80.69%when fed into the KNN model, shown in Fig. 10b.

5.4.2 Estimating the effectiveness of prediction phase

This section evaluates the proposed prediction model for the crop yield. In order to evaluate the proposed ICOA to optimize the parameters of different machine learning models, ICOA is applied for the five utilized machine learning models and the results are captured. Tables 9 and 10 present the obtained results applied as a result of using ICOA to optimize the parameters of different ML models. This section tests the ICOA parameter optimization with and without using the proposed hybrid feature selection approach. Tables 9 and 10 test the ICOA as a parameter optimizer for many ML models without and with using the proposed hybrid feature selection approach, respectively. According to Table 9 it is obvious that the performance of ML models increased after utilizing ICOA compared with Table 5 and 7. This is because the utilized chaotic levy Crayfish algorithm utilized the search process for the vest set of parameters that adapt the ML models. Moreover, Table 10 evaluates the proposed ICOA to adapt the parameters of ML models in the presence of the hybrid feature selection approach. It is clear that the parameter optimization process along with the hybrid feature selection approach largely increased the performance of the ML to predict the best results. From Table 10, SVR obtained the best results, least prediction errors, compared with other algorithms. The performance presented by SVR is superior to all other algorithms indicating a robust prediction result can be obtained.

For further evaluation of the proposed ICOA as a parameter optimizer approach, several optimization algorithms are compared with ICOA. Table 11 presents the captured results of applying different optimization algorithms to SVM ML model. The MAE, MSE, RMSE, R2 and MedAE are used to differentiate between different approaches. According to Table 11 it is clear that ICOA-SVM is superior to all other algorithms particularly COA-SVM which indicates the original COA. The MAE and RMSE of ICOA-SVM is 0.151 and 0.228 which is the minimum among all optimizers. Therefore, the proposed ICOA-SVM is a promising algorithm to optimize the parameters of ML models particularly SVM model. Figure 11 visualizes the comparison reported at Table 11 for further analysis and visualization.

5.4.3 Comparison with some recent state-of-art approaches

To undertake a more extensive study of the suggested framework, it is compared to various current state-of-the-art techniques to evaluate. In this experiment, multiple recently published methodologies are combined to demonstrate the innovative model as a possible solution to the crop yield forecast problem. The compared state-of-the-art techniques include RF [62], 1DCNN [25], LSTM-DBN [31] and CYPA [32]. The results show that the suggested model outperformed other prediction models, indicating a robust model. Table 12 records the captured results in terms of MAE, MSE, R2 and MedAE. According to Table 12, it is clear that the MAE and RMSE obtained by the proposed framework is the minimum among all state-of-art works while CYPA is ranked as the second best one. The obtained results indicate that the proposed framework presents an excellent contribution to the literature work.

5.4.4 Statistical analysis on accuracy

Table 12 shows the statistical comparison of the proposed model to the RF, 1DCNN, CYPA, and LSTM-DBN for agricultural yield prediction. Furthermore, it is evaluated for accuracy. Because metaheuristic procedures are unreliable, each method is rigorously tested to assure improved estimation. Furthermore, five different types of statistical measures are investigated: the mean, maximum, Wilcoxon test with p-value is 0.05 [63], Friedman rank [64], median, standard deviation, and minimum. Furthermore, the proposed framework has a maximum statistical metric accuracy of 0.949, while the RF has an accuracy of 0.853, LSTM-DBN has an accuracy of 0.914, CYPA has maximum accuracy of 0.939 and 1DCNN [24] has an accuracy of 0.924. The average accuracy obtained by the proposed framework is the best among all literature, which is 0.943. It is also analyzed that the obtained p-value of Proposed framework versus all other approaches is less than 0.05 indicating the significance of the obtained results between

the proposed framework compared to other approaches. Finally, the Friedman rank is used to rank the obtained results among all approaches where the proposed framework is ranked first with rank 1.42 while CYPA is ranked second with 2.29.

5.4.5 Comparative analyses of regression performance

Specifically, diagnostic regression plots [61] are built using features from the conventional hybrid feature selection approach to validate the regression findings of the machine learning models. By providing a convenient set of tools for assessing the model's validity, diagnostic regression charts boost the exploratory performance of the regression model. This examination may involve looking into the model's unstated statistical assumptions or analyzing the model's framework by thinking about alternatives that use fewer or more varied illustrative components. They are also helpful when searching for outliers or samples that have an outsized influence on the regression model's predictions but are not well represented by the data. When a model is fit to data, it often leaves behind something called a residual. But residuals may show how poorly a model represents the data. They also use the tested model to find previously unknown patterns in the data. These numbers will let us check if our regression hypotheses hold up and allow us to make cultivated guesses about improving the model. Figure 12 shows the four diagnostic plots that depict residuals in varying ways. In this section, we summarize the results from our evaluations of the forecasting models employing the proposed hybrid feature selection approach.

5.4.6 Features of data distributions

The probability density function of both the experimental models and actual data of crop yield is observed to ascertain if the proposed model retains the original distributional attributes of the data. An analytical expression, the Probability Density function (PDF), compares the distribution of one random variable to that of another, either continuous or discrete. The area graphically represents the range over which the expected variable is found under the PDF curve. The probability of observing the constant random variable is equal to the absolute area in the graph interval. We can use it to determine the probability of specific outcomes. Different probability density graphs for the raw data and tried ML models are shown in Fig. 13. To better match the distribution features of the actual crop production data, the random forest model, as specified clearly in Fig. 13, outperforms the other tested machine learning algorithms.

6 Discussion

Results from the proposed model are discussed, and the future implications of this research are briefly outlined. Nonlinear residual patterns can be seen clearly in the residuals vs fitted graph. Nonlinear relationships between the real and the predictor variable may manifest in these graphs if the model fails to do so initially. Nonlinear relationships are shown by randomly scattered residuals about the zero line.

Scale location plots check if residuals are spread out typically across the predictor's scale. It allows us to test for homoscedasticity [65] or the assumption of equal variance. A horizontal line with points placed at random is preferable. As can be seen in Fig. 12a, the residuals are distributed as possible to isolate the most critical information. All the outliers can't matter significantly for the regression line. Margin can be established thanks to Cook's distance. Those outliers who fall beyond the norm but have a significant impact are those that have a high Cook's distance score. Because of this, there are no significant outliers, as shown in Fig. 12b. Therefore, the improved model performance with the proposed feature selection technique is defined by the regression diagnostic charts. Based on Fig. 12c, it appears that the model data have well met the linear regression assumptions. No unique data pattern emerges when considering the linear distribution of information. The Q-Q plot will reveal whether or not the residuals are normally distributed with a constant standard deviation. The best-case scenario is where the residuals interline smoothly on a straight line with minimal variance. Suppose the residual deviates significantly from what would be expected under a normal distribution. In that case, the confidence intervals and p-values do not correctly reflect the true extent of the variation in the data. Figure 12d depicts the normal distribution of the residual, with residuals drawn almost perfectly along the diagonal line.

Factor analysis is an exploratory data analysis approach performed in addition to feature extraction techniques to discover the essential or hidden variables. It reduces the number of potential factors, making it easier to conclude the data. Factor analysis, a linear statistical model that explains the variance among the observed variables, refers to the unobserved variables as "factors". Multiple variables with consistent response patterns are linked to the same factors. It's the process of seeing if the factors f1, f2,… fn can explain the relationships between the relevant variables ($\times$ 1, $\times$ 2,… xn). To isolate latent factors, factor analysis seeks to reduce the number of variables that can be measured. In addition, factor extraction or rotation can be used to accomplish this. The factor analyzer library in Python is also used to actualize the proposed work factor

analysis. Evaluation of the dataset's factorability is required before running the factor analysis. Aside from that, the Kaiser–Meyer–Olkin (KMO) test is used to determine if the data is suitable for factor analysis. It details whether or not the overall model and set of data are adequate. The KMO values might range from 0 to 1, with anything below 0.1 deemed insufficient.

In general, the KMO for the crop dataset is 0.83, which is a good fit for moving forward with factor analysis. The eigenvalues define the number of factors in a scree plot. A straight line represents each factor and its eigenvalue in the scree plot procedure. The variables with eigenvalues more significant than one are regarded to be independent. The scree plot shown in Fig. 14 reveals 32 eigenvectors with squared values larger than 1. These factors together account for 55% of the total variance. By analyzing massive datasets, factor analysis can uncover hidden relationships and identify groups of connected variables.

In any case, the same data components might be used to support competing explanations. Our proposed feature selection technique yields 32 deciding factors close to the number of features. The overall performance and comparison findings demonstrate that the proposed hybrid feature selection approach yields superior performance results compared to the other feature selection process. As a result, the frameworks' prediction ability and efficiency are enhanced, as evidenced by a lower mean squared error (MSE), root mean square error (RMSE), mean absolute error (MAE), median absolute error (MedAE) and higher value of determination coefficient. The diagnostic graphs additionally detail the improved exploratory performance of the models.

7 Conclusion and future work

Agriculture is one of the most challenging departments to incorporate analytical results. Weather, soil, crop diseases, and pest infestations affect agricultural productivity and precision agriculture. Machine learning can change agribusiness by incorporating yield forecasting components. Machine learning models assess facts, translate data, and provide in-depth process knowledge. Feature selection using statistical measurements is critical for streamlining the predictive model's learning process and efficiently representing the dataset. This paper proposes a novel framework with a new hybrid feature selection strategy for machine learning models. The models predict paddy crop production based on soil, climate, and groundwater hydrochemical parameters. The FMIG-RFE-SVM method determines an intriguing study area's most crucial agricultural yield feature. The proposed approach combines the filter and RFE-SVM wrapper approaches. The filter approach eliminates redundant and non-essential features utilizing information gain and fisher score, resulting in a smaller subgroup. These features can be used to build an intelligent agricultural model for crop prediction. Future research may focus on new cutting-edge fuzzy-based clustering algorithms that can provide more helpful information for yield prediction. Second, we may include additional features in the dataset to improve the accuracy of the prediction model.

Author contributions All authors contributed equally to the research by conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft, review & editing.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Data availability All data used and required are mentioned in the manuscript.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relations that could have appeared to influence the work reported in this paper.

References

1. Holzman ME, Carmona F, Rivas R, Niclòs R (2018) Early assessment of crop yield from remotely sensed water stress and solar radiation data. ISPRS J Photogramm Remote Sens 145:297–308
2. Singh A, Ganapathysubramanian B, Singh AK, Sarkar S (2016) Machine learning for high-throughput stress phenotyping in plants. Trends Plant Sci 21(2):110–124
3. Xing L, Li L, Gong J, Ren C, Liu J, Chen H (2018) Daily soil temperatures predictions for various climates in United States using data-driven model. Energy 160:430–440
4. Liu S, Wang X, Liu M, Zhu J (2017) Towards better analysis of machine learning models: A visual analytics perspective. Visual Informatics 1(1):48–56
5. Johnson MD, Hsieh WW, Cannon AJ, Davidson A, Bédard F (2016) Crop yield forecasting on the Canadian Prairies by

remotely sensed vegetation indices and machine learning methods. Agric For Meteorol 218:74–84
6. Y.-H. Kuo, Z. Li, and D. Kifer, "Detecting outliers in data with correlated measures," in *Proceedings of the 27th ACM* International Conference on Information and Knowledge Management, 2018, pp. 287–296
7. Irita K (2011) Risk and crisis management in intraoperative hemorrhage: Human factors in hemorrhagic critical events. Korean J Anesthesiol 60(3):151–160
8. Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28
9. Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M (2020) Benchmark for filter methods for feature selection in high-dimensional classification data. Comput Stat Data Anal 143:106839
10. Askr H, Abdel-Salam M, Hassanien AE (2024) Copula entropy-based golden jackal optimization algorithm for high-dimensional feature selection problems. Expert Syst Appl 238:121582
11. Mielniczuk J, Teisseyre P (2019) Stopping rules for mutual information-based feature selection. Neurocomputing 358:255–274
12. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97(1–2):273–324
13. Taher F, Abdel-salam M, Elhoseny M, El-hasnony IM (2023) Reliable Machine Learning Model for IIoT Botnet Detection. IEEE Access 11:49319–49336
14. Chen G, Chen J (2015) A novel wrapper method for feature selection and its applications. Neurocomputing 159:219–226
15. Pourpanah F, Lim CP, Wang X, Tan CJ, Seera M, Shi Y (2019) A hybrid model of fuzzy min–max and brain storm optimization for feature selection and data classification. Neurocomputing 333:440–451
16. Paudel D et al (2021) Machine learning for large-scale crop yield forecasting. Agric Syst 187:103016
17. Becker-Reshef I, Vermote E, Lindeman M, Justice C (2010) A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. Remote Sens Environ 114(6):1312–1323
18. Qader SH, Dash J, Atkinson PM (2018) Forecasting wheat and barley crop production in arid and semi-arid regions using remotely sensed primary productivity and crop phenology: A case study in Iraq. Sci Total Environ 613:250–262
19. Van Ittersum M, Donatelli M (2003) Modelling cropping systems: highlights of the symposium and preface to the special issues. Eur J Agron 18(3–4):187–197
20. Kasampalis DA, Alexandridis TK, Deva C, Challinor A, Moshou D, Zalidis G (2018) Contribution of remote sensing on crop models: a review. Journal of Imaging 4(4):52
21. Vani PS, Rathi S (2023) Improved data clustering methods and integrated A-FP algorithm for crop yield prediction. Distributed and Parallel Databases 41(1):117–131
22. Xu J et al (2021) Estimation of Frost Hazard for Tea Tree in Zhejiang Province Based on Machine Learning. Agriculture 11(7):607
23. Jui SJJ et al (2022) Spatiotemporal Hybrid Random Forest Model for Tea Yield Prediction Using Satellite-Derived Variables. Remote Sensing 14(3):805
24. Reyana A, Kautish S, Karthik PS, Al-Baltah IA, Jasser MB, Mohamed AW (2023) Accelerating Crop Yield: Multisensor Data Fusion and Machine Learning for Agriculture Text Classification. IEEE Access 11:20795–20805
25. Paudel D, de Wit A, Boogaard H, Marcos D, Osinga S, Athanasiadis IN (2023) Interpretability of deep learning models for crop yield forecasting. Comput Electron Agric 206:107663
26. Khaki S, Wang L (2019) Crop yield prediction using deep neural networks. Front Plant Sci 10:621
27. J. You, X. Li, M. Low, D. Lobell, and S. Ermon (2017) "Deep gaussian process for crop yield prediction based on remote sensing data" in Thirty-First AAAI conference on artificial intelligence
28. Khanali M, Mobli H, Hosseinzadeh-Bandbafha H (2017) Modeling of yield and environmental impact categories in tea processing units based on artificial neural networks. Environ Sci Pollut Res 24(34):26324–26340
29. Khaki S, Wang L, Archontoulis SV (2020) A cnn-rnn framework for crop yield prediction. Front Plant Sci 10:1750
30. Iqbal U, Shahbaz M, Khalid A (2015) Development of a Decision Support System to increase the Tea Crops yield. Bahria University Journal of Information & Communication Technologies (BUJICT) 8:2
31. Boppudi S, Jayachandran S (2024) Improved feature ranking fusion process with Hybrid model for crop yield prediction. Biomed Signal Process Control 93:106121
32. Talaat FM (2023) Crop yield prediction algorithm (CYPA) in precision agriculture based on IoT techniques and climate changes. Neural Comput Appl 35(23):17281–17292
33. Alharbi A, Equbal K, Ahmad S, Rahman HU, Alyami H (2021) Human gait analysis and prediction using the levenberg-marquardt method. J Healthcare Eng 2021:1–11
34. Garg H (2020) Neutrality operations-based Pythagorean fuzzy aggregation operators and its applications to multiple attribute group decision-making process. J Ambient Intell Humaniz Comput 11(7):3021–3041
35. Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N (2019) Deep learning and process understanding for data-driven Earth system science. Nature 566(7743):195–204
36. Kern A et al (2018) Statistical modelling of crop yield in Central Europe using climate data and remote sensing vegetation indices. Agric For Meteorol 260:300–320
37. Azzari G, Jain M, Lobell DB (2017) Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. Remote Sens Environ 202:129–141
38. Cai Y et al (2019) Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agric For Meteorol 274:144–159
39. A. Masjedi *et al.*, "Sorghum biomass prediction using UAV-based remote sensing data and crop model simulation," in *IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018: IEEE, pp. 7719–7722
40. Hammer RG, Sentelhas PC, Mariano JC (2020) Sugarcane yield prediction through data mining and crop simulation models. Sugar Tech 22(2):216–225
41. Sun J, Di L, Sun Z, Shen Y, Lai Z (2019) County-level soybean yield prediction using deep CNN-LSTM model. Sensors 19(20):4363
42. Alhnaity B, Pearson S, Leontidis G, Kollias S (2019) Using deep learning to predict plant growth and yield in greenhouse environments. In International Symposium on Advanced Technologies and Management for Innovative Greenhouses GreenSys2019 1296:425–432
43. Alhnaity B, Kollias S, Leontidis G, Jiang S, Schamp B, Pearson S (2021) An autoencoder wavelet based deep neural network with attention mechanism for multi-step prediction of plant growth. Inf Sci 560:35–50
44. Jia H, Rao H, Wen C, Mirjalili S (2023) Crayfish optimization algorithm. Artif Intell Rev 56(Suppl 2):1919–1979
45. X.-S. Yang and S. Deb, "Cuckoo search via Lévy flights," in 2009 World congress on nature & biologically inspired computing (NaBIC), 2009: Ieee, pp. 210–214.
46. Reynolds AM, Frye MA (2007) Free-flight odor tracking in Drosophila is consistent with an optimal intermittent scale-free search. PLoS ONE 2(4):e354

47. Barthelemy P, Bertolotti J, Wiersma DS (2008) A Lévy flight for light. Nature 453(7194):495–498
48. R. Kohavi and G. H. John, "The wrapper approach," in Feature extraction, construction and selection: Springer, 1998, pp. 33–50.
49. Elavarasan D, Vincent PD (2020) Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. IEEE access 8:86886–86901
50. E. d. n. i. (2016). "Directorate Of Economics And Statistics, Ministry Of Agriculture, Government Of India." http://eands.dacnet.nic.in (accessed 21–12–2022.
51. "Agriculture Marketing." http://agmarknet.gov.in/PriceTrends/ (accessed 12/21/2022.
52. M. n. i. 2016. "Ministry Of Statistics And Program Implementation, Government Of India." http://mospi.nic.in/ (accessed 21–12–2022.
53. Prasad R, Deo RC, Li Y, Maraseni T (2018) Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition. Geoderma 330:136–161
54. Oh H-J, Pradhan B (2011) Application of a neuro-fuzzy model to landslide-susceptibility mapping for shallow landslides in a tropical hilly area. Comput Geosci 37(9):1264–1276
55. Peterson LE (2009) K-nearest neighbor. Scholarpedia 4(2):1883
56. Kari D, Mirza AH, Khan F, Ozkan H, Kozat SS (2018) Boosted adaptive filters. Digital Signal Processing 81:61–78
57. Pal M (2005) Random forest classifier for remote sensing classification. Int J Remote Sens 26(1):217–222
58. Ali M, Deo RC, Downs NJ, Maraseni T (2018) Multi-stage committee based extreme learning machine model incorporating the influence of climate parameters and seasonality on drought forecasting. Comput Electron Agric 152:149–165
59. Deepa N, Ganesan K (2019) Hybrid rough fuzzy soft classifier based multi-class classification model for agriculture crop selection. Soft Comput 23(21):10793–10809
60. Torres AF, Walker WR, McKee M (2011) Forecasting daily potential evapotranspiration using machine learning and limited climatic data. Agric Water Manag 98(4):553–562
61. S. D. Brown, R. Tauler, and B. Walczak, *Comprehensive chemometrics: chemical and biochemical data analysis*. Elsevier, 2020.
62. Van Klompenburg T, Kassahun A, Catal C (2020) Crop yield prediction using machine learning: A systematic literature review. Comput Electron Agric 177:105709
63. Cuzick J (1985) A Wilcoxon-type test for trend. Stat Med 4(1):87–90
64. S. Siegel and N. Castellan, "The Friedman two-way analysis of variance by ranks," *Nonparametric statistics for the behavioral sciences,* pp. 174–184, 1988, https://doi.org/10.1201/9781420036268.ch25.
65. R. Srinivasan and C. Lohith, "Main study—detailed statistical analysis by multiple regression," in Strategic marketing and innovation for Indian MSMEs: Springer, 2017, pp. 69–92.