e-Prime - Advances in Electrical Engineering, Electronics and Energy 8 (2024) 100611

Contents lists available at ScienceDirect

e-Prime - Advances in Electrical Engineering, Electronics and Energy

journal homepage: www.elsevier.com/locate/prime

Crop yield prediction using effective deep learning and dimensionality reduction approaches for Indian regional crops

Leelavathi Kandasamy Subramaniam [a,*], Rajasenathipathi Marimuthu [b]

[a] Research Scholar, Department of Computer Science, Nallamuthu Gounder Mahalingam College (NGMC), Affiliated to Bharathiar University, Pollachi, Tamilnadu, India
[b] Associate Professor, Department of Computer Science, Nallamuthu Gounder Mahalingam College (NGMC), Affiliated to Bharathiar University, Pollachi, Tamilnadu, India

ARTICLE INFO

Keywords:
Crop yield prediction
Machine learning
Deep learning
Principal component analysis
Deep convolutional neural network

ABSTRACT

Crop yield prediction (CYP) at the field level is crucial in quantitative and economic assessment for creating agricultural commodities plans for import-export strategies and enhancing farmer incomes. Crop breeding has always required a significant amount of time and money. CYP is developed to forecast higher crop production. This paper proposes an efficient deep learning (DL) and dimensionality reduction (DR) approaches for CYP for Indian regional crops. This paper comprised '3' phases: preprocessing, DR, and classification. Initially, the agricultural data of the south Indian region are collected from the dataset. Then preprocessing is applied to the collected dataset by performing data cleaning and normalization. After that, the DR is performed using squared exponential kernel-based principal component analysis (SEKPCA). Finally, CYP is based on a weight-tuned deep convolutional neural network (WTDCNN), which predicts the high crop yield profit. The simulation outcomes shows that the proposed method attains superior performance for CYP compared to exiting schemes with an improved accuracy of 98.96 %. The novelty of the proposed approach lies in the combination of DL, DR, and WTDCNN techniques for accurate crop yield prediction, specifically tailored for Indian regional crops.

1. Introduction

Crop yield prediction is the process of forecasting the amount of crop that will be produced in a given area. It is an important tool for farmers, governments, and businesses to make informed decisions about agricultural production. In India, crop yield prediction is a challenging task due to the country's diverse climate, terrain, and agricultural practices. However, there are a number of factors that can be used to predict crop yield, including:

Weather: The weather is one of the most important factors affecting crop yield. Rainfall, temperature, and humidity all play a role in plant growth.

Soil: The type of soil and its fertility also affect crop yield.

Crop variety: The choice of crop variety can also affect yield. Some varieties are more resistant to pests and diseases than others.

Farm management practices: The way that a farm is managed can also affect crop yield. Good farm management practices, such as irrigation and pest control, can help to improve yield.

Crop yield prediction models can be used to take into account these factors and forecast crop yield. These models can be based on statistical methods, machine learning, or a combination of both.

In recent years, there has been a significant increase in the amount of agricultural data collected from various sources, such as remote sensing, weather stations, and soil sensors. This data can be used to develop predictive models for crop yield estimation and management. However, the large volume and complexity of agricultural data pose significant challenges for data analysis and modeling. Therefore, there is a need for efficient and effective methods for analyzing and modeling agricultural data to improve crop yield prediction and management.

Agriculture is the primary food source for India's enormous population and a substantial source of economic support. Due to India's rapid population growth and critical climate changes, the food supply and demand chain must be maintained [1]. To maximize agricultural productivity, agronomic (agriculture scientist) experts have performed an important study to map, monitor, analyze, and manage yield variability. Crop production forecasts are one strategy that can help with crop management [2]. CYP is critical in food production [3]. CYP for strategic plants such as rice, maize and wheat are a fascinating field of research for agrometeorologists because it is significant in national and international programming. As a result, there exist systems that estimate

* Corresponding author.
E-mail address: leelavathi@ngmc.org (L.K. Subramaniam).

https://doi.org/10.1016/j.prime.2024.100611
Received 17 June 2023; Received in revised form 14 May 2024; Accepted 18 May 2024
Available online 19 May 2024

accuracy based on meteorological data [4]. The crop yield forecast is currently a difficult task for decision-makers at all global and local levels. Farmers can use a reliable crop production prediction model to determine what and when to sow. Crop production prediction can be accomplished through various methods [5].

Machine learning (ML) is one of the methods used to forecast agricultural yields, along with SVM, RF, DT, and others [6,7]. Calibration crop models are more easily implemented than simulation crop models because they do not need expert knowledge or user skills, have shorter execution times, and have less storage for data limits [8]. Despite developing numerous ML models to increase prediction accuracy, spatial and temporal non-stationarity, inherent in many geographical phenomena, is rarely incorporated in agricultural production modelling [9]. Recently, DL has been used to develop a variety of successful computations since it is used to select the best suitable crop when several options are available [10]. It is an ML class with multiple layers of neural networks capable of learning from data [11]. It seeks to produce predictions by establishing relationships between input and response variables. However, a critical difficulty with DL is its reliance on hyper-parameters, which can be avoided to improve the effectiveness of the results. Previously proposed architectures for predicting crop yields are frequently hand-designed, with DL approach professionals investigating challenges. They are unable to develop ideal structures because they do not comprehend agriculture. Hence, this paper proposed a practical deep-learning approach with optimal hyperparameters tuning for CYP for Indian regional crops. The main contributions of the work are as follows:

- The pre-processing is performed based on data cleaning and normalization to remove the noise and normalize the dataset.
- The DR is performed using the SEKPCA method to reduce higher dimensional data into lower dimensional data.
- The most profitable crop yield is predicted using the WTDCNN model, and the weights of DCNN are optimally selected using the enhanced whale optimization algorithm (EWOA).

2. Related works

Farhat Abbas et al. [12] presented a CYP system through proximal sensing and ML algorithms. Four publicly available datasets such as PE-2017, PE-2018, NB-2017, and NB-2018, were collected to perform training. The collected data were trained on the ML models such as elastic net (EN), linear regression (LR), support vector regression (SVR), and k-nearest neighbor (KNN) for predicting crop yields. The SVR achieved better results for all four tested datasets with lower RMSE than other existing schemes. Martin Kuradusenge et al. [13] presented several machine-learning models for CYP. Initially, the Irish potato and maize datasets were collected, and the pre-processing operations, like removal of null values and correlation determination, were carried out to enhance the system's performance further. After that, the classification of the pre-processed data was performed using three ML models, such as random forest (RF), polynomial regression (PR), and support vector machine (SVM), for CYP. Results showed that the RF model attained bets results than the SVM and PR in predicting the crop yields of potato and maize with an RMSE of 510.8 and 129.9 on the tested datasets.

Liyun Gong et al. [14] recommended hybrid DL approaches such as recurrent neural networks and temporal convolutional networks for CYP. The data was collected from multiple real greenhouse sites for tomato growing. The collected data was pre-processed by performing data normalization, and the normalized data was given to the RNN to process the normalized sequence data. Finally, the output of the RNN was passed to TCN for tomato CYP. The method achieved better results than the existing related schemes for the collected datasets with lower RMSE. Dhivya Elavarasan and P. M. Durai Raj Vincent [15] presented a hybrid approach called reinforced RF for CYP with agrarian parameters. Initially, the system collected the crop data from the agrarian dataset and the collected data was fed into the hybrid DL model, namely reinforced RF. The reinforced RF used the reinforcement learning approach in every internal node to determine the significance of the collected input data. The RF then used the most significant data determined using the reinforcement model to classify crop yield. The hybrid approach achieved better results than the existing ML models for CYP, such as SVM, LR, and KNN.

Aghila Rajagopal et al. [16] presented an optimal deep-learning model for CYP. The collected data were pre-processed, and the relevant features were extracted from the pre-processed dataset using principal component analysis. Then the selected features were further optimized using an improved chicken swarm algorithm to enhance the classifier's performance. Finally, classification was done using a discrete DBN-VGGNet classifier. The system achieved 97 % accuracy and 0.01 % MSE, which was superior to the previous state-of-the-art models. Dilli Paudel et al. [17] proffered an ensemble of machine-learning models for large-scale CYP. Initially, the system collected the crop yield data such as crop growth simulation outputs, weather observations and yield statistics from various sources. The collected data were cleaned for classification processes. Then feature design was applied to some of the input data, and they were fed into the classifier. The ML classifiers such as SVM, KNN, ridge regression, and gradient-boosted decision trees were used for CYP.

Using a remote sensing and geospatial analysis techniques Zamani et al. [18] suggested a weighted linear combination model to study saffron cultivation potential in Miyaneh City, Iran. The analysis is simple but does not capture complex nonlinear relationships between variables influencing suitability. Sharifi and Hosseingholizadeh [19] used Sentinel-1 Synthetic Aperture Radar (SAR) data to estimate height and biomass of rice crops in Astaneh-ye Ashrafiyeh, Iran. Machine learning algorithms say Multiple Linear Regression (MLR), Support vector Regression (SVR) and relevance vector regression (RVR) were applied and RVR achieved the best results. Sharifi A develops and evaluates a specific flood detection method based on Sentinel-1 images and classifier algorithms in [20]. The accuracy of the flood detection method is validated using ground truth data, such as data collected from ground sensors or high-resolution optical imagery. In [20], the proposed method for rapid performance sizing of ADCS in EO-satellites using matching diagram technique significantly reduces the complexity and time duration of the performance sizing process with an acceptable level of accuracy. An algorithm [21] was proposed to extract optimized features from POLSAR images that are required for estimation. Adespeckling [22] approach based on fast independent component analysis (Fast ICA) algorithm is proposed for improving of the results when polarimetric channels are added. To improve accuracy [23] in the case of limited training samples, the researcher proposed a multiscale dual-branch residual spectral–spatial network (MDBRSSN) with attention to the hyper spectral images (HIS) classification model. The study in [24] identified suitable soil types and locations for saffron cultivation in Miyaneh City using the Weighted Linear Composition (WLC) method and remote sensing data

The previous research highlights using traditional machine-learning algorithms for CYP. Classical ML models are built with specific quantities of training data to forecast agricultural yields depending on specific criteria. However, it has several limitations. For example, features collected from data for creating traditional ML models could not be the most accurate or most representative, resulting in lower yield performance. It must be able to successfully handle data of great volume or complexity. As a result, the authors directed to suggest DL algorithms, although it still requires improvement in the model's prediction rate and computing complexity. Furthermore, previous attempts should have focused on DR, which directly predicts crop production from the dataset, which reduces the interpretation of the DL parameters and requires more storage space. As a result, the proposed system employs optimal DL and DR methodologies to estimate crop yields for Indian regional crops

Table 1
Summary and Limitations of Existing CYP Approaches.

Authors	Approach	Techniques/Classifiers Used	Issues Addressed	Limitations
Kavitha et al. [1]	ML based approach	Combination of DT and regression classifiers used	Less Accuracy and precision while prediction	Does not focus pre-processing technique
Burdett et al. [2]	ML based approach	RF, ANN, Regression classifiers used	Accuracy prediction for Crop Yield Prediction	Does not study data imbalance and noise
Pant et al. [3]	ML based approach	DT, Gradient Boosting, SVM	Accuracy,precision and RMSE	Does not focus on ensemble learning errors
Saranya et al. [4]	Neural Network and Population based approach	ANN,CNN	To predict higher accuracy and other confusion matrix metrics for crop yielding	Does not focus bias, variance and oversampling errors
Jhajharia et al. [5]	ML based approach	SVM,RF,LSTM, Gradient Descent, and Lasso regression	Predict higher accuracy and precision rate for crops	Does not handle various homogeneous data
Gavahi et al. [6]	DL based approach	CNN,ConvLSTM	Predict higher accuracy for crop yield prediction using feature extraction	Does not focus on mis classification errors
Nishant et al. [7]	ML based approach	Kernel Ridge, Lasso and ENet algorithms	Predict higher accuracy for crops using stacking regression	Does not focus on ensemble learning errors.
Shahhosseini et al. [8]	ML based approach	linear regression, LASSO, LightGBM, random forest, and XGBoost	To improve the accuracy prediction for the corn crops	Does not focus on ensemble learning errors which leads to misclassification
Feng et al. [9]	DL based approach	ANN and GTWR	To improve the accuracy prediction of the crops based on geographically.	Does not consider various datasets and oversampling.
Abbas [11]	ML based approach	Four ML algorithms, namely linear regression (LR), elastic net (EN), k-nearest neighbor (k-NN), and support vector regression (SVR), were used to predict potato (*Solanumtuberosum*) tuber	The impact of climate variability, soil properties, and management practices on potato tuber yield prediction and the potential for site-specific management zones to enhance food security initiatives.	The models may not be applicable to other crops or regions without appropriate modifications and validation
Gong et al. [13]	DL based approach	Recurrent Neural Network (RNN) and Temporal Convolutional Network (TCN), for predicting greenhouse crop yield based on historical yields	emphasizing the importance of accurately predicting crop yield in greenhouses	Limited to specific dataset.
Elavarasan et al. [14]	ML based approach	RF and DT	the proposed approach requires less parameter tuning, reduces over-fitting, faster calculation and more transparent	Does not focus on data imbalance and noise.

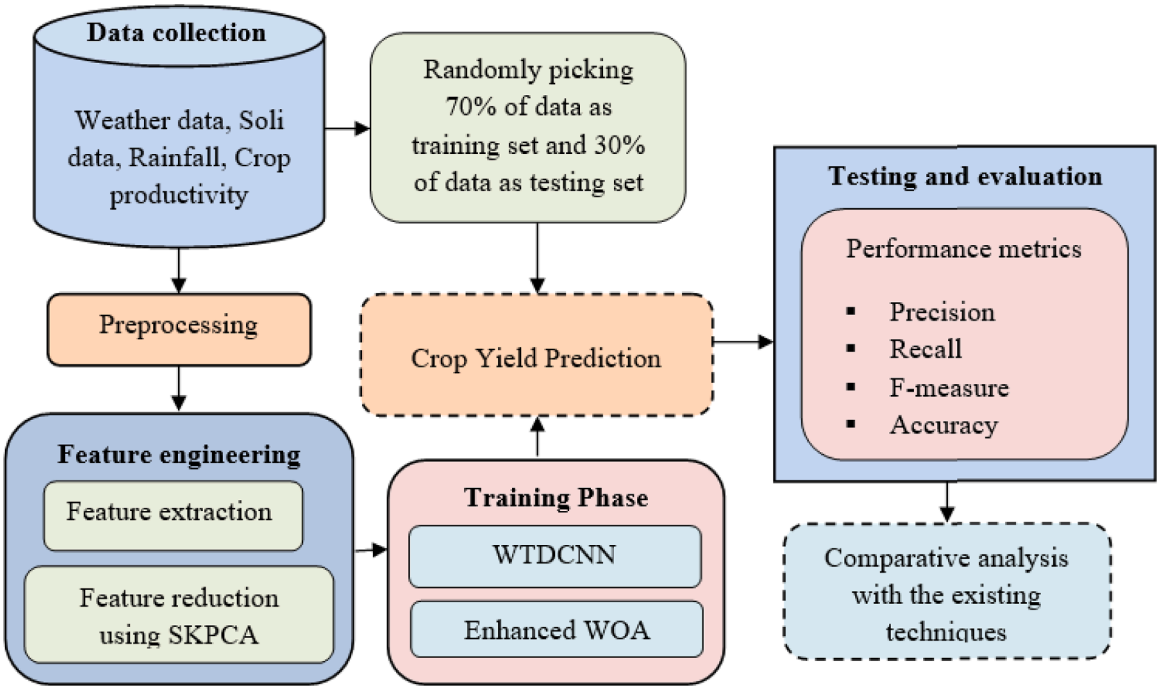

Fig. 1. Workflow of the proposed system.

are tabulated as Table 1.

3. Proposed methodology

This paper proposes an optimal DL model with DR approaches for CYP for Indian regional crops. Initially, the agricultural data of the south Indian region, such as rainfall, crop productivity, soil type, and weather data, are collected from publicly available data sources. Then pre-processing of the data is done by applying data cleaning and data normalization. After that, DR is made using SEKPCA, which results in lower dimensional features for CYP. Finally, CYP is made using WTDCNN. The workflow of the proposed work is shown in Fig. 1.

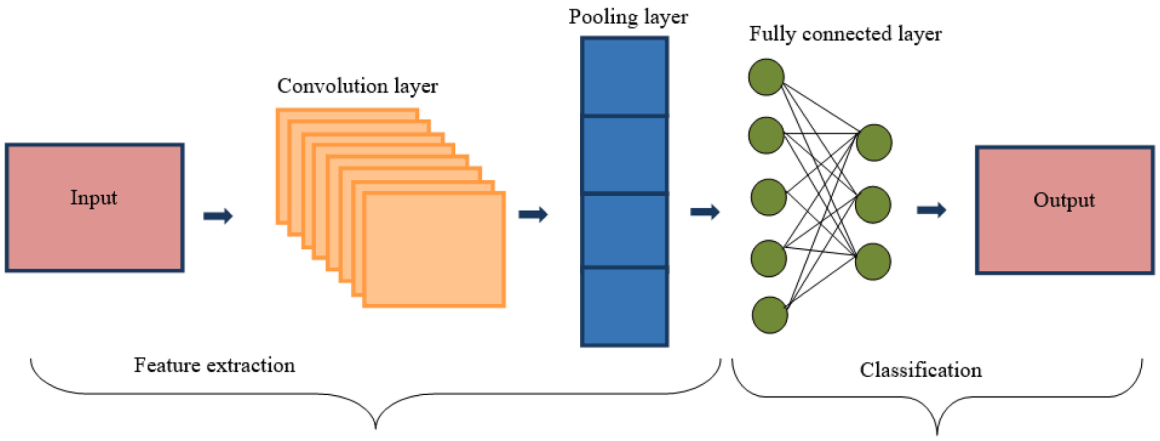

Fig. 2. Structure of DCNN.

3.1. Preprocessing

Initially, the agricultural data of the south Indian region, such as rainfall, crop productivity, soil type, and weather data, are collected from publicly available data sources. Following that, data preparation or preprocessing is undertaken since data is acquired from various sources. It is collected in raw format, which is not suitable for analysis. So preprocessing is most important before predicting the crop yield to improve the prediction rate. The preprocessing steps are explained as follows.

Step 1: Data cleaning

After gathering data from repositories, data cleaning is performed through missing value imputation and outliers' elimination. Missing values influence the model's accuracy in the data. As a result, the missing values are replaced with the mean or median values of the entire dataset or some other summary statistic. Outlier removal is conducted after missing value imputation to reduce noise from the dataset. The most straightforward technique to eliminate outliers from the data set is to delete them, which improves data quality. The dataset contains 3355 rows (input samples) and 22 columns (features), so the dimension of the matrix fed into the CNN is (3355, 1, 22), which indicates the number of data samples using a kernel size of 1 with 22 features

Step 2: Normalization

After performing data cleaning, normalization of the dataset is done. Normalization aims to convert data to be dimensionless and have similar distributions. It is mathematically expressed as follows:

$$\underset{\rightarrow}{C}'_{Norm} = \frac{\underset{\rightarrow}{C}' - \underset{\rightarrow}{C}'_{\min}}{\underset{\rightarrow}{C}'_{\max} - \underset{\rightarrow}{C}'_{\min}} \tag{1}$$

Where, $\underset{\rightarrow}{C}'_{Norm}$ refers to the normalized data, $\underset{\rightarrow}{C}'$ indicates the original data, $\underset{\rightarrow}{C}'_{\min}$ and $\underset{\rightarrow}{C}'_{\max}$ signifies the minimum and maximum value from the data set. The dataset values are between 0 and 1 using this min-max normalization.

Step 3: Splitting the datasets

The preprocessed dataset is portioned into training and testing datasets to implement the proposed system. The proposed system randomly picks 70% data for training and 30% data for testing.

3.2. Dimensionality reduction

After preprocessing, the DR of the dataset is made using squared exponential kernel-based principal component analysis (SEKPCA), which transforms the higher-dimension data into a lower dimension. PCA operates by computing the principal components and changing the basis. It solves the variable's correlation and can significantly improve crop yield detection and diagnosis of high-dimensional data in the actual production process. Even so, PCA is only effective if the variables are all highly uncorrelated. In addition, the PCA has difficulty recognizing nonlinear data models. Because the relationship between different features in the preprocessed dataset is nonlinear, the proposed system incorporates squared exponential kernels (SEK) in conventional PCA, which improves the system's performance by recognizing the nonlinear data and DR of the dataset in an effective manner. Initially, consider the preprocessed dataset with dimensions and analyze the mean vector [2] for each dimension using the following Eq. (2):

$$\underset{\rightarrow}{V}_{sv} = \frac{\underset{\rightarrow}{P}_{ds} - \mu}{\sigma} \tag{2}$$

Where, $\underset{\rightarrow}{V}_{sv}$refers to the scaled value, $\underset{\rightarrow}{P}_{ds}$indicates the preprocessed dataset, μ and σrepresents mean and standard deviation. Then the covariance matrix is computed using SEK, the popular kernel function for the covariance matrix estimation. Using the SEK will result in a smooth prior on functions sampled from the covariance calculation process. To summarize, the SEK function, $SEK(v_x, v_y)$ models the covariance between each pair in $\underset{\rightarrow}{V}_{sv}$. It is expressed as follows:

$$\sum = SEK(v_x, v_y) \tag{3}$$

$$SEK(v_x, v_y) = \sigma^2 \exp\left(-\frac{\| v_x - v_y \|^2}{2r^2}\right) \tag{4}$$

Where, σ^2indicates the overall variance, $r$signifies the length scale. Next, the eigenvalues and eigenvectors of the covariance matrix can be computed as follows,

$$\sum = \underset{\rightarrow}{M} \Lambda \left(\underset{\rightarrow}{M}\right)^T \tag{5}$$

Where,$\underset{\rightarrow}{M}$indicates the matrix composed of eigenvectors and Λrefers to an eigenvalue diagonal matrix. These eigenvectors are unit eigenvectors whose lengths are both 1. The eigenvectors from the covariance matrix are then ordered by eigenvalue, from highest to lowest. This lists the components in descending order of importance. As a result, the less essential components must be addressed. It is mathematically expressed as follows:

$$\underset{\rightarrow}{DR}_s = \left\{\underset{\rightarrow}{m}_1, \underset{\rightarrow}{m}_2, \underset{\rightarrow}{m}_3, \ldots\ldots\ldots, \underset{\rightarrow}{m}_n\right\} \tag{6}$$

Where, $\underset{\rightarrow}{DR}_s$indicates a dimensionality-reduced feature set which consists of significant eigenvectors and $n-$ refers to a total number of selected dimensions.

Table 2
Results of the classifiers regarding detection metrics.

Techniques/Metrics (%)	Accuracy	Precision	Recall	F-Measure
Proposed WTDCNN	98.96	98.67	99.03	98.87
DCNN	96.98	96.27	97.03	96.78
DBN	94.43	94.16	94.66	94.35
ELM	90.34	90.02	90.46	90.27
RF	89.21	89.05	89.32	89.19

3.3. Crop yield prediction

After DR, a weight-tuned deep convolutional neural network (WTDCNN) is used for CYP. DCNN comprises several layers, each of which computes convolutional transforms before moving on to non-linearities and pooling1 operators. In DCNN, the random weight and bias values are utilized for backpropagation training, which increases the chances of getting sup optimal results and higher loss in the prediction process. So proper tuning of weight and biases in the network is essential to enhance the detection accuracy and reduces the loss of the network. As a result, the proposed system employs a EWOA to determine the network's weights and bias values, which produces optimal results by minimizing the vanishing gradient saturation and prediction loss of the network for CYP. Fig. 2 depicts the general structure of DCNN.

The structure of DCNN comprises '4' layers such as convolution, pooling, activation, and a fully connected layer. In DCNN, the network weight and biases are chosen randomly for backpropagation training. Instead of choosing them randomly, in the proposed system, they are selected optimally using EWOA to enhance the network's performance in yield prediction. The WOA is a new type of swarm-based optimization algorithm that mimics the humpback foraging behavior of whales. WOA employs three operators to find prey: encircling, researching, and attacking prey. The random population initialization of whales in its initial stages decreases its convergence efficiency and algorithm's quality to get optimal global solutions. In addition, in the later stages of the search process, the algorithm gets stuck into the optimal local issues, which degrades the algorithm's performance. So, the proposed system uses Tent chaotic map to initialize the population, which improves the algorithm's population diversity and convergence efficiency. In addition, the levy flight mechanism is employed for updating whales' position in the later stages of the algorithm, which prevents the system from

Table 3
Analysis of classification error.

Classifiers/Metrics	MSE	RMSE	FPR	FNR	FRR
Proposed WTDCNN	0.034	0.219	0.029	0.065	0.061
DCNN	0.095	0.298	0.089	0.194	0.187
DBN	0.124	0.367	0.121	0.258	0.223
ELM	0.345	0.412	0.334	0.322	0.305
RF	0.398	0.483	0.379	0.423	0.402

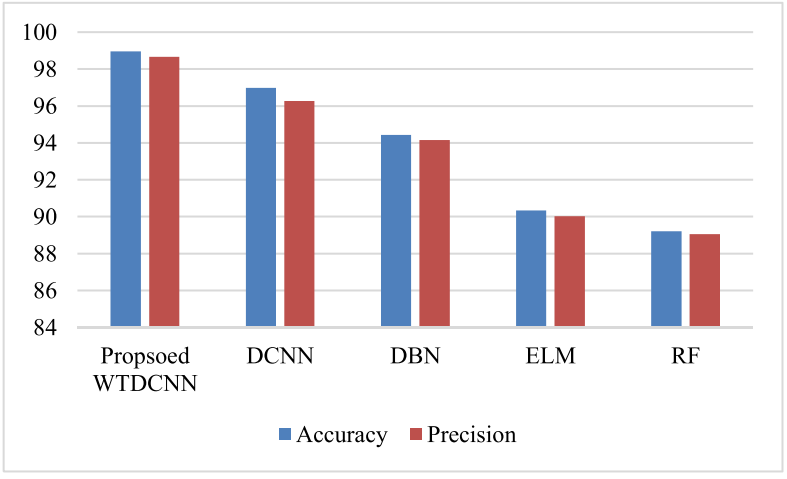

Fig. 3. Analysis of PR and AC.

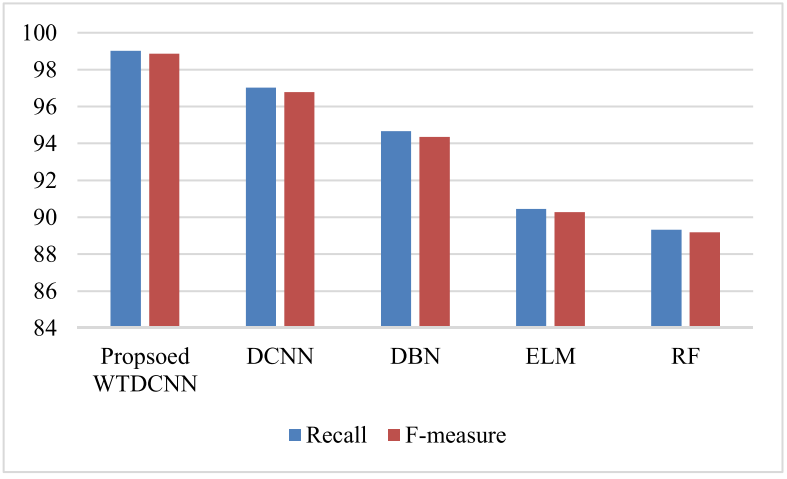

Fig. 4. RC and FM analysis.

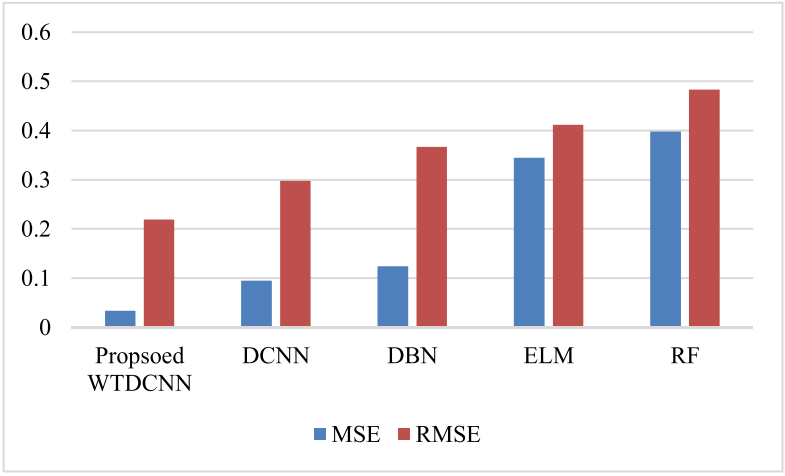

Fig. 5. MSE and RMSE analysis.

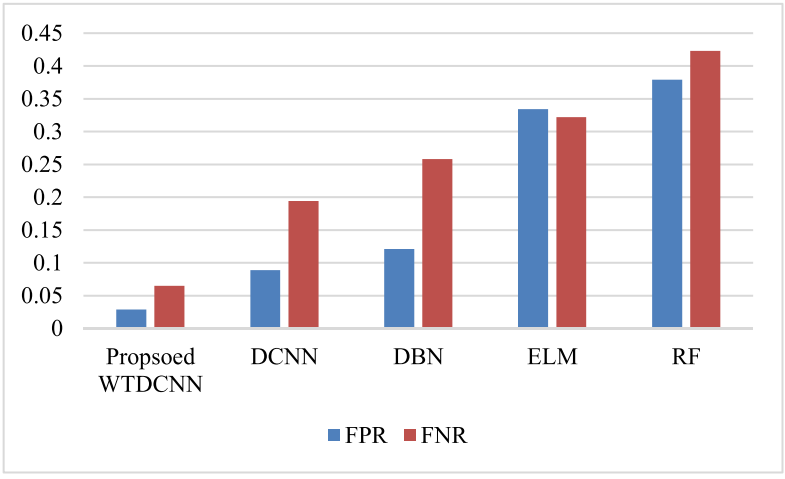

Fig. 6. FPR and FNR analysis.

finding locally optimal solutions. These two enhancements in conventional WOA are termed EWOA.

The feature extraction method based on deep learning is implemented by VWCNN. The network consists of five convolution layers, five pooling layers and three fully connected layers, and each of the convolution layers is connected to a activation layer using Relu. The input image size is 200 × 200 × 3, and the output size is 6 × 6 × 128 after the fifth convolution layer and the pooling layer, that is, the size of each feature map is 6 × 6, with 128 feature maps. After the first two fully connected layers, the output is a 1024- dimensional vector. After the last fully connected layer, a 5- dimensional vector is output and input to the Softmax classifier for classification.

The algorithm starts by initializing the population of the individuals in the search space using a chaotic tent map. Tent chaotic maps have improved population distribution uniformity and search speed, reducing the influence of the initial population distribution. It is written as follows:

$$\underrightarrow{Z}_{\tau+1} = \begin{cases} 2\underrightarrow{Z}_{\tau} & 0 \le \underrightarrow{Z} \le 0.5 \\ 2\left(1-\underrightarrow{Z}_{\tau}\right), & 0.5 < \underrightarrow{Z} \le 1 \end{cases} \tag{7}$$

Where, $\underrightarrow{Z}_{\tau+1}$ refers to the whales' initial population using a tent map and $\underrightarrow{Z}_{\tau}$ indicates the random population. Then the fitness $\left(\underrightarrow{FN}_{cal}\right)$ of the whales in the initialized population is estimated using the classifier's mean square error (MSE). MSE is computed by taking the difference between the actual output and the predicted output of the classifier in yield prediction. It is expressed as follows:

$$\underrightarrow{FN}_{cal} = Min\,(MSE) \tag{8}$$

$$MSE = \frac{1}{d}\sum_{p=1}^{d}\left(q_{val} - q_{val}^{*}\right) \tag{9}$$

Where, q_{val} and q_{val}^{*} indicates the actual and predicted value of the classifier and $d-$ refers to the number of samples in the training dataset. Then the position of the whales can be detected to surround them. The whale close to prey location is considered the best whale $\underrightarrow{Z}^{*}$ in the current population and the position of other whales is updated based on $\underrightarrow{Z}^{*}$ as follows.

$$\underrightarrow{DT} = \left|\beta \times \underrightarrow{Z}^{*}(\tau) - \underrightarrow{Z}(\tau)\right| \tag{10}$$

$$\underrightarrow{Z}(\tau+1) = \underrightarrow{Z}^{*}(\tau) - \alpha \times \underrightarrow{DT} \tag{11}$$

Where, τ indicates the current iteration and $\underrightarrow{DT}$ refers to the distance betwixt the prey $\underrightarrow{Z}^{*}(\tau)$ and the whale $\underrightarrow{Z}(\tau)$. In addition, α and β represents the coefficient vectors computed using Eqs. (12) and (13)

Algorithm 1
Proposed algorithm for weight-tuned deep convolutional neural network.

Require: Input of crop yield images image
Ensure: Class level for each image
Initialize with weight and bias of the pre-trained weight-tuned cnn model.
Initialize the population of whales by using Eq. (5.7).
Compute the fitness of each individual by using Eq. (5.8).
While
For each search agent
Calculate the coefficients and
If
If
Update the position of the current whale using Eq. (5.11).
Else
Update the position of the current whale using Eq. (5.16).
End if
Else if
Update the position of the current whale using Eq. (5.14).
End if
End for
Update if there is a better solution
End while
End
Define the learning rate, batch size and number of epochs for training.
Rescale the input image to a fixed size of 224 × 224.
for each rescaled image x^i do
Rotate (x^i, θ); where $\theta \epsilon [-30^{\circ}, 30^{\circ}]$
Add a randomly initialized new head to the weight-tuned architecture
Freeze the body of the network and train the newly added head of previous step using RMSProp with learning rate 10^{-3}.
Unfreeze the body of the network and continue the training process using SGD optimizer with a learning rate 10^{-3}.
end for
Evaluate the fine-tuned network and serialize the weights to disk.

$$\alpha = 2 \times l_d \times R_{num} - l_d(\tau) \tag{12}$$

$$\beta = 2 \times R_{num} \tag{13}$$

Where, R_{num}indicates a random number ranges between [0, 1] and l_d is decreased linearly from 2 to 0 over the number of iterations, Then the humpback whales' bubble-net [3,4] behavior is updated using the following equation:

$$\underrightarrow{Z}(\tau+1) = \begin{cases} \underrightarrow{Z}^{*}(\tau) - \alpha \times \underrightarrow{DT} & if\, p < 0.5 \\ \left(\underrightarrow{DT}\right)' \times e^{h_k g_{lk}} \times \cos(2 \times \pi \times g_{lk}) + \underrightarrow{Z}^{*}(\tau) & if\, p \geq 0.5 \end{cases} \tag{14}$$

Whereas, $p$represents a random integer between [0, 1] and shows the likelihood of updating the position of whales according to the spiral updating position $(if\, p \geq 0.5)$or shrinking encircling technique$(if\, p < 0.5)$,g_{lk}denotes an arbitrary integer between [−1, 1], and h_k defines the spiral movement shape. Then the global search process (exploration) of the whales is executed, and completed when the absolute vector value is greater or equal to one. Otherwise, the algorithm implements the exploitation phase. Instead of considering the best whale $\underrightarrow{Z}^{*}$, the random value $\underrightarrow{Z}_{rand}$is considered in the exploration phase to update whales' positions, which is expressed as follows:

$$\underrightarrow{DT} = \left|\beta \times \underrightarrow{Z}_{rand} - \underrightarrow{Z}(\tau)\right| \tag{15}$$

$$\underrightarrow{Z}(\tau+1) = \underrightarrow{Z}_{rand} - \alpha \times \underrightarrow{DT} \times \mathrm{L_f}(\xi) \tag{16}$$

Where $\underrightarrow{Z}_{rand}$refers to the arbitrarily chosen whale from the current population, and $\mathrm{L_f}(\xi)$ indicates a levy flight mechanism, which enhances the exploration and exploitation capabilities of the algorithm. It is arandom walk where the steps are denoted regarding step lengths with a given probability distribution and is written as follows:

$$\mathrm{L_f}(\xi) \sim |\xi|^{-1-\lambda} \tag{17}$$

$$\xi = \frac{N_{ud}}{|N_{vd}|^{1/\lambda}} \tag{18}$$

Where, $\lambda\,(0 < \lambda \leq 2)$ signifies an index, ξindicates the step length, $N_{ud}$and N_{vd}represents drawn from normal distributions. After optimally chosen weights and biases, the convolution layer extracts relevant features from the dimensionality-reduced dataset. As the biases and weights accept the best values at each iteration, the likelihood of an improved model gradually rises.

$$FeatureVector = \sum \left(\underrightarrow{DR}_s + \breve{O}^{*}_{d\times d}\right) + \breve{B}^{*} \tag{19}$$

Where, $\underrightarrow{DR}_s$refers to the dimensionality reduced dataset on which the convolutional operation is performed, $\breve{O}^{*}_{d\times d}$and $\breve{B}^{*}$indicates the optimal filter weights and bias selected by EWOA, and $d-$ denotes the kernel size. The final feature vector obtained is then inputted into the activation layer to increase the nonlinearity in the output. ReLU is used as an activation function in the activation layer that outputs the input directly if the input value is positive. Otherwise, it will output zero. The output of the convolution layer with activation is fed into the pooling layer to minimize the input data size. The polling layers use smaller rectangular boxes of the convolution layer and produce the output by sampling the convolution's rectangular boxes. The output of the polling layers is given to the fully connected layer for performing CYP, which uses the SoftMax activation function to perform the classification. The classifier's output shows the productivity of different crops in Indian regions under various seasons that help the farmers to plant the crops according to their productivity level in future.

4. Results and discussion

This section looks at the experimental findings of the suggested yield prediction for Indian regional crops utilizing efficient DL and DR methodologies. The proposed methodology is compared with existing schemes for CYP regarding classification metrics. The predictions were

made in Python with an Intel Core i7–8550 CPU, an NVIDIA GEFORCE MX130 graphics card, and 8.0 GB of RAM.

4.1. Dataset descriptions

The proposed system collected the crop production data from the publicly available data source using https://data.world/thatzprem/agriculture-india, which consists of State Name, District Name, Crop, Year, Season, Crop, Crop class, Area, and Production Yield. Also, the weather data are collected from the Indian website, which consists of minimum temperature (°C), maximum temperature (°C), average temperature (°C), precipitation (mm), humidity (%), pressure, dew point (°C), wind (m/s).

4.2. Performance analysis

Here the outcomes of the proposed classification model (WTDCNN) are compared with the existing classification schemes namely with the existing DCNN, Random Forest (RF) models Deep Belief Network (DBN), and Extreme Learning Machine (ELM).The techniques are compared based on precision (PR), recall (RC), f-measure(FM) and accuracy (AC), MSE, Root Mean Square Error (RMSE), False Positive Rate (FPR), and False Negative Rate (FNR). The equations for the above metrics are given as follows

$$\mathrm{PR} = \frac{Tp}{Tp + Fp}$$

$$\mathrm{RC} = \frac{Tp}{Tp + Fn}$$

$$\mathrm{AC} = \frac{Tp}{Tp + Tn + Fp + Fn}$$

$$\mathrm{FM} = \frac{Tp}{Tp + 1/2(Fp + Fn)}$$

$$\mathrm{FPR} = \frac{Fp}{Fp + Tn}$$

$$\mathrm{FNR} = \frac{Fn}{Fn + Tp}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(A - B)^2$$

$$RMSE = \sqrt{\mathrm{MSE}}$$

Where, Tp, Tn, Fp, *and* Fn indicates the true positive, true negative, false positive, and false negative values of the classifier, and A and B denotes the original and predicted values of the dataset. The outcomes of the models regarding PR, RC, FM, and AC are tabulated in Table 2.

Table 1 has demonstrated that DL can play an important role in CYP, and our results confirmed the same. Despite being based on essential performance criteria, the outcomes are compared with other state-of-the-art methodologies. The suggested WTDCNN produces better results than the existing ones. For example, the existing DCNN achieves accuracy, precision, recall, and f-measure of 96.98 %, 96.27 %, 97.03 %, and 96.78 %, respectively. Also, the existing RF attains minimal 89.21 % accuracy, 89.05 % precision, 89.32 % recall, and 89.19 % f-measure, which is lower than the proposed one, because the proposed one achieves maximum accuracy of 98.96 % along with 98.67 % precision, 99.03 % recall, and 98.87 %f-measure. Similarly, considering other existing methods (DBN and ELM), the proposed one achieves more excellent performance. Thus, the outcomes proved that the proposed one outperformed the conventional methodology. The diagrammatic representation of the Table 1 is shown in Fig. 3.

Fig. 3 shows the outcomes of the models regarding PR, RC, FM, and AC. From the figure it was clear that the proposed model attains better results than the existing schemes. The proposed WTDCNN attains the PR of 98.67, which is higher than DCNN (96.27), DBN (94.16), ELM (90.02), and RF (89.21). Likewise, the proposed method attains highest accuracy than other existing schemes i.e., the WTDCNN attains an AC of 98.96, whereas the existing schemes such as DCNN, DBN, ELM, and RF attains an AC of 96.98, 94.43, 90.34, and 89.21, which are lower than the proposed scheme.

Fig. 4 shows the outcomes of the models regarding RC and FM. From the figure it was clear that the proposed model attains better results than the existing schemes. The proposed WTDCNN attains the FM of 98.87, which is higher than DCNN (96.78), DBN (94.35), ELM (90.27), and RF (89.19). Likewise, the proposed method attains highest RC than other existing schemes i.e., the WTDCNN attains an RC of 99.03, whereas the existing schemes such as DCNN, DBN, ELM, and RF attains an RC of 97.03, 94.66, 90.46, and 89.32, which are lower than the proposed scheme.

Next, the outcomes of the proposed one are investigated based on error metrics, namely, MSE, RMSE, FPR, FNR, and FRR metrics. This could be given in Table 3.

Table 2 demonstrates the outcomes of the proposed one is investigated against the existing DCNN, DBN, ELM, and RF methods in terms of MSE, RMSE, FPR, FNR, and FRR. The results showed that the proposed method obtains better performance than the existing models by achieving lower error values in classification. The proposed has MSE, RMSE, FPR, FNR, and FRR of 0.034, 0.219, 0.029, 0.065, and 0.061, respectively, which showed more excellent performance than the existing methods because the existing method has higher error values. However, the system is considered a sound system if the system has lower error values. Henceforth, it proved that the proposed system achieved superior performance than the previous existing schemes for accurate CYP. The diagrammatic representation of the Table 2 is given in Figs. 5 and 6.

Fig. 5 shows the MSE and RMSE of the proposed and existing classifiers. It was clear that the proposed model attains better results than the existing schemes. The proposed WTDCNN attains the MSE of 0.034, which is lower than DCNN (0.095), DBN (0.124), ELM (0.345), and RF (0.398). Likewise, the proposed method attains lowest RMSE than other existing schemes i.e., the WTDCNN attains an RMSE of 0.219, whereas the existing schemes such as DCNN, DBN, ELM, and RF attains an RMSE of 0.298, 0.367, 0.412, and 0.483 which are lower than the proposed scheme (Algorithm 1).

Fig. 6 shows the FPR and FNR of the proposed and existing classifiers. It was clear that the proposed model attains better results than the existing schemes. The proposed WTDCNN attains the FPR of 0.029, which is lower than DCNN (0.089), DBN (0.121), ELM (0.334), and RF (0.379). Likewise, the proposed method attains lowest FNR than other existing schemes i.e., the WTDCNN attains an FNR of 0.065, whereas the existing schemes such as DCNN, DBN, ELM, and RF attains an RMSE of 0.194, 0.258, 0.322, and 0.423 which are lower than the proposed scheme. The results of our outperforming WTDCNN model demonstrate the novelty of the proposed work by using the proper data preprocessing methods, architecture, and hyperparameters values. Clearly, the proposed one first preprocesses the dataset before prediction and efficiently utilizes the DR method. So, these approaches are more efficient in making predictions.

5. Conclusion and future works

This paper suggests efficient DL and DR approaches for CYP for Indian regional crops. The proposed system comprises '3' main phases: preprocessing, DR, and classification. The results of the proposed work are weighted against the conventional DCNN, DBN, ELM, and RF concerning the accuracy, precision, recall, f-measure, MSE, RMSE, FPR, FNR, and FRR. The outcomes of the proposed one have significant

performance because it achieves maximum accuracy of 98.96 % along with 98.67 % precision, 99.03 % recall, and 98.87 % f-measure. In addition, the proposed one attains lower error values of 0.034 MSE, 0.219 RMSE, 0.029 FPR, 0.065 and FNR. The outcomes concluded that the proposed optimal DL approach with a practical DR approach achieves superior results than the existing state-of-the-art schemes for CYP. The authors suggest that the proposed method can be extended to other regions and crops to improve the generalizability of the model. They also suggest that the proposed method can be further improved by incorporating additional data sources, such as satellite imagery and remote sensing data. Finally, the authors suggest that the proposed method can be used to develop decision support systems for farmers to help them make informed decisions about crop management and production. In future, the proposed methodology can be integrated with precision agriculture technologies to provide real-time crop yield prediction and decision-making support to farmers. The proposed methodology can be used to develop a web-based application that can be accessed by farmers and policymakers to make informed decisions about crop production and import-export strategies.

CRediT authorship contribution statement

Leelavathi Kandasamy Subramaniam: Visualization, Validation, Project administration. **Rajasenathipathi Marimuthu:** Validation, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financialinterestsor personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] M. Kavita, P. Mathur, Crop yield estimation in India using machine learning, in: Proceedings of the IEEE 5th International Conference on Computing Communication and Automation (ICCCA), IEEE, 2020, pp. 220–224.

[2] H. Burdett, C. Wellen, Statistical and machine learning methods for crop yield prediction in the context of precision agriculture, Precision Agriculture 23 (5) (2022) 1553–1574.

[3] J. Pant, R.P. Pant, M.K. Singh, D.P. Singh, H. Pant, Analysis of agricultural crop yield prediction using statistical techniques of machine learning, Mater. Today Proc. 46 (2021) 10922–10926.

[4] C.P. Saranya, N. Nagarajan, Efficient agricultural yield prediction using metaheuristic optimized artificial neural network using Hadoop framework, Soft Comput. 24 (2020) 12659–12669.

[5] M. Annamalai, P. Muthiah, An early prediction of tumor in heart by cardiac masses classification in echocardiogram images using robust back propagation neural network classifier, Braz. Arch. Biol. Technol. 65 (2022), https://doi.org/10.1590/1678-4324-2022210316.

[6] K. Gavahi, P. Abbaszadeh, H. Moradkhani, DeepYield: a combined convolutional neural network with long short-term memory for crop yield forecasting, Expert Syst. Appl. 184 (2021) 115511.

[7] P.S. Nishant, P.S. Venkat, B.L. Avinash, B. Jabber, Crop yield prediction based on Indian agriculture using machine learning, in: Proceedings of the International Conference for Emerging Technology (INCET), IEEE, 2020, pp. 1–4.

[8] S. Pokhariyal, N.R. Patel, A. Govind, Machine learning-driven remote sensing applications for agriculture in India–a systematic review, Agronomy 13 (9) (2023) 2302.

[9] M. Shahhosseini, G. Hu, I. Huber, S.V. Archontoulis, Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt, Sci. Rep. 11 (1) (2021) 1–15.

[10] R. Ali, A. Manikandan, J. Xu, A novel framework of adaptive fuzzy-GLCM segmentation and fuzzy with capsules network (F-CapsNet) classification, Neural Comput. Appl. (2023), https://doi.org/10.1007/s00521-023-08666-y.

[11] A. Manikandan, M. Ponni Bala, Intracardiac mass detection and classification using double convolutional neural network classifier, J. Eng. Res. 11 (2A) (2023) 272–280, https://doi.org/10.36909/jer.12237.

[12] P. Muruganantham, S. Wibowo, S. Grandhi, N.H. Samrat, N. Islam, A systematic literature review on crop yield prediction with deep learning and remote sensing, Remote Sens. 14 (9) (2022) 1990.

[13] A. Joshi, B. Pradhan, S. Gite, S. Chakraborty, Remote-sensing data and deep-learning techniques in crop mapping and yield prediction: a systematic review, Remote Sens. 15 (8) (2023) 2014.

[14] F. Abbas, H. Afzaal, A.A. Farooque, S. Tang, Crop yield prediction through proximal sensing and machine learning algorithms, Agronomy 10 (7) (2020) 1046.

[15] M.M. Islam, M.A.A. Adil, M.A. Talukder, M.K.U. Ahamed, M.A. Uddin, M.K. Hasan, S.K. Debnath, S. Sharmin, M. Rahman, DeepCrop: deep learning-based crop disease prediction with web application, J. Agric. Food Res. 14 (2023) 100764.

[16] M. Kuradusenge, E. Hitimana, D. Hanyurwimfura, P. Rukundo, K. Mtonga, A. Mukasine, C. Uwitonze, J Ngabonziza, A. Uwamahoro, Crop yield prediction using machine learning models: case of Irish Potato and Maize, Agriculture 13 (1) (2023) 225.

[17] Palaniappan, M. & Annamalai, M.. (2019). Advances in signal and image processing in biomedical applications. 10.5772/intechopen.88759.

[18] D. Elavarasan, P.D.R. Vincent, A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters, J. Ambient Intell. Humaniz. Comput. 12 (2021) 10009–10022.

[19] A. Rajagopal, S. Jha, M. Khari, S. Ahmad, B. Alouffi, A. Alharbi, A novel approach in prediction of crop production using recurrent cuckoo search optimization neural networks, Appl. Sci. 11 (21) (2021) 9816.

[20] Kolli, S. & V., Praveen & John, A. & Manikandan, A.. (2023). Internet of things for pervasive and personalized healthcare: architecture, technologies, components, applications, and prototype development. 10.4018/978-1-6684-8913-0.ch008.

[21] A.R. Venmathi, S. David, E. Govinda, K. Ganapriya, R. Dhanapal, A. Manikandan, An automatic brain tumors detection and classification using deep convolutional neural network with VGG-19, in: Proceedings of the 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 2023, pp. 1–5, https://doi.org/10.1109/ICAECA56562.2023.10200949.

[22] A. Sharifi, M. Hosseingholizadeh, Application of Sentinel-1 data to estimate height and biomass of rice crop in Astaneh-ye Ashrafiyeh, Iran, J. Indian Soc. Remote Sen. 48 (2020) 11–19.

[23] A. Sharifi, Development of a method for flood detection based on Sentinel-1 images and classifier algorithms, Water Environ. J. 35 (3) (2021) 924–929.

[24] A. Kosari, et al., Remote sensing satellite's attitude control system: rapid performance sizing for passive scan imaging mode, Aircr. Eng. Aerosp. Technol. 92 (7) (2020) 1073–1083, https://doi.org/10.1108/aeat-02-2020-0030.

K. S. Leelavathi was born in 1983 at Coimbatore District, Tamilnadu. She Completed her M.Phil in 2010,passed UGC NET and TN SET in July 2018 and pursuing her Ph.D under the guidance of Dr. M. Rajasenathipathi, Associate professor, Department of Computer Science, NallamuthuGounderMahalingam College(Autonomous), Pollachi, TamilNadu. Currently she is working as Assistant Professor in Computer Technology in NallamuthuGounderMahalingam College, Pollachi. She is having 17 years of teaching experience. Her area of research interest includes Data Mining and Machine Learning. She has published 2 papers in UGC CARE Journals and nearly 5 papers in National and International Journals.

Dr. M. Rajasenathipathi was born in 1973 at Coimbatore District, TamilNadu. He has completed his Ph.D from Bharathiar University, Coimbatore in the year 2014. Currently he is working as Associate Professor in Department of Computer Science in NallamuthuGounderMahalingamCollege(Autonomous), Pollachi, TamilNadu. He has more than 25 years of teaching experience. His area of research interest is Networking and Network Security. He has published more than 10reasearch papers in Scopus Indexed Journals and 2 papers in UGC CARE Journal.