The Scientific Temper (2024) Vol. 15 (spl): 394-403
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2024.15.spl.46
https://scientifictemper.com/

RESEARCH ARTICLE

Stochastic kernelized discriminant extreme learning machine classifier for big data predictive analytics

Anita M, Shakila S

Abstract

Predictive analytics has appeared as a dominant tool to improve crop yield in the agriculture field by leveraging big data. Soil is a vital aspect in determining the growth of crop production, and its attributes considerably influence crop growth, nutrient availability, and overall crop yield. Predictive analytics involves the combination of big soil data with weather information for crop yield estimation. By utilizing chronological data on crop performance, different machine learning (ML) and deep learning (DL) models have been developed to forecast crop yield outcomes under different scenarios. However, accurate prediction in the shortest possible time is a major challenging issue. A novel model called stochastic kernelized discriminant extreme learning machine classifier (SKDELMC) is introduced for crop yield forecast by analyzing large amounts of soil as well as weather big data. This SKDELMC model typically includes feature selection and classification to identify the most relevant features and classify the data into different categories. A number of data is gathered from a dataset. This data includes a range of soil parameters and the weather features that influence crop growth. After the data collection, with a huge number of features, it's important to choose mainly relevant ones for predictive analytics. The stochastic kernelized quadratic discriminant analysis is applied to identify the main informative features to minimize time complexity prediction. Once relevant features are chosen, the next step is to classify data into different categories using the qualitative indexed extreme learning classifier. It is a feed-forward neural network having a straightforward solution without requiring any iteration. A network includes different layers, such as the input layer, multiple hidden layers, as well as output layer. Relevant features are provided to the input layer. Then Baroni–Urbani–Buser coefficient is applied in the hidden layer by analyzing testing as well as training data is the qualitative index used to analyze the similarity between the data. After that, the Hardlimit activation function is utilized for evaluating similarity value as well as providing classification results. Based on the classification results, accurate prediction outcomes are attained at the output layer. Experimental evaluation is carried out by dissimilar quantitative parameters, namely disease prediction accuracy, sensitivity, false-positive rate, prediction time, and space complexity. Discussed performance outcomes illustrate that the SKDELMC model improves the accuracy of prediction and decreases the time consumption as well as space complexity than existing prediction techniques.

Keywords: Big data, predictive analytics, stochastic kernelized quadratic discriminant analysis, qualitative indexed extreme learning classifier, Baroni–Urbani–Buser coefficient, Hardlimit activation function.

Introduction

Prediction analytics with big data is a process of combining the capabilities of advanced analytics techniques with a huge volume of data to make accurate predictions. With the exponential growth of data from diverse sources, extracting meaningful insights and actionable intelligence from this data is a major challenging task without efficient and effective techniques. Advanced analytics methods such as statistical modeling, ML, and DL models have been used to analyze the data and construct predictive models.

An integrated ConvLSTM layer through 3-Dimensional CNN (3DCNN) for crop yield prediction technique called 'DeepYield' was developed for accurate and reliable spatiotemporal feature extraction and classification. But, the prediction accuracy level was not enhanced by lesser time utilization. A novel Bayesian model averaging (BMA) model was developed for predicting crop yield by measuring the uncertainty of model parameters as well as inputs concurrently. But high complexity of crop yield forecast was a major issue.

For effectively predicting soil moisture with better accuracy, a long short-term memory network (LSTM) was designed. The averaging method was applied to the outputs

Department of Computer Science, Government Arts College (Affiliated to Bharathidasan University, Tiruchirappalli), Tiruchirappalli, Tamilnadu, India.

***Corresponding Author:** Anita M, Department of Computer Science, Government Arts College (Affiliated to Bharathidasan University, Tiruchirappalli), Tiruchirappalli, Tamilnadu, India., E-Mail:

How to cite this article: Anita, M., Shakila, S. (2024). Stochastic kernelized discriminant extreme learning machine classifier for big data predictive analytics. The Scientific Temper, **15**(spl):394-403.
Doi: 10.58414/SCIENTIFICTEMPER.2024.15.spl.46

Source of support: Nil

Conflict of interest: None.

Received: 14/09/2024 **Accepted:** 27/09/2024 **Published:** 16/10/2024

of individual LSTM methods to enhance forecast accuracy. But soil moisture attributes such as temperature, humidity, pH, and electrical conductivity, were not considered to enhance the prediction.

Machine learning techniques were developed to predict popular yields of crops with higher accuracy. The model's prediction performance was not enhanced since it failed to select some more relevant features. A Relief algorithm was designed to select the significant feature for efficient agricultural crop yield forecast with the help of ML techniques. However, the time-efficient prediction was not obtained, Gavahi, K., Abbaszadeh, P., & Moradkhani, H. (2021), Bazrafshan, O., Ehteram, M., Latif, S. D., Huang, Y. F., Teo, F. Y., Ahmed, A. N., & El-Shafie, A. (2022), Datta, P., & Faroughi, S. A. (2023), Pant, J., Pant, R. P., Singh, M. K., Singh, D. P., & Pant, H. (2021), Gupta, S., Geetha, A., Sankaran, K. S., Zamani, A. S., Ritonga, M., Raj, R., Ray, S., & Mohammed, H. S. (2022).

For multi-layer soil moisture forecast, Integration of support vector machines (SVM) and ensemble Kalman filter was developed. However it failed to consider more data to further estimate the spatial performance of soil moisture prediction. An artificial neural network was developed to measure the yield prediction with paddy crops by using climatic data. However, it failed to use the soil features for accurate prediction. Coupling Delphi durum wheat method was designed using climate seasonal forecasts information for early crop yield prediction. But a higher sensitivity was not attained in the crop yield prediction, Zhu, Q., Wang, Y., & Luo, Y. (2021), Amaratunga, V., Wickramasinghe, L., Perera, A., Jayasinghe, J., & Rathnayake, U. (2020), Dainelli, R., Calmanti, S., Pasqui, M., Rocchi, L., Di Giuseppe, E., Monotti, C., Quaresima, S., Matese, A., Di Gennaro, S. F., & Toscano, P. (2022).

A classifier ensemble-based prediction method was developed in [9] for rice yield prediction by using climatic datasets. But ensemble-based forecast method was not extended for the prediction of dissimilar crop yields with higher accuracy. For soil moisture prediction based on the correlation between meteorological features, a back propagation (BP) neural network regression method optimized through a genetic algorithm (GA) was developed. However, analysis of soil moisture difference at dissimilar soil depths was not performed, Mishra, S., Mishra, D., Mallick, P. K., Santra, G. H., & Kumar, S. (2021), Liu, D., Liu, C., Tang, Y., & Gong, C. (2022).

Contributing remarks

- A novel SKDELMC model is developed to improve the prediction analysis through feature selection and classification.
- To minimize prediction time and space complexity, stochastic kernelized quadratic discriminant analysis is performed to choose relevant features from the big dataset.
- To design an algorithm named qualitative indexed extreme learning classifier for the forecast with selected features through the Baroni–Urbani–Buser coefficient. Then, the Hardlimit activation function is also employed in the learning process to categorize data into dissimilar classes. This in turn, increases accuracy sensitivity and minimizes the false positive rate.
- An extensive experimental evaluation is performed through an assortment of performance metrics to demonstrate the enhancement of the SKDELMC model over existing techniques.

Structure of the manuscript

The rest of this manuscript is organized as below. Section 2 provides related works on several prediction methods. Section 3 narrates the principle behind our research via the SKDELMC model for prediction. Section 4 describes the experimental setup. Section 5 explains the results as well as their discussions in detail. Section 6 presents the conclusion.

Related works

A deep recurrent Q-network approach was introduced for predicting crop yield. The designed approach minimized error and maximized forecast accuracy, but the computing efficiency of the training process was not minimized. A Gaussian processes (GPs) model was developed for the evaluation of crop yield prediction. However, it failed to assess the model's transportability through multitask GPs for higher crop diversity. An ensembling classifier system was developed for crop yield forecasts depending on soil classification. However feature selection process was not performed to enhance forecast performance with minimum time, Elavarasan, D., & Vincent, P. M. D. (2020), Martínez-Ferrer, L., Piles, M., & Camps-Valls, G. (2021), Waikar, V. C., Thorat, S. Y., Ghute, A. A., Rajput, P. P., & Shinde, M. S. (2020).

ML and AI models were designed for enhanced forecasts of crop yield. But, the performance of crop yield forecast with the minimum error was not attained. The artificial neural network (ANN) method was designed to provide predictions of cotton yield. However, environmental-related factors were not considered for cotton yield prediction. ML model was developed in [16] for better prediction accuracy by using proficient feature selection techniques to preprocess raw data., the prediction time was not reduced, Kundu, S. G., Ghosh, A., Kundu, A., & Girish, G. P. (2022), Yildirim, T., Moriasi, D. N., Starks, P. J., & Chakraborty, D. (2022), Raja, S. P., Sawicka, B., Stamenkovic, Z., & Mariammal, G. (2022).

An XGBoost model was designed for maize yield prediction accuracy by merging soil parameters as well as environmental variables. However, sensitivity analysis was not carried out. The random forest algorithm was designed for accurate crop yield forecasts based on the environment as well as weather data. But complexity of the crop yield

prediction was not reduced, Nyéki, A., Kerepesi, C., Daróczy, B., Benczúr, A., Milics, G., Nagy, J., Harsányi, E., Kovács, A. J., & Neményi, M. (2021), Jhajhariaa, K., Mathura, P., Jaina, S., & Nijhawan, S. (2023).

A Multiscale Extrapolative Learning Algorithm (MELA) was developed for predicting crop yields depending on soil moisture data. The designed algorithm failed to include validation of consistent extensibility of time series of various data sorts than the soil moisture. Crop yield prediction through remotely sensed data, Deep Learning Multi-Layer Perceptron (DLMLP) neural networks were introduced. But it failed to obtain better as well as more precise yield data, Chakraborty, D., Başağaoğlu, H., Alian, S., Mirchi, A., Moriasi, D. N., Starks, P. J., & Verser, J. A. (2023), Tripathi, A., Tiwari, R. K., & Tiwari, S. P. (2022).

Methodology

Crop yield forecast is a crucial task for agricultural experts to make informed decisions about planting, harvesting, as well as administration of their crops. By using climate as well as soil data, with corresponding crop yield data, the aim is to develop a predictive method for accurately estimating crop yields. Conventional methods have some significant challenges to performing accurate predictive analytics in a time-efficient manner. Therefore, the SKDELMC model is developed for crop yield forecast by evaluating the big data with minimum time consumption.

The SKDELMC model consists of feature selection and classification that offer several advantages in crop yield forecast. The feature selection process of the SKDELMC model improves the model performance by the dimensionality of the data and focuses on the variables that have the most significant impact on the target. Also significantly reduces the computational complexity and training time required for predictive models to be more time-efficient and scalable. Overall, feature selection and classification processes play a vital role in improving model performance, computational efficiency, and robustness in prediction tasks. The architecture diagram of the SKDELMC model is shown in figure 1.

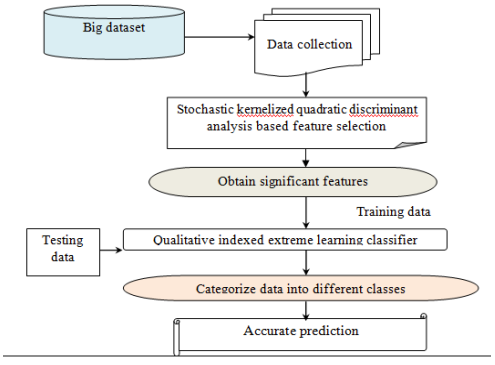

Figure1: architecture diagram of SKDELMC model

Figure 1, given above illustrates the architecture design of SKDELMC that includes "Stochastic kernelized quadratic discriminant analysis-based feature selection for enhancing prediction accuracy through big data." Initially, a data collection process is carried out that involves gathering and storing large volumes of structured data—the key steps involved in collecting and managing the big data collected from the dataset.

When dealing with big data, dimensionality reduction techniques are employed in the SKDELMC model to decrease the number of features while preserving important information. Stochastic kernelized quadratic discriminant analysis is commonly used for dimensionality reduction. Once the relevant features from your dataset are present, the SKDELMC model proceeds with data classification. The dataset is divided into two subsets called training set and test set to train a classification model. Selected features from the training set are fed into the chosen classification algorithm called the Qualitative indexed extreme learning classifier. The algorithm learns the patterns and analyzes the relationships among training and testing data as well as finally provides the corresponding labels or classes. A detailed explanation of the proposed SKDELMC model is given below.

Stochastic kernelized quadratic discriminant analysis-based feature selection

Feature selection is a method of choosing a subset of significant features or attributes from a bigger set of features in a dataset. The main aim is to identify the discriminative features that contribute the most to the prediction analysis. As the number of features or dimensions in the dataset enhances, the amount of data needed to efficiently symbolize and analyze that data efficiently also increases rapidly. By eliminating irrelevant or redundant features, the proposed SKDELMC model aims to improve model performance, optimize computational efficiency, and obtain dimensionality reduction.

The SKDELMC model uses stochastic kernelized quadratic

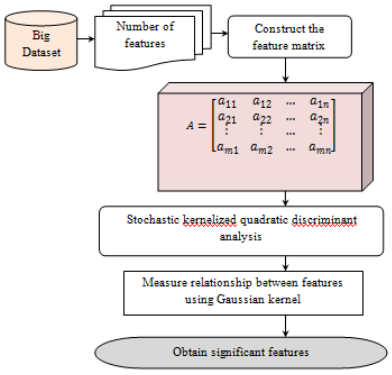

Figure 2: Flow diagram of stochastic kernelized quadratic discriminant analysis-based feature selection

discriminant analysis for relevant feature selection from a big dataset. Discriminant analysis is a method employed to measure likelihood estimation through the help of Gaussian kernel functions. The likelihood method is a measure of the relationship among features.

Figure 2 illustrates a flow diagram of stochastic kernelized quadratic discriminant analysis-based feature selection. The big crop yield dataset is provided as input. To begin with, raw input big dataset ' D ' originated in the structure of matrix as below.

$$A = \left[a_{11}\, a_{12} \ldots a_{1n}\, a_{21}\, a_{22} \ldots a_{2n} \vdots \ldots \vdots a_{m1}\, a_{m2} \ldots a_{mn} \right] \quad (1)$$

Where A denotes an input feature matrix with 'n' column and m row. Indicates a number of instances. Column denotes a number of features. Through the above set of matrices, initial relevant features are identified. To obtain relevant features from a huge dataset, the Gaussian kernel is applied to analyze the relationship among columns of features in the matrix.

Let us consider a_i and a_j be column vectors of features. The Gaussian kernel is expressed as below,

$$K\left(a_i, a_j\right) = \frac{1}{\sqrt{2\pi}\, R} \exp \exp \left[-0.5 \left[\frac{\left| a_i - a_j \right|}{R} \right]^2 \right] \quad (2)$$

Where $K\left(a_i, a_j\right)$ denotes Gaussian kernel Quadratic discriminant analysis output, R indicates deviation. The kernel function provides an output score from 0 to 1. On the scores obtained from equation (2), select a subset of features that are most relevant to the classification task. The highest score value is used for selecting the top relevant features. The output of the kernel is typically a decision function that provides two outputs to each input feature such as relevant or irrelevant.

$$Y = \{ K\left(a_i, a_j\right) > 0.5 \text{ ;relevant features subset Otherwise; irrlevant feature} \quad (3)$$

Along with the decision output ($K\left(a_i, a_j\right) > 0.5$), the relevant feature subset is used for the next classification task. Other irrelevant or redundant features are removed. This process reduces the time complexity of prediction.

The overall Gaussian kernel quadratic discriminant analysis algorithm is given below.

Algorithm 1 illustrates a procedure of relevant feature selection as well as redundant feature removal using Gaussian kernel Quadratic discriminant analysis. The raw dataset is chosen and constructs the input feature matrix. The gaussian kernel is employed for measuring the relationship between features. After that, based on the estimated score value, relevant features are chosen, as well as eliminating redundant features. This assists in enhancing accurate forecasts in a timely manner.

// Algorithm 1: Gaussian kernel Quadratic discriminant analysis based feature selection

Input: Dataset 'D', features or attributes $A = \{a_1, a_2, a_3, \ldots a_n\}$

Output: selected attributes

Begin

1. **For each** dataset 'D' with attributes 'A'
2. Create input feature matrix 'A' as given in (1)
3. Measure relationship between the features using (2)
4. **If** $(K(a_i, a_j) > 0.5)$ **then**
5. **Select** relevant features
6. **else if** '$(K(a_i, a_j) < 0.5)$' **then**
7. **Irrelevant** features
8. **End if**
9. **Select relevant** features and **remove redundant** features
10. **End for**

End

Qualitative indexed extreme learning classifier-based prediction

Once relevant features are chosen, the next step is to categorize the data into different classes using the qualitative indexed extreme learning classifier. It is a feed-forward neural network having a straightforward solution without requiring any iteration. The network comprises numerous layers. Selected relevant features are provided to the input layer. Then Baroni–Urbani–Buser coefficient is applied in the hidden layer by analyzing testing and training data. It is the qualitative index used to analyze the similarity between the data. After that, the Hardlimit activation function is used for evaluating similarity value as well as providing classification results.

Figure 3 illustrates the structure of a qualitative indexed extreme learning classifier for accurate data classification. It is a sort of feed-forward neural network employed for data classification as well as feature learning by a single layer or multiple layers of hidden. In Figure 3, let us assume which training set {T, Z} where T indicates training data through selected features.' $\left\{a_1, a_2, \ldots, a_k\right\}$ ' and label or output 'Z' denoting its type that belongs to dissimilar classes.

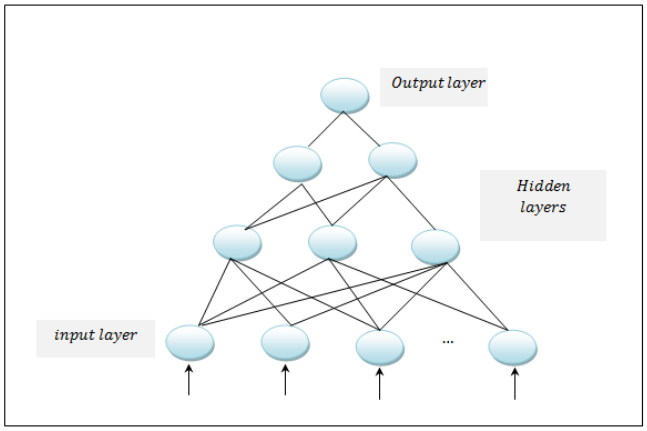

Figure 3: constructions of qualitative indexed extreme learning classifier

In Figure 2, the classifier receives 'n' training data ($T_i = T_1, T_2, \ldots, T_n$), and arbitrarily set a weight matrix between the input and hidden layer.

$$\beta_{ij} = \left[\beta_{11} \beta_{12} \ldots \beta_{1n} \beta_{21} \beta_{22} \ldots \beta_{2n} \vdots \ldots \vdots \beta_{m1} \beta_{m2} \ldots \beta_{mn} \right] \quad (4)$$

Where, β_{ij} denotes a weight matrix among input as well as hidden layer, and bias is added as follows,

$$X_i = \sum_{i=1}^{n} \left(T_i * \beta_{ij} \right) + W \quad (5)$$

Where activity of neurons at the input layer.' X_i ', ' β_{ij} ' indicate weight between input as well as a hidden layer, added bias function'W' that stored value is '1'. Layer receives only training data. However, it did not carry out any mathematical process,

In hidden layers, Baroni–Urbani–Buser coefficient is applied for analysing testing and training data. is the qualitative index used to evaluate the similarity between the training and testing data.

$$B_c = 1 - \left[\frac{N - \sum_{j=1}^{m} \sum_{i=1}^{n} \left| T_i - Ts_j \right|}{N + \sqrt{T_i * Ts_j} - Ts_j} \right] \quad (6)$$

Where, B_c indicates a Baroni–Urbani–Buser coefficient, N indicates a number of data, T_i denotes training data, Ts_j indicates testing data. The coefficient returns the output values as 0 or 1. coefficient outcomes are provided to the Hardlimit activation function to provide classification results.

Output of the hidden layer is expressed as below,

$$Q = \sum_{i=1}^{b} f_a \left(\beta_{jk} h_o + W \right) \quad (7)$$

Where, Q ' indicates an output of hidden layer, f_a denotes an activation function, ' β_{jk} ' denotes weight among j^{th} hidden layer neuron and k^{th} output layer neuron, h_o represent output of previously hidden layer, b indicate the number of hidden units, W indicates a bias

The hardlimit activation function ' f_a ' pushes the neuron to generate the output 1 if the Baroni–Urbani–Buser coefficient reaches a maximum value, otherwise it outputs 0. This allows a neuron to make a decision or classification

$$f_a = \{1, \text{ if} B_c \ 0, \text{ otherwise} \quad (8)$$

Where, f_a indicates an activation function returns '1' if the coefficient reaches a maximum value ' B_c ', otherwise ' f_a ' returns' 0'. Finally, output of final classification at output layer is given below,

$$Z = Q\beta_j \quad (9)$$

Where, Z indicates an output of the classifier, Q indicate output of hidden layer, β_j represent a weight of output layer. Finally, the classified results are obtained at the output layer. Depend on accurate classification outcomes, prediction is carried out by higher accuracy and lesser error rate.

The above algorithmic steps are employed for classifying input data to different categories using a Qualitative indexed extreme learning classifier. Extreme learning classifier receives training data as input. The classifier uses the training data to construct a hidden layer. Similarity coefficients between training data and testing data are measured in the hidden layer. Similarity coefficient values obtained in the previous step are analyzed using a hardlimit activation function. The hardlimit function applies a threshold to the similarity coefficients, transforming them into binary values (0 or 1). This analysis helps determine which data are classified into particular classes. The analysis is performed, and final classified outcomes are obtained at the output layer. These results indicate the class or category to which each input data belongs based on the activation of hidden layer neurons. With classified results obtained in the previous step, the effective prediction results are obtained with minimum error.

Experimental evaluation

Experimental evaluations of the SKDELMC and existing DeepYield and BMA are implemented using JAVA with SMART FASAL (Smart Irrigation and Fertilization System for Precision Agriculture using Internet of Things and Cloud Infrastructure) dataset taken from http://smartfasal.in/ftp-dataset-portal/. Portal stores real-time soil data for three crops namely Capsicum, Wheat Dataset, and Rice Dataset. Among three crops, the rice dataset is considered to perform the experiment. The dataset comprises 13 attributes or features and 42666 instances. First, soil moisture data and weather conditions are collected for Precision Agriculture, Gavahi, K., Abbaszadeh, P., & Moradkhani, H. (2021), Bazrafshan, O., Ehteram, M., Latif, S. D., Huang, Y. F., Teo, F. Y., Ahmed, A. N., & El-Shafie, A. (2022).

// Algorithm 2: Qualitative indexed extreme learning classifier

Input: selected features (i.e. training data) T_i and testing data Ts_i

Output: Increase the prediction accuracy

Begin

1. Number of relevant features with training data given to **input layer**
2. Assign the random weight and bias $X_i = \Sigma_{i=1}^{n} \left(T_i * \beta_{ij} \right) + W$
3. **For each** training data T_i **// [hidden layer 1]**
4. **For each** testing data T_j
5. Measure the similarity $B_c = 1 - \left[\frac{N - \Sigma_{j=1}^{m} \Sigma_{i=1}^{n} \left| T_i - Ts_j \right|}{N + \sqrt{T_i * Ts_j} - Ts_j} \right]$
6. **end for**
7. **end for**
8. **Apply** activation function to estimate similarity value
9. **if** (max B_c) **then// [hidden layer 2]**
10. f_a provides the results '1'
11. **else**
12. f_a provides the results '0'
13. **end if**
14. Obtain the classification results at the **output layer**

End

Table 1: Feature Description

S. No	*Feature*	*Description*
1	Sensor ID	
2	Soil_moisture 1	Acquires information from the sensors installed within the soil at a depth level 15cms
3	Soil_moisture 2	Acquires information from the sensors installed within the soil at a depth level 45cms
4	Soil_moisture 3	Acquires information from the sensors installed within the soil at a depth level 80cms
5	TEMP	Soil temperature
6	HUMD	Soil humidity
7	PRSR	Soil pressure
8	LMNS	Soil Luminosity
9	Rainfall	Rainfall per day (mm)
10	week cycle count	Week cycle count of recording
11	Day	day of recording
12	Date	Date of recording (DD:MM:YY)
13	Time IST	Time of recording

Experimental Results for Model Comparison

Experimental results of SKDELMC and conventional DeepYield and BMA are discussed through dissimilar evaluation parameters, namely prediction accuracy, sensitivity, false-positive rate, prediction time, and space complexity, Gavahi, K., Abbaszadeh, P., & Moradkhani, H. (2021), Bazrafshan, O., Ehteram, M., Latif, S. D., Huang, Y. F., Teo, F. Y., Ahmed, A. N., & El-Shafie, A. (2022).

Comparison of prediction accuracy

It is measured as a ratio of a number of data taken from the dataset correctly classified to dissimilar classes to total number of data taken for experimentation. Performance of overall accuracy is evaluated as given below,

$$\mathrm{Pre}_a = \left[\frac{\mathrm{ACd}_i}{\mathrm{d}_n}\right] * 100 \qquad (10)$$

Where 'Pre_a' denotes the prediction accuracy, 'ACD_1' represents a number of data properly classified and 'd_n' indicates a total number of data. It is measured in percentage (%).

Figure 4 given above illustrates the graphical analysis of prediction accuracy with number of data related to weather and soil taken from the dataset. This figure, the x-axis indicates a number of data, and y-axis denotes prediction accuracy of crop yield. The graph shows both upward and down trend, suggesting that as the number of data increases, the prediction accuracy improves or decreases based on the complexity of the problem quality of data. on the information provided, it appears that the SKDELMC model has shown enhanced prediction accuracy than the existing DeepYield and BMA methods. The SKDELMC model utilizes a qualitative indexed extreme learning classifier and applies a Baroni–Urbani–Buser coefficient to examine provided training data samples by testing data for crop yield forecast. Depending on analyzed results, the crop yield prediction is attained., examined outcome indicates that performance of crop yield prediction accuracy by SKDELMC model has shown a 10% increase compared to the DeepYield method and a 6% increase compared to the BMA method. This suggests that the SKDELMC model has outperformed other two techniques in terms of prediction accuracy for crop yield, Gavahi, K., Abbaszadeh, P., & Moradkhani, H. (2021), Bazrafshan, O., Ehteram, M., Latif, S. D., Huang, Y. F., Teo, F. Y., Ahmed, A. N., & El-Shafie, A. (2022).

Table 2: Prediction Accuracy versus Number of Data

Number of data	*Prediction Accuracy (in %)*		
	SKDELMC (%)	*DeepYield (%)*	*BMA (%)*
4000	96.4	88.05	90.3
8000	95.68	89	90.82
12000	97.12	87.12	91.54
16000	96.52	86.86	91.03
20000	95.61	87.27	90.61
24000	95.88	86.89	89.21
28000	97.18	86.26	91.62
32000	96.95	87.04	91
36000	96.14	87.93	90.7
40000	95.6	86.41	89.97

Comparison of sensitivity:

It is measured as ratio of number of true positives i.e. proportion of correct predictions in predictions of positive

Table 3: Sensitivity (in %) Versus Number of Data

Number of data	*Sensitivity (in %)*		
	SKDELMC (%)	*DeepYield (%)*	*BMA (%)*
4000	95.27	86.45	89.7
8000	94.56	87.77	89.43
12000	95.45	85.18	88.79
16000	94.44	82.84	88.26
20000	94.94	85.62	87.77
24000	93.19	84.85	87.59
28000	95.03	83.71	86.11
32000	94.4	85.79	88.78
36000	95.03	84.04	86.81
40000	93.91	83.13	85.38

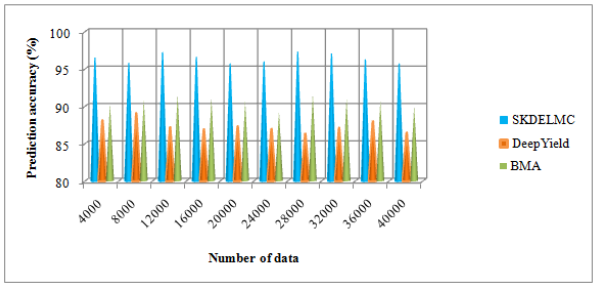

Figure 4: Graphical analysis of prediction accuracy

class to total number of data taken for experimentation. It is evaluated as given below,

$$S = \left[\frac{TPd_i}{d_n}\right] * 100 \quad (11)$$

Where 'S denotes the sensitivity, 'TPd_{ii}' indicate number of data true positively classified and 'd_n' denotes total number of data. It is measured in percentage (%).

Figure 5 shows the graphical analysis for sensitivity with number of data in crop yield prediction with big data. Sensitivity refers to different aspects of the prediction model, such as its ability to analyze the weather and solid data for a rice crop. In this figure 5, x-axis represents number of data, and y-axis indicates sensitivity of prediction model. With the big data, as number of data increases, SKDELMC model potentially measures a similarity and variations present in the data. This increased sensitivity allows the SKDELMC model to make accurate predictions. It's important to note that the relationship between testing and training data depends on the Baroni–Urbani–Buser coefficient. Then the activation function analyzes the results with higher true positive rate. The overall comparison results information provided, the proposed SKDELMC model has shown a sensitivity rate of 12% improved in crop yield prediction when compared and 8% when compared, Gavahi, K., Abbaszadeh, P., & Moradkhani, H. (2021), Bazrafshan, O., Ehteram, M., Latif, S. D., Huang, Y. F., Teo, F. Y., Ahmed, A. N., & El-Shafie, A. (2022) .

False positive rate

It is measured as ratio of number of data falsely or wrongly classified to various classes to total number of data taken for experimentation. Performance of overall accuracy is evaluated as given below,

$$FPR = \left[\frac{NICd_i}{d_n}\right] * 100 \quad (12)$$

Where 'FPR' indicates a false positive rate, '$NICd_i$' denotes the number of data wrongly classified 'd_n' be total number of data. It is measured in percentage (%).

Figure 6 illustrates performance analysis of the false

Table 4: False Positive Rate (in %) Versus Number of Data

Number of data	*False Positive Rate (in %)*		
	SKDELM	*DeepYield*	*BMA*
4000	3.6	11.95	9.7
8000	4.31	11	9.17
12000	2.87	12.87	8.45
16000	3.47	13.13	8.96
20000	4.39	12.72	9.39
24000	4.11	13.1	10.78
28000	2.81	13.73	8.37
32000	3.04	12.95	8.99
36000	3.85	12.06	9.29
40000	4.39	13.58	10.03

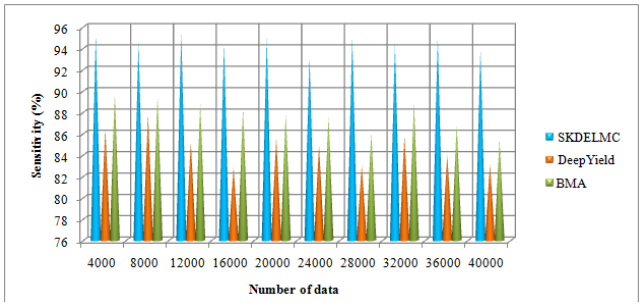

Figure 5: Graphical analysis of sensitivity

positive rate with a number of data by SKDELMC and existing DeepYield and BMA. As shown in Figure 6, the performance of the false positive rate is considerably reduced by the SKDELMC model than the existing methods. This is because the data accurately classified through enhanced accuracy and a true positive rate. Improved performance in terms of false positive rate is attained through accurate classification of data with higher accuracy and true positive rate. This

Table 5: Prediction time versus Number of Data

Number of data	*Prediction Time (in ms)*		
	SKDELM	*DeepYield*	*BMA*
4000	16.8	22	18
8000	18.4	22.4	20.8
12000	25.2	31.2	27.6
16000	28.8	38.4	32
20000	30	40	36
24000	33.6	43.2	37.2
28000	40.6	47.6	42
32000	45.44	51.2	48
36000	50.76	57.6	55.08
40000	54.4	64	60.4

Table 6: Space Complexity versus Number of Data

Number of data	*Space complexity (in MB)*		
	SKDELMC (MB)	*Deepyield (MB)*	*BMA (MB)*
4000	17.2	24	20
8000	20	26.4	22.4
12000	25.2	33.6	30.6
16000	28.8	38.4	33.6
20000	32	40	36
24000	34.8	39.6	37.2
28000	38.08	42.56	40.6
32000	41.6	48	44.8
36000	50.4	57.6	54
40000	58	64	60

accuracy is achieved by applying an extreme machine classifier, which accurately categorizes the data and by minimizing incorrect classifications through the use of a similarity coefficient and activation functions. In the experiment conducted with 4000 data, the false positive rate was observed to be 3.6% using the SKDELMC model, while it was 11.95 and 9.7% using the existing methods, respectively. This indicates a significant reduction in the false positive rate when by SKDELMC model. average of ten results further supports the conclusion that the SKDELMC model reduces false positive rate by 71% and 61% in crop yield prediction, Gavahi, K., Abbaszadeh, P., & Moradkhani, H. (2021), Bazrafshan, O., Ehteram, M., Latif, S. D., Huang, Y. F., Teo, F. Y., Ahmed, A. N., & El-Shafie, A. (2022).

Prediction time

It is formulated as amount of time taken for accurate prediction of future results through the data classification. The prediction time is formulated as given below,

$$T_p = \left[d_n\right] * T\left(d_i\right) \quad (13)$$

Where 'T_p' denotes prediction time, d_n represents number of data and '$T\left[d_i\right]$' denotes time for classifying single data. It is measured in milliseconds (mms).

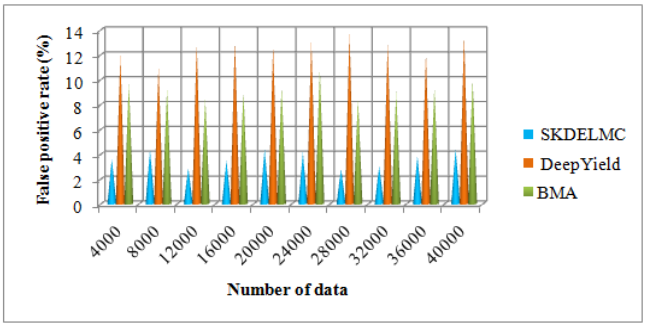

Figure 6: Graphical analysis of false positive rate

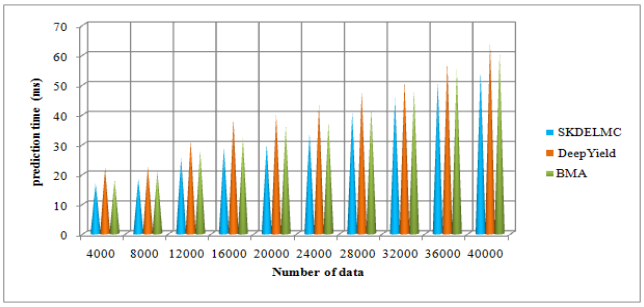

Figure 7: Graphical analysis of prediction time

Figure 7 depicts graphical representation of prediction time regarding a number of data. The graph shows that the prediction time of all three techniques generally increases as number of data enhances. However, comparatively, SKDELMC model demonstrates a decreased prediction time compared to the conventional methods. Reduction in prediction time in the SKDELMC model is selection of significant features from dataset. Technique applies Stochastic Kernelized Quadratic Discriminant Analysis, utilizing a Gaussian kernel, to identify the most informative features. By focusing on these significant features, the SKDELMC model reduces the computational burden and improves the prediction process. According to the validation results, the prediction time using the SKDELMC model is reported to be reduced by 19% compared to DeepYield and 9% compared to BMA. This suggests that the SKDELMC model offers improved efficiency in terms of prediction time compared to conventional methods, Gavahi, K., Abbaszadeh, P., & Moradkhani, H. (2021), Bazrafshan, O., Ehteram, M., Latif, S. D., Huang, Y. F., Teo, F. Y., Ahmed, A. N., & El-Shafie, A. (2022).

Comparison of space complexity

It is defined as amount of memory space consumed through algorithm to perform accurate big data prediction. The memory consumption is calculated using given formula,

$$S_{com} = \left[d_n\right] * MS\left[d_i\right] \quad (14)$$

Where, 'S_{com}' denotes the space complexity, 'd_n' represents the number of data and '$MS\left[d_i\right]$' is the memory consumed for single data. It is measured in terms of Megabytes (MB).

The performance analysis of space complexity for a proposed technique called SKDELMC in comparison to two other methods, DeepYield and BMA. The analysis is conducted using a range of data sizes from 4000 to 40000. According to the results obtained, the space complexity of SKDELMC is minimized compared to DeepYield and BMA. Figure 8 shows that as the number of data increases, the space consumption of all three methods, including SKDELMC, also increases. But the SKDELMC model employed Gaussian kernel Quadratic discriminant analysis to select a reduced number of features for predictive analytics. This feature

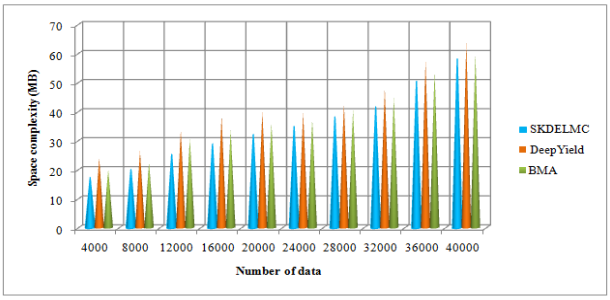

Figure 8 Graphical analysis of space complexity

Figure 7: Graphical analysis of space compexity

selection approach resulted in a lesser amount of storage space required during the prediction process. On average, based on ten results, it is found that the space complexity of the SKDELMC model is reduced by 18% compared to DeepYield and 10% compared to BMA. These findings imply that the SKDELMC model offers a more efficient use of storage space for predictive analytics compared to the other two methods, Gavahi, K., Abbaszadeh, P., & Moradkhani, H. (2021), Bazrafshan, O., Ehteram, M., Latif, S. D., Huang, Y. F., Teo, F. Y., Ahmed, A. N., & El-Shafie, A. (2022).

Conclusion

Big data refers to enormous volume, variety, in addition to velocity of data that collects from a variety of sources. This data is often distinguished by its complexity and entails advanced techniques to process, analyze, and extract significant information from it. Predictive analytics is the significant approach used to analyze a large volume of data. Agriculture is a well-known and improved application in big data analytics. Health as well as output are monitored through the application of big data predictive analytics in precision agriculture with soil quality and weather-conditional data. It leverages a deep learning technique called the SKDELMC model introduced to analyze large datasets and predict future behavior or events. To explore these large datasets and make predictions, the SKDELMC model, a deep learning technique, is introduced. This technique enables the feature selection and analysis of big data in precision agriculture. The feature selection process is a crucial step in data analysis when dealing with large datasets. It involves identifying the relevant features to enhance accuracy and reduce the complexity of prediction. Finally, the classification is done with the relevant features using an extreme learning classifier to learn patterns and relationships and make predictions or classify different data. Comprehensive experimental evaluation is performed through different performance metrics with respect to a number of data. Overall performance metric analysis illustrates that the presented SKDELMC model achieves higher prediction accuracy and sensitivity with lesser time, space complexity, and false positive rate than the conventional methods.

References

Amaratunga, V., Wickramasinghe, L., Perera, A., Jayasinghe, J., & Rathnayake, U. (2020). Artificial neural network to estimate the paddy yield prediction using climatic data. *Mathematical Problems in Engineering*, 2020, 1–11.

Bazrafshan, O., Ehteram, M., Latif, S. D., Huang, Y. F., Teo, F. Y., Ahmed, A. N., & El-Shafie, A. (2022). Predicting crop yields using a new robust Bayesian averaging model based on multiple hybrid ANFIS and MLP models. *Ain Shams Engineering Journal*, 13(5), 1–21.

Chakraborty, D., Başağaoğlu, H., Alian, S., Mirchi, A., Moriasi, D. N., Starks, P. J., & Verser, J. A. (2023). Multiscale extrapolative learning algorithm for predictive soil moisture modeling and applications. *Expert Systems with Applications*, 213, 1–11.

Dainelli, R., Calmanti, S., Pasqui, M., Rocchi, L., Di Giuseppe, E., Monotti, C., Quaresima, S., Matese, A., Di Gennaro, S. F., & Toscano, P. (2022). Moving climate seasonal forecasts information from useful to usable for early within-season predictions of durum wheat yield. *Climate Services*, 28, 1–14.

Datta, P., & Faroughi, S. A. (2023). A multihead LSTM technique for prognostic prediction of soil moisture. *Geoderma*, 433, 1–13.

Elavarasan, D., & Vincent, P. M. D. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. *IEEE Access*, 8, 86886–86901.

Gavahi, K., Abbaszadeh, P., & Moradkhani, H. (2021). DeepYield: A combined convolutional neural network with long short-term memory for crop yield forecasting. *Expert Systems with Applications*, 184, 1–11.

Gupta, S., Geetha, A., Sankaran, K. S., Zamani, A. S., Ritonga, M., Raj, R., Ray, S., & Mohammed, H. S. (2022). Machine learning- and feature selection-enabled framework for accurate crop yield prediction. *Journal of Food Quality*, 2022, 1–7.

Jhajhariaa, K., Mathura, P., Jaina, S., & Nijhawan, S. (2023). Crop yield prediction using machine learning and deep learning techniques. *Procedia Computer Science*, 218, 406–417.

Kundu, S. G., Ghosh, A., Kundu, A., & Girish, G. P. (2022). A ML-AI enabled ensemble model for predicting agricultural yield. *Cogent Food & Agriculture*, 8, 1–21.

Liu, D., Liu, C., Tang, Y., & Gong, C. (2022). A GA-BP neural network regression model for predicting soil moisture in slope ecological protection. *Sustainability*, 14, 1–14.

Martínez-Ferrer, L., Piles, M., & Camps-Valls, G. (2021). Crop yield estimation and interpretability with Gaussian processes. *IEEE Geoscience and Remote Sensing Letters*, 18(12), 2043–2047.

Mishra, S., Mishra, D., Mallick, P. K., Santra, G. H., & Kumar, S. (2021). A classifier ensemble approach for prediction of rice yield based on climatic variability for the coastal Odisha region of India. *Informatica*, 45, 367–380.

Nyéki, A., Kerepesi, C., Daróczy, B., Benczúr, A., Milics, G., Nagy, J., Harsányi, E., Kovács, A. J., & Neményi, M. (2021). Application of spatio-temporal data in site-specific maize yield prediction with machine learning methods. *Precision Agriculture*, 22, 1397–1415.

Pant, J., Pant, R. P., Singh, M. K., Singh, D. P., & Pant, H. (2021). Analysis of agricultural crop yield prediction using statistical techniques of machine learning. *Materials Today: Proceedings*, 46, 10922–10926.

Raja, S. P., Sawicka, B., Stamenkovic, Z., & Mariammal, G. (2022). Crop prediction based on characteristics of the agricultural environment using various feature selection techniques and

classifiers. *IEEE Access*, 10, 23625–23641.

Tripathi, A., Tiwari, R. K., & Tiwari, S. P. (2022). A deep learning multi-layer perceptron and remote sensing approach for soil health-based crop yield estimation. *International Journal of Applied Earth Observation and Geoinformation*, 113, 1–12.

Waikar, V. C., Thorat, S. Y., Ghute, A. A., Rajput, P. P., & Shinde, M. S. (2020). Crop prediction based on soil classification using machine learning with classifier ensembling. *International Research Journal of Engineering and Technology (IRJET)*, 7(5), 4857–4861.

Yildirim, T., Moriasi, D. N., Starks, P. J., & Chakraborty, D. (2022). Using artificial neural network (ANN) for short-range prediction of cotton yield in data-scarce regions. *Agronomy*, 12, 1–19.

Zhu, Q., Wang, Y., & Luo, Y. (2021). Improvement of multi-layer soil moisture prediction using support vector machines and ensemble Kalman filter coupled with remote sensing soil moisture datasets over an agriculture dominant basin in China. *Hydrological Processes*, 35(4), 1–22.