International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I9P211

Volume 69 Issue 9, 80-91, September, 2021
©2021 Seventh Sense Research Group®

# Predictive Analytics in Soil for Agriculture Using Kendall Normalized Feature Selection Based Jaccarized Rocchio Boyer-Moore Bootstrap Aggregative Mapreduce Classifier for Predictive Analytics with Big data

Anita M[1], Dr. Shakila S[2]

*PG & Research, Department of Computer Science, Government Arts College, Trichy. Tamilnadu, India.*

[1]anitarajkumar040908@gmail.com, [2]shakilamuthusamy@gmail.com

***Abstract*** *- Big Data investigation is the method of collecting, arranging, and examining a huge amount of raw data, which extracts useful information. Big data investigation is a very difficult practice of investigating large datasets for taking future decisions. The conventional techniques failed to look up the prediction accurateness and also diminishes the prediction time while processing the large volumes of data. A novel Kendall Normalized feature selection based Jaccarized Rocchio Boyer-Moore Bootstrap Aggregative Mapreduce classifier (KNFS-JRBMBAMC) method is the preamble for advance predicting the potential outcomes with elevated prediction accurateness and get with smaller time. The KNFS-JRBMBAMC methods encompass two techniques, namely data-based feature selection and its related classification for prediction. In the KNFS-JRBMBAMC method, Kendall Ranking Correlative Normalized Discriminant feature selection is agreed to identify the linear combination of features and select the relevant features for performing the classification task. After feature selection, the Jaccarized Rocchio Boyer-Moore Bootstrap Aggregative Mapreduce classification method is applied for classifying the raw input data into dissimilar classes with higher classification accuracy using the Boyer-Moore voting scheme. Then, the map () and reduce () function is used for the classifier result to perform an accurate prediction. Exploratory assessment is completed utilizing agricultural soil data collection set on factors like expectation exactness, bogus positive rate, forecast time, and space intricacy regarding various information. The talked about outcomes investigation shows that the KNFS-JRBMBAMC strategy gives better execution as far as accomplishing higher expectation exactness and lesser time just as space intricacy when contrasted with the cutting edge works*

**Keywords** *- Intelligent Data Prediction(IDP), Extreme Learning Machine (ELM) Kendall Normalized feature selection (KNFS), Jaccarized Rocchio Boyer-MooreBootstrap aggregation, Jaccarized similarity MapReduce, (JRBMBAMC), Cuckoo–Grey wolf-based Correlative Naive Bayes classifier and MapReduce Model (CGCNB-MRM)*

## I. INTRODUCTION

Big Data is a set of a large volume of information stored in a large database for further analysis. Enormous information is steadily being utilized in different applications like industry, monetary managing, farming, medication, etc., since it handles a lot of information. In huge information, the prescient examination is the huge cycle of mining the significant data from enormous datasets to discover future results. Several methods have been developed for taking care of the huge volume of information. An Intelligent Data Expectation (Prediction) (IDP) technique was created in [1] to anticipate soil hefty metal substance utilizing MBPSO and LSSVM and limit the blunder of information prediction. However, the performance of prediction exactness of the model was not expanded. Cuckoo–Grey wolf-based Correlative Naive Bayes classifier and MapReduce Model (CGCNB-MRM) was presented in [2] for the Classification of big data. However, the performance of accurate prediction with minimum time consumption was not obtained.

An improved online bagging algorithm was designed in [3] for the characterization of the developing huge information stream. The designed algorithm improves the better accuracy. However, it failed to point in on further developing the execution productivity of the calculation algorithm.

Markov random fields (MRF) approach was developed in [4] for predictive analytics using big data. But, the data analytics was not performed using efficient, innovative machine learning techniques to extract useful information.

A big data processing approach was introduced in [5] using climate and health data. But the efficient machine learning technique was not applied to further improve the prediction accuracy. Different Machine Learning


This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Techniques were developed in [6] for predicting the soil quality based on chemical, physical and biological compositions. But it failed to perform the feature selection to minimize the time complexity.

Optimizing the extreme learning machine technique was introduced in [7] for predicting the soil quality. However, the designed technique failed to analyze more soil parameters for accurate prediction. Advanced machine learning models were developed in [8] to accurately predict the accuracy of soil temperature. However, the designed models failed to handle more data for predictive analytics. Mainly this type of model will help the mining analyst to predict the effective ranking in the market and their social reachability and also the analyst to identify their needs in various ranges. Most of the industries need the big data analyst to predict their outcomes

A five-layer-fifteen level (FLFL) satellite distant detecting information the executive's structure was portrayed in [9] for management and precision agriculture. But, the higher prediction accuracy was not obtained using FLFL satellite far off detecting information the board structure. The partially autoregressive coordinated moving normal (FARIMA) strategy was created in [10] to gauge the day-by-day mud high temperature (soil). The designed method reduces the error rate, but it failed to use the machine learning model for improving the accuracy.

### *A. New Technique to Upgrade Existing Flaws*

The existing techniques have a few limitations, such as lesser precision, additional time utilization, higher blunder rate, etc. To beat such sorts of issues, an original method called KNFS-JRBMBAMC is presented. The significant commitment of the KNFS-JRBMBAMC method is summed up as given below. To further develop the forecast exactness, a clever procedure called KNFS-JRBMBAMC is presented using a MapReduce function.

The KNFS-JRBMBAMC uses the Kendall Ranking Correlative Normalized Discriminant feature selection to identify relevant feature subsets for performing the classification task. This helps to limit the expectation reality intricacy.

The KNFS-JRBMBAMC practice uses the map function in the Jaccarized Rocchio Boyer-Moore Bootstrap Aggregative classifier. The ensemble classifier uses the Rocchio classifier for measuring the Jaccard similarity between training data and testing data. Based on similarity value, the data is classified into different classes.

After that, the Boyer-Moore voting scheme is applied in reduce phase for identifying the majority voting samples to display the last classification results. This, in turn, builds the forecast exactness and limits the space complexity.

Finally, an experiment is conducted to evaluate the analysis of the KNFS-JRBMBAMC technique against the existing methods with various performance metrics.

The paper is prepared into diverse sections. Section II reviews the related work in the field of big data prediction and classification. Section 3 describes the detailed explanation of the KNFS-JRBMBAMC technique with the help of an architecture outline and calculation.

In segment 4, the test setting of the proposed techniques is depicted. Segment 5 examines the results of proposed and existing methods. Section 6 depicts the conclusion of the paper.

## II. EXISTING METHODS IN PREDICTION

A convolution long transient memory (convLSTM) framework was created in [11] for precipitation forecast utilizing spatial-fleeting examples. In any case, the time intricacy of expectation was not diminished. An LSTM neural organization was created in [12] to further develop the forecast execution and limit the blunder esteem. However, it neglected to deal with the huge size of preparing information for prescient examination.

A profound learning model was created in [13] for soil dampness expectation. Be that as it may, it neglected to break down the effect of different meteorological highlights on the precision of soil dampness forecast. Extreme Learning Machine (ELM) algorithm was designed in [14] using a feature selection process for improving the forecasting accuracy of soil strength. But the lesser prediction time was not achieved.

Two new combination models were presented in [15], dependent on Elman neural organization (ENN) coordinated with gravitational pursuit calculation algorithm (GSA) for working on the spatial and worldly assessment at different soil profundities. Be that as it may, the execution time of mixture models was somewhat more than the typical strategies. A Fluffy Correlative Innocent naïve Bayes Classifier with Map Reduce approach was presented in [16] for Large data order classification. The planned calculation algorithm expands the exactness, yet the bogus positive rate was not limited.

Chi calculation-based Large Information arrangement was introduced in [17] for handling the enormous data. Yet, it neglected to play out the element determination measure for limiting the time intricacy of large information forecast. Diverse AI calculations were created in [18] to work on the arrangement of soil types. However, the soil’s physical and chemical properties were not considered to improve the model execution.

A clever circulated learning calculation was created in [19] to develop a careful and conservative fluffy principle-based classification of Enormous Data. However, the characterization execution was not improved. An all-extended belief rule-based system (EBRBS) was presented in [20] for the characterization of enormous information. The planned framework decreases the time intricacy and further developing the workout efficiency for multi-class classification problems.

## III. ENHANCEMENT OF PREDICTION ANALYTICS

### *A. PROPOSED SOLUTION*

Big Data has become used for several organizations to collect and analyze huge amounts of domain-specific information. Mining and extracting the significant patterns from large input data for prediction is at the core of big data analytics. Analyzing such kinds of big data using machine learning algorithms faces several problems like more time-consuming and complexity in big data analytics. These problems need to be solved by presenting a clever method called KNFS-JRBMBAMC.

Figure 1 given below delineates a design outline of the KNFS-JRBMBAMC procedure for prescient investigation utilizing large data. Initially, the quantity of highlights set {k_1,k_2,k_3,… .k_(n ) } and number of information d_1,d_2,d_3,… .d_(m ) are gathered from the huge dataset. Then, Kendall Ranking Correlative Normalized Discriminant analysis is performed to evaluate features and finds the relevant features for classification. Finally, classification is done with the selected feature set using Jaccarized Rocchio Boyer-Moore Bootstrap Aggregative classifier.

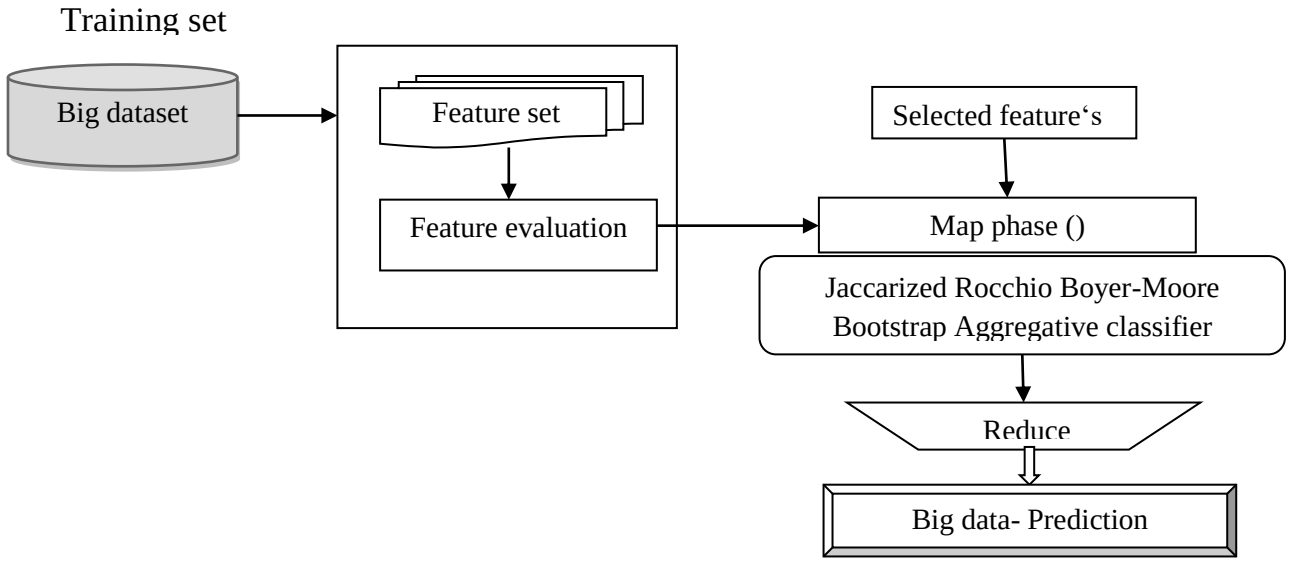


**Figure 1 Architecture diagram of KNFS-JRBMBAMC technique**

Then the Map-reduce function is applied for prediction. These different processes of the KNFS-JRBMBAMC technique are explained in the following subsections.

### *B. Kendall Ranking Correlative Normalized Discriminant Feature Selection*

The big data platform is capable of processing a large amount of data, as a result of which, data analysis becomes very complex. In order to minimize the complexity of big data analysis, feature selection is important in several big data applications. The significant challenges strictly related to big data are the dimensionality is extremely high. It is attractive to decrease the dimensionality of data resulting in increases computational and storage costs.

Feature selection is a dimension reduction technique used for efficiency in handling high-dimensional data. It is directly used to select the subset of significant features for the model construction. The selected features are used for maximizing the accuracy of the classification and reducing the computational time and storage space. Therefore, the proposed KNFS-JRBMBAMC technique uses the Kendall ranking correlative Normalized Discriminant Analyses for relevant feature selection. The Kendall ranking correlative Normalized Discriminant Analyses is a machine learning technique used for detecting the relevant features from the total feature set using a linear discriminant vector.

The discriminant vector is separating the features into two subsets, namely, based on the correlation measure. Thus, the projection increases the variance between the subset and minimizing the variance within the subset. The separation function is definite as the relation of the variance between the two subsets.

$$Q = \frac{\sigma_B}{\sigma_W} = \frac{W\,A_B(s)D}{WA_w(s)D} \qquad (1)$$

From (1), ‘Q denotes a separation function, $\sigma_B$ denotes the variance between the subset and $\sigma_W$ denotes a variance within the subset, W signifies a discriminant vector to projects the highlights into the specific class dependent on ideal projection heading 'D', $A_B(s)$symbolizes a scatter matrix between the subset, $A_w(s)$symbolizes the scatter matrix within the subset. The discriminant vector projects the features based on the correlation measure using Kendall rank correlation.

$$\tau = \frac{2\,(k_i - k_j)}{n(n-1)} \qquad (2)$$

From the above mathematical equation (2),'τ' denotes a Kendall rank correlation coefficient, '$k_i$', $k_j$ represents the features in the dataset 'n' indicates a number of features. The Relationship coefficient gives the yield runs between '- 1' and '+1'. These qualities are utilized to recognize the ideal level of connection between's the two highlights. '+1 Shows a positive Relationship between's the highlights, -1' indicates a negative connection between's the highlights. The decidedly related highlights are selected projected into the relevant subset, and other features are projected into the irrelevant subset.

After projecting the features, the scatter matrix is constructed to determine whether the features are positively or negatively correlated within the subset. The scatter matrix within the subset is measured as follows.

$$A_w(s) = \sum\sum(k_i - \mu_s)(k_i - \mu_s)^T \qquad (3)$$

Where, $A_w(s)$indicates a scatter matrix within the subset and. '$k_i'$ denotes features, $\mu_s$ represents a mean of the subset T denotes a transpose of a matrix. Similarly, the scatter matrix between the subset is measured as follows,

$$A_B(s) = \sum n * (\mu_{s1} - \mu_{s2})(\mu_{s1} - \mu_{s2})^T \qquad (4)$$

Where,$A_B(s)$ denotes a scatter matrix between the subset, $'n'$ denotes the number of highlights, μ_s1 signifies a mean of the split 1, μ_s1 indicates a mean of split two. Subsequently, the discriminant examination is partitioned into the list of capabilities into two subsets. Hence, the projection vectors in the division work limit the difference between and boost the connection within the split. Finally, the selected features are used for classification to minimize the prediction time just as space intricacy. The algorithmic process of the Kendall ranking correlative Normalized Discriminate feature selection is described as given below.

**// Algorithm 1: Kendall ranking correlative Normalized Discriminant feature selection**

**Input**: dataset, number of features$k_1, k_2, k_3, \ldots . k_n$

**Output**: Select relevant features

**Begin**

**Step 1:For each** input feature$\boldsymbol{k_i}$

**Step 2: Define separation function**

**Step 3:** Measure correlation '$\tau$'

**Step 4: if (**$\tau = +1$**)then**

**Step 5**: Features are positively correlated

**Step 6**: **else**

**Step 7**: Features are negatively correlated

**Step 8: end if**

**Step 9:** Divide the features into two different subsets using discriminant vector *'W'*

**Step 10**: Measure scatter matrix of $A_w(s)$ and $A_B(s)$

**Step 11: Return (**selected positively correlated features**)**

**Step 12: End for loop**

**Stop**

Calculation 1 depicts the bit-by-bit cycle of Kendall ranking correlative Normalized Discriminant feature selection. For each feature in the dataset, the correlation is measured. Based on the correlation value, the projection vector projects the features into two splits, namely pertinent or inappropriate feature subset. The positively correlated features are projected into subset 1, whereas the negatively correlated features are projected into subset 2. After that, the scatter matrix is measured within the subset and between the subsets. This shows that the positively correlated features are projected into subset 1, and the negatively correlated features are projected into subset 2. This helps to minimize the prediction time and memory complexity of prediction.

Figure 2 indicates a block diagram of the bootstrap ensemble classification technique to get an exact forecast. The bootstrap gathering strategy considers the training set $\{d_i, Z\}$ where$d_i = d_1, d_2, \ldots, d_m$' denotes the data, and 'Z 'indicates the ensemble classification outcomes. As shown in figure 2, the bootstrap ensemble classification technique initially constructs 'b' weak

learners$R_1, R_{2,}R_{3,} \ldots . R_b$and results are combined to obtain strong classification results. MapReduce is a simple programming model used for processing a huge volume of information in an equal way. This model comprises the Map phase and Reduce phase. The number of input data. '$d_i = d_1, d_2, \ldots, d_m$' are collected in the map phase in the form of rows and columns. The troupe strategy utilizes the powerless

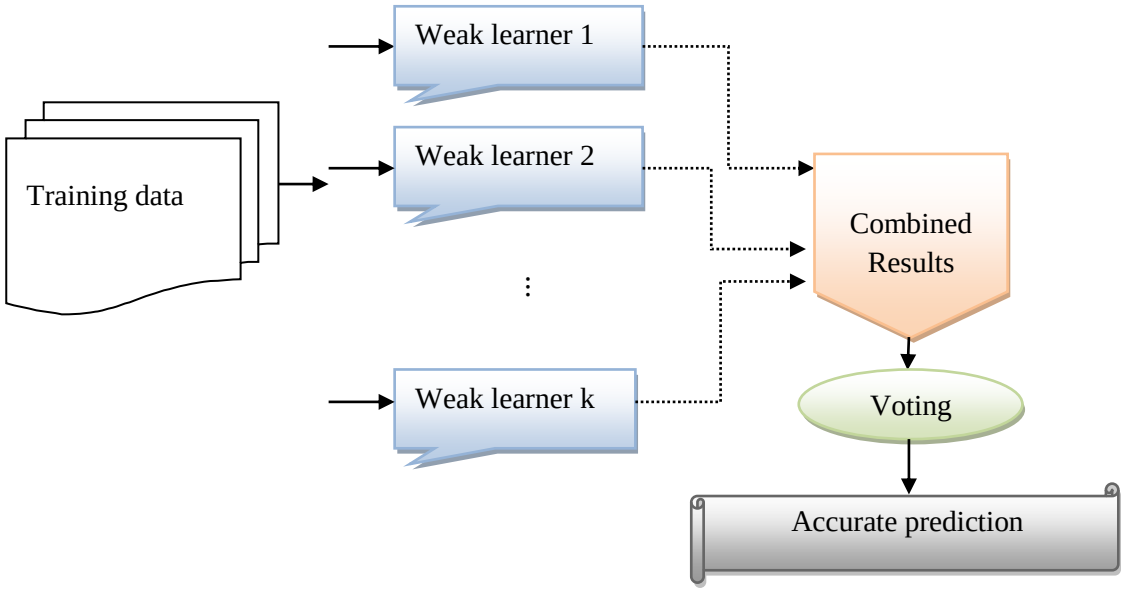


**Figure 2 Block diagram of Jaccarized Rocchio Boyer-Moore Bootstrap aggregative data classification**

### *C. Jaccarized Boyer-Moore Bootstrap Aggregative Data Classification*

After the component choice, the expectation is performed through the characterization with the assistance of the Bootstrap accumulating procedure. Bootstrap accumulating is a group strategy that assists with working on the precision of arrangement and relapse. The outfit classifier changes powerless students over to solid ones. A frail student is a base classifier that is not proficient to furnishes exact grouping with the least blunder. Despite what might be expected, a solid student is likewise a classifier that all around corresponded with the genuine grouping.

Learners as a Jaccard ordering classifier to arrange the information into various classes. Allow us to think about the quantity of information $d_1, d_2, d_3, \ldots . d_m$ for predictive analytics. The Jaccard index Rocchio classifier is used to find the class of training samples whose similarity is closest to the observation. The Rocchio classifier measures the Jaccard similarity between the training and testing data

$$\beta = \frac{d_{tr} \cap d_{ts}}{\sum d_{tr} + \sum d_{ts} - d_{tr} \cap d_{ts}} \qquad (5)$$

From (5),βdenotesa Jaccard Similarity coefficient, $d_{tr}$ denotes training data, $d_{ts}$ indicates a testing data, the crossing point image '⋂' assigns common freedom between the preparation and testing information are genuinely reliant, ∑dtr is the amount of dtr score, ∑dts is the amount of dts score. The Jaccard Comparability coefficient gives the yield esteems from 0 to 1. The high similitude of testing and preparing information are characterized into a specific class. Along these lines, the powerless student groups the information into various classes.

Yet, the feeble classifier makes them train blunder in the order results. To work on the exactness of order and limit the blunder, the feeble student results are joined into solid.

$$Z = \sum_{i=1}^{k} R_i\,(d_n) \qquad (6)$$

Where'Z'signifies, the yield of a solid learner,Ri (dn), represents a yield of the frail learns. For each frail learner, the training mistake is assessed to track down the precise grouping results. The blunder rate is determined as the distinction between the real outcomes and noticed grouping results.

$$\text{Error} = [R_a - R_o]^2 \qquad (7)$$

Where,$R_a$ symbolizes the actual output of the weak learner,$R_o$ represents the observed consequences of the frail classifier. Subsequent to computing the blunder(error) rate, the feeble learners are sorted. The classifier that has minimum error is ranked first, then the other. Similarly, all the weak learners are sorted. At last, the feeble student with the least blunder is chosen as the last prediction result. Then the proposed technique uses the Boyer-Moore voting scheme in reduce phase to find the majority of the classified data. Boyer-Moore voting

schema computes the majority vote of classified data. A majority vote in a sequence of 'l' data appears more than $m/2$ times in the sequence.

$$F = \arg\max\ [\text{Count (l)} > m/2] \quad (8)$$

Where F denotes an output of voting results at reducing phase, argmax denotes an argument of the maximum function, Count (l) denotes a sequence counts of the data appeared, m denotes a total length of the sequence. The majority votes of the data are obtained at the final classification results. The algorithmic process of the Jaccarized Boyer-Moore Bootstrap aggregative data classification is described as given below

**// Algorithm 2: Jaccarized Rocchio Boyer-Moore Bootstrap aggregative data classification**

**Input**: Data $d_i = d_1, d_2, \dots, d_m$

**Output**: Increase prediction accuracy

**Begin**

**Steps 1:foreach** data. '$d_i$'

**Steps 2:**Construct 'k' number of weak learners

**Steps 3**: **Measure** Jaccard Similarity coefficient '$\beta$'

**Steps 4**: **If [**'$\beta$' $= +1$)**then**

**Steps 5**: Data are classified into a particular class

**Steps 6**: **else**

**Steps 7:** Data are classified into another class

**Steps 8**: **end if**

**Steps 9**: Obtain weak learner results '$R_i(d_n)$'

**Steps 10**: Combine all weak learners $Z = \sum_{i=1}^{k} R_i\ (d_n)$

**Steps 11:For** each $R_i(d_n)$

**Steps 12**: Calculate error '$Error$'

**Steps 13**: Sorting weak learners in ascending order

**Steps 14**: Select the frail learners with the least error

**Steps15**: Find the majority votes $\arg\max \rightarrow Count\ (m) > m/2$

**Steps 16**: Obtain a well-built classification outcome

**Steps 17**: **end for**

**End**

Calculation describes the bit-by-bit cycle of Jaccarized Rocchio Boyer-Moore Bootstrap aggregate information characterization to further develop the forecast precision and decrease the mistake rate. In the map phase, the input training data are mapped into the testing data using Jaccarized Rocchio Boyer-Moore Bootstrap aggregative classifier. Initially, 'k's number of frail learners as a Jaccard indexive Rocchio classifier to analyze testing and training data and perform classification. If the similarity rate is high, then data is secret into a particular class. Or else, the data are classified into another class. Then the frail learner outcomes are combined and measure the error. Then the weak learners are sorted in ascending order according to the error rate. In other words, the weak learners with a minimum error are ordered first than the other. Followed by, another weak learner is arranged. Then, the Boyer-Moore majority voting scheme is applied in reduce phase for finding the majority votes of the samples. In this way, data are correctly classified. Based on classification results, the prediction is performed at higher accuracy.

## IV. INVESTIGATIONAL SETTINGS

The investigational assessment of the planned KNFS-JRBMBAMC method and obtainable [1] and [2] are implemented in Java programming language using the Soil dataset for prediction.

The soil dataset is collected from the: https://soilhealth.dac.gov.in/during the period2016–2017 for Erode region Tamilnadu. This webpage shows the soil testing report taken from Erode district https://soilhealth.dac.gov.in/PublicReports/NSVW. The Soil informational index has 15 credits and an aggregate of in excess of 5000 cases from the Soil example test report. The credits portrayals for each dirt example test result are listed in table 1

**Table 1 facial appearance description**

| S. No | Features | Description |
|---|---|---|
| 1 | Model No | Soil Testing Report Identification Number |
| 2 | pH | Soil pH value |
| 3 | EC | Electrical conductivity |
| 4 | OC | Organic carbon |
| 5 | N | Nitrogen |
| 6 | P | Phosphorus |
| 7 | K | Potassium |
| 8 | S | Sulphur |
| 9 | Zn | Zinc |
| 10 | Fe | Symbol of Iron |
| 11 | Cu | Symbol of Copper |
| 12 | Mn | Symbol of Manganese |
| 13 | Ca | Symbol of Calcium |
| 14 | B | Symbol of Boron |
| 15 | division | Very tall/ tall/ average/ short/Very short |

***B. False Positive Rate***: It is estimated as the quantity of information inaccurately arranged to the complete number of information taken as information. The bogus positive rate is determined using the given formula,

$$\mathrm{FPR} = \left[\frac{\mathrm{NICd_i}}{\mathrm{d_n}}\right] * 100 \qquad (10)$$

Where 'FPR' indicates a false positive rate, '$\mathrm{NICd_i}$' denotes the number of data incorrectly classified. '$\mathrm{d_n}$' be the absolute number of information. FPR is estimated as far as rate (%).

***C. Expectation time*:** prediction time is characterized as the measure of time used for future result prediction through the classification process. The formula for prediction is given by,

$$\mathrm{T_p} = \mathrm{d_n} \times \mathrm{Td_i} \qquad (11)$$

In the above condition (11), 'Tp' shows the expectation time, dn demonstrates the quantity of information, and 'Tdi' indicates the time taken for a single piece of information. Forecast time is estimated as far as milliseconds (ms).

***D. Space complexity:*** It is defined as the amount of storage space required by the algorithm to perform big data prediction. Space complexity is calculated by,

Table 1 shows the example informational index utilized for experimentation in our work that was put away as CSV (Comma Separated Values) document format that was obtained over the blocks of Erode District.

## V. PERFORMANCE RESULTS ANALYSIS

The exhibition of the KNFS-JRBMBAMC strategy is resolved as far as expectation precision, forecast time, fake positive rate, and space intricacy as for different quantities of information.

### *A. Prediction Accuracy*

Prediction accuracy is measured as the proportion of the quantity of information that is precisely arranged to the all-out number of information. The general expectation exactness is planned as given below,

$$Pre_a = \left[\frac{ACd_i}{d_n}\right] * 100 \qquad (9)$$

Where, '$Pre_a$' indicates prediction accuracy, '$ACd_i$' indicates the quantity of information precisely grouped and dn' determines the complete number of information. The forecast precision is measured in percentage (%).

$$\mathrm{S_{com}} = \mathrm{d_n} \times \mathrm{MSd_i} \qquad (12)$$

In the above equation (12), '$\mathrm{S_{com}}$' denotes the space complexity, '$\mathrm{d_n}$' represents the number of data and '$\mathrm{MSd_i}$' is the memory required for putting away single information. Space intricacy is estimated as far as Megabytes (MB). Like the above parameter, another type of parameter calculation is available in image filtering methods [20]

As shown in the above table 2, the prediction accuracy of three different methods, namely KNFS-JRBMBAMC, IDP [1], CGCNB-MRM [2], regarding various soil information taken from the dataset to predict the health status. In order to conduct the experiment, the performance of prediction accuracy is higher using KNFS-JRBMBAMC than the current techniques. Allow us to consider the 500 data, and the KNFS-JRBMBAMC correctly predicts the 450data, and the prediction accuracy is 90%. Similarly, the prediction accuracy of the other two methods, namely IDP [1], CGCNB-MRM [2], is 88% and 85%, respectively. After that, the nine results are obtained for each method with a number of data. Therefore, the overall prediction accuracy of the proposed KNFS-JRBMBAMC technique is compared to the prediction accuracy of existing methods. The comparison results prove that the soil health prediction accuracy in the agriculture sector gets improved by 4% and 7% when compared to existing methods.

**Table 2 Number of data versus Prediction accuracy**

| Number of data | Prediction accuracy (%) | | |
|---|---|---|---|
| | **KNFS-JRBMBAMC** | **IDP** | **CGCNB-MRM** |
| 500 | 90 | 88 | 85 |
| 1000 | 88 | 85 | 82 |
| 1500 | 85 | 83 | 80 |
| 2000 | 88 | 84 | 82 |
| 2500 | 87 | 82 | 80 |
| 3000 | 88 | 85 | 82 |
| 3500 | 92 | 89 | 86 |
| 4000 | 91 | 88 | 85 |
| 4500 | 90 | 86 | 83 |
| 5000 | 92 | 89 | 85 |

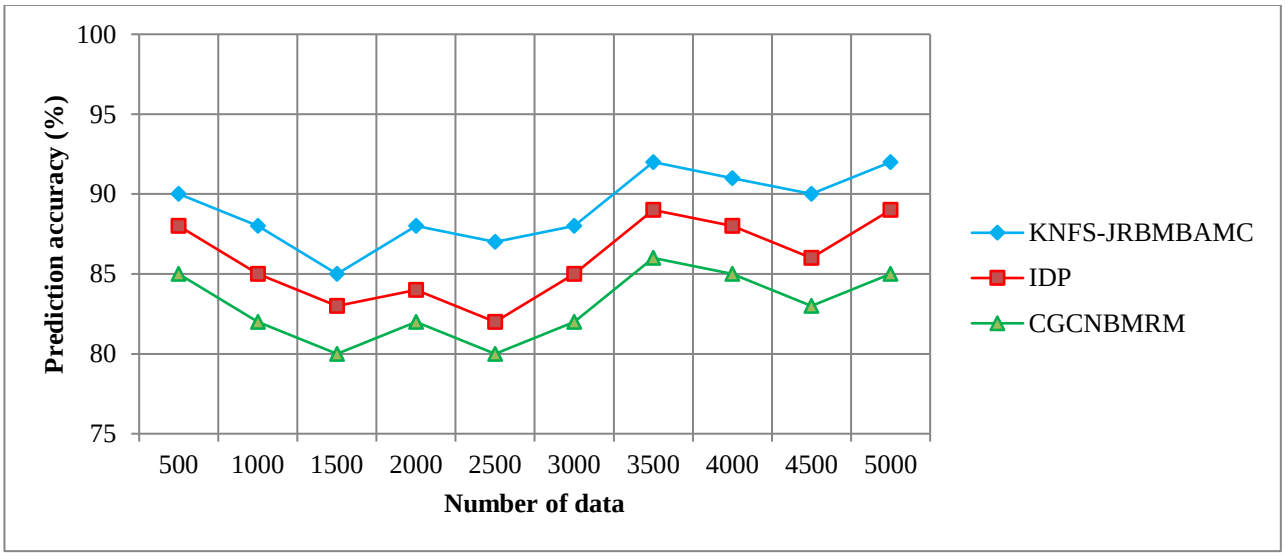


**Figure 3 Graphical portrayal of Prediction (expectation) exactness**

Figure 3 illustrates the graphical representation of expectation exactness versus various information in the range of 500 to 5000 is taken from the big dataset. The numbers of data are given as the input in the horizontal direction, and the prediction accuracy is obtained at the vertical axis.

The prediction accuracy of three methods KNFS-JRBMBAMC, IDP [1], CGCNB-MRM [2], is represented by three different colors of lines such as blue, red, and green, respectively. The above graph visibly reveals that the KNFS-JRBMBAMC technique increases the forecast accuracy than the existing prediction techniques.

This reason is to apply the Jaccarized Rocchio Boyer-Moore Bootstrap aggregative classifier. The ensemble classifier uses the number of weak learners as a Jaccard indexive Rocchio classifier to analyze the testing soil data and training data and perform classification.

If the similarity value is higher, then the data is accurately classified into very high, high, medium, low, and very low. As a result, the KNFS-JRBMBAMC technique achieves higher prediction accuracy.

Table 3 describes the bogus positive pace of three strategies to be specific KNFS-JRBMBAMC, IDP [1],

CGCNB-MRM [2] individually. In soil well-being prescient information examination, the bogus positive rate is the significant boundary to acquire a precise forecast. While dealing with an enormous number of information, the above table unmistakably shows that the KNFS-JRBMBAMC strategy diminishes the bogus positive pace of forecast than

**Table 3 Number of data versus false (bogus)-positive rate**

| Integer of data | False(bogus)-positive rate (%) | | |
|---|---|---|---|
| | **KNFS-JRBMBAMC** | **IDP** | **CGCNBMRM** |
| 500 | 10 | 12 | 15 |
| 1000 | 12 | 15 | 18 |
| 1500 | 15 | 17 | 20 |
| 2000 | 12 | 16 | 18 |
| 2500 | 13 | 18 | 20 |
| 3000 | 12 | 15 | 18 |
| 3500 | 8 | 11 | 14 |
| 4000 | 9 | 12 | 15 |
| 4500 | 10 | 14 | 17 |
| 5000 | 8 | 11 | 15 |

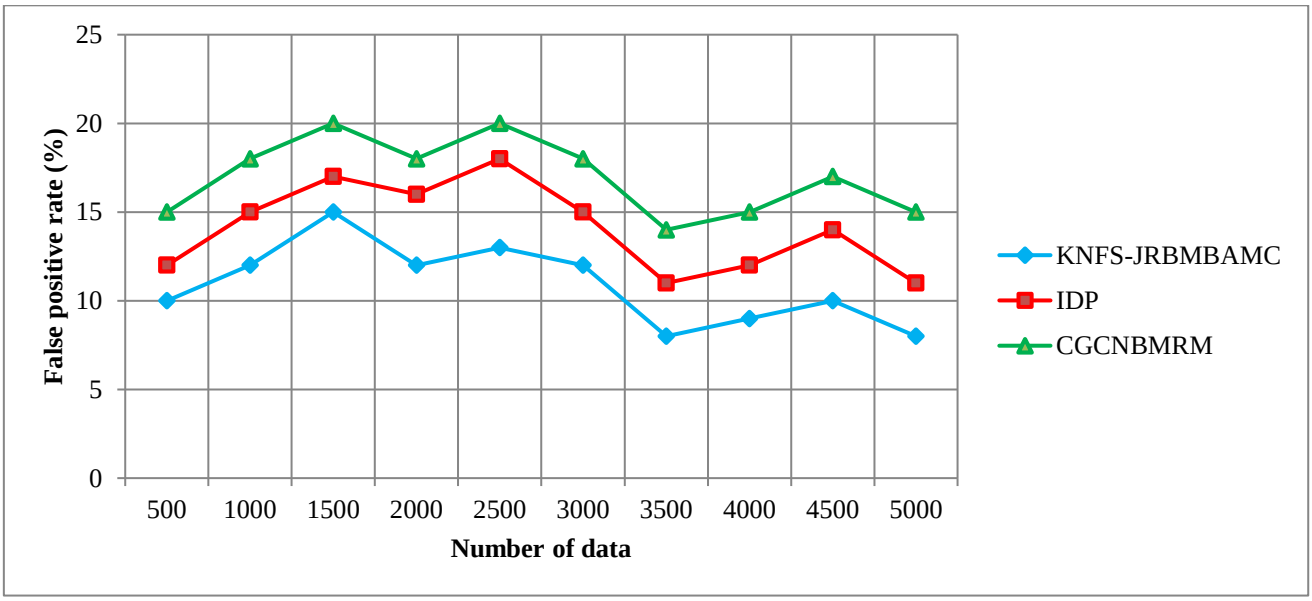


**Figure 4 Graphical representation of the false (bogus)-positive rate**

The current techniques. Let us consider 500 data in use as of the dataset to measure the false positive rate in the first iteration. By applying the KNFS-JRBMBAMC technique, 50 data are incorrectly classified, and the bogus positive rate is 10%. Likewise, 60 and 70 data are erroneously grouped by applying IDP [1], CGCNB-MRM [2], and the false-positive rates are 12% and 15%. Similarly, various results are observed with various counts of the information. The bogus positive paces of the proposed KNFS-JRBMBAMC strategy are contrasted with the current strategies. The average of ten results indicates that the false positive rate is significantly reduced by 24% and 35% when evaluated to existing methods. Finally, the ten consequences of the proposed strategy are contrasted with existing strategies. The examination after-effects of the proposed JRBMBAMC technique are considerably reduced by 23%, 36% when compared to existing [1], [2], respectively.

The performance analysis of bogus-positive rate in predictive analytics using soil data is demonstrated in figure4, based on the three methods, with the number of data varied from 500 to 5000. In figure4, the false positive rate analysis chart is shown. To evaluate the performance level of the proposed technique with that of the current techniques, a near examination is finished. The examination is made in the proposed KNFS-JRBMBAMC technique with two methods, namely IDP [1], CGCNB-MRM [2], for different inputs.

From figure 4, it is shown that the bogus positive rate is minimized by using KNFS-JRBMBAMC

technique. The reason is to apply the Jaccarized Rocchio Boyer-Moore Bootstrap Aggregative Map-reduce classifier. The classifier combines the weak learner results in the reduce phase. For each weak learner, the error rate is measured and sorted in ascending order along with the error rate. Then, the Boyer-Moore majority voting scheme is applied to find majority votes of the classified samples. The data has maximum votes classified into a particular class coming about in limits the bogus positive rate.

**Table 4 Quantity of data versus Prediction time**

| Number of data | Prediction time (ms) | | |
|---|---|---|---|
| | **KNFS-JRBMBAMC** | **IDP** | **CGCNBMRM** |
| 500 | 20 | 23 | 25 |
| 1000 | 23 | 26 | 28 |
| 1500 | 26 | 30 | 33 |
| 2000 | 30 | 33 | 35 |
| 2500 | 33 | 35 | 38 |
| 3000 | 36 | 39 | 42 |
| 3500 | 40 | 44 | 46 |
| 4000 | 42 | 46 | 50 |
| 4500 | 43 | 50 | 54 |
| 5000 | 45 | 52 | 55 |

Table 4 presents the relative examination aftereffects of forecast time acquired utilizing the Soil dataset by varying the training data from 500 to 5000. The experimental outcomes illustrate the proposed KNFS-JRBMBAMC technique has preferable execution over the current techniques. For 500 data, soil health prediction time is attained by the KNFS-JRBMBAMC technique is 20ms, and the time consumption of the other two existing methods, IDP [1], CGCNB-MRM [2], are 23ms and 25ms, respectively. Consequently, from the similar examination, obviously, the proposed approach has further developed execution in terms of minimizing the prediction time than the other comparative methods. Therefore, the overall examination results show that the KNFS-JRBMBAMC procedure limits the forecast time by 11% when contrasted with [1] and 17% when contrasted with [2], respectively

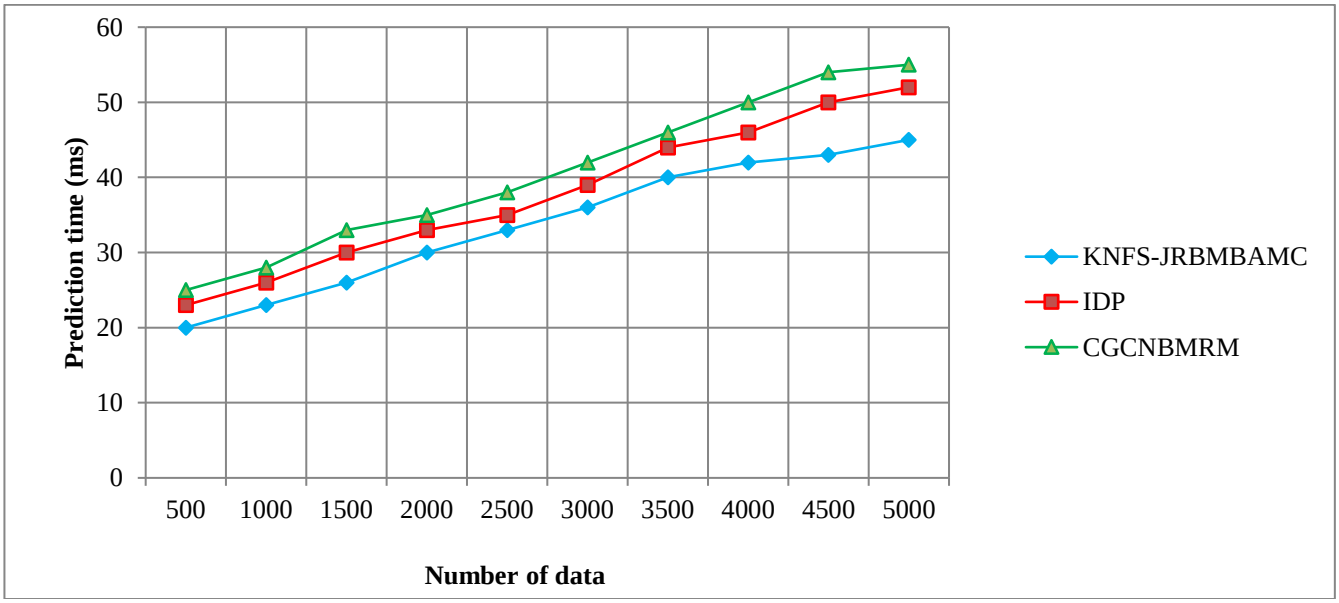


**Figure 5 Graphical demonstration of Prediction (forecast) time**

Figure 5 addresses the presentation consequences of expectation time alongside the quantity of information. As displayed in the graphical outline, the dirt wellbeing forecast season of the multitude of techniques is bit by bit expanded while expanding the quantity of information for various runs. Be that as it may, similarly, the forecast time gets limited utilizing the KNFS-JRBMBAMC strategy. The huge explanation is to apply the Kendall ranking correlative Normalized Discriminant feature selection. The cKendall rank correlation is measured between the features in the dataset. Discriminant projection vectors separate the relevant or irrelevant features. The positively correlated features are called relevant features and are used for classification instead of using entire features in the dataset. As a result, minimizes the time consumption of accurate prediction

**Table 5 Number of data versus Space complexity**

| Number of data | Space complexity (MB) | | |
|---|---|---|---|
| | **KNFS-JRBMBAMC** | **IDP** | **CGCNBMRM** |
| 500 | 13 | 15 | 18 |
| 1000 | 15 | 17 | 20 |
| 1500 | 18 | 21 | 24 |
| 2000 | 22 | 24 | 26 |
| 2500 | 23 | 26 | 30 |
| 3000 | 26 | 28 | 32 |
| 3500 | 28 | 32 | 35 |
| 4000 | 31 | 34 | 37 |
| 4500 | 34 | 36 | 41 |
| 5000 | 37 | 40 | 43 |

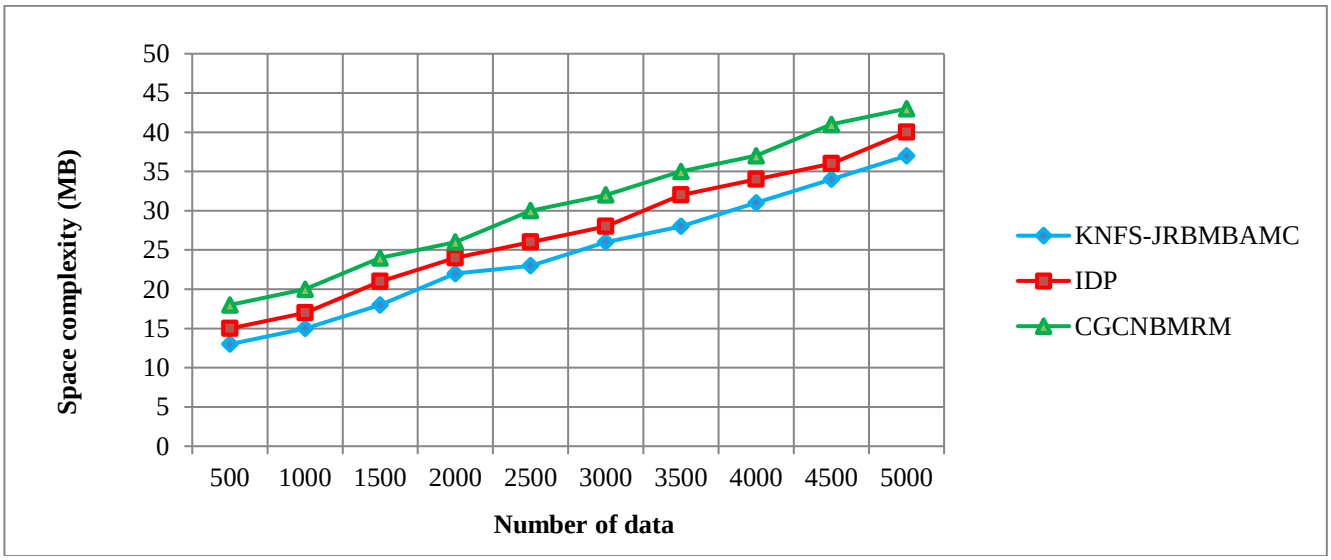


**Figure 6 Graphical representation of Space complexity**

Table 5 and figure 6 illustrate the graphical illustration of the space complexity of predicting the student grade level using three different methods, namely the KNFS-JRBMBAMC technique and existing IDP [1] CGCNB-MRM [2]. As shown in figure 6 and table 5, an increasing linear trend is to be observed for all the three prediction techniques while expanding the quantity of information. Among the three methods, the KNFS-JRBMBAMC provides superior performance than the other two methods. Besides, from the sample numerical computation provided using table 5, with '500' data are considered to perform the experimentation, the space complexity for predicting the soil status using KNFS-JRBMBAMC technique was found to be '13MB' and memory consumption of existing [1], [2] was found to be '13MB', 15MB' separately. Also, the leftover nine runs are done to break down the exhibition of the proposed procedure against the current techniques. The general examination results determine that the average value of space intricacy of the KNFS-JRBMBAMC procedure is impressively diminished by 8% and 15% when contrasted with regular strategies. This is on the grounds that by applying Kendall positioning correlative Standardized Discriminant significant element determination for projecting the pertinent highlights. Accordingly, the proposed KNFS-JRBMBAMC procedure utilizes a lesser memory space for predicting the soil status.

## VI. CONCLUSION

Accurate prediction models play a most important role in Big Data analytics. An effective disease prediction model called the KNFS-JRBMBAMC technique is introduced for accurate prediction by integrating feature selection and classification. At first, the KNFS-JRBMBAMC technique model uses the Kendall ranking correlative Normalized Discriminant method for finding the relevant feature and irrelevant features for classification. Based on the analysis, the positively correlated features are selected from the dataset to limit the time utilization and space intricacy of the forecast. Secondly, the Jaccarized Rocchio Boyer-Moore Bootstrap Aggregative Map-reduce classifier is applied for investigating the preparation and testing of information with the help of a map-reduce classifier. The Boyer-Moore voting scheme accurately finds the classification results furthermore, limits the bogus positive rate. The exploratory assessment is conducted with the agriculture soil datasets. The experimental results and conversation of different measurements show that the KNFS-JRBMBAMC method achieved better performance by achieving higher prediction accuracy and lesser bogus positive, expectation time, and space intricacy than that of state-of-the-art models

## VII. REFERENCES


[1] Fang Chen, Cong Zhang, Junjie Zhang, Wenqi Cao, IDP: An Intelligent Data Prediction Scheme Based on Big Data and Smart Service for Soil Heavy Metal Content Prediction, IEEE Access, 9(2021) 32351 – 32367

[2] Chitrakant Banchhor and N. Srinivasu, Integrating Cuckoo search-Grey wolf optimization and Correlative Naive Bayes classifier with Map Reduce model for big data classification, Data & Knowledge Engineering, Elsevier, 127 (2020) 1-3.

[3] Yanxia Lv, Sancheng Peng, Ying Yuan, Cong Wang, Pengfei Yin, Jiemin Liu, Cuirong Wang, A classifier using online bagging ensemble method for big data stream learning, Tsinghua Science and Technology, 24(4) (2019) 379 – 388.

[4] Ryan H.L.Ip, Li-MinnAng, Kah PhooiSeng, J.C.Broster, J.E.Pratley, Big data and machine learning for crop protection, Computers and Electronics in Agriculture, Elsevier, 151 (2018) 376-383

[5] Gunasekaran Manogaran, Daphne Lopez, Naveen Chilamkurti In-Mapper Combiner based Map-Reduce Algorithm for Big Data Processing of IoT based Climate Data, Future Generation Computer Systems, Elsevier, 86 (2018) 433-445

[6] T. Venkat Narayana Rao, Prediction of Soil Quality Using Machine learning techniques, International Journal of Scientific & Technology Research, 8(11) (2019) 1309-1313

[7] ]M.S. Suchithra, Maya L. Pai, Improving the prediction accuracy of soil nutrient classification by optimizing extreme learning machine parameters, Information Processing in Agriculture, Elsevier, 7(1) (2020) 72-82

[8] Meysam Alizamir, Ozgur Kisi, Ali Najah Ahmed, Cihan Mert, Chow Ming Fai, Sungwon Kim, Nam Won Kim, Ahmed El-Shafie, "Advanced machine learning model for better prediction accuracy of the soil temperature at different depths", PLoS ONE, Volume 15, 2020, Pages 1-25

[9] Yanbo Huang, Chen Zhong-xin, Yu Tao, Huang Xiang-Zhi, Gu Xing-fa, Agricultural remote sensing big data: Management and applications, Journal of Integrative Agriculture, Elsevier, 17(9) (2018) 1915-1931.

[10] Saeid Mehdizadeh, Farshad Fathian, Mir Jafar Sadegh Safari, Ali Khosravi, Developing novel hybrid models for estimation of daily soil temperature at various depths, Soil and Tillage Research, Elsevier, 197 (2020) 1-12

[11] Oswalt Manoj S, Ananth J P, MapReduce and Optimized Deep Network for Rainfall Prediction in Agriculture, The Computer Journal, 63(1) 2020) 900 – 912.

[12] Kiran M. Sabu, T. K. Manoj Kumar, Predictive analytics in Agriculture: Forecasting prices of Arecanutsin Kerala, Procedia Computer Science, 171 (2020) 699-708

[13] Yu Cai, Wengang ZhengID, Xin Zhang, Lili Zhangzhong, Xuzhang Xue, "Research on soil moisture prediction model-based on deep learning", PLoS ONE 14, 4 (2019) 1-19

[14] Binh Thai Pham, Trung Nguyen-Thoi, Hai-Bang Ly, Manh Duc Nguyen, Nadhir Al-Ansari, Van-Quan Tran and Tien-Thinh Le, Extreme Learning Machine Based Prediction of Soil Shear Strength: A Sensitivity Analysis Using Monte-Carlo Simulations and Feature Backward Elimination, Sustainability, 12 (2020) 1-29

[15] Saeid Mehdizadeh, Babak Mohammadi, Quoc Bao Pham, Dao Nguyen Khoi, Pham Thi Thao Nhi, Implementing novel hybrid models to improve indirect measurement of the daily soil temperature: Elman neural network coupled with gravitational search algorithm and ant colony optimization, Measurement, Elsevier, 165 (2020) 1-57.

[16] Chitrakant Banchhor and N. Srinivasu, FCNB: Fuzzy Correlative Naive Bayes Classifier with MapReduce Framework for Big Data Classification, Journal of Intelligent System, 29(1) (2020) 994–1006

[17] Mikel Elkano, Mikel Galar, Jose Sanz, Humberto Bustince, CHI-BD: A Fuzzy Rule-Based Classification System for Big Data classification problems, Fuzzy Sets and Systems, Elsevier, 348 (2018) 75-101

[18] Kingsley John, Isong Abraham Isong, Ndiye Michael Kebonye, Esther Okon Ayito , Prince Chapman Agyeman, and Sunday Marcus Afu, Using Machine Learning Algorithms to Estimate soil organic Carbon Variability with Environmental Variables and Soil Nutrient Indicators in an Alluvial Soil, Land, 9 (2020) 1-20

[19] Mikel Elkano, Jose Antonio Sanz, Edurne Barrenechea, Humberto Bustince, Mikel Galar, CFM-BD: A Distributed Rule Induction Algorithm for Building Compact Fuzzy Models in Big Data Classification Problems, IEEE Transactions on Fuzzy Systems, 28(1) (2020) 163 – 177.

[20] A.Suresh&Dr.P.Malathi,. An improved cellular automata (ca) based image denoising method for biometric applications. Biomedical Research 2017; Special Issue: ISSN 0970-938X (2017)

[21] Long-Hao Yang, Jun Liu, Ying-Ming Wang, Luis Martínez, A Micro-Extended Belief Rule-Based System for Big Data Multi-class Classification Problems, IEEE Transactions on Systems, Man, and Cybernetics: Systems, 51(1) (2021) 420 – 440.

[22] Fatimah Bibi Hamzah, Firdaus Mohd Hamzah, Siti Fatin Mohd Razali, Juanita Zainudin, Bidirectional Recurrence Neural Network Imputation For Recovering Missing Daily Streamflow Data International Journal of Engineering Trends and Technology 69(8) (2021).

[23] Ramesh Babu Palepu1, Rajesh Reddy Muley, An Analysis of Agricultural Soils by using Data Mining Techniques, International Journal of Engineering Science and Computing, (2017) 1516715177

[24] Luke Bornn and James V. Zidek, Efficient stabilization of crop yield prediction in the Canadian Prairies, International Journal of Agricultural and forest meteorology, 152 (2012) 223-232.

[25] Dimitrios Voloudakis, Andreas Karamanos, Garifalia Economou, Dionissios Kalivas, Petros Vahamidis, Vasilios Kotoulas, John Kapsomenakis and Christos Zerefos, Prediction of climate change impacts on cotton yields in Greece under eight climatic models using the AquaCrop crop simulation model and discriminant function analysis. Agricultural and Water Management, 147 (2015) 116-128.


.