

Research

Computational intelligence based sustainable computing with classification model for big data visualization on map reduce environment

Zheng Xu¹

Received: 15 March 2022 / Accepted: 28 April 2022

Published online: 09 May 2022

© The Author(s) 2022 **OPEN**

Abstract

In recent years, the researchers have perceived the modifications or transformations motivated by the presence of big data on the definition, complexity, and future direction of the real world optimization problems. Big Data visualization is mainly based on the efficient computer system for ingesting actual data and producing graphical representation for understanding large quantity of data in a fraction of seconds. At the same time, clustering is an effective data mining tool used to analyze big data and computational intelligence (CI) techniques can be employed to solve big data classification process. In this aspect, this study develops a novel Computational Intelligence based Clustering with Classification Model for Big Data Visualization on Map Reduce Environment, named CICC-BDVMR technique. The proposed CICC-BDVMR technique intends to perform effective BDV using the clustering and data classification processes on the Map Reduce environment. For clustering process, a grasshopper optimization algorithm (GOA) with kernelized fuzzy c-means (KFCM) technique is used to cluster the big data and the GOA is mainly utilized to determine the initial cluster centers of the KFCM technique. GOA is a recently proposed metaheuristic algorithm inspired by the swarming behaviour of grasshoppers. This algorithm has been shown to be efficient in tackling global unconstrained and constrained optimization problems. Based on the modified GOA, an effective kernel extreme learning machine model for financial stress prediction was created. Besides, big data classification process takes place using the Ridge Regression (RR) and the parameter optimization of the RR model is carried out via the Red Colobuses Monkey (RCM) algorithm. The design of GOA and RCM algorithms for parameter optimization processes for big data classification shows the novelty of the study. A wide ranging simulation analysis is carried out using benchmark big datasets and the comparative results reported the enhanced outcomes of the CICC-BDVMR technique over the recent state of art approaches. The broad comparison research illustrates the CICC-BDVMR approach's promising performance against contemporary state-of-the-art techniques. As a result, the CICC-BDVMR technique has been demonstrated to be an effective technique for visualising and classifying large amounts of data.

Keywords Data visualization · Big data · Data classification · Computational intelligence · Evolutionary algorithm

✉ Zheng Xu, xuzheng@sspu.edu.cn | ¹School of Computer and Information Engineering, Shanghai Polytechnic University, 2360 JinHai Road, Pudong District, Shanghai 201209, China.

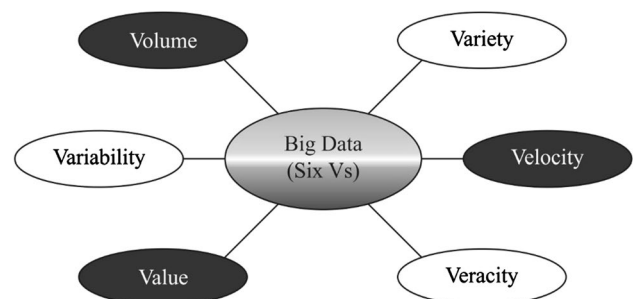


1 Introduction

Recently, Big Data has gained considerable interest in a wide-ranging application which generates Big Data, e.g., social networking profiles, health care services, MapReduce scientific experiments, cloud applications, e-government services, and transportation [1]. Data processing massive volumes of data in parallel across multiple nodes is possible using the MapReduce programming paradigm. MapReduce is an analytics framework for large-scale complicated data analysis. To process large volumes of data across multiple clusters, MapReduce is a Hadoop framework. It is also a programming model that allows large datasets to be processed over numerous computer clusters. Distributed data storage is enabled by this programme. This is achieved by splitting petabytes of data into smaller parts and processing them on commodity Hadoop servers. It then consolidates data from many servers and returns it to the application. Data are rapidly generated, and application produces increasing amount of structured and unstructured data consisting of several variables that should be analyzed in a shorter period of time. The National Institute of Standard and Technology has specified that Big Data has four common features (4Vs): veracity, volume, variety, and velocity [2]. Veracity denotes a measure of understandability and quality of the data. Volume represents the data size that could be very large to be produced by the existing generation of techniques or systems. Variety refers to the most fascinating of the four Vs since it includes data of different kinds, namely audio, video, text, and images for a provided object. Velocity refers to data that is streaming at faster speed when compared to traditional algorithms and systems [3]. Data mining (DM) and Data analysis are difficult processes since the quantity of data is significant and this data can be polluted with noise and might be stored by different processes. Such data are classified by the four Vs of Big Data. Figure 1 illustrates the types of Vs involved in big data.

The major problem in research is data analytics viz. implemented on the basis of DM and machine learning (ML) methods [4]. Usually, big data mining (BDM) method has difficulty in handling DM software tools and presentation techniques because the size of the information is complex and large. Executing DM method through largescale data sets with a single Personal Computer (PC) necessitates higher cost of computation. Therefore, it is important to utilize efficient computing environment for big data processing and analyzing [5]. Big data increases the demand for smart data analytic models such as automatic classification, image processing, data fusion, and multi-temporal processing. Parallel processing is a computing technique that involves running two or more processors (CPUs) simultaneously to perform separate pieces of a larger operation. Parallel processing is a technique that is widely used to conduct complex activities and computations in parallel. Parallel processing will be widely used by data scientists for compute- and data-intensive activities. The parallelization method is technologically advanced for scaling with the data available by increasing the computation significantly. In order to manage the problem based on largescale data sets, Google presented the MapReduce architecture [6]. The MapReduce approach along with distributed file system (DFS) offers robust and simple environments to handle largescale data processing. In DM, this approach is currently being considered than other parallelization methods, i.e., Message Passing Interface (MPI), because of its fault tolerance system, i.e., needed for the task which consumes significant amount of time, and because of their MPI of [7]. In general, the MapReduce architecture is implemented by an effective parallel programming method named Hadoop [8]. The MapReduce techniques involve map and reduce function. The mapping process is utilized for sorting and filtering, where the reduce functions perform a summary process for generating a result. Several researches-based methods are presented for BDM methods such as for instance selection, attribute reduction, and class imbalance. Therefore, by using MapReduce technique and traditional distributed approach, BDM is efficiently implemented by several computer nodes or processors to simultaneously perform the task [9]. In the study, Decision Tree (DT), ML methods, optimization algorithms are utilized for classifying big data.

Fig. 1 Six Vs of Big data



This study introduces an efficient Computational Intelligence based Clustering with Classification Model for Big Data Visualization on Map Reduce Environment, named CICC-BDVMR technique. The proposed CICC-BDVMR technique involves the design of grasshopper optimization algorithm (GOA) with kernelized fuzzy c-means (KFCM) technique to group the big data and the GOA is applied to effectively compute the initial cluster centers of the KFCM technique. KFCM is abbreviated as kernel fuzzy c-means clustering algorithm (KFCM) derives from the fuzzy c-means clustering approach (FCM). Comparing the KFCM approach to the standard fuzzy c-means technique, the former allows for more accurate clustering and has a higher accuracy. The latter also allows for more accurate clustering. Moreover, big data classification process takes place using the Ridge Regression (RR) and the parameter optimization of the RR model is carried out via the Red Colobus Monkey (RCM) algorithm. In order to demonstrate the enhanced performance of the CICC-BDVMR technique, a comprehensive result analysis is made using benchmark datasets. A grasshopper optimization algorithm (GOA), the Red Colobus Monkey (RCM) algorithm, the design of GOA and RCM algorithms for parameter optimization processes for large data categorization demonstrates the study's uniqueness. Parameter optimization procedures for big data categorization are being designed using GOA and RCM algorithms.

2 Literature review

Abukhodair et al. [10] developed a meta heuristic optimization based on big data classification in MapReduce (MOBDC-MR) architecture. The presented method focuses on selecting optimum features and efficiently categorizing big data. Additionally, the suggested techniques involve the proposal of BPOA based FS method for increasing the accuracy and reducing the difficulty. Beetle antenna search (BAS) with LSTM is applied for classifying big data. Brahmane and Krishna [11] introduced an approach to handle big data with Spark architecture. The presented method undergoes two stages to classify the big data that includes feature classification and selection, i.e., implemented in the primary nodes of Spark framework. The presented optimization method is called rider chaotic biography optimization (RCBO) method, i.e., combination of chaotic biogeography-based optimization (CBBO) and rider optimization algorithm (ROA). The presented RCBO-DSAE method with Spark architecture efficiently handles the big data to attain an efficient big data classification.

Qin et al. [12] the DEEPEYE method has been presented for addressing this challenge. The scheme resolves the problem by training a binary classification for deciding either a certain visualization is effective for a provided data set, and utilizing supervised learning to rank method for ranking the abovementioned visualization. Also, it considers common visualization processes, namely binning and grouping, that could manipulate the data, also describe the searching space.

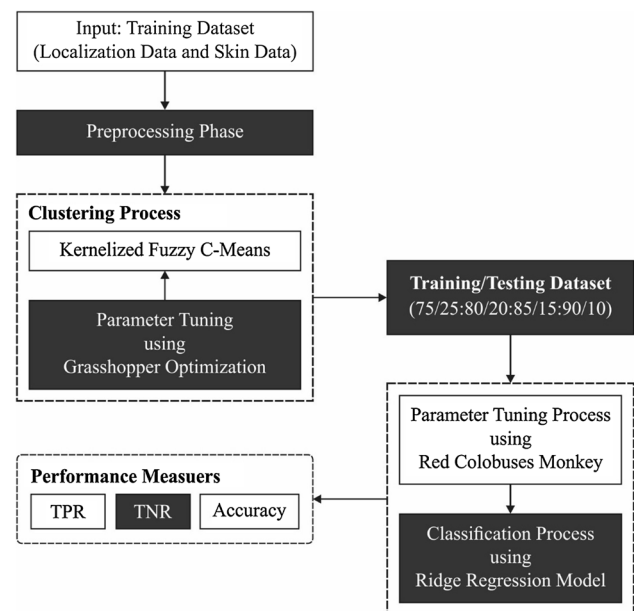
Galletta et al. [13] proposed a graphical tool for the visualization of healthcare data, which is simply used to monitor health condition of person remotely. The tool is easy to use, and assist medical doctor to understand fast the existing condition of patient by observing a coloured circle. Cui et al. [14], proposed a Big Data Visualization enables Multi-modal Feedback Framework (BDVMFF) for boosting motivation, student confidence, and self-consciousness in the online learning environment. The presented method provides the teacher a digital task to efficiently exchange input and writing to employ multi-modal feedback. Those systems provide students and teachers with straightforward and effective digital learning platforms.

Lakshmanprabu et al. [15], developed big data analytics on IoT based medical systems with the MapReduce and Random Forest Classifier (RFC). The e-health information is gathered from the patient affected by various diseases is taken into account for analysis. The optimum attribute is selected by an Improved DA (IDA) from the databases for effective classification. At last, RFC method is utilized for classifying the e-health data using optimum features. Dubey et al. [16], proposed an effective ACO and PSO-based architecture for data preprocessing and classification in big data. It has shown that content part is fetched and collaborated for analyzing the integration of velocity and volume. Next, weight marking is performed by the variety and volume of data. At last, the ranking is performed by the variety and velocity features of big data.

3 The proposed model

In this study, a novel CICC-BDVMR technique has been developed for accomplishing effectual BDV by the use of clustering and data classification process on the Map Reduce environment. The proposed CICC-BDVMR technique encompasses several subprocesses namely KFCM based clustering, GOA based initial cluster center selection, RR based classification, and RCM based parameter tuning. Figure 2 illustrates the overall process of CICC-BDVMR technique.

Fig. 2 Block diagram of CICC-BDVMR technique



3.1 Map reduce

The MR method is applied for parallel and distributed processing of massive amounts of unstructured and structured information, whereby Hadoop is generally stored in HDFS, clustered with a large computer [17]. Therefore, scaling in small steps is feasible (scale-out). The architecture consists of (a) reduce—an aggregation/consolidation stage, whereby all the related records are processed in single entity. (b) Map—a key transformation, and recording stage, whereby individual-input record is simultaneously processed. Correctly configure the cluster using the appropriate diagnostic tools. When writing intermediate data to disc, utilise compression. Adjust the amount of Map and Reduce tasks in accordance with the aforementioned recommendations. Whenever possible, incorporate Combiner. MapReduce uses the input data to pass each data element to the mapper during the mapping phase. The reducer process all of the mapper’s outputs and arrives at the result during the reducing step. Simply put, the mapper’s job is to filter and change the input into something that the reducer can accumulate over. The two great advantages are interrelated with: map task and Logical block. The key idea is that the input data is separated into logical blocks. Each block is processed via map task. The results from functioning block are divided into dissimilar sets and then arranged. All the sorted blocks are transported to the reduced task (RT). The RT: a map task could run-in cluster node, and map task could run in parallel that is responsible to transform the input record to value or key pair. The output from all the maps is split and later arranged. But there is a separate division for each RT.

3.2 Design of GOA-KFCM based clustering technique

In recent years, the kernel method [18] is the most researched subject within ML community and has been extensively employed to function approximation and pattern recognition. The key motivation of utilizing the kernel method consists of: (1) enhances strength of original clustering algorithm to outliers and noise, (2) induces a class of strong non-Euclidean distance measure for the novel data space to derive objective function and thereby cluster the non-Euclidean structure in data; and (3) still retain computation simplicity. This procedure can be realized by changing the objective function in the traditional FCM method with a kernel-induced distance rather than Euclidean distance in the FCM, and thereby the respective process is acquired and known as the kernelized FCM (KFCM) model, that is very powerful when compared to FCM:

$$\sum_{i=1}^C u_{ij} = 1, \forall j = 1 \dots N \tag{1}$$

In the study, the kernel function $K(x, C)$ is considered as a Gaussian radial basic function (GRBF):

$$K(x, c) = \exp\left(\frac{-\|x - c\|^2}{\sigma^2}\right), \quad (2)$$

whereas σ represent an adjustable variable:

$$j_m = 2 \sum_{i=1}^C \sum_{j=1}^N u_{ij}^m [1 - K(X_j, C_i)]. \quad (3)$$

The fuzzy membership matrix u is attained by:

$$u_{ij} = \frac{(1 - K(x_j, c_i))^{-1/(m-1)}}{\sum_{k=1}^C (1 - K(x_j, c_k))^{-1/(m-1)}}. \quad (4)$$

The cluster center c_i is attained by:

$$C_i = \frac{\sum_{j=1}^N u_{ij}^m K(x_j, c_i) x_j}{\sum_{j=1}^N u_{ij}^m K(x_j, c_i)} \quad (5)$$

As the K -means model focus on minimizing the sum of squared distance from each point to the cluster center, it leads to compact cluster. Then employ the intra-cluster distance measures, i.e., median distance between a cluster center and point [19]. The following equation can be used:

$$intra = midean\left(\sum_{i=1}^C \sum_{x \in c_i} \|x - v_j\|^2\right) \quad (6)$$

Thus, the clustering provides minimal values for the validity measure shows the ideal value of cluster. Next, the amount of cluster is known beforehand evaluating the membership matrix.

For determining the initial cluster centers of the KFCM technique, the GOA is utilized. The GOA method is an evolutionary model proposed by the simulation behavior of swarm of grasshoppers while searching for food. Typically, they are insects of, destructive nature; cause harm to agricultural produce and harvest production [20]. The growth of a full-grown grasshopper drives as egg, nymph, and adults. It can be mathematically modelled by the following equation for resolving different optimization issues.

$$Y_i^d = cx \left\{ \sum_{\substack{j=i \\ j \neq i + \hat{T}_d}}^n cx(ul_d - ll_d/2) sf(|Y_j^d - Y_i^d|)^{Y_j - Y_i} / D_{ij} \right\} \quad (7)$$

Here, Y_j, Y_i represents the location of j th and i th grasshopper. The j th and i th locations of the grasshopper in D th dimension are represented by Y_j^d and Y_i^d , correspondingly. The distance, number of grasshoppers, and social interaction between j th and i th grasshoppers are denoted as, sf , and D_{ij} respectively. \hat{T}_d indicates the value of the target in the D th dimension, while ul_d and ll_d denotes the upper and lower limits in D th dimension. According to the coefficient cx , the comfort zone is reduced in proportion to the number of iterations. The adoptive variable cz is utilized for reducing the comfort zone. To balance exploitation and exploration of the grasshopper swarm near the optimal global solution, the initial cx value is used. Moreover, repulsion zone, comfort zone, and attraction amongst the grasshoppers are reduced by using the second cx value [21]. The coefficient cx reduces the comfort zone proportionate to the amount of iterations as follows

$$cz = cz_{\max} - t(t_{\max}) \tag{8}$$

In which cz_{\max} means the maximal value, cz_{\min} shows the minimal value, t represent the existing iteration, and r_{\max} indicates the maximal amount of iterations.

Algorithm 1: Pseudocode of GOA

```

Start
Set the swarm size,  $cz_{\max}$ ,  $cz_{\min}$  and
maximal amount of iterations  $t_{\max}$ ,
Random generation of population  $y$ ;
Calculation of the fitness of all the grasshopper agents,
 $B$  =optimal search agent,
While ( $t < t_{\max}$  )
Upgrade  $cz$  based on Eq. (8),
For  $i = 1:n$ 
    Normalize the distance between Grasshoppers  $y$ ,
    Upgrade the present agent location  $y_i$  by Eq. (7);
end For
     $t = t + 1$ 
end while
return B
end
    
```

3.3 Design of RCM-RR based classification technique

Once the big data is clustered into different groups based on the class labels that exist in it, the next stage is to perform classification process using the RR technique. The RR [22] is an SLFN system where the weights between the hidden and input layers are selected in an arbitrary way. RR is computational free from iteration that makes RR very faster by considerably minimizing the computational time needed for training the SLFN. The SLFN frequently needs large amount of hidden layers when creating optimum solutions. The output function of SLFN with L hidden node is determined as follows:

$$f_L(x) = \sum_{i=1}^L \beta_i g_i x = \sum_{i=1}^L \beta_i G(a_i, b_i, x), x \in R^d, \beta_i \in R^m \tag{9}$$

For additive nodes with activation function g , g is determined by

$$g_i = G(a_i, b_i, x) = g(a_i x + b_i), a_i \in R^d, b_i \in R \tag{10}$$

$$\sum_{i=1}^L \beta_i G(a_i, b_i, x) = t_j, j = 1, \dots, N \tag{11}$$

The above equations are updated by the following equation

$$H\beta = T \quad (12)$$

now

$$H(w_1 \cdots w_L, b_1 \cdots b_L, x_1 \cdots x_N) \\ = \begin{pmatrix} g(w_1 \cdot x_1 + b_1) & \cdots & g(w_L \cdot x_1 + b_L) \\ \vdots & \ddots & \vdots \\ g(w_1 \cdot x_N + b_1) & \cdots & g(w_L \cdot x_N + b_L) \end{pmatrix} \quad (13)$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_N^T \end{bmatrix} \quad (14)$$

$$T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix} \quad (15)$$

H signifies the hidden neuron output matrix of the NN system. The SLFN is trained to resolve a linear optimization issue as follows:

$$\|H\hat{\beta} - T\| = \min_{\beta} \|H\beta - T\| \quad (16)$$

Now $\hat{\beta}$ is represented by

$$\hat{\beta} = H^{\dagger}T = (H^T H)^{-1}H^T T \quad (17)$$

is the minimum norm least square solution of $w_i = T$ and H^T characterizes the Moore Penrose generalized inverse of H [23].

The process of RR is described as follows.

Step 1 Arbitrarily Select the input weight w_i and hidden layer bias b_i .

Step 2 Evaluate the hidden neuron output matrix H .

Step 3 Attain the output weight $\hat{\beta}$ by utilizing $\hat{\beta} = H^{\dagger}T$

For properly tuning the parameters involved in the RR technique, the RCM algorithm has been utilized and thereby achieved improved classification outcomes. The RCM approach stimulates the red monkey behavior. In order to model this interaction, every cluster in the monkey area unit needed maneuvering through the searching region [24]. Young males must quickly go out because of the territorial aspects related to the *Cercopithecus mitis* to be very effective since they are entering challenges with dominant males from other families. As well, there is no specific interaction among young ones and male *Cercopithecus mitis*. Once they defeated that male, they will be leader in the family and offers food supplies, place to live, and socialization for the young male. The location update about each one of the red monkeys in a group is depending on the location of the optimal red monkey of the group has been delineated by the succeeding equations:

$$PB_{i+1} = (PA_i \times PB_i) + (W_{leader} - W_i) \times rand \times (X_{best} - X_i) \quad (18)$$

$$X_{i+1} = X_i + PB_{i+1} \quad (19)$$

$$PB_{i+1} = PA_i \times rand \quad (20)$$

While

- PA signifies the monkey combat power (an arbitrarily selected value in the range of $[0, 1]$);
- PB denotes the monkey body power (an arbitrarily selected value in the range of $[-5, 5]$);

- W_i characterizes the monkey weight (an arbitrarily selected value between [4, 6]);
- W_{leader} indicates the leader weight;
- X_{best} refers to the location of the leader.
- rand has shown any number in the range of [0, 1].
- X illustrates the location of the red monkey;

In order to upgrade the location associated with the children of red monkey, the following equation has been used:

$$PBch_{i+1} = (PA_i \times PBch_i) + (Wch_{leader} - Wch_i) \times rand \times (Xch_{best} - Xch_i) \tag{21}$$

$$Xch_{i+1} = Xch_i + PBch_{i+1} \tag{22}$$

$$PAch_{i+1} = PAch_i \times rand \tag{23}$$

where $PAch$ indicates the child combat power; $PBch$ denotes the power rate of child body, and Wch_i represents the child weight in which each weight was stated for being arbitrary number within [4, 6]. It is noteworthy that each parameter of RCM depending on the problem’s nature or set by experiment that should be resolved. RCM is considered as a parameter which makes it easier to execute; also RCM balances between exploration and exploitation stages, which makes it applicable to resolve optimization problems.

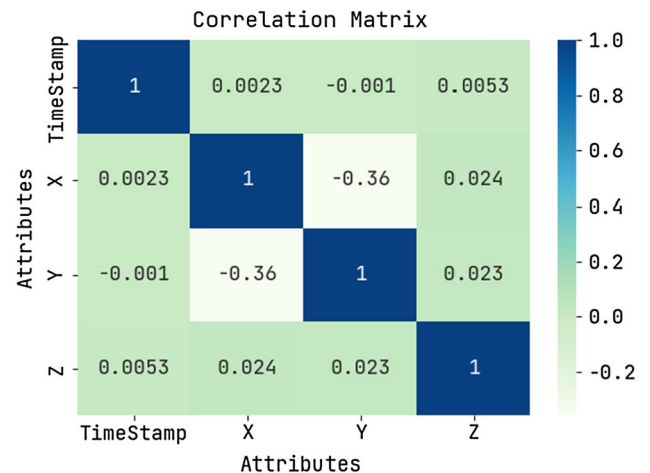
The RCM approach derives a FF for attaining enhanced classification performance. It defines a positive integer for representing the optimum efficiency of the candidate solutions. During this analysis, the minimization of the classification error rate is regarded as the FF, as offered in Eq. (24). An optimum solution has a lesser error rate and the worst solution attains an enhanced error rate.

$$Fitness(x_i) = Classifier\ Error\ Rate(x_i) = \frac{number\ of\ misclassified\ instances}{Total\ number\ of\ instances} \times 100 \tag{24}$$

4 Performance validation

This section assesses the performance of the CICC-BDVMR approach using two standard datasets [25] namely localization data and skin data. The first localization data involves 8 attributes and 164,860 instances. Besides, the skin dataset includes 245,057, amongst which 50,859 are skin samples, and the remaining 194,198 are non-skin samples. A correlation matrix is simply a table that shows the correlation. The measure is optimally utilized in variables that illustrate linear relations among others. The fit of the data is visually characterized in a scatterplot. Figure 3 demonstrates the correlation matrix of n the test localization dataset.

Fig. 3 Confusion matrix of CICC-BDVMR technique on localization dataset



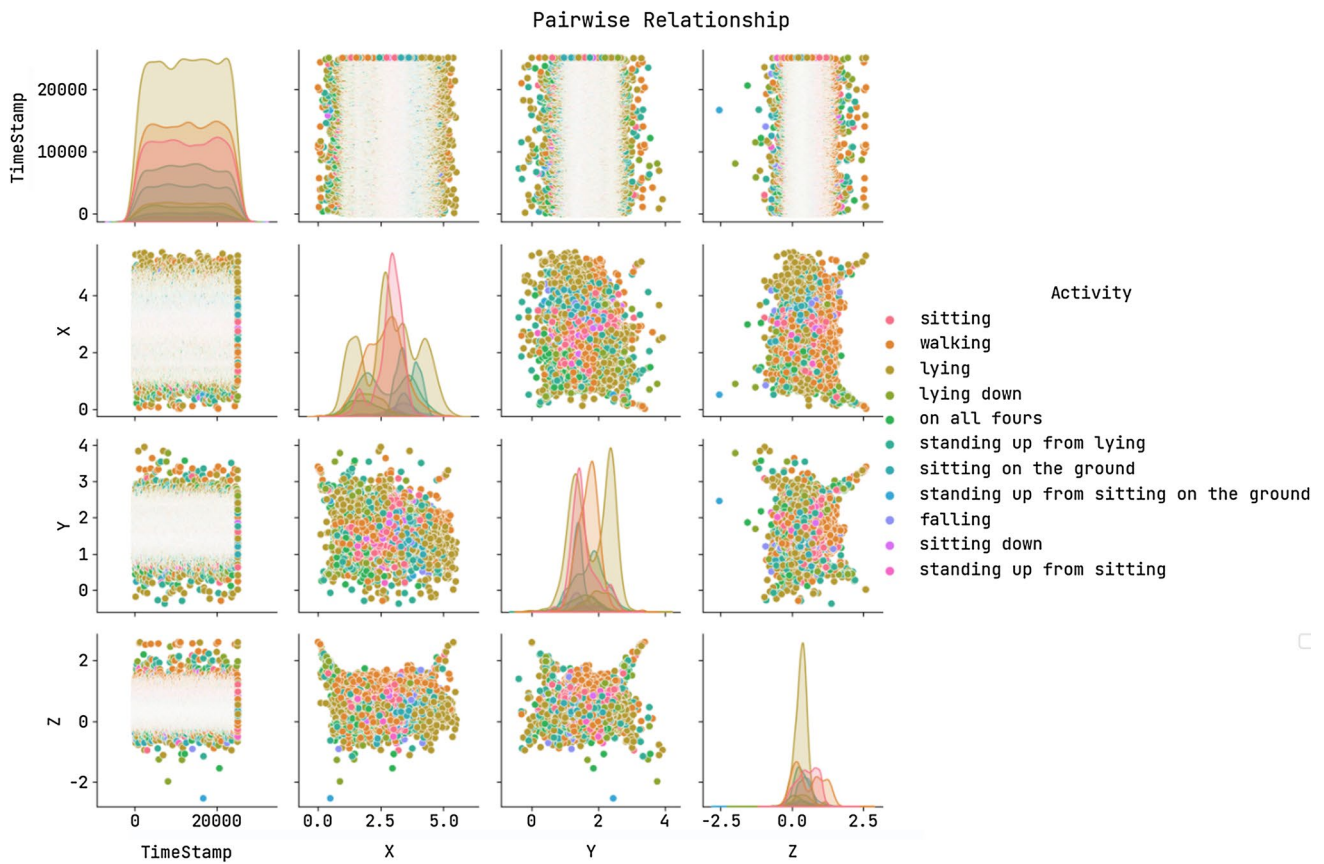


Fig. 4 Pairwise relationship of class labels in localization dataset

Fig. 5 Confusion matrix of CICC-BDVMR technique on localization dataset

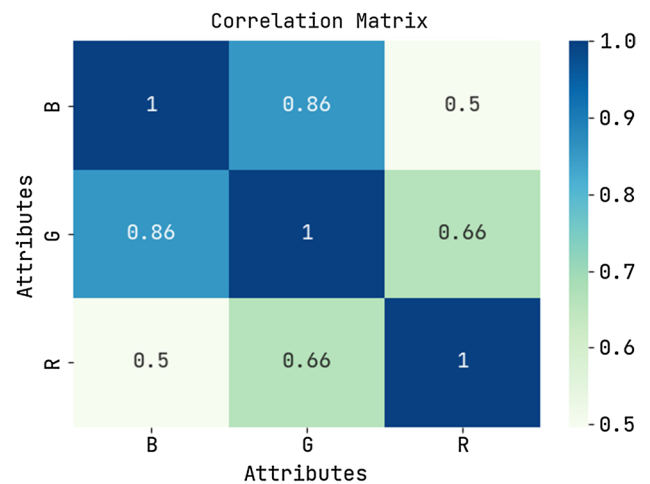


Figure 4 shows the pairwise relationship plot of the class labels involved in the localization dataset such as sitting, walking, lying, lying down, on all fours, standing up from lying, sitting on the ground, standing up from sitting on the ground, falling, sitting down, and standing up from sitting.

Figure 5 proves the correlation matrix attained by the CICC-BDVMR method on the test skin data set. The correlation matrix proved that our CICC-BDVMR method has gained enhanced performance on the test localization data set. Figure 6 displays the pairwise relation plot of the class label included in the skin dataset namely skin and Non-skin.

Table 1 provides the comparative classification result analysis of the CICC-BDVMR technique with other techniques on the test localization dataset under different mappers (M). The experimental results indicated that the CICC-BDVMR

Fig. 6 Pairwise relationship of class labels in skin dataset

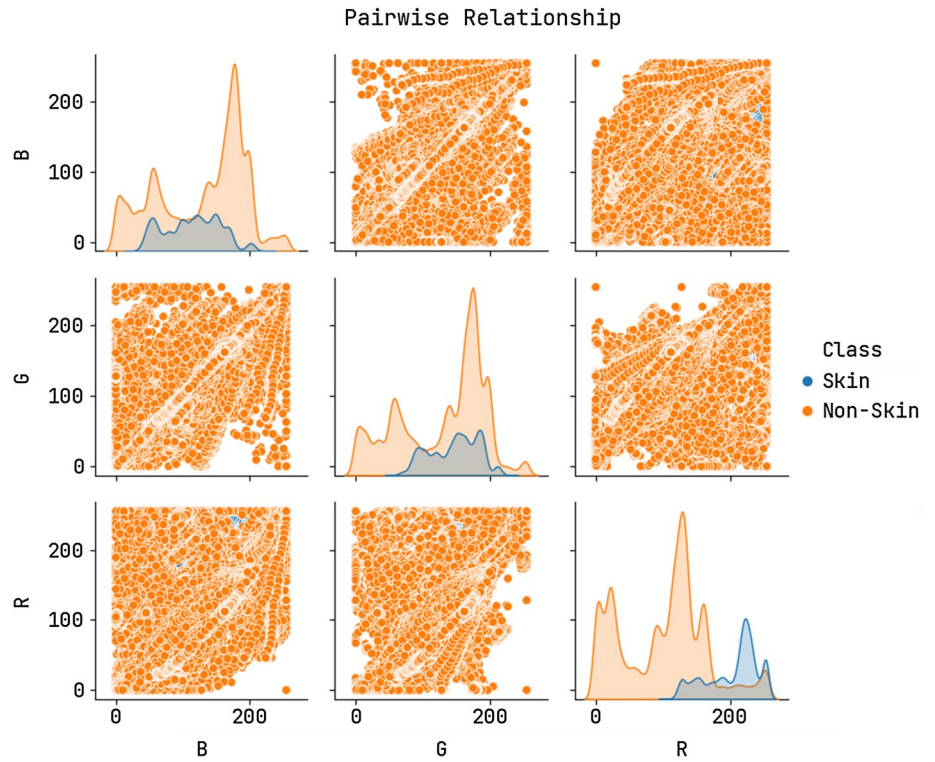


Table 1 Result analysis of CICC-BDVMR technique under different mappers in localization dataset

Mappers	CNB			GWOCNB			CGCNB			CICC-BDVMR		
	Accuracy	TPR	TNR	Accuracy	TPR	TNR	Accuracy	TPR	TNR	Accuracy	TPR	TNR
Training/testing (75:25)												
M=2	77.92	80.24	72.60	77.96	81.53	74.42	79.06	83.32	75.72	81.60	85.71	79.11
M=3	76.81	82.40	73.81	78.47	81.68	75.17	81.01	82.17	75.69	84.10	85.54	78.16
M=4	76.96	81.46	73.20	78.22	83.25	76.07	80.76	83.16	75.38	83.39	85.93	79.39
M=5	76.15	81.45	73.06	80.18	82.14	75.85	81.05	84.36	76.45	84.35	87.25	78.51
Average	76.96	81.39	73.17	78.71	82.15	75.38	80.47	83.25	75.81	83.36	86.11	78.79
Training/testing (80:20)												
M=2	76.84	81.64	73.16	78.44	81.74	75.10	79.66	83.96	76.38	82.40	86.60	79.57
M=3	77.57	82.31	72.42	78.71	84.03	76.48	79.13	83.42	75.04	81.66	86.64	79.14
M=4	78.41	80.51	74.79	78.84	82.82	76.33	80.68	83.24	77.59	83.02	85.34	80.07
M=5	76.39	81.93	73.51	79.96	81.40	74.41	80.93	83.60	75.12	84.11	86.50	77.44
Average	77.30	81.60	73.47	78.99	82.50	75.58	80.10	83.56	76.03	82.80	86.27	79.06
Training/testing (85:15)												
M=2	76.34	82.47	72.70	78.32	83.24	74.20	81.03	83.60	77.70	83.19	86.21	80.53
M=3	76.55	82.00	74.71	80.35	83.44	74.68	79.68	84.30	75.79	83.49	87.63	78.10
M=4	76.77	82.28	73.50	79.24	83.28	76.80	79.98	82.73	76.44	82.83	85.76	80.22
M=5	76.54	81.55	73.93	79.51	83.12	73.76	79.00	83.75	75.26	82.15	86.89	78.42
Average	76.55	82.08	73.71	79.36	83.27	74.86	79.92	83.60	76.30	82.92	86.62	79.32
Training/testing (90:10)												
M=2	78.00	81.11	72.72	79.32	81.86	76.30	80.17	82.82	76.96	83.00	85.95	79.71
M=3	78.71	81.23	74.81	79.31	82.34	75.20	79.73	83.70	77.32	82.34	86.96	80.29
M=4	77.70	82.25	73.83	79.33	82.49	77.05	80.72	85.13	77.87	83.44	87.26	80.38
M=5	77.73	81.61	72.61	78.71	83.17	76.69	81.21	85.62	77.49	84.62	88.38	80.18
Average	78.04	81.55	73.49	79.17	82.47	76.31	80.46	84.32	77.41	83.35	87.14	80.14

technique has obtained effective classification performance under all sizes of M . For instance, with a training/testing dataset of 75:25 and $M=2$, the CICC-BDVMR technique has achieved higher accuracy of 81.60% whereas the CNB, GWOCNB, and CGCNB techniques have obtained lower accuracy of 77.92%, 77.96%, and 79.06% respectively. Besides, with $M=5$, the CICC-BDVMR technique has achieved higher accuracy of 84.35% whereas the CNB, GWOCNB, and CGCNB techniques have obtained lower accuracy of 76.15%, 80.18%, and 81.05% respectively. Besides, with training/testing dataset of 80:20 and $M=2$, the CICC-BDVMR system has accomplished high accuracy of 82.40% while the CNB, GWOCNB, and CGCNB systems have attained minimum accuracy of 76.84%, 78.44%, and 79.66% correspondingly. In addition, with $M=5$, the CICC-BDVMR method has reached maximum accuracy of 84.11% whereas the CNB, GWOCNB, and CGCNB methods have attained less accuracy of 76.39%, 79.96%, and 80.93% correspondingly.

Table 2 and Fig. 7 showcases the average classification results obtained by the CICC-BDVMR with recent methods under distinct sizes of training/testing data. With training/testing data of 75:25, the CICC-BDVMR technique has achieved better performance with the maximum average accuracy, TPR, and TNR of 83.36%, 86.11%, and 78.79% whereas the CNB, GWOCNB, and CGCNB techniques have resulted in ineffective outcomes with the lower accuracy of 76.96%, 78.71%, and 80.47% respectively. Simultaneously, with training/testing data of 85:15, the CICC-BDVMR system has accomplished improved performance with the maximal average accuracy, TPR, and TNR of 82.92%, 86.62%, and 79.32% while the CNB, GWOCNB, and CGCNB systems have resulted in inefficient outcomes with the less accuracy of 76.55%, 82.08%, and 73.71% correspondingly.

Concurrently, with training/testing data of 80:20, the CICC-BDVMR system has accomplished good performance with the highest average accuracy, TPR, and TNR of 82.80%, 86.27%, and 79.06% while the CNB, GWOCNB, and CGCNB systems have resulted in inefficient outcomes with the less accuracy of 77.30%, 81.60%, and 73.47% correspondingly. Furthermore, with training/testing data of 90:10, the CICC-BDVMR method has accomplished effective performance with the highest average accuracy, TPR, and TNR of 83.35%, 87.14%, and 80.14% while the CNB, GWOCNB, and CGCNB systems have resulted in inefficient outcomes with the minimum accuracy of 78.04%, 81.55%, and 73.49% correspondingly.

The overall accuracy outcome analysis of the CICC-BDVMR technique on localization data is portrayed in Fig. 8. The results demonstrated that the CICC-BDVMR technique has accomplished improved validation accuracy compared to training accuracy. It is also observable that the accuracy values get saturated with the epoch count of 1000. The overall loss outcome analysis of the CICC-BDVMR technique on localization data is Table 3 offers the relative analysis of CICC-BDVMR system with other approaches on the test skin dataset under dissimilar mappers (M). The experiment result indicates that the CICC-BDVMR method has attained good classification performance under each size of M . For example, with training/testing dataset of 75:25 and $M=2$, the CICC-BDVMR method has accomplished high accuracy of 83.44% while the CNB, GWOCNB, and CGCNB approaches have attained less accuracy of 76.04%, 75.95%, and 79.70% correspondingly. In addition, with $M=5$, the CICC-BDVMR system has realized high accuracy of 81.54% while the CNB, GWOCNB, and CGCNB methods have gained less accuracy of 76.70%, 77.27%, and 78.34% correspondingly.

Besides, with training/testing dataset of 80:20 and $M=2$, the CICC-BDVMR method has accomplished high accuracy of 82.09% while the CNB, GWOCNB, and CGCNB methods have attained less accuracy of 75.96%, 77.66%, and 79.46%

Table 2 Average analysis of CICC-BDVMR technique under distinct sizes of training/testing data in localization data

Methods	Training/testing (75:25)			Training/testing (85:15)		
	Accuracy	TPR	TNR	Accuracy	TPR	TNR
CNB	76.96	81.39	73.17	76.55	82.08	73.71
GWO+CNB	78.71	82.15	75.38	79.36	83.27	74.86
CGCNB	80.47	83.25	75.81	79.92	83.6	76.30
CICC-BDVMR	83.36	86.11	78.79	82.92	86.62	79.32
Methods	Training/testing (80:20)			Training/testing (90:10)		
	Accuracy	TPR	TNR	Accuracy	TPR	TNR
CNB	77.30	81.60	73.47	78.04	81.55	73.49
GWO+CNB	78.99	82.50	75.58	79.17	82.47	76.31
CGCNB	80.10	83.56	76.03	80.46	84.32	77.41
CICC-BDVMR	82.80	86.27	79.06	83.35	87.14	80.14

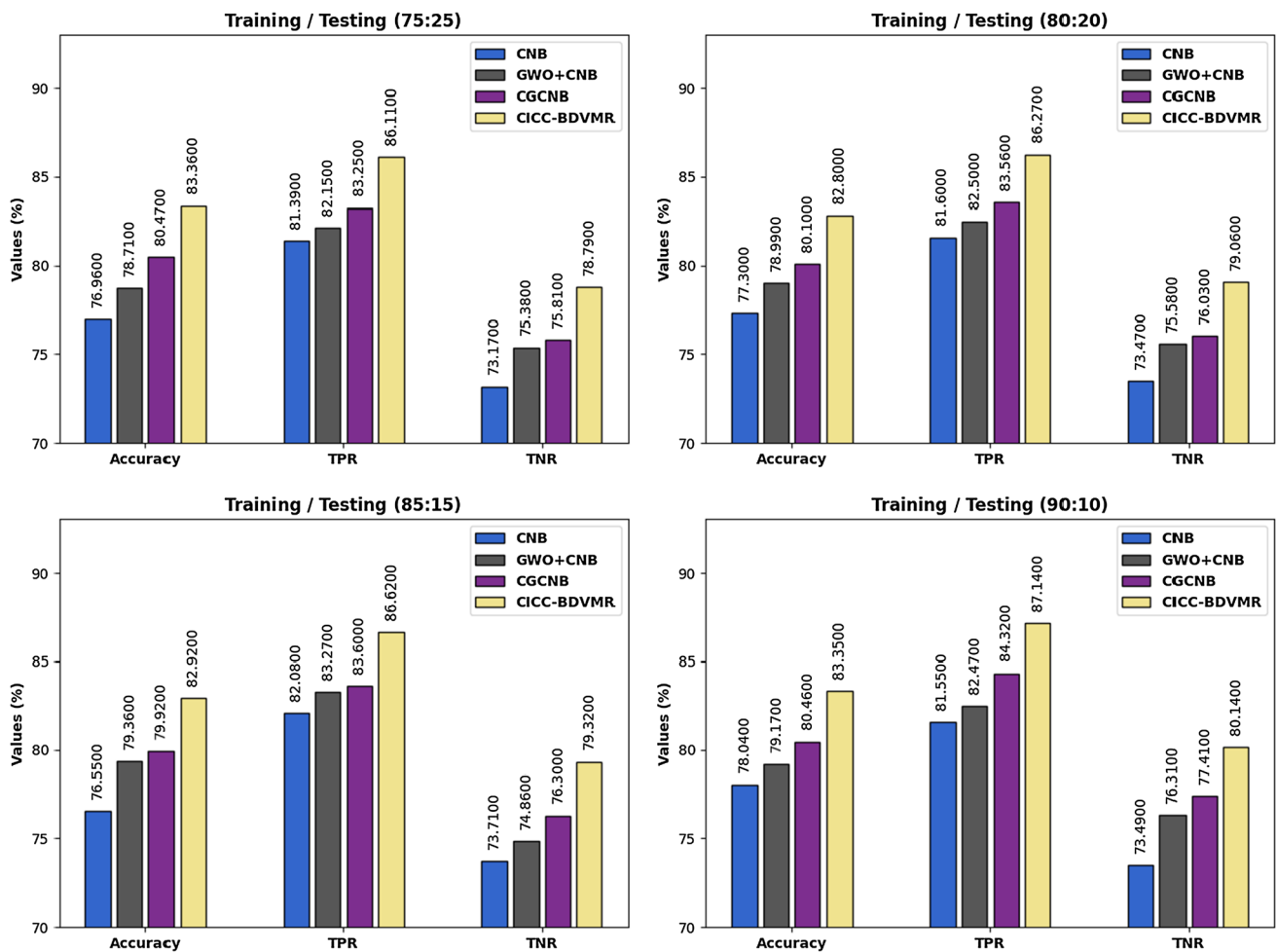
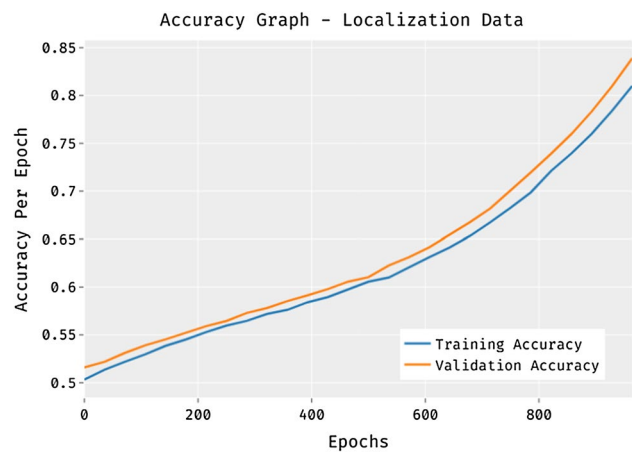


Fig. 7 Average analysis of CICC-BDVMR technique under localization data

Fig. 8 Accuracy analysis of CICC-BDVMR technique under localization data



correspondingly. In addition, with $M=5$, the CICC-BDVMR system has reached high accuracy of 80.63% while the CNB, GWOCNB, and CGCNB methods have attained less accuracy of 76.33%, 77.34%, and 77.34% correspondingly.

Table 4 and Fig. 9 show the average classification outcome attained by the CICC-BDVMR with existing models under dissimilar sizes of testing or training data [26]. With training/testing data of 75:25, the CICC-BDVMR system has accomplished improved performance with the maximal average accuracy, TPR, and TNR of 81.89%, 85.71%, and 76.75% where the CNB, GWOCNB, and CGCNB methods have resulted in inefficient outcomes with the less accuracy of 76.19%, 81.02%,

Table 3 Result analysis of CICC-BDVMR technique under different mappers in skin dataset

Mappers	CNB			GWOCNB			CGCNB			CICC-BDVMR		
	Accuracy	TPR	TNR	Accuracy	TPR	TNR	Accuracy	TPR	TNR	Accuracy	TPR	TNR
Training/testing (75:25)												
M=2	76.04	81.22	71.33	75.98	82.43	73.06	79.70	84.51	73.38	83.44	88.44	76.95
M=3	75.44	80.34	72.19	77.07	82.93	71.82	77.97	82.56	73.14	80.75	84.97	75.47
M=4	76.59	81.80	70.32	75.64	81.70	73.25	79.35	82.22	74.66	81.83	84.88	77.02
M=5	76.70	80.70	70.58	77.27	82.15	73.37	78.34	82.12	73.62	81.54	84.56	77.54
Average	76.19	81.02	71.11	76.49	82.30	72.88	78.84	82.85	73.70	81.89	85.71	76.75
Training/testing (80:20)												
M=2	75.96	80.09	69.88	77.66	81.11	71.58	79.46	83.70	72.44	82.09	86.78	74.71
M=3	77.63	82.06	71.48	76.39	82.36	71.38	78.45	84.30	73.14	81.97	87.94	75.34
M=4	76.80	81.08	70.10	78.12	81.96	73.47	78.88	84.44	72.27	81.40	86.72	77.40
M=5	76.33	81.56	71.73	77.34	82.17	71.60	77.34	84.77	75.31	80.63	87.37	78.35
Average	76.68	81.20	70.80	77.38	81.90	72.01	78.53	84.30	73.29	81.52	87.20	76.45
Training/testing (85:15)												
M=2	77.63	80.83	71.65	78.25	81.23	73.67	77.63	84.45	74.68	82.22	87.71	78.40
M=3	75.18	81.64	70.31	75.77	82.65	74.13	78.72	84.47	72.87	81.43	88.31	76.62
M=4	77.59	80.93	71.66	76.62	82.81	73.57	77.89	83.25	72.90	80.84	86.76	76.51
M=5	76.86	81.26	71.37	78.29	82.71	71.36	79.80	84.84	75.29	82.84	87.79	79.04
Average	76.82	81.17	71.25	77.23	82.35	73.18	78.51	84.25	73.94	81.83	87.64	77.64
Training/testing (90:10)												
M=2	77.55	80.68	72.07	77.11	81.48	72.59	78.56	83.67	75.10	80.68	86.90	79.02
M=3	77.23	81.35	71.82	77.66	84.11	71.95	79.28	83.80	74.26	81.34	87.47	77.00
M=4	77.66	82.37	72.32	76.96	83.46	71.26	78.36	83.84	74.21	80.56	87.17	76.48
M=5	77.49	82.70	72.50	77.81	82.58	71.87	78.73	84.67	74.31	82.24	88.09	77.20
Average	77.48	81.78	72.18	77.39	82.91	71.92	78.73	84.00	74.47	81.21	87.41	77.43

Table 4 Average analysis of CICC-BDVMR technique under distinct sizes of training/testing data in skin data

Methods	Training/testing (75:25)			Training/testing (85:15)		
	Accuracy	TPR	TNR	Accuracy	TPR	TNR
CNB	76.19	81.02	71.11	76.82	81.17	71.25
GWO+CNB	76.49	82.30	72.88	77.23	82.35	73.18
CGCNB	78.84	82.85	73.70	78.51	84.25	73.94
CICC-BDVMR	81.89	85.71	76.75	81.83	87.64	77.64
Methods	Training/testing (80:20)			Training/testing (90:10)		
	Accuracy	TPR	TNR	Accuracy	TPR	TNR
CNB	76.68	81.20	70.80	77.48	81.78	72.18
GWO+CNB	77.38	81.90	72.01	77.39	82.91	71.92
CGCNB	78.53	84.30	73.29	78.73	84.00	74.47
CICC-BDVMR	81.52	87.20	76.45	81.21	87.41	77.43

and 71.11% correspondingly. At the same time, with training/testing data of 85:15, the CICC-BDVMR method has reached improved performance with the maximal average accuracy, TPR, and TNR of 81.83%, 87.64%, and 77.64% while the CNB, GWOCNB, and CGCNB systems have resulted in inefficient outcomes with the less accuracy of 76.82%, 81.17%, and 71.25% correspondingly.

Simultaneously, with training/testing data of 80:20, the CICC-BDVMR system has accomplished good performance with the maximal average accuracy, TPR, and TNR of 81.52%, 87.20%, and 76.45% while the CNB, GWOCNB, and CGCNB

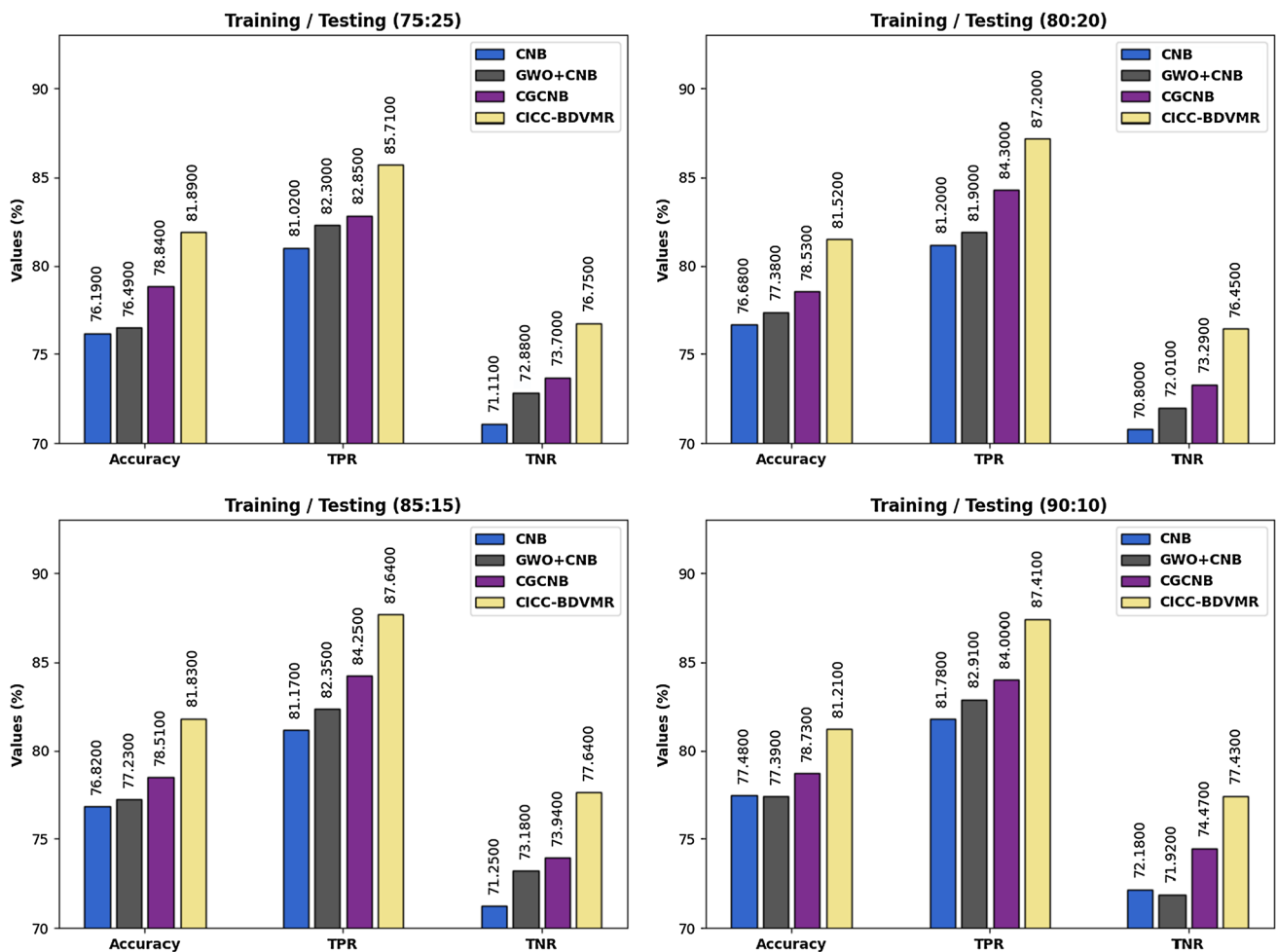
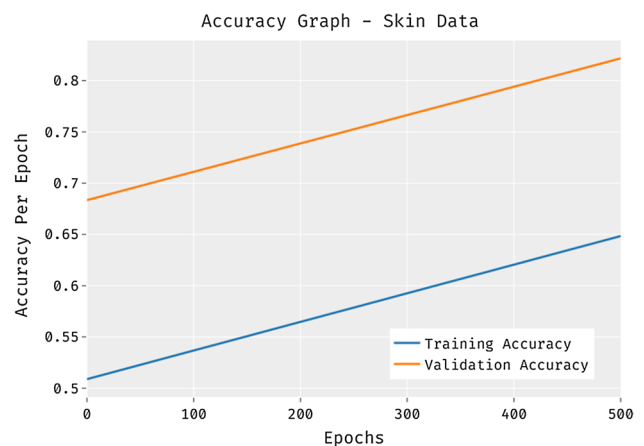


Fig. 9 Average analysis of CICC-BDVMR technique under skin data

systems have resulted in inefficient outcomes with the less accuracy of 76.68%, 81.20%, and 70.80% correspondingly. Additionally, with training/testing data of 90:10, the CICC-BDVMR procedure has attained effective performance with the maximal average accuracy, TPR, and TNR of 81.21%, 87.41%, and 77.43% while the CNB, GWOCNB, and CGCNB methods have resulted in inefficient outcomes with the less accuracy of 77.48%, 81.78%, and 72.18% correspondingly.

The overall accuracy analysis of CICC-BDVMR method on skin data is depicted in Fig. 10. The result demonstrates that the CICC-BDVMR approach has attained enhanced validation accuracy than training accuracy. Also, it is noticeable that

Fig. 10 Accuracy analysis of CICC-BDVMR technique under skin data



the accuracy value gets saturated with the epoch count of 1000. The above mentioned tables and figures ensured that the proposed model has accomplished effectual outcome over the other techniques.

5 Conclusion

In this study, a novel CICC-BDVMR technique has been developed for accomplishing effectual BDV by the use of clustering and data classification process on the Map Reduce environment. The proposed CICC-BDVMR technique encompasses several subprocesses namely KFCM based clustering, GOA based initial cluster center selection, RR based classification, and RCM based parameter tuning. The utilization of the GOA and RCM algorithms helps to effectually improve the overall big data classification outcomes. In order to demonstrate the enhanced performance of the CICC-BDVMR system, a comprehensive comparative result analysis is made with the benchmark datasets. The extensive comparison study demonstrates the promising performance of the CICC-BDVMR approach on the recent state of art approaches. Therefore, the CICC-BDVMR technique has been found to be a proficient tool to visualize and classify big data. In future, feature selection and feature reduction methodologies can be integrated into the proposed model to improve the classification outcomes. Our final study direction is text clustering metaheuristic optimization. Text clustering performance can be improved by combining these strategies. Text clustering difficulties can be solved via hybrid and updated methods. New meta-heuristic optimization methods for clustering problems have recently been proposed. Text clustering difficulties can be solved via hybrid and updated methods. New meta-heuristic optimization methods for clustering problems have recently been proposed. Others include Salp Swarm Optimization, Harris Hawks Optimization, and Henry Gas Solubility Optimization.

Author contributions ZX did all work of this paper. The author read and approved the final manuscript.

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests The author declares no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Abd Elaziz M, Li L, Jayasena KN, Xiong S. Multiobjective big data optimization based on a hybrid salp swarm algorithm and differential evolution. *Appl Math Model*. 2020;80:929–43.
2. Shelke N. An efficient low complexity compression based optimal homomorphic encryption for secure fiber optic communication. *Optik*. 2022;252: 168545. <https://doi.org/10.1016/j.ijleo.2021.168545>.
3. Niu C, Wang L. Big data-driven scheduling optimization algorithm for Cyber-Physical Systems based on a cloud platform. *Comput Commun*. 2022;181:173–81.
4. Aslan S, Karaboga D. A genetic Artificial Bee Colony algorithm for signal reconstruction based big data optimization. *Appl Soft Comput*. 2020;88: 106053.
5. Zhou Z-H, Chawla NV, Jin Y, Williams GJ. Big data opportunities and challenges: discussions from data analytics perspectives. *IEEE Comput Intell Mag*. 2014;9:62–74.
6. Gudivada VN, Baeza-Yates R, Raghavan VV. Big data: promises and problems. *Computer*. 2015;48(3):20–3.
7. Snijders C, Matzat U, Reips UD. "Big Data": big gaps of knowledge in the field of internet science. *Int J Internet Sci*. 2012;7:1–5.
8. Tsai C-W, Lai C-F, Chao H-C, Vasilakos AV. Big data analytics: a survey. *J Big Data*. 2015;2(1):21.
9. Kaisler S, Armour F, Espinosa JA, Money W. Big data: issues and challenges moving forward. *Int Conf Syst Sci*. 2013. <https://doi.org/10.1109/HICSS.2013.645>.

10. Abukhodair F, Alsaggaf W, Jamal AT, Abdel-Khalek S, Mansour RF. An intelligent metaheuristic binary pigeon optimization-based feature selection and big data classification in a mapreduce environment. *Mathematics*. 2021;9(20):2627.
11. Brahmane AV, Krishna CB. Rider chaotic biography optimization-driven deep stacked auto-encoder for big data classification using spark architecture: rider chaotic biography optimization. *Int J Web Serv Res*. 2021;18(3):42–62.
12. Qin X, Luo Y, Tang N, Li G. Deepeye: an automatic big data visualization framework. *Big Data Min Anal*. 2018;1(1):75–82.
13. Galletta A, Carnevale L, Bramanti A, Fazio M. An innovative methodology for big data visualization for telemedicine. *IEEE Trans Industr Inf*. 2018;15(1):490–7.
14. Cui Y, Song X, Hu Q, Li Y, Shanthini A, Vadivel T. Big data visualization using multimodal feedback in education. *Comput Electr Eng*. 2021;96:107544.
15. Hardas BM, Ch T, et al. An automated word embedding with parameter tuned model for web crawling. *Intell Autom Soft Comput*. 2022;32(3):1617–32.
16. Dubey AK, Kumar A, Agrawal R. An efficient ACO-PSO-based framework for data classification and preprocessing in big data. *Evol Intel*. 2021;14(2):909–22.
17. Wang HB, Gao YJ. Research on C4. 5 algorithm improvement strategy based on MapReduce. *Proc Comput Sci*. 2021;183:160–5.
18. Zhang DQ, Chen SC. A novel kernelized fuzzy c-means algorithm with application in medical image segmentation. *Artif Intell Med*. 2004;32(1):37–50.
19. Chen L, Chen CP, Lu M. A multiple-kernel fuzzy c-means algorithm for image segmentation. *IEEE Transact Syst Man Cybern Part B*. 2011;41(5):1263–74.
20. Mirjalili SZ, Mirjalili S, Saremi S, Faris H, Aljarah I. Grasshopper optimization algorithm for multi-objective optimization problems. *Appl Intell*. 2018;48(4):805–20.
21. Sulaiman M, Masihullah M, Hussain Z, Ahmad S, Mashwani WK, Jan MA, Khanum RA. Implementation of improved grasshopper optimization algorithm to solve economic load dispatch problems. *Hacet J Math Stat*. 2019;48(5):1570–89.
22. Paulraj D. 'A gradient boosted decision tree-based sentiment classification of twitter data. *Int J Wavelets Multiresolution Inf Process*. 2020;18(4):205027. <https://doi.org/10.1142/S0219691320500277>.
23. Jain DK, SahTyagi SKK, Neelakandan S, Prakash M, Natrayan L. Metaheuristic optimization-based resource allocation technique for cyber-twin-driven 6G on IoE environment. *IEEE Transact Ind Inf*. 2022. <https://doi.org/10.1109/TII.2021.3138915>.
24. Craig S, Gammerman A, Vovk V. Ridge regression learning algorithm in dual variables. In: *Proceedings of the 15th International Conference on Machine Learning, ICML-1998*. Burlington: Morgan Kaufmann; 1998.
25. Rayen SJ, Arunajsmine J. Social media networks owing to disruptions for effective learning. *Proc Comput Sci*. 2020;172:145–51. <https://doi.org/10.1016/j.procs.2020.05.022>.
26. Al-Kubaisy WJ, Yousif M, Al-Khateeb B, Mahmood M, Le DN. The red colobuses monkey: a new nature-inspired metaheuristic optimization algorithm. *Int J Comput Intell Syst*. 2021;14(1):1108–18.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.