Research Article

Hanumanthu Bhukya* and Sadanandam Manchala

Design of metaheuristic rough set-based feature selection and rule-based medical data classification model on MapReduce framework

https://doi.org/10.1515/jisys-2022-0066
received November 15, 2021; accepted May 07, 2022

Abstract: Recently, big data analytics have gained significant attention in healthcare industry due to generation of massive quantities of data in various forms such as electronic health records, sensors, medical imaging, and pharmaceutical details. However, the data gathered from various sources are intrinsically uncertain owing to noise, incompleteness, and inconsistency. The analysis of such huge data necessitates advanced analytical techniques using machine learning and computational intelligence for effective decision making. To handle data uncertainty in healthcare sector, this article presents a novel metaheuristic rough set-based feature selection with rule-based medical data classification (MRSFS-RMDC) technique on MapReduce framework. The proposed MRSFS-RMDC technique designs a butterfly optimization algorithm for minimal rough set selection. In addition, Hadoop MapReduce is applied to process massive quantity of data. Moreover, a rule-based classification approach named Repeated Incremental Pruning for Error Reduction (RIPPER) is used with the inclusion of a set of conditional rules. The RIPPER will scale in a linear way with the number of training records utilized and is suitable to build models with data uncertainty. The proposed MRSFS-RMDC technique is validated using benchmark dataset and the results are inspected under varying aspects. The experimental results highlighted the supremacy of the MRSFS-RMDC technique over the recent state of art methods in terms of different performance measures. The proposed methodology has achieved a higher *F*-score of 96.49%.

Keywords: MapReduce, big data analytics, medical data classification, rough set, feature selection, rule-based classification

1 Introduction

Big data analytics (BDA) becomes a hot research area among research communications and finds its applicability in several application areas encompassing healthcare, business, and industrial sector [1]. BDA refers to the application of artificial intelligence (AI) techniques to much massive, heterogeneous huge datasets that comprise processed, semi-structured, and unstructured information from many resources with sizes ranging from terabytes to zettabytes. It is a phrase used to describe data sets that are too large or too complex for typical database systems to gather, maintain, and analyze with less delay.

* **Corresponding author: Hanumanthu Bhukya,** Department of Computer Science and Engineering, KUCE&T, Kakatiya University, Warangal-506 009, Telangana, India, e-mail: bhcsekits@gmail.com
Sadanandam Manchala: Department of Computer Science and Engineering, KUCE&T, Kakatiya University, Warangal-506 009, Telangana, India, e-mail: msadanandam@kakatiya.ac.in

Big data analysis enables investigators, academics, and enterprise customers to generate faster and more effective choices utilizing unprecedented or unsuitable information. The large quantity of data generated at maximum velocity in healthcare poses a challenging issue. It results in repetitive data, which leads to being expensive and consumes more time. At the same time, the massive quantity of data from disease diagnosis, meaningful data inspection, prediction, and optimization approaches offer insights into healthcare applications [2]. Hence, the healthcare association is finding an effective information technology (IT) artifact which authorized to consolidate authoritative resources to carry over a high-quality patient involvement [3]. Recently, MapReduce-based BDA techniques have been developed to handle massive quantities of data [4,5]. The advantages of BDA are enhanced customer satisfaction, quality growth and creativity, complicated source connections, risk administration, faster and good decision making, and so on.

Behavior patterns are sometimes known as chain of activities, which emphasizes their origin as a complicated connection of shorter sections of behavior. They can be produced by behavior modification of several parts delivered in the proper sequence, usually known as a pattern of behavior. In the business sector, the essential worth of large data is proficiently used to identify the behavioral pattern of the consumer to design novel business services and solutions [6]. In the medical field, the implication of big data acts as a prediction model and machine learning (ML) model to provide useful solutions to implement treatment plans and customized medical care. Owing to the advancements of BDA and the combined technologies, the healthcare sector has identified pragmatic transformation at distinct levels from the perspective of existing stakeholders [7]. The influence of big data in medical field leads to detect new data resources such as social networking, telematics, and wearables, to analyze the legacy resources which comprise patient's data, diagnosis, clinical trial data, and so on. If the data sources and analysis are integrated, it offers a valued source of data for medical community in achieving effective healthcare solutions [8].

Although BDA using AI is useful, an extensive challenging issue is data uncertainty [9]. For example, every V feature includes diverse resources of uncertainty, like unstructured, incomplete, or noisy data. In addition, it can be included in the whole analytic process such as collection, organization, and investigation of big data [10,11]. Moreover, the ML model might not result in optimum outcomes when the training dataset is biased. Previous studies [12,13] presented six major challenging issues in BDA, comprising data uncertainty. They have concentrated majorly on the way the uncertainty affects the efficiency of learning from big data, while a distinct concern exists in the mitigation of uncertainty inheritance in the huge dataset. They generally exist in data mining as well as ML techniques. Thus, resolving uncertainty in BDA should be at the front of any automated solutions, since uncertainty has major impact on the overall accuracy.

This study focuses on the design of metaheuristic rough set-based feature selection with rule-based medical data classification (MRSFS-RMDC) technique on MapReduce framework. It is utilized to reduce the number of semantic similarity selections and RIPPER-based classifications. It is utilized in MapReduce environments to manage large amounts of information. The proposed MRSFS-RMDC technique designs a butterfly optimization algorithm (BOA) for minimal rough set selection. The butterfly (BF) optimization method is a revolutionary population-based swarm intelligence system that replicates the foraging behavior of BFs. BOA has been used in a variety of disciplines. It mimics the grazing activity of insects. Enzymatically, every insect has sensory receptors all throughout its body. These receptors are known as chemoreceptors, and they are responsible for inhaling the aroma of foodstuff and flowers. Also, Hadoop MapReduce is applied to process massive quantity of data. Furthermore, a rule-based classification approach named Repeated Incremental Pruning for Error Reduction (RIPPER) is used with the inclusion of a set of conditional rules. For inspecting the enhanced performance of the MRSFS-RMDC technique, a wide range of simulations take place on benchmark PIMA Indians diabetes dataset, and the results are examined under varying aspects.

In this article, the information based on BDA and some other techniques are given in Section 1. Section 2 shows the related work which are related to the BDA process and neural networks (NNs). The proposed methodology is explained in Section 3. The analyzed techniques are experimentally validated in Section 4.

Finally, the efficiency and analysis process are summarized in the conclusion part, which is illustrated in Section 5.

2 Related works

In the study of Wang and He [14] and Ali et al. [15], an intelligent healthcare model is presented to predict heart disease by the use of ensemble DL and feature fusion models. Fusion centers undertake analytics and promote data exchange, supporting federal authorities and border patrol authorities in the prevention, detection, and response to extremism and violence. First, the feature fusion model integrates the derived attributes from sensors and electronic health records to produce meaningful healthcare data. Second, the information gain (IG) approach removes the irrelated and repetitive attributes and chooses the essential ones that reduce the computation complexity and improve the system efficiency. Moreover, the conditional probability method determines a particular feature weight for all classes that additionally enhance the efficiency of the system. Several possible uses of probability density functions include calculating likelihood function for variables and crucial areas for theory testing. It is frequently beneficial to construct a plausible distribution of income theory for discrete variables. At last, an ensemble DL model undergoes training to predict heart disease.

Ramani et al. [16] presented an improved artificial neural network (ANN) classification model on a MapReduce model to predict diabetes. Initially, min–max normalization approach is involved to preprocess the healthcare data, and the MapReduce is utilized to offer an effective model in the predictive programming algorithms for the map and reduce functions. It is an easier programming interface that assists to solve the prediction problem. Chrimes et al. [17] established a novel BDA that is effectively designed in the Hadoop/MapReduce technology developed in the HBase environment and created hospital-oriented metadata at high volume. Generally, the model over generated HBase data files takes a week or a month for 1 billion (10TB) and 3 billion (30TB), respectively. Furthermore, the evaluation test results from the patient data with Apache tools in Hadoop ecosystem.

Selvi and Muthulakshmi [18] developed an effective map reduce-based optimal data classifier technique for proficiently diagnosing diabetes. It encompasses Hadoop tool, data collection, and classification using gradient boosting tree (GBT). For enhancing the classification performance of the GBT, an improved k-means clustering method is combined together. After obtaining training dataset, the k-means clustering is used, which is a sort of unsupervised learning (for instance, information in the absence of characterized classifications and sets). The purpose of this method is to locate relationships among data, with its parameter k representing the number of communities found. The clustering of sample points is dependent on characteristic resemblance. AlZubi [19] introduced an effective big data classification model such as proficient MapReduce technology to identify diabetes. Primarily, the data are gathered from a massive dataset and the MapReduce concept is employed for composing the smaller chunks of data proficiently. In line with this, data normalization is carried out and the features are chosen by the ant bee colony (ABC) algorithm. The ABC method is an optimization approach which replicates honey bee feeding behavior and has been effectively utilized in a variety of real scenarios. The percentage of working or observer honeybees in the swarm is equivalent to the total of remedies in the colony. Finally, the elected features are processed by the use of support vector machine (SVM) with multi-layer NN.

SVMs are a type of learning algorithms used for categorization, prediction, and anomaly analysis. It is also easy to learn and understand since it employs a selection of training images in the classification model (named training set). It works best when there is a decent margin of distance. It works well in three-dimensional areas. It works well when the dimensionality is more than the amount of data.

Syed et al. [20] offered an effective smart healthcare model for ambient-assisted living (AAL) for monitoring the physical actions of old people by the use of sensors and ML models for rapid examination and recommendation. AAL is described as the application of data as well as communication technique (ICT) in a human's everyday housing and living surroundings to help individuals to keep busy more, maintain

talkative, as well as lead a normal life beyond older years. Primarily, wearables are used for data collection and sent to the cloud and data analytics layer. For managing large quantity of data in a simultaneous way, Hadoop MapReduce tool is employed. Finally, the multi-nominal Naïve Bayes classification model fitted into the MapReduce tool is utilized for the motion classification process.

3 The proposed model

In this study, a new MRSFS-RMDC technique has been developed for medical data classification with data uncertainty. The proposed MRSFS-RMDC technique encompasses preprocessing BOA-based minimal rough set selection and RIPPER-based classification. In addition, the MRSFS-RMDC technique is executed in the MapReduce environment to handle big data. Figure 1 demonstrates the overall block diagram of proposed MRSFS-RMDC model. The detailed working of these three modules is offered in the succeeding sections.

3.1 Preprocessing

At the initial stage, the data preprocessing takes place to transform the data into a compatible format. Primarily, the pre-processed step was implemented to change non-traditional into traditional datasets that improve the accuracy of presented approach. Here, the min–max normalization manner was applied. Among the most prevalent methods for normalizing information is min–max normalization. For each attribute, the significance level is converted to a 0, the highest value is converted to a 1, and all other values are converted to a fraction within 0 and 1. In this approach, the feature is being rescaled to the range of 0 to 1 or −1 to 1. This rescaling is done by the utilization of linear interpretation equation defined as follows:

$$y' = (y_{\max} - y_{\min}) \times \frac{(x_i - x_{\min})}{(x_{\max} - x_{\min})} + x_{\min}, \tag{1}$$

where $(y_{\max} - y_{\min}) = 0$; if $(x_{\max} - x_{\min}) = 0$ to a feature, it defines the continuous rate of feature from the data. Once the value of feature has been identified with constant value from the data, it needs the normalized value as it does not deliver some data to NN. Once the min–max normalization was implemented, all the features lie from the novel range of values that in fit remain similar. The normalization utilizing min–max is an advantage of maintaining every connection from the data correctly. The min–max normalization approach ($y = (x - \min)$) is utilized; however, there are additional choices. The original picture information will be changed in the range of 0 to 1 by using min–max normalization (inclusive).

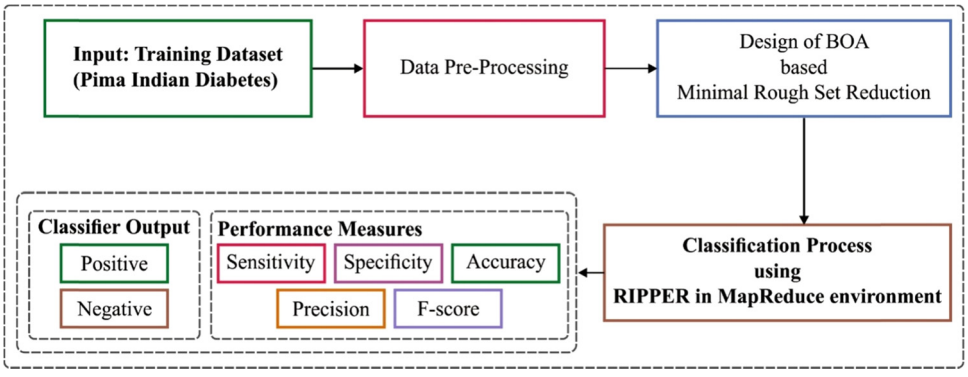

Figure 1: Overall process of MRSFS-RMDC model.

3.2 Design of BOA-based minimal rough set selection

Once the medical data are preprocessed, the feature selection process is carried out by the use of BOA-based minimal rough set selection. The BOA has been metaheuristic technique, which is simulated as foraging and mating nature of BFs. An essential feature of BOA from other metaheuristics is that all BFs hold their individual scent. The fragrance is represented using the following equation:

$$f_i = cI^a, \tag{2}$$

where f_i denotes the apparent order of fragrance, c implies the sensory modality, I is the stimulus intensity, and a signifies the power exponent that depends on degree of fragrance absorption. In practice, the values of sensory morphology coefficient c from the interval of $[0, \infty]$ are assumed. Then, the value is decided utilizing the discrimination of optimized problem from the iterated system of BOA [19]. The sensory modality c from the optimal searching stage of the BOA is defined using the following equation:

$$c_{t+1} = c_t + [0.\ 025/(c_t \cdot T_{\max})], \tag{3}$$

where $T_{\max}$ denotes higher iteration count and the beginning value of c is fixed as 0.01. Also, two major processes are involved in the BOA, namely global search and local search stages. The global searching movement of the BFs can be defined using the following equation:

$$x_i^{t+1} = x_i^t + (r^2 \times g_{\text{best}} - x_i^t) \times f_i, \tag{4}$$

where x_i^t signifies the solution vector x_i of the ith BF in t round and r denotes an arbitrary number from the range of $[0, 1]$. In this case, g_{best} denotes the present optimum solution attained from every solution attained in the present stage. Specifically, f_i signifies the fragrance of the ith BF. The local searching process can be defined using the following equation:

$$x_i^{t+1} = x_i^t + (r^2 \times x_i^k - x_j^t) \times f_i, \tag{5}$$

where x_j^t and x_i^k are jth and kth BFs arbitrarily selected from the solution area. When x_j^t and x_i^k belong to the identical round, they denote that the BF will become a local random walk. Else, this type of arbitrary movement diversified the solutions. The global as well as local searching process for food and mating partners by the BF takes place naturally. Therefore, a switching probability p has been considered for transforming the common global and intensive local searching process. At all rounds, the BOA arbitrarily makes a number from the range of $[0, 1]$ that has undergone comparison with the switch probability p in deciding to perform a global or local searching process [21].

The concept of BOA can be used for the minimal rough set selection issue. Assume a huge feature space with entire feature subsets. Every feature subset can be considered as a point or location in the space. When a total of N features exist, a set of 2^N types of subsets occur varying from one another in the length and features exist in every subset. The optimum location is the subset with minimal length and maximum classification accuracy. There were multiple BFs in the feature space, and each BF was assigned a position. If the BF searches the food in the space, the goal is to move toward the optimum location. Later, the position gets changed, they communicate with one other and search for the local as well as global optimum position. The process involved in the BOA-based rough set selection process is detailed in the following.

Initially, the BF position is considered as the binary bit string of length N, where N denotes the feature count. Every bit denotes a feature, and particularly, the value "1" implies that the respective feature is chosen whereas "0" denotes unselected. Every position implies a feature subset. Consider X, Y as the binary bit strings which denote the position of two BFs.

The Hamming distance is a statistic that may be used to compare two binary value sequences. When two discrete sequences of similar duration are compared, the Hamming distance is the amount of bit locations where the data pairs disagree. d is indicated as hamming distance and it can be illustrated as $d(a,b)$. For every binary bit string X and Y, the Hamming distance is equivalent to the number of 1's in

X *XOR* Y. Assume X, Y are the two binary bit strings denoting the position of two BFs, and the Hamming distance can be equated using the following equation:

$$h(X, Y) = \sum_{i=1}^{N} x_i \oplus y_i, \tag{6}$$

where $\oplus$ denotes a modulo-2 addition, $x_i, y_i \in \{0, 1\}$. The parameter x_i signifies a binary bit in X. Consider $X^1, X^2, \ldots, X^n$ are the binary bit strings representing the position of n BFs. The midpoint position of n BFs can be formulated as follows:

$$\begin{aligned} X_c = \{c_1, \ldots, c_N| \text{ if } \frac{1}{n}\sum_{j=1}^{n} x_i^j > 0.5, \text{ then} \\ c_i = 1; \quad \text{otherwise } c_i = 0\}, \end{aligned} \tag{7}$$

where x_i^j characterizes the ith bit of BF position X^j, and X_c is the middle position of n BFs.

Every BF begins with an arbitrary position in each round [22]. Every BF aims to modify each step in the searching space based on the behavior of searching, swarming, and following. A fitness function can be derived based on the three behaviors and the one with higher fitness value can be chosen for updating the succeeding position. It can be represented as follows:

$$\text{Fitness} = \alpha^* \gamma_R(D) + \beta^* \frac{|C| - |R|}{|C|}, \tag{8}$$

where $\gamma_R(D)$ denotes the classification quality (also called dependency), $|R|$ indicates the total number of "1" numbers in a binary BF position, indicating the chosen feature count, also termed as the feature subset length. $|C|$ denotes the total feature count. α and β are two variables implying the significance of classifier quality and length of the subsets, $\alpha \in [0, 1]$, and $\beta = 1 - \alpha$. It indicates that the classification quality and feature length might hold distinct significance to the feature selection process. The fitness function (sometimes referred to as the analytical solution) determines how near a particular approach is to the optimization method of a particular topic. It assesses how appropriate a proposal is. The importance of every position is determined using the fitness function.

Once the BF reached a maximum fitness value, it is perished with getting a rough set select. It indicates that the BF constructs the local optimum solution. The succeeding round beings when every BF gets perished. The termination condition is set as the maximum number of iterations or attaining an identical set of feature selects under three succeeding rounds. [23]

3.3 Design of RIPPER in MapReduce environment

During the classification process, the RIPPER technique gets executed on the MapReduce platform to classify the healthcare data. MapReduce performs two important capabilities as it classifies and distributes activity to different devices in the system or mapping, a service known as the mapper, and it organizes and combines the data for every server into a coherent response to a question, known as the reduction. The RIPPER has been extremely utilized as a rule induction technique. It scales linearly with the training sample count utilized and is suitable to structure techniques with data uncertainty. Also, it utilizes a validation set (VS) for preventing the technique from overfitting. The RIPPER orders the classes based on the frequency. When $(y_1, y_2, \ldots, y_c)$ denotes the class label and y_1 refers to the minimum frequency and y_c the maximum frequency, afterward, RIPPER primarily forms rules to y_1 taking residual class record as negative record. Then, RIPPER removes principles to y_2. This procedure has been repeated till y_c is left, which is considered as default class.

In order to generate rules, RIPPER utilizes an approach which primarily considers all rules are empty and afterward it can be constructed with more conjuncts to it consecutively. It utilizes FOIL IG for adding conjunct to rule. Assume that rule R: $A \rightarrow$ *class* cover p_0 positive record and n_0 negative record. Then, a

more novel conjunct B, the rule $R' : A \wedge B \rightarrow class$ cover p_1 positive record and n_1 negative record. Next, the FOIL IG is computed as:

$$\text{FOIL IG} = p_1 \times \left(\log \frac{p_1}{p_1 + n_1} - \log \frac{p_0}{p_0 + n_0} \right). \tag{9}$$

The conjunct is increasing till the rule begins covering negative instances. The rule was pruned dependent upon their efficiency on the VS utilizing the subsequent metric $(p - n)/(p + n)$, where p refers to the amount of positive record concealed with rule from the VS and n implies the amount of negative record supporting the rule under the VS.

After creating the rule, every record under the rule is removed. The minimum description length (MDL) concept is a strong inductive assessment approach that serves as the foundation for data analysis, analytical thinking, and computer vision. It asserts that the optimal interpretation, provided a restricted collection of observable facts, is the one which allows for the most quantization. This technique then continues with constructing a novel rule. The rule has been made if the rule set does not violate the MDL rule and the error on the VS was lesser than 50%.

RIPPER can be applied in Java using Hadoop Java library. The dataset has been separated horizontally for supporting the Hadoop MapReduce structure and making sure parallel implementation of code. The mapper–reducer is used for three purposes: one in one to rule generating, rule pruning, and computing accuracy [20]. Therefore, all the mappers implement their code on some of the datasets, and the reducer aggregates on the outcome of mapper for producing one general output. Figure 2 illustrates the framework of MapReduce.

In order to rule generating, the mapper–reducer function computes the values of p_1 and n_1 to compute the FOIL's IG (p_0 and n_0 values are p_1 and n_1 values to the old rule, respectively). Altogether, all feasible values to every attribute are regarded as conjuncts that rule more. The FOIL's IG to all of these values are computed, and the value to that IG is maximal, that is, more conjunct added to rule. The rule pruning was complete utilizing the VS as situation. The mapper–reducer function to pruning computes the p and n values to metric $(p - n)/(p + n)$. According to the value of metric, the rule was pruned along with the rule set. Afterward, every rule has made the rule set require that validation on test record. The accuracy mapper–reducer function computes the amount of positive as well as negative records supported by all the rules and entire rule set. These values have been utilized for calculating the accuracy of all rules and entire accuracy as well.

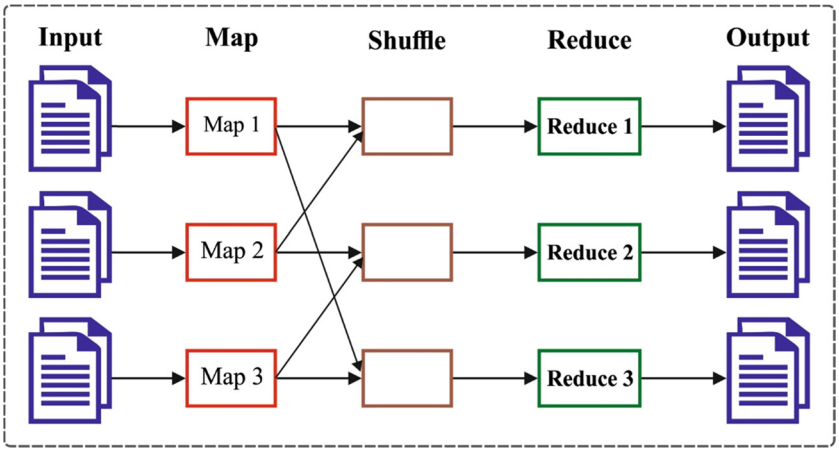

Figure 2: Structure of MapReduce.

3.4 Rule growing phase

The rule has been adjusted as an empty rule, for instance, it covers every record. Then, the conjunct is added further one by one to the rule. The conjuncts with the value of FOIL's IG measure have been chosen more. The parameter of measure was computed utilizing MapReduce functions, and the <key, value> pairs have the values of p_0, p_1, n_0, and n_1. The conjunct is further added to rule if it does not cover negative record.

3.5 Rule pruning phase

The rule created in one is before pruned utilizing $(p - n)/(p + n)$ metric. For calculating the parameters p and n of this metric, the MapReduce model was named as <key, value> pairs.· Stages 1 and 2 are repeated till increasing a novel rule violates the MDL rule.

3.6 Model evaluation phase

Then the rule set is created, and the rules have been utilized for classifying the test record. The MapReduce purpose is named for classifying the record and computing the accuracy of this technique. The <key, value> pairs comprise the value of entire positive and negative records supporting the method. The rule set is returned, and the correctness of the rules and the rule set on the test record are assessed.

4 Experimental validation

The performance validation of the MRSFS-RMDC technique takes place using PIMA Indians diabetes dataset from Kaggle repository [24–27]. Customers may use Kaggle to search and post statistical models, study and construct models in a web-based network infrastructure, collaborate with other data professionals as well as supervised learning experts, and compete to accomplish ML tasks. The dataset comprises 768 samples with 8 attributes and 2 classes namely positive/negative. The results are examined under varying sizes of data. Table 1 provides the sensitivity and specificity analysis of the MRSFS-RMDC technique under varying data sizes in GB.

Table 1: Result analysis of MRSFS-RMDC model with different data sizes

Data size (GB)	DNN	SVM	DCD-ANN	MRSFS-RMDC
Sensitivity (%)				
2	89.03	91.81	93.34	94.92
4	89.77	92.27	93.71	95.49
6	90.00	92.70	94.77	96.18
8	90.95	93.02	95.95	97.27
10	91.81	94.37	96.93	97.96
Specificity (%)				
2	81.00	80.46	82.68	84.91
4	81.72	80.76	83.40	85.21
6	82.56	80.94	84.06	86.17
8	82.98	80.88	84.73	87.13
10	84.49	81.18	86.89	90.01

Figure 3 illustrates the sensitivity analysis of the MRSFS-RMDC technique with existing ones under distinct sizes of dataset. The results show that the MRSFS-RMDC technique has accomplished effective outcomes with the maximum sensitivity values. For instance, with 2 GB data, the MRSFS-RMDC technique has attained an increased sensitivity of 94.92%, whereas the DNN, SVM, and DCD-ANN techniques have obtained reduced sensitivity of 89.03, 91.81, and 93.34%, respectively. At the same time, with 10 GB data, the MRSFS-RMDC technique has achieved a higher sensitivity of 97.96%, whereas the DNN, SVM, and DCD-ANN techniques have accomplished lower sensitivity of 91.81, 94.37, and 96.93%, respectively.

Figure 4 shows the specificity analysis of the MRSFS-RMDC system with existing ones under distinct sizes of dataset. The outcomes demonstrated that the MRSFS-RMDC approach has accomplished effective outcomes with the maximum specificity values. For example, with 2 GB data, the MRSFS-RMDC manner has attained an increased specificity of 84.91%, whereas the DNN, SVM, and DCD-ANN techniques have obtained reduced specificity of 81, 80.46, and 82.68%, respectively. Also, with 10 GB data, the MRSFS-RMDC approach has reached an increased specificity of 90.01%, whereas the DNN, SVM, and DCD-ANN methodologies have accomplished minimum specificity of 84.49, 81.18, and 86.89%, respectively.

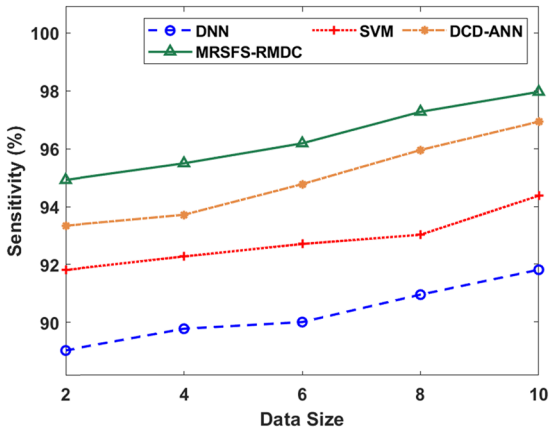

Figure 3: Sensitivity analysis of MRSFS-RMDC model with distinct data size.

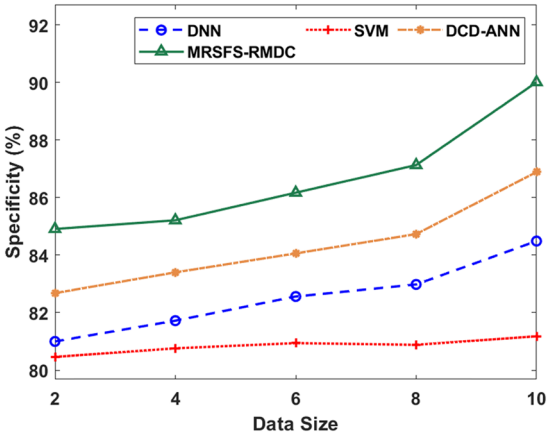

Figure 4: Specificity analysis of MRSFS-RMDC model with distinct data size.

Table 2 and Figure 5 depict the precision analysis of the MRSFS-RMDC approach with existing ones in varying sizes of dataset. The results portrayed that the MRSFS-RMDC methodology has accomplished effective outcomes with the maximal precision values. For example, with 2 GB data, the MRSFS-RMDC manner has attained an enhanced precision of 85.09%, whereas the DNN, SVM, and DCD-ANN techniques have obtained reduced precision of 82.04, 81.37, and 83.71%, respectively. Simultaneously, with 10 GB data, the MRSFS-RMDC algorithm has obtained a maximum precision of 90.81%, whereas the DNN, SVM, and DCD-ANN approaches have accomplished lower precision of 86.62, 84.19, and 88.90%, respectively.

Table 2: Precision analysis of MRSFS-RMDC model with varying data size

Precision (%)				
Data size (GB)	**DNN**	**SVM**	**DCD-ANN**	**MRSFS-RMDC**
2	82.04	81.37	83.71	85.09
4	83.38	81.80	83.95	85.66
6	84.47	82.09	85.47	86.85
8	84.71	83.38	86.76	88.24
10	86.62	84.19	88.90	90.81

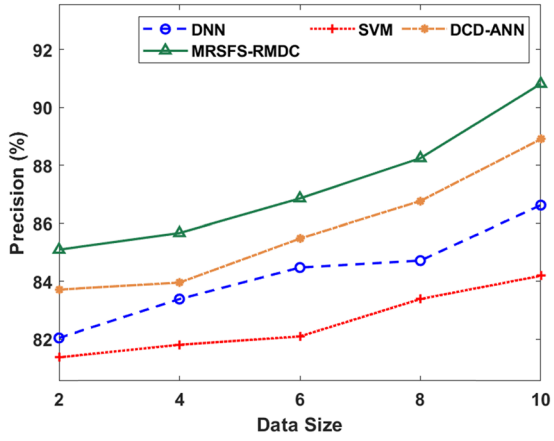

Figure 5: Precision analysis of MRSFS-RMDC model with distinct data size.

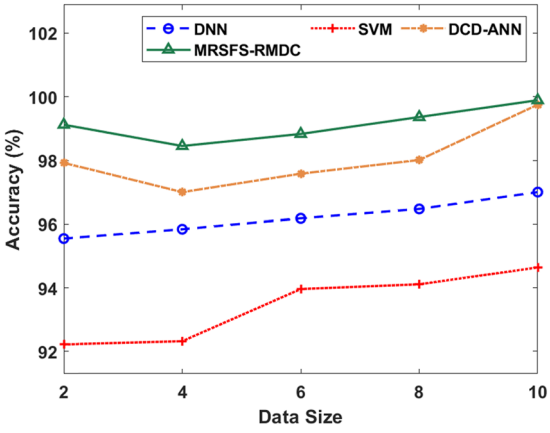

Figure 6: Accuracy analysis of MRSFS-RMDC model with distinct data size.

Table 3 offers the accuracy and *F*-score analysis of the MRSFS-RMDC approach under varying data sizes in GB. Figure 6 demonstrates the accuracy analysis of the MRSFS-RMDC method with existing ones under varying sizes of dataset. The outcome outperformed that the MRSFS-RMDC algorithm has accomplished effectual results with the maximal accuracy values. For example, with 2 GB data, the MRSFS-RMDC approach has gained a superior accuracy of 99.12%, whereas the DNN, SVM, and DCD-ANN systems have achieved decreased accuracy of 95.55, 92.22, and 97.92%, respectively. Along with that, with 10 GB data, the MRSFS-RMDC methodology has achieved an improved accuracy of 99.89%, whereas the DNN, SVM, and DCD-ANN techniques have accomplished reduced accuracy of 97, 94.64, and 99.75%, respectively.

Figure 7 shows the *F*-score analysis of the MRSFS-RMDC technique with existing ones under different sizes of dataset. The outcomes exhibited that the MRSFS-RMDC technique has accomplished effective outcomes with the higher *F*-score values. The *F*-score or *F*-measure is an estimate of a test's efficiency in descriptive statistics of binary categorization. An *F*-score can have a maximum benefit of 1.0, signifying flawless accuracy or recollection, and a minimum value of 0 if the accuracy or the recollection is 0. For example, with 2 GB data, the MRSFS-RMDC algorithm has reached an enhanced *F*-score of 95.29%, whereas the DNN, SVM, and DCD-ANN techniques have obtained minimal *F*-scores of 83.77, 82.86, and 93.66%, respectively. Besides, with 10 GB data, the MRSFS-RMDC methodology has achieved a higher *F*-score of 96.49%, whereas the DNN, SVM, and DCD-ANN manners have accomplished lower *F*-scores of 88.24, 85.74, and 94.62%, respectively.

Table 3: Comparative analysis of MRSFS-RMDC model in terms of accuracy and *F*-score

Data size (GB)	DNN	SVM	DCD-ANN	MRSFS-RMDC
Accuracy (%)				
2	95.55	92.22	97.92	99.12
4	95.84	92.32	97.00	98.45
6	96.18	93.96	97.58	98.83
8	96.47	94.11	98.01	99.36
10	97.00	94.64	99.75	99.89
***F*-Score (%)**				
2	83.77	82.86	93.66	95.29
4	84.16	83.20	93.75	95.14
6	85.55	84.01	93.90	95.58
8	86.37	84.59	94.14	95.77
10	88.24	85.74	94.62	96.49

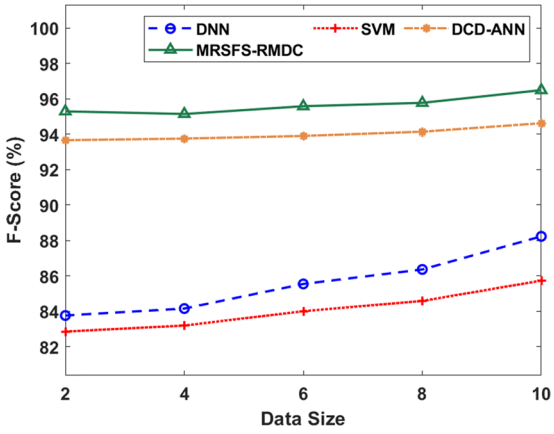

Figure 7: *F*-score analysis of MRSFS-RMDC model with distinct data size.

By looking into the aforementioned tables and figures, it can be ensured that the MRSFS-RMDC technique is found to be an effective tool to medical data classification.

5 Conclusion

In this research, a new MRSFS-RMDC technique has been developed for medical data classification with data uncertainty. The proposed MRSFS-RMDC technique encompasses preprocessing BOA-based minimal rough set selection and RIPPER-based classification. In addition, the MRSFS-RMDC technique is executed in the MapReduce environment to handle big data. The design of BOA technique and RIPPER helps to handle data uncertainty in medical data classification. For examining the improved performance of the MRSFS-RMDC technique, a wide range of simulations take place on benchmark PIMA Indians diabetes dataset and the results are inspected under varying aspects. The experimental results showcased that the MRSFS-RMDC technique has accomplished effectual outcomes over the other recent approaches in terms of different performance measures. In future, the medical data classification performance of the MRSFS-RMDC technique can be boosted by the inclusion of clustering and outlier detection approaches.

Conflict of interest: The author declares no conflict of interest.

References

[1] Hariri RH, Fredericks EM, Bowers KM. Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data. 2019;6(1):1–16.
[2] Rahini S. Large scale optimization to minimize network traffic using MapReduce in big data applications. International Conference on Computation of Power, Energy Information and Communication (ICCPEIC); April 2016. p. 193–9.
[3] Ma C, Zhang HH, Wang X. Machine learning for big data analytics in plants. Trends Plant Sci. 2014;19(12):798–808.
[4] Kumar S, Kumar-Solanki V, Choudhary SK, Selamat A, Gonzalez-Crespo R. Comparative study on ant colony optimization (ACO) and K-means clustering approaches for jobs scheduling and energy optimization model in internet of things (IoT). Int J Interact Multimed Artif Intell. 2020;6(1):107.
[5] Zhou L, Pan S, Wang J, Vasilakos AV. Machine learning on big data: opportunities and challenges. Neurocomputing. 2017;237:350–61.
[6] Wang L, Alexander CA. Big data in medical applications and health care. Am Med J. 2015;6:1–8.
[7] Paulraj D. An automated exploring and learning model for data prediction using balanced CA-Svm. J Ambient Intell Humanized Comput. 2020;Springer 1–12. ISSN 1868-5137 (online), Published Online: April 2020.

[8] Tsai CW, Lai CF, Chao HC, Vasilakos AV. Big data analytics: a survey. J Big Data. 2015;2(1):21.
[9] Neelakandan S, Berlin MA, Tripathi S, Devi VB, Bhardwaj I, Arulkumar N. IoT-based traffic prediction and traffic signal control system for smart city. Soft Comput. 2021;25:12241–48. doi: 10.1007/s00500-021-05896-x.
[10] Palanisamy V, Thirunavukarasu R. Implications of big data analytics in developing healthcare frameworks–A review. J King Saud Univ-Computer Inf Sci. 2019;31(4):415–25.
[11] Slagter K, Hsu CH, Chung YC, Zhang D. An improved partitioning mechanism for optimizing massive data analysis using MapReduce. J Supercomputing. 2013;66(1):539–55.
[12] Dineshkumar M. Decentralized access control of data in cloud services using key policy attribute based encryption. Int J Sci Res Dev. APRIL 2015;3(2):2016–20. ISSN 2321-0613.
[13] Chen M, Li Y, Zhang Z, Hsu CH, Wang S. Real-time, large-scale duplicate image detection method based on multi-feature fusion. J Real-Time Image Process. 2016;13(3):557–70.
[14] Wang X, He Y. Learning from uncertainty for big data: future analytical challenges and strategies. IEEE Syst Man Cybern Mag. 2016;2(2):26–31.
[15] Ali F, El-Sappagh S, Islam SR, Kwak D, Ali A, Imran M, et al. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Inf Fusion. 2020;63:208–22.
[16] Ramani R, Devi KV, Soundar KR. MapReduce-based big data framework using modified artificial neural network classifier for diabetic chronic disease prediction. Soft Comput. 2020;24(21):16335–45.
[17] Chrimes D, Zamani H, Moa B, Kuo A. Simulations of Hadoop/MapReduce-based platform to support its usability of big data analytics in healthcare.
[18] Selvi RT, Muthulakshmi I. Modelling the map reduce based optimal gradient boosted tree classification algorithm for diabetes mellitus diagnosis system. J Ambient Intell Humanized Comput. 2021;12(2):1717–30.
[19] AlZubi AA. Big data analytic diabetics using map reduce and classification techniques. J Supercomputing. 2020;76(6):4328–37.
[20] Syed L, Jabeen S, Manimala S, Alsaeedi A. Smart healthcare framework for ambient assisted living using IoMT and big data analytics techniques. Future Gener Computer Syst. 2019;101:136–51.
[21] Wang L, Wu Y, Xie J, Wu S, Wu Z. Energy-efficient Hadoop for big data analytics and computing: A systematic review and research insights. Future Gener Computer Syst. 2018;86:1351–67.
[22] Reshma G, Al-Atroshi C, Nassa VK, Geetha B, Sunitha G, Galety MG, et al. Deep learning-based skin lesion diagnosis model using dermoscopic images. Intell Autom Soft Comput. 2022;31(1):621–34.
[23] Kamalraj R, Neelakandan S, Kumar MR, Rao VC, Anand R, Singh H. Interpretable filter based convolutional neural network (IF-CNN) for glucose prediction and classification using PD-SS algorithm. Measurement. 2021;183:109804. doi: 10.1016/j.measurement.2021.109804.
[24] Zhang M, Long D, Qin T, Yang J. A chaotic hybrid butterfly optimization algorithm with particle swarm optimization for high-dimensional optimization problems. Symmetry. 2020;12(11):1800.
[25] Chen Y, Zhu Q, Xu H. Finding rough set reducts with fish swarm algorithm. Knowl Syst. 2015;81:22–9.
[26] Gugnani S, Khanolkar D, Bihany T, Khadilkar N. Rule based classification on a multi node scalable Hadoop cluster. In International Conference on Internet and Distributed Computing Systems. Cham: Springer; 2014, September. p. 174–83.
[27] https://www.kaggle.com/uciml/pima-indians-diabetes-database.