e-ISSN: 2583-195X

Technoarete Transactions on Intelligent Data Mining and Knowledge Discovery
Vol-3, Issue-1, February 2023

Local Search Based Genetic Algorithm for Feature Selection in Health Big Data Classification

K.H. Vani[1]*, J.Rathika[2]

[1, 2] Assistant Professor, Department of Computing, Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India
*Corresponding Author Email: khvani@cit.edu.in

Abstract

Feature selection and classification have been historically utilized in a variety of domains like business, media and medical. Mining data and its classification is extremely difficult due to the nature of high dimensional. There is a high level of complexity in big data, which makes it challenging to achieve a standard feature selection approach. The irrelevant and redundant characteristics will adversely affect the computational complexity and workflow of classification algorithms. The most common existing classification algorithms intakes all the features. However, all features are not useful in the classifier and it leads the results to subpar. Hence, there exists a need for optimization in selection features for performing classification. In this paper, Local Search based Genetic Algorithm for Feature Selection (LSGNFS) is proposed for performing classification with health big data. Genetic algorithm is modified to perform a local search. Using the local search strategy, the calculated correlation information yields unique and significant input characteristics. The purpose is to help direct the search process in such a way that freshly generated features may be fine-tuned by the features that are characterized by general and specific qualities. This helps to limit the amount of duplicated information the LSGNFS possesses by supplying just the requested features. Performance of LSGNFS is analyzed using standard data mining metrics Accuracy and F-measure with 3 health big data set namely (i) coronary heart disease dataset (ii) diabetes disease dataset and (iii) bronchial tuberculosis disease dataset. Results make an indication that LSGNFS performs better than the existing classifier and well suited for performing classification in big data.

Keywords

Big Data, Classification, Feature Selection, Genetic Algorithm, Health.

INTRODUCTION

Big data produces new ideas about illness and its risk factors. More options are there to interact directly with (i) specific patients (ii) input data from mobile health applications [29]. Physicians use these data in real-time to (i) encourage better health behavior (ii) lower hazardous exposures (iii) improve health outcomes [15]. Big data assist healthcare professionals in identifying increased health risks on individuals and treating them swiftly [5]. Efficient approaches in data management consequently assist in gauging therapy responses and helping patients meet their particular health demands. These characteristics eventually contribute to reduced inefficiencies and improved healthcare system cost control [17].

E-Healthcare System (EHS) is a multi-dimensional way of establishing, preventing, diagnosing and treating health-related issues [25]. The main components of EHS are (i) healthcare professionals (HP) (ii) available facilities (iii) financial support for the previous two things. Nursing, dental, psychology, medicine, physiotherapy and other health professions are represented among the health professionals (HP). Various levels of healthcare are required depending on the severity of the condition [7]. E-Health Record (EHR) is an electronic (digital) representation of an individual's medical information, saved on a computer. An EHR contains an individual’s (or patient's) medical history, such as diagnoses, prescriptions, medications, procedures, allergies, and prevention/treatments [6]. With appropriate permissions, healthcare professionals can utilize the same to aid the care of patients. EHR can be alternatively called EMR (i.e., Electronic Medical Record). On the internet, more than tens of gigabytes of data are created and collected every second [9]. The big data paradigm has arrived and the plentitude of data has made it more amenable to new interpretations [2]. Data from new sources allows for [13] [22]: (i) improved analysis of motivation and behavior (ii) identification of instant reactions to a particular offering and (iii) use of new signals to drive interest. Having ample and diverse data may be utilized to identify patient's needs and to optimize those needs. When data is sparse, the analysis will be difficult or meaningless [16]. Since non-systematic information in EHR is often gets missed due to different causes, it's a common occurrence that the patient and the subject refuse to provide information when it is gathered. All methods for dealing with missing data have their pitfalls and assumptions [2].

The complete-case analysis incorporates complete data that are collected from patients and skips those that are incomplete. On the other hand, the available-case analysis relies on all known cases and adopts EHR even if exist missing data is present. But these procedures are rarely applicable and this makes the option for other ways to be used and enhances the value of multiple imputations.

Problem Statement

(i). Medical data might include errors at different levels that are close to zero.

(ii). Increasingly, large data sets expand while outdated measuring techniques are inadequate, which motivates the need for feature selection techniques that are well-suited to large data sets.

(iii). Feature selection approaches for historical data are comparable to an offline approach. The net result is a massive number of irrelevant and duplicated features, resulting in a significant drop in classification accuracy.

(iv). Extracting relevant information from big data necessitates efficient feature selection algorithms.

(v). Feature selection methods might be challenging when it comes to dealing with dynamically changing data.

(vi). Non-optimized selected features lead classifier to poor results.

Objective

The main intention of this paper is to propose a Local Search based Genetic Algorithm for Feature Selection that lead to effective classification of health big data for the prediction of deadly diseases namely: (i) diabetes (ii) coronary heart disease (iii) bronchial tuberculosis.

LITERATURE REVIEW

"Heterogeneous Ensemble Method (HEM)" [24] is proposed to classify CAD by combining 3 different ML algorithms namely random forest, support vector machine and k-nearest neighbor. Features are selected using boruta-wrapper method. The importance of the features is identified using SVM. Before evaluating the performance of HEM, the dataset is balanced over-sampling strategy. "Knowledge-Driven Model (KDM)" [10] is proposed for the possible prediction of the risk level of diabetes in the early stage. It follows the association rule mining strategy to choose better features for performing the classification process. It intends to act as a robust classifier to estimate the likelihood of disease. Further, it introduces an epidemiology library to determine the risk prediction. "Multi-task Learning Network (MLM)" [28] is proposed for the rapid diagnosis and classification of cardiovascular diseases. It makes utilization of weighted fusion attention technique to address the extraction of discriminative features. Weights are assigned to increase the learning of features. "Decision Tree Methods (DTM)" [8] is proposed to predict CAD by using classification and regression tree algorithm (CART). Significance of the features is identified before classification and 3 different CART models were developed based on count (i.e., 18, 10 and 5) of feature selected for classification. Except for 5 feature CART model, the other 2 CART models have more than medium performance in classification "Principal Component Analysis based Convolutional Network (PCACN)" [27] is proposed to classify CAD by resolving the issues that arise in heart rate variability. To mine the features deeply, the convolutional network works with two layers and extracts the relevant features of heartbeats. Linear SVM is applied to handle the high-dimensional features.

"Ensemble of fuzzy logic and data mining methods (EFLDM)" [23] is proposed to predict diabetes. A strategy of data mining is utilized to locate patterns in a high-dimensional dataset. Fuzzy logic is utilized to overcome the uncertainties that are detected in medical diagnosis. The fuzzy expert system mines the dataset and provides linguistic concept-based suggestions. Feature extraction is performed for improving the accuracy. "eXtreme Gradient Boosting (XGB)" [26] is proposed to analyze the electronic health record and differentiate healthy people and diabetic affected people. Logistic regression is applied to estimate the risk level of diabetic people and assist them in managing their lifestyles. "Ensemble Soft Voting Classifier (ESVC)" [14] is proposed to classify and predict diabetes. It attempts to provide results. For performing the classification, three different machine learning-oriented algorithms are combined which are (i) logistic regression (ii) random forest and (iii) naïve Bayes. Random forest assists in performing feature selection. "Deep Neural Network Framework (DNNF)" [12] is proposed to classify diabetes-affected persons with better accuracy. It makes use of stacked auto-encoders to extract the features where backpropagation methodology is applied to fine-tune the network. Dataset was trained using a supervised learning approach. "Particle Swarm Optimization (PSO) based Fuzzy Clustering Means (PSO-FCM)" [19] is proposed to predict type 2 diabetes (T2B) disease. Initially, different clusters are randomly created and fuzzy logic is applied on every cluster for the prediction of T2B. PSO is utilized to optimize the records present in every cluster which will result in enhanced classification accuracy.

"Instance Selection Mechanism (ISM)" [18] is proposed to predict tuberculosis by minimizing the effort spend for labeling during the classification. This method attempts to identify a meaningful label that will be most suitable for application-oriented domains. The active learning framework is applied for detecting abnormalities that are highly related to tuberculosis. "Bayesian-based Convolutional Neural Network (BCNN)" [1] is proposed to identify tuberculosis. It analyzes the uncertain cases that have decreased discernibility. It validates the results using filtering concepts. The softmax layer of CNN lacks in providing trusted probability values leading to better classification. "Exploratory Machine Learning Strategy (EMLS)" [21] is proposed to predict tuberculosis. It attempts to uncover the complexity-oriented relationships and identifies the risk levels. An analysis was made on the classification tree to validate the classification accuracy. Logistic regression is applied for binary classification towards the best results. "Enhanced Logistic Regression (ELR)" [4] is proposed to classify and predict tuberculosis based on the symptoms. Statistics-oriented tests are conducted to ensure the relationship between symptoms. Association scores are calculated between symptoms and tuberculosis. Further, it attempted to prove the level of relationship between predicted-variable and target-variable. Chronic Heart Failure Prediction [3] is proposed to identify the historical medical data to enhance the prediction. Initially, features are selected dynamically and then features are selected in a static manner. The rate of classification accuracy is analyzed to identify the effect of dynamic and static feature selection.

MODIFIED GENETIC ALGORITHM BASED FEATURE SELECTION (LSGAFS)

The GA helps in identifying an optimum solution to a given research issue i.e., feature selection (FS). FS is the task with the advantage of GA producing superior solutions, but it suffers from two major issues, which are: (i) premature convergence and (ii) reduced efficacy at local optima. Recently, incorporating (or hybridizing) domain-specific information is a common approach to defeat the shortcoming faced in GA. Feature selection indicates limiting the number of 1-bits in individual strings which is combined with the local search method to form LSGAFS. Using fitness function values, it combines the NN (i.e., neural network) performance and feature correlation. By making use of local search optimization, LSGAFS highlights features that have the most utility for reducing redundancy. To increase comprehensibility, the following sections provide detailed information on each component of LSGAFS.

A basic probabilistic random function is used in the subset size determination method to return a lower result whenever the subset size is requested. In this situation, the smaller value of the random number causes GA to seek out the most important traits with a decreased count. This preprocessing step is done in LSGAFS before evolution begins. LSGAFS is, therefore, able to be led throughout the evolutionary process in a certain direction, which results in the generation of a subset of features that are critical in nature. It is important to note that, this method runs automatically rather than being manually tuned. This technique depends on hand-tuning that determines the range of j depending on the dataset's information. When dealing with datasets with a large number of input characteristics, it is better to set the range-wide. After making a few trial iterations, it should be assessed whether the range should be extended. As a result, LSGAFS's ability to recognize relevant characteristics is no longer ideal. A bad choice of value j may lessen the impact of the overall outcome. To illustrate how the LSGAFS process functions, it is express in the following manner.

The subset size of feature $j\ (< ft)$ is decided using a probabilistic algorithm, which is given by

$$Lin_j = \frac{ft - j}{\sum_{p=1}^{t}(ft - p)} \tag{1}$$

By varying Lin in Eq. (1), it obtains greater improvement while limiting the value of j, which is controlled using constraint $4 \leq j \leq w$. Hence, $j = 4,5,6, \dots, , w-1, w$, where $w = \varepsilon . ft$ and $t = ft - j$. A fundamental observation to have in mind is that the value of j in the constraint is assigned here, and begun, at 4. Crucially, solutions created as a result of imperfect search functionality are likely to be in a state of incorrect. This also describes the notion ε is a user-configurable parameter. The value of the metric relies on the number of rows in a particular dataset. The search space for important characteristics widens, and as a result, a greater feature subset is created. It is known that the purpose of LSGAFS is to provide a salient features subset that is smaller, this research work suggests the subset's length falls anywhere between 4 and 16, based on the dataset. Consequently, ε is a set where it lies in $[0.45, 0.22]$. Thus, all the potential values of Lin_j are added together, and then normalized so that the sum of all possible outcomes is equivalent to 1.

Second, the methodology employed by LSGAFS uses all of the values of Lin_j in order to attain the subset size that is eventually decided via the process illustrated in Algorithm 1. As it turns out, this approach offers a much like the usual roulette wheel selection process.

Algorithm 1: Pseudocode of random selection

```
$select\_randomly()$
{
    $randomly\ generate\ value\ v[0,1];$
    $total = 0;$
    $for(j = 4\ to\ s)$
    {
        $total = total + Lin_j;$
        $if(v \leq total)$
            $break;$
    }
    $return\ j;$
}
```

Local Search Operation

Special search operations that are performed locally refine the unique and general properties of the dataset. This search operation will integrate these refined features into a model. It means that a classifier can obtain all relevant information about the dataset, which ultimately enables classifiers to better demonstrate their capacity to generalize. This is a very important aspect of this algorithm. In order to do local search optimization in LSGAFS, feature grouping is necessary to be carried out.

The primary purpose of a grouping of features in LSGAFS is to establish connections between characteristics in order for the algorithm to distribute the features that are distinctive and informative. LSGAFS measures correlation between distinct aspects of a training set using a familiar Pearson correlation coefficient strategy. CC_{pq} is a measure of how closely two features (p and q) are related to each other and it is expressed in Eq. (2)

$$CC_{pq} = \frac{\sum_b (y_p - \bar{y}_p)(y_q - \bar{y}_q)}{\sqrt{\sum_b (y_p - \bar{y}_p)^2}\sqrt{\sum_b (y_q - \bar{y}_q)^2}} \tag{2}$$

where y_p and y_q indicates the value significance of p and q features, $\bar{y}_p$ and $\bar{y}_q$ indicates the mean of y_p and y_q that are computed from b samples.

To determine the correlation between any two features, LSGAFS uses the correlation coefficient Cor_p to calculate the correlation between a feature and it is expressed in Eq. (3).

$$Cor_p = \frac{\sum_{q=1}^{ft} \|CC_{pq}\|}{ft - 1} \quad if\ p \neq q \tag{3}$$

Then, the characteristics are ordered based on their correlation values, with ascending order for subsequent groups.

Fitness Calculation

An efficient GA fitness computation for single characters has importance in FS as well. This method of doing fitness calculation tasks usually employs a filter or wrapper technique. We were encouraged to solve the problem by utilizing hybrid techniques in this work since they both provide unique benefits and advantages derived from one another. In this research, the percentage of acquired classification accuracy ($CAcc$) is calculated by summing the NN's classification accuracy with the distinct value of p that was specified as follows:

$$y(s) = CAcc(s) + a(s) \tag{4}$$

Here, $a(s)$ is used to describe the distinct value of the population of strings s that contains a single selection of features from the available list of Cor_p. This estimation is accomplished by the inverse of the sum of the correlation information of those individual selected features s that make up the population of strings S.

In general, Eq. (3) is used to calculate the correlations among the different dataset features. The goal of incorporating the distinct value of each string is to harness the value of each string in order to aid the overall performance of the system. As discussed previously, the LSO in LSGAFS serves the aim of distributing different features to the population. Since the LSO and fitness function works together to achieve the same target, this enables the system to work optimally. The proportion of the training system's classifier's successful outcomes defined the successful subset of salient features, and it is measured by testing CA.

In LSGAFS, the training of NN plays a significant role. Training involves setting the hidden layer size by the construction of a fixed input layer. Conversely, layer growth occurs by simultaneous determination of both input and hidden layers.

Termination of genetic process

The likelihood of strings getting the same feature vector (i.e. converging to the optimal or near-optimal subset of features) increases as the evolutionary process gets progresses and when 0- and 1-bit strings reach a state where they only differ in their number of features. The system doesn't entirely eliminate genetic drift, but rather requires it to pause after some generations so that the population can go on to its final solution.

To fulfill the stated purpose, certain specifications have been included in the genetic algorithm for LSGAFS: (1) Fitness value is flattened for best strings, as such: it was not increasing appreciably with few generations. (2) After r generations, LSGAFS calculates fitness value for the best string, termed strip. If the fitness value has been decreasing over R strips, it is essentially terminating the genetic process. While fitness values reduce R times during the course of one or more cycles, the number of reductions can be interpreted to signify that genetic optimization has ended at local maxima. The requirement to be met is defined as

$$0 > y(r' + r) - y(r'), \qquad r = 1,2,3, \dots, R-1, R \tag{5}$$

where $y(r)$ indicates the best string's fitness value in a specific generation and after threshold time duration value of fitness gets decayed where the best string's fitness value gets progressively decayed. R represent a user-specified positive integer. Stopping criteria 1 and 2 both include an absolute need that the genetic process is terminated when at least one of them is met.

Computational Complexity

The proper computational analysis goes a long way in showing the true cost of a given method. Each string in LSGAFS indicates a subset of the candidate features. Amongst the primary purposes of LSO in HGAFS is to improve the overall quality of the features that are generated. The newly generated features provide features with readjusted special or general characteristics. When j features are used, then the newly created features become accessible. It takes r_j computational time to meet all the requirements therefore the LSO must be applied to each of them.

$$r_j = \begin{cases} (S - |Y_m|) & |Y_m| < S \\ (|Y_m| - S) & |Y_m| > S \\ (E - |Y_g|) & |Y_g| < E \\ (|Y_g| - E) & |Y_g| > E \end{cases} \tag{6}$$

where $|Y_m|$ and $|Y_g|$ are equivalent to the features that are present in the dataset. ft_{gt}, S and E are the sum of j. Hence, Eq. (6) gets modified to Eq. (7).

$$r_j = \begin{cases} (ft_{gt} - k), & ft_{gt} > j \\ (j - ft_{gt}), & j < ft_{gt} \end{cases} \tag{7}$$

The complexity of time consumption of LSO in LSGAFS is expressed as Eq. (8)

$$r_g = \begin{cases} (ft_{gt} - j)(2h - 1) & j < ft_{gt} \\ 2(2h - 1) & ft_{gt} = j \\ (j - ft_{gt})(2h - 1) & j > ft_{gt} \end{cases} \tag{8}$$

In Eq. (8), h initially eliminates low-level features from the subset of the previously chosen features (i.e $h \in 1,2,3, \dots$) and $h - 1$ refers to the number of times a new set of most important unselected features is introduced. In other words, the removal of one characteristic yields the addition of another. It conducts both adding and removing operations h times, respectively, if the current subset has the same number of features as the needed number. Finally, concerning the computational cost, it may be seen in this technique which performs the significant calculation towards trains and tests classifiers.

DATASET

To evaluate the performance of the proposed classifier, this research work uses 3 different datasets used in RBFNN [11] namely coronary heart disease dataset, diabetes dataset and bronchial tuberculosis dataset. The count of instances present in the datasets is provided in Table 1.

Table 1. Count of Instances

Dataset Name	Instances Count	Features
Coronary Heart Disease (CHD)	3231	56
Diabetes (DB)	9628	17
Bronchial Tuberculosis (BT)	5628	

PERFORMANCE METRICS

This research work makes use of accuracy and f-measure metrics to measure the performance of proposed classifier against RBFNN and WMEKL.

Accuracy:

It is the percentage of cases that are accurately classified.

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} \qquad (11)$$

F-Measure:

It is the percentage of weighted average or harmonic mean of recall and precision.

$$F - Measure = \frac{2T_P}{2T_P + F_P + F_N} \qquad (12)$$

Variables used in Eq. (11) and Eq. (12) are defined as follows:

- **True Positive (T_P):** It represents the result of accurately predicted positive class.
- **False Positive (F_P):** It represents the result of inaccurately predicted positive class.
- **True Negative (T_N):** It represents the result of accurately predicted negative class.
- **False Negative (F_N):** It represents the result of inaccurately predicted negative class.

RESULTS AND DISCUSSION

Performance Metric Variable Analysis

Figure 1 to Figure 4 discusses the results obtained for the proposed classifier LSGNFS and existing classifiers SMLM [20] and RBFNN [11] for the performance metric variables (i.e., T_P, F_P, T_N, F_N). From the figures it possible to make a better understanding that the proposed classifier outperforms the other two classifiers. Proposed classifier attains better results due to the optimization technique that it follows during the classification where SMLM and RBFNN simply performs classification alone.

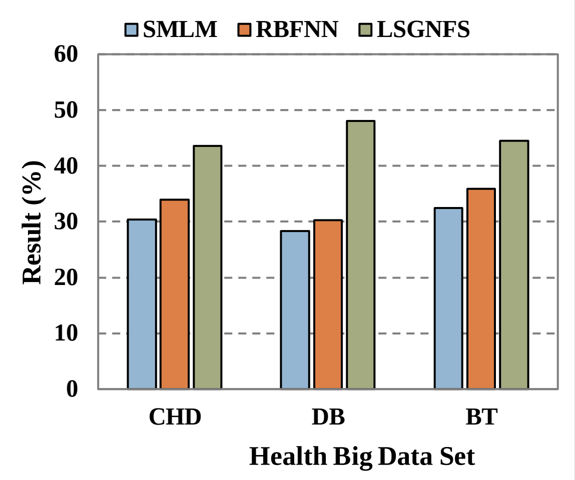

Figure 1. True Positive Analysis

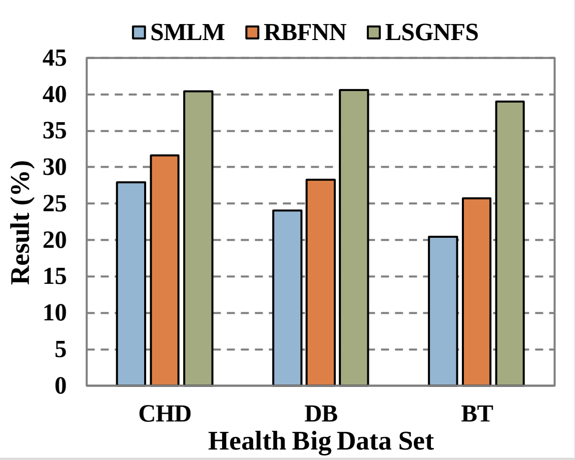

Figure 2. True Negative Analysis

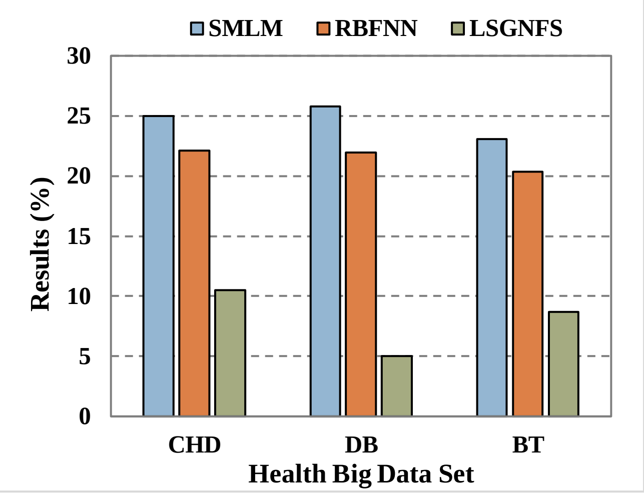

Figure 3. False Positive Analysis

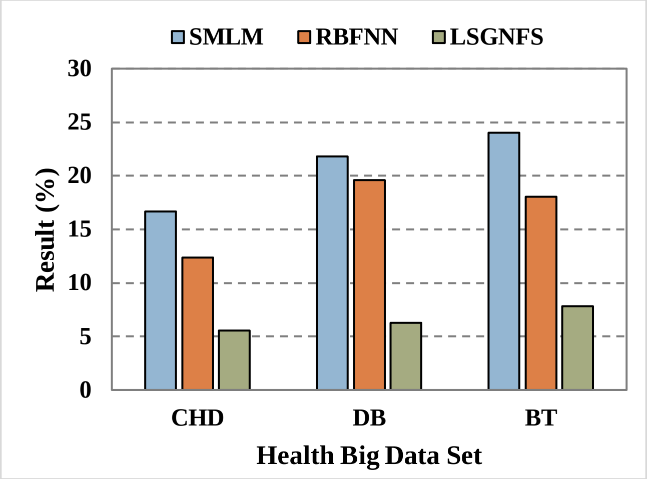

Figure 4. False Negative Analysis

Accuracy Analysis

In Figure 5, Health Big Datasets namely CAD, DB and BT are plotted in the x-axis where the y-axis is marked with accuracy in percentage. From Figure 5, it is clear to make a better understanding that the proposed classifier has attained better results with all three different health big datasets. Optimization and local search in LSGAFS assist in achieving better results than existing classifiers SMLM [20] and RBFNN [11]. Also, it is a notable thing that LSGAFS has performed well with the diabetes disease dataset with the accuracy of 88.70%. Averagely LSGSFS has achieved 85.39% of classification accuracy where SMLM and RBFNN have achieved 54.53% and 61.87% respectively. The numerical values of Figure 5 is provided in Table 2.

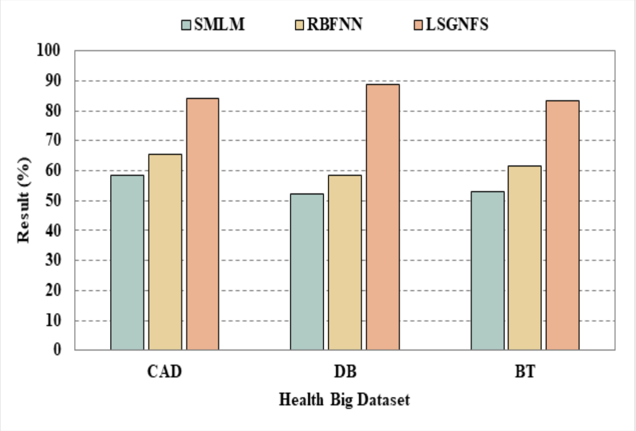

Figure 5. LSGSFS Vs Accuracy

Table 2. Accuracy

Classifiers / Datasets	SMLM	RBFNN	LSGSFS
CAD	58.31%	65.52%	83.97%
DB	52.37%	58.48%	88.70%
BT	52.91%	61.62%	83.49%

F-Measure Analysis

In Figure 6, Health Big Data Sets namely CAD, DB and BT are plotted in the x-axis where the y-axis is marked with f-measure in percentage. From Figure 6, it is indisputable that LSGAFS achieved better f-measure than existing classifiers SMLM [20] and RBFNN [11]. Enhanced genetic algorithm for performing the better local search assist proposed model in achieving better results than existing classifiers. Averagely LSGSFS has achieved 86.10% of f-measure where SMLM and RBFNN have achieved 57.20% and 63.58% respectively. The numerical values of Figure 6 are provided in Table 2.

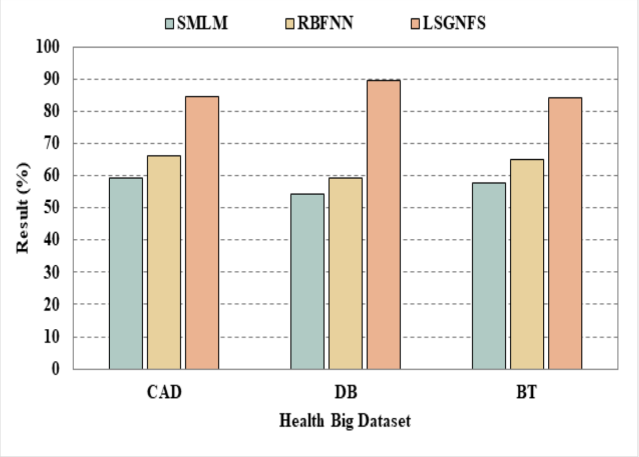

Figure 6. LSGSFS Vs F-Measure

Table 3. F-Measure

Classifiers / Datasets	SMLM	RBFNN	LSGSFS
CAD	59.34%	66.30%	84.46%
DB	54.33%	59.29%	89.49%
BT	57.94%	65.15%	84.35%

CONCLUSIONS

Feature selection is a strategy for reducing the number of features towards increasing accuracy of classification. This has been demonstrated to be useful and efficient for managing high-dimensional big data. It focuses on group of features that will be useful for model creation. In this paper, Local Search based Genetic Algorithm for Feature Selection (LSGNFS) is introduced for health big data classification for the prediction of deadly diseases namely: (i) diabetes (ii) coronary heart disease (iii) bronchial tuberculosis. LSGNFS focuses on the most significant aspects, leaving unessential and redundant ones behind. LSGNFS retains the original feature definitions with interpretability. Optimization for selecting the features are done via local search using genetic algorithm. LSGNFS is evaluated using three benchmark datasets with the metrics accuracy and f-measure. With the coronary heart disease, dataset LSGNFS has attained 83.97% of accuracy where SMLM and RBFNN have attained 58.31% and 65.52% respectively. With the diabetes disease dataset, LSGNFS has attained 88.70% accuracy where SMLM and RBFNN have attained 52.37% and 58.48% respectively. With the bronchial tuberculosis disease dataset, LSGNFS attained 83.49% classification accuracy where SMLM and RBFNN have attained 52.91% and 61.62% respectively. Averagely, LSGNFS has attained 85.39% of classification accuracy where SMLM and RBFNN have attained 54.53% and 61.87% respectively. The future dimension of this research work can be focused on increasing the classification accuracy and classifying more diseases.

REFERENCES

[1] Abideen, Z. Ul, M. Ghafoor, K. Munir, M. Saqib, A. Ullah, T. Zia, S. A. Tariq, G. Ahmed, and A. Zahra. 2020. “Uncertainty Assisted Robust Tuberculosis Identification With Bayesian Convolutional Neural Networks.” IEEE Access 8:22812–25. doi: 10.1109/ACCESS.2020.2970023.

[2] Ang, Kenneth Li Minn, Feng Lu Ge, and Kah Phooi Seng. 2020. “Big Educational Data & Analytics: Survey, Architecture and Challenges.” IEEE Access 8:116392–414. doi: 10.1109/ACCESS.2020.2994561.

[3] Balabaeva, Ksenia, and Sergey Kovalchuk. 2019. “Comparison of Temporal and Non-Temporal Features Effect on Machine Learning Models Quality and Interpretability for Chronic Heart Failure Patients.” Procedia Computer Science 156:87–96. doi: https://doi.org/10.1016/j.procs.2019.08.183.

[4] Bridget, Odu Nkiruka, Rajesh Prasad, Clement Onime, and Adamu Abubakar Ali. 2021. “Drug Resistant Tuberculosis Classification Using Logistic Regression.” International Journal of Information Technology (Singapore) 13(2):741–49. doi: 10.1007/s41870-020-00592-9.

[5] Cai, Zhuoran, Jidong Wang, and Minghuan Ma. 2021. “The Performance Evaluation of Big Data-Driven Modulation Classification in Complex Environment.” IEEE Access 9:26313–22. doi: 10.1109/ACCESS.2021.3054756.

[6] Cavallaro, Gabriele, Morris Riedel, Matthias Richerzhagen, Jón Atli Benediktsson, and Antonio Plaza. 2015. “On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8(10):4634–46. doi: 10.1109/JSTARS.2015.2458855.

[7] Fong, Simon, Raymond Wong, and Athanasios V. Vasilakos. 2016. “Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data.” IEEE Transactions on Services Computing 9(1):33–45. doi: 10.1109/TSC.2015.2439695.

[8] Ghiasi, Mohammad M., Sohrab Zendehboudi, and Ali Asghar Mohsenipour. 2020. “Decision Tree-Based Diagnosis of Coronary Artery Disease: CART Model.” Computer Methods and Programs in Biomedicine 192:105400. doi: 10.1016/j.cmpb.2020.105400.

[9] He, Weina, Yafei Wang, and Dongliang Xia. 2021. “Fuzzy Integration Algorithm of Big Data in Peer-to-Peer Communication Network Based on Deep Learning.” Wireless Personal Communications 1–17. doi: 10.1007/s11277-021-08581-2.

[10] Hossain, M. Anwar, Rahatara Ferdousi, and Mohammed F. Alhamid. 2020. “Knowledge-Driven Machine Learning Based Framework for Early-Stage Disease Risk Prediction in Edge Environment.” Journal of Parallel and Distributed Computing 146:25–34. doi: 10.1016/j.jpdc.2020.07.003.

[11] Jiang, Congshi, and Yihong Li. 2019. “Health Big Data Classification Using Improved Radial Basis Function Neural Network and Nearest Neighbor Propagation Algorithm.” IEEE Access 7:176782–89. doi: 10.1109/ACCESS.2019.2956751.

[12] Kannadasan, K., Damodar Reddy Edla, and Venkatanareshbabu Kuppili. 2019. “Type 2 Diabetes Data Classification Using Stacked Autoencoders in Deep Neural Networks.” Clinical Epidemiology and Global Health 7(4):530–35. doi: https://doi.org/10.1016/j.cegh.2018.12.004.

[13] Kumar, Dinesh, and Mihir Narayan Mohanty. 2019. “A Survey: Classification of Big Data.” Pp. 299–306 in Advances in Intelligent Systems and Computing. Vol. 768. Springer Verlag.

[14] Kumari, Saloni, Deepika Kumar, and Mamta Mittal. 2021. “An Ensemble Approach for Classification and Prediction of Diabetes Mellitus Using Soft Voting Classifier.” International Journal of Cognitive Computing in Engineering 2:40–46. doi: https://doi.org/10.1016/j.ijcce.2021.01.001.

[15] L’Heureux, Alexandra, Katarina Grolinger, Hany F. Elyamany, and Miriam A. M. Capretz. 2017. “Machine Learning with Big Data: Challenges and Approaches.” IEEE Access 5:7776–97. doi: 10.1109/ACCESS.2017.2696365.

[16] Maillo, Jesus, Isaac Triguero, and Francisco Herrera. 2020. “Redundancy and Complexity Metrics for Big Data Classification: Towards Smart Data.” IEEE Access 8:87918–28. doi: 10.1109/ACCESS.2020.2991800.

[17] Manogaran, Gunasekaran, P. Mohamed Shakeel, S. Baskar, Ching-H. Hsien Hsu, Seifedine Nimer Kadry, Revathi Sundarasekar, Priyan Malarvizhi Kumar, and Bala Anand Muthu. 2021. “FDM: Fuzzy-Optimized Data Management Technique for Improving Big Data Analytics.” IEEE Transactions on Fuzzy Systems 29(1):177–85. doi: 10.1109/TFUZZ.2020.3016346.

[18] Melendez, Jaime, Bram Van Ginneken, Pragnya Maduskar, Rick H. H. M. Philipsen, Helen Ayles, and Clara I. Sánchez. 2016. “On Combining Multiple-Instance Learning and Active Learning for Computer-Aided Detection of Tuberculosis.” IEEE Transactions on Medical Imaging 35(4):1013–24. doi: 10.1109/TMI.2015.2505672.

[19] Raja, J. Beschi, and S. Chenthur Pandian. 2020. “PSO-FCM Based Data Mining Model to Predict Diabetic Disease.” Computer Methods and Programs in Biomedicine 196:105659. doi: https://doi.org/10.1016/j.cmpb.2020.105659.

[20] Reyes, Oscar, Eduardo Pérez, Raúl M. Luque, Justo Castaño, and Sebastián Ventura. 2020. “A Supervised Machine Learning-Based Methodology for Analyzing Dysregulation in Splicing Machinery: An Application in Cancer Diagnosis.” Artificial Intelligence in Medicine 108:101950. doi: https://doi.org/10.1016/j.artmed.2020.101950.

[21] Romero, M. Pilar, Yu-Mei Chang, Lucy A. Brunton, Jessica Parry, Alison Prosser, Paul Upton, Eleanor Rees, Oliver Tearne, Mark Arnold, Kim Stevens, and Julian A. Drewe. 2020. “Decision Tree Machine Learning Applied to Bovine Tuberculosis Risk Factors to Aid Disease Control Decision Making.” Preventive Veterinary Medicine 175:104860. doi: https://doi.org/10.1016/j.prevetmed.2019.104860.

[22] Shokrzade, Amin, Fardin Akhlaghian Tab, and Mohsen Ramezani. 2020. "ELM-NET, a Closer to Practice Approach for Classifying the Big Data Using Multiple Independent ELMs." Cluster Computing 23(2):735–57. doi: 10.1007/s10586-019-02957-7.

[23] Thakkar, Harshil, Vaishnavi Shah, Hiteshri Yagnik, and Manan Shah. 2021. "Comparative Anatomization of Data Mining and Fuzzy Logic Techniques Used in Diabetes Prognosis." Clinical EHealth 4:12–23. doi: https://doi.org/10.1016/j.ceh.2020.11.001.

[24] Velusamy, Durgadevi, and Karthikeyan Ramasamy. 2021. "Ensemble of Heterogeneous Classifiers for Diagnosis and Prediction of Coronary Artery Disease with Reduced Feature Subset." Computer Methods and Programs in Biomedicine 198:105770. doi: https://doi.org/10.1016/j.cmpb.2020.105770.

[25] Xu, Chenhan, Kun Wang, Yanfei Sun, Song Guo, and Albert Y. Zomaya. 2020. "Redundancy Avoidance for Big Data in Data Centers: A Conventional Neural Network Approach." IEEE Transactions on Network Science and Engineering 7(1):104–14. doi: 10.1109/TNSE.2018.2843326.

[26] Yang, Hui, Yamei Luo, Xiaolei Ren, Ming Wu, Xiaolin He, Bowen Peng, Kejun Deng, Dan Yan, Hua Tang, and Hao Lin. 2021. "Risk Prediction of Diabetes: Big Data Mining with Fusion of Multifarious Physical Examination Indicators." Information Fusion 75:140–49. doi: https://doi.org/10.1016/j.inffus.2021.02.015.

[27] Yang, Weiyi, Yujuan Si, Di Wang, Gong Zhang, Xin Liu, and Liangliang Li. 2020. "Automated Intra-Patient and Inter-Patient Coronary Artery Disease and Congestive Heart Failure Detection Using EFAP-Net." Knowledge-Based Systems 201–202:106083. doi: https://doi.org/10.1016/j.knosys.2020.106083.

[28] Zhang, Weiwei, Guang Yang, Nan Zhang, Lei Xu, Xiaoqing Wang, Yanping Zhang, Heye Zhang, Javier Del Ser, and Victor Hugo C. de Albuquerque. 2021. "Multi-Task Learning with Multi-View Weighted Fusion Attention for Artery-Specific Calcification Analysis." Information Fusion 71:64–76. doi: https://doi.org/10.1016/j.inffus.2021.01.009.

[29] Zhu, Minjun, and Qinghua Chen. 2020. "Big Data Image Classification Based on Distributed Deep Representation Learning Model." IEEE Access 8:133890–904. doi: 10.1109/ACCESS.2020.3011127.