Egyptian Informatics Journal 19 (2018) 179–189

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Review

Feature selection and classification systems for chronic disease prediction: A review

Divya Jain *, Vijendra Singh

The NorthCap University, Gurugram 122017, India

ARTICLE INFO

Article history:
Received 1 February 2017
Revised 30 December 2017
Accepted 20 March 2018
Available online 5 April 2018

Keywords:
Chronic disease
Feature selection
Traditional systems
Disease diagnosis
Parallel classification systems
Adaptive classification systems

ABSTRACT

Chronic Disease Prediction plays a pivotal role in healthcare informatics. It is crucial to diagnose the disease at an early stage. This paper presents a survey on the utilization of feature selection and classification techniques for the diagnosis and prediction of chronic diseases. Adequate selection of features plays a significant role for enhancing accuracy of classification systems. Dimensionality reduction helps in improving overall performance of machine learning algorithm. The application of classification algorithms on disease datasets yields promising results by developing adaptive, automated and intelligent diagnostic systems for chronic diseases. Parallel classification systems can be used to expedite the process and to enhance the computational efficiency of results. This work presents a comprehensive overview of various feature selection methods and their inherent pros and cons. We then analyze adaptive classification systems and parallel classification systems for chronic disease prediction.

Contents

1. Introduction . . . 179
2. Feature selection for chronic disease prediction . . . 180
   2.1. Feature selection approaches . . . 180
3. Traditional classification systems and adaptive classification systems for chronic disease prediction . . . 182
4. Parallel classification systems for chronic disease prediction . . . 185
5. Comparisons of traditional classification systems, adaptive classification systems and parallel classification systems . . . 186
6. Performance metrics used in medical diagnosis systems to measure the performance of the classification methods . . . 187
7. Conclusion and future perspectives . . . 187
   References . . . 188

1. Introduction

Diagnosis of chronic diseases is very essential in the medical field as these diseases persist for long time. The leading chronic diseases include diabetes, stokes, cardiovascular disease, arthritis, cancer, hepatitis C. Early detection of chronic disease helps in taking preventive actions and effective treatment at an initial stage has always been found to be helpful for patients.

Currently, maintenance of clinical databases has become a crucial task in medical field. The patient data consisting of various features and diagnostics related to disease should be entered with utmost care to provide quality services. As the data stored in medical databases may contain missing values and redundant data, mining of the medical data becomes cumbersome. As it can affect the results of mining, it is essential to have good data preparation and data reduction before applying data mining algorithms. Prediction of disease becomes quick and easier if data is precise and consistent and free from noise.

* Corresponding author.
E-mail addresses: divyajain1890@gmail.com (D. Jain), vsingh.fet@gmail.com (V. Singh).
Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.eij.2018.03.002
1110-8665/

Feature Selection is an efficient data preprocessing technique in data mining for reducing dimensionality of data [1–3]. In medical diagnosis, it is very important to identify most significant risk factors related to disease. Relevant feature identification helps in the removal of unnecessary, redundant attributes from the disease dataset which, in turn, gives quick and better results.

Classification and prediction [4,5] is a data mining technique which first uses training data to develop a model and then the resulted model is applied on testing data to get results of prediction. Various classification algorithms have been applied on disease datasets for the diagnosis of chronic disease and the results have been found to be very promising. There is an utmost need to develop a novel classification technique which can expedite and simplify the process of diagnosis of chronic disease.

In this age of data explosion, voluminous amount of medical data is generated and updated daily. Healthcare data includes Electronic Health Records (EHR) which comprises of clinical reports of patients, diagnostic test reports, doctor's prescription, information related to pharmacy, information related to patient's health insurance, posts on social media such as blogs, tweets [7,41]. There is an utmost need of an efficient parallel data processing technique which is capable to manage and analyze the huge volumes of healthcare data.

Chronic Disease Diagnosis (CDD) systems [8] can be used as valuable tools for proper control and management of the chronic disease. It monitors the health of patients and assist physicians and medical professions to provide 24/7 healthcare services.

This paper is organized as follows. Firstly, a short description of feature selection for chronic disease prediction is presented. Secondly, various feature selection methods and related work on various feature selection approaches is presented. Along with that, a tabular study consisting of various feature selection algorithms, characteristics, merits, demerits and their assessment criteria is also given. Thirdly, a survey on traditional classification systems and adaptive classification systems for chronic disease prediction is shown in Section 3. Section 4 gives a brief overview of the related work based on parallel classification systems for the prediction of chronic diseases. Lastly, a summarized comparison of different classification systems results is shown in Section 7 followed by a conclusion.

2. Feature selection for chronic disease prediction

Feature selection, also known as Variable Selection [9], is an extensively used data preprocessing technique in data mining which is basically used for reduction of data by eliminating insignificant and superfluous attributes from any dataset [3]. Moreover, this technique enhances the comprehensibility of data, facilitates better visualization of data, reduces training time of learning algorithm and improves the performance of prediction.

There exist numerous applications of relevant feature identification techniques in healthcare sector. Filter methods, wrapper methods, ensemble methods and embedded methods are some of the popularly used techniques used for variable selection. In recent years, most of the authors are focusing on hybrid approaches used for feature selection.

Before any model is applied to the data, it is always better to remove noisy and inconsistent data to get more accurate results in less time. Reducing the dimensionality of a dataset is of paramount importance in real-world applications. Moreover, if most important features are selected, the complexity decreases exponentially.

In recent years, various feature identification approaches have been applied on healthcare datasets to get more valuable information. The utilization of feature selection methods is done on clinical databases for the prediction of numerous chronic diseases like diabetes, heart disease, strokes, hypertension, thalassemia etc. Various learning algorithms work efficiently and give more accurate results if the data contains more significant and non-redundant attributes. As the medical datasets contains large number of redundant & irrelevant features, an efficient feature selection technique is needed to extract interesting features relevant to the disease.

A highly accurate diagnostic system for the detection of knee joint disorders using VAG signals was proposed by the authors in Ref. [10]. The methodology was developed using a novel feature selection and classification technique. For the identification of most significant and stable features, apriori algorithm and genetic algorithm were used. To evaluate their performance, random forest and LS-SVM classifiers were used. Additionally, the concept of wavelet decomposition was used to classify normal VAG signals from abnormal ones. A comparison of the results based on evaluation metrics revealed that the performance of LS-SVM using the apriori algorithm was the best with an accuracy of 94.31%. The proposed approach could be of great help for early diagnosis of knee-joint disorders so that treatment can be provided to patients at an early stage.

A basic taxonomy of feature selection and various gene selection methods were reviewed by authors in Ref. [11]. Authors classified these approaches under three divisions – supervised, semi-supervised and unsupervised feature selection. Various challenges and obstacles in extracting knowledge from gene expression data were also addressed. Some of the basic issues raised were (1) reducing dimensionality of data with hundreds of thousands of features (2) how to handle mislabeled, imprecise data (3) how to deal with extremely imbalanced data (4) determining the gene relevancy/redundancy and extracting relevant biological information from the gene expressions. It was revealed through comparative study on gene selection that the classification accuracy of semi-supervised and unsupervised approaches was as promising as supervised feature selection.

A novel feature selection approach using SVM ranking with backward search method was presented by authors in Ref. [12] to final the ideal subset of features on type II diabetes dataset. With the proposed approach, the predictive accuracy of Naïve Bayes classifier got significantly increased. The methodology used was very simple yet effective which would definitely help the physicians and medical professions for the diagnosis of Type 2 diabetes.

ModifiedFAST is a fast and efficient feature identification algorithm which was proposed by authors in Ref. [13]. An ideal value of threshold with the inclusion of symmetric uncertainty (SU) was appropriately found. The minimum spanning tree was constructed after applying symmetric uncertainty (SU). The comparison of the results of the proposed algorithm was made with other algorithms like FAST, FCBF, Relief and CFS based on classification accuracy and percentage of features selected and it was demonstrated that ModifiedFAST was the best algorithm among all.

2.1. Feature selection approaches

Traditional feature selection approaches for machine learning are broadly classified into three categories:

(a) Filter method
(b) Wrapper method
(c) Embedded method

Currently, hybrid methods consisting of combination of these approaches is also used by many authors the results of which are also promising.

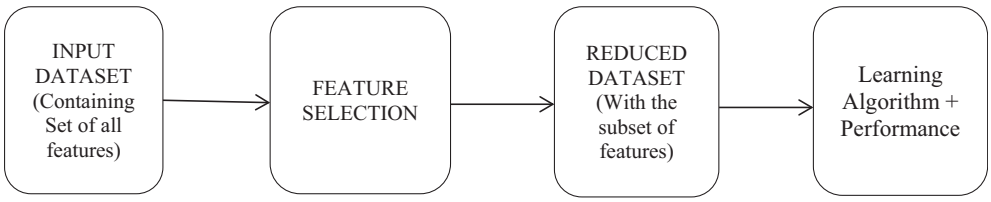

Fig. 1. Feature selection process.

Fig. 1 shows the feature selection process that can be applied on input dataset to get reduced dataset which is then passed to the learning algorithm.

A. **Filter Method** – It is one of the oldest methods of feature selection. In variable selection using filter approach, filtering of features is done before the implementation of any learning algorithm. It ranks features based on a certain evaluation criteria [2,3]. As it is not dependent on the classifier applied, it tends to give varied performance on prediction. These methods give fast and efficient results on execution. Consequently, they are preferred for voluminous databases over wrapper methods. The limitation of these approaches is that they ignore interaction among classifiers and dependency of one feature over other and may fail to select the most "useful" features [11].

MIFS (Mutual Information based Feature Selection), proposed by Battiti [14], is a feature selection approach based on the concept of mutual information that does "greedy" selection of features. This method does feature identification in such a way that it extracts maximum mutual information. However, due to the presence of large number of errors in its implementation, MIFS is less preferred.

MIFS-U [15] is a modified feature selection approach over MIFS method which was developed to make considerable use of the mutual information. The performance of MIFS-U is like an "ideal greedy selection algorithm" when there exist uniform distribution of information. This approach can be efficiently used to solve large problems. The performance of MIFS-U degrades if information distributions of features deviate from the uniform distribution [15].

MICC (Mutual Information-based Constructive Criterion) [16] is a greedy filter feature selection approach based on the concept of mutual information that was developed to overcome the limitations of its precedents. The most significant aspect is that it considers relevance as well as non-redundancy of the features to the output classes. The major advantage over its precedent algorithms MIFS [14] and MIFS-U [15] is that it selects features without using any parameters such as B (Beta). Consequently the results were more promising with MICC as compared to its precedents.

Correlation based feature selection approach was applied for the diagnosis of Coronary Artery Disease (CAD) though a hybridized model in Ref. [17]. Most significant risk factors related to CAD disease were identified using correlation feature selection approach along with particle swam optimization method followed by a clustering algorithm. In order to construct diagnostic models for CAD disease, C4.5 algorithm, multi-layer perceptron (MLP), multinomial logistic regression (MLR) and fuzzy unordered rule induction algorithm (FURIA) were used. The CAD model was validated with 10-fold cross validation method. The predictive accuracy of MLR algorithm was the highest while it was lowest with MLP algorithm on both clinical data and Cleveland heart disease data. The results of the proposed methodology were very promising which significantly improved the accuracy of classifier. Consequently, this method can be used as a valuable tool for clinical decisions related to CAD disease diagnostics.

Yu and Liu [18] designed a correlation based filter approach to deal with the issues of high dimensionality. Authors introduced the concept of 'predominant correlation' for the identification and removal of irrelevant and redundant features and implemented fast correlation-based feature selection (FCBF) algorithm. The results of the experiments revealed that the proposed algorithm executed with less quadratic time complexity and was very efficient to deal with high-dimensional data.

B. **Wrapper Method** – Wrapper methods does selection of features by giving due consideration to the learning algorithm to be used. The major advantage over filter methods is that it finds the most "useful" features and does optimal selection of features for the learning algorithm [19]. Moreover, it considers dependencies among features and gives more accurate results in comparison to filter methods [11]. However, it has a problem that if another learning algorithm has to be utilized, this method needs to be re-executed. Moreover, this method is very complex and more prone to over-fitting on small training datasets.

A detailed analysis and comparison of wrapper feature selection approach and relief algorithm (a filter feature selection approach) was done by the authors in Ref. [20]. Authors explored the strengths and limitations of the wrapper approach for optimal feature subset selection. The experiments were conducted with both real and artificial datasets along with two induction algorithms namely, Naïve Bayes classifier and decision trees. It was revealed from results that the error rate was significantly reduced when wrapper approach was used with Naïve Bayes classifier.

Maldonado et al. [21] applied wrapper approach using sequential backward elimination. The method used support vector machines and kernel functions for implementation. The proposed methodology presented an effective validation error measure for the elimination of features. Moreover, the key aspect was that it could be used with any of the kernel functions. The significant aspect of the algorithm was that each run of algorithm selected different set of features. The comparison of the results revealed that the proposed wrapper algorithm exhibited better performance than existing filter and wrapper methods. However, due to backward elimination of features, it was expensive to use this approach if there were large number of input features.

C. **Embedded Method** – In embedded feature selection approach, search is usually guided by the learning process. This approach, also known as nested subset method [22], usually measures the "usefulness" of feature subsets and performs feature selection as a part of the training process [9]. They usually work according to specific learning algorithm which helps in optimizing the performance of a learning algorithm. This method make better usage of available data and provides faster solution as they do not require splitting of training data into training set and validation set. They are computationally inexpensive and less prone to over-fitting than wrapper techniques. Moreover, the computational complexity of embedded methods is better than wrapper methods [23]. The major limitation with these methods is that it takes decisions depending on the classifier. Hence, selection of features can be affected by the hypothesis that the classifier makes which might not work with some other classifier [24].

An embedded method based on backward feature selection was proposed by Maldonado et al. [25]. The purpose was to select most significant features from imbalanced data for applying binary classification using support vector machines. The proposed method

was very flexible and facilitated to be used with several objective functions. With the use of embedded feature selection process, the proposed strategy achieved very good results on highly imbalanced data sets.

ESFS (Embedded Sequential Forward Selection), proposed by Xiao et al. [26], is a novel embedded selection method which was simply based on incrementally adding the most relevant features. This method was concerned with the use of mass functions introduced from the concept of evidence theory that facilitated the merging of information provided by features. The proposed method significantly improved the classification accuracy and was able to select the most discriminative features when applied to emotional classification (speech and music samples). With the experimental results, the proposed embedded method (ESFS) was found to give lower computational cost than the wrapper method (Sequential Forward Selection).

D. **Hybrid Method** – In recent times, it is one of the widely used approaches used by the researchers for applying feature selection technique. The method aggregates one or more approaches together to take advantage of the merits of different approaches to get optimal results. These methods usually achieve higher accuracy compared to wrapper methods and high computational efficiency compared to filter methods.

A hybrid feature selection approach based on improved particle swarm optimization algorithm was presented by the authors in Ref. [27]. Researchers applied filter and wrapper methods together for image steganalysis. It was found from the experimental results that the proposed hybridized approach significantly reduced the number of features and enhanced the classification accuracy as compared to other previous feature selection algorithms. Furthermore, computational cost and time also got reduced with the proposed methodology.

BBHFS (Boosting Based Hybrid Feature Selection), proposed by Das [28], is a fast and scalable hybrid algorithm which included the concept of boosting and advantages of both filter and wrapper methods. Authors presented a more informed filter method by incorporating forward selection algorithm and some of the benefits of wrapper method such as natural stopping criterion. This algorithm yielded fast and better results than wrapper methods when applied on DNA dataset using Naive Bayes classifier and on the Chess dataset using ID3 algorithm. The approach significantly enhanced the performance of these classifiers. The proposed hybrid approach was found to be very scalable on datasets consisting of large number of features. In future, the proposed algorithm can be applied on multi-class datasets to get more interesting results. Also, usage of k-level decision trees instead of decision stumps can also yield promising results in the further extension to this work.

A hybrid genetic algorithm based feature selection method consisting of both filter and wrapper methods was proposed by the authors in Ref. [29]. Selection of optimal subset of features was done with the incorporation of the benefits of both filter and wrapper approaches in two optimization stages. Wrapper approach for global search and filter approach for local search were applied to get best subset of features in outer and inner stages respectively. When both optimizations were applied together, the results achieved were produced with very high predictive accuracy and very high local search efficiency.

A hybrid feature selection approach combining filter and wrapper methods for biomedical data classification was developed by authors in Ref. [30]. The proposed approach introduced a feature pre-selection step and used Receiver Operating Characteristics (ROC) curves to deal with the issues of high dimensional biomedical data and to improve the performance of SVM classifier. The experimental results with biomedical databases revealed that the proposed approach significantly improved the classification performance and the results of this approach were found to be much better than the results with Sequential Forward Floating Search (SFFS) method. Moreover, the added pre-selection step helped in solving the problems of over-fitting. The proposed method holds great potential for the classification of biomedical data.

Table 1 depicts some feature selection techniques, algorithms and their characteristics.

Table 2 depicts categorization of feature selection algorithms based on search strategy, evaluation criteria, and data mining tasks [44].

3. Traditional classification systems and adaptive classification systems for chronic disease prediction

In medical data mining [4–6], literature shows that many researchers used different classifier systems for chronic disease prediction to get good diagnostic results and prediction accuracy. Various classifiers like support vector machines, neural networks, decision trees, naïve bayes etc. [31,6] have been implemented and used in the past for the prognosis and diagnosis of chronic diseases.

Fig. 2 depicts how classification process is applied on processed data to get predictive results. In many research and scientific applications, intelligible and automated systems [35] are required for proper diagnosis and prediction of numerous diseases. However, most of the traditional systems are not adaptive which decrease the success rate and increase the decision-making time in the effective diagnosis of disease. Due to these limitations of existing traditional classification systems like lower success rate and more decision-making time, an adaptive classification technique is required for disease prediction. The adaptive classifier would predict the disease more accurately and efficiently. Moreover, parallel classification systems can also be used to enhance the accuracy of results and for execution of multiple units in parallel.

Fig. 3 depicts how adaptive classification process and parallel classification process can be applied on processed data to get predictive results.

Diagnosis of diseases was done by adding the feature of adaptvity to support vector machines in Ref. [32]. The objective was to propose a fast, automatic and adaptive diagnostic system with the help of adaptive SVM. The bias value used in standard SVM was modified to get better results. The output provided by the proposed classifier was in the form of 'if-then' rules. The designed system was utilized for the diagnosis of diabetes and breast cancer and it gave 100% correct classification rates for both diseases. Future work should include finding more efficient methods to change bias value in standard SVM.

A hybrid model based on clustering followed by classification was presented in Ref. [33] for the prediction of type-2 diabetes. The proposed model used K-means clustering and C4.5 classification algorithm with *k*-fold cross-validation for prediction. With the hybrid approach, the model obtained promising results with the classification accuracy of 92.38% which can very helpful for the physicians for taking effective clinical decisions related to diabetes. In future, we can work on developing more refined models for the diagnosis of diseases.

Diagnosis of erythemato-squamous diseases was done using the implementation of multiclass support vector machines (SVM) with error correcting output codes (ECOC) in Ref. [34]. The recurrent neural network and multilayer perceptron neural network approaches were also used for the same. The objective was to detect the six erythemato-squamous diseases using the optimum approach of classification. With the results obtained, it was found that the performance of 'Multiclass SVM' was the best with 98.3% accuracy, recurrent neural network was good with 96.6% accurate

Table 1
Feature selection algorithms, characteristics and techniques.

Feature selection methods	Evaluation criteria/algorithms used/search strategy	Characteristics	Benefits/limitations	Assessment
Filter method	• Relief and ReliefF algorithm • Correlation-based Feature Selection • Pearson product-moment correlation coefficient • Fisher score • Mutual information based methods • Inter/intra class distance • Chi squared test, • Information gain • Markov blanket filter • Fast correlation-based feature selection • Gain ratio • Symmetrical uncertainty • Inconsistency criterion • Minimum redundancy, maximum relevance • *t*-test feature selection • FOCUS algorithm • χ^2 statistic	• Selects subsets of features • before applying model learning algorithm • Relies on the intrinsic characteristics of data • Ranks features based on certain evaluation criteria • The variable selection process needs to • be performed only once using this approach • Selects features independent of any classifier • Uses statistical measures for assigning a score to each feature • Robust against over-fitting compared to other techniques • Measures the "relevance" of feature subsets • May fail to select the most "useful" features	**Benefits** • (1) Works well with voluminous databases • (2) Computationally less expensive • (3) Tends to give fast results on execution • (4) Scalable approach to large datasets • (5) Computationally efficient • (6) Good Generality • (7) Works with low computational complexity **Limitations** • Ignores interaction among classifiers • Does not consider dependency of one feature over other • Can miss some features that are not useful by themselves but can be useful when combined with others [19] • Doesn't guarantee accuracy	Using statistical tests (*t*-test, *F*-test)
Wrapper method	• Sequential forward selection • Backward elimination Method • Randomized Hill-Climbing • Best-First Search • Branch-and-Bound Method • Genetic Algorithms • Stepwise regression • Recursive feature elimination method • Simulated annealing • plus L minus R • Beam search method • Estimation of distribution algorithms	• Searches for the best subset of features considering the learning algorithm to be applied • Utilize a specific classifier to evaluate the quality of selected features [3] • Classifier runs many times for assessing the quality of features • Score is assigned based on accuracy of model • Does optimal selection of features for the learning algorithm • This method calculates • the estimated accuracy of the learning algorithm • for each feature • (7) Use the bias of the induction algorithm to select features	**Benefits** • Better performance in terms of predictive accuracy • Detects dependencies among features • Better classifier interaction • Finds the most "useful" features • Optimize the classifier performance **Limitations** • More computational • Complexity • Greater execution time • Prone to over-fitting on small training dataset • Requires greater computational resources • Computationally more expensive than filter methods • Less scalable for large datasets • (7) Lacks generality	Using cross – validation
Embedded method	• Decision Trees including ID3 algorithm, C4.5 algorithm, CART algorithm and random forest algorithm • Support vector machines- Recursive Feature Elimination approach • Least absolute shrinkage and selection operator (LASSO) method • Elastic Net • Ridge Regression • Artificial neural networks • Weighted naïve Bayes • Sequential Forward • Selection(SFS) • Feature selection using the weighted vector of SVM	• In embedded feature selection approach, search is usually guided by the learning process • Performs feature selection in the process of training • (3) Aggregates the benefits of filter and wrapper approaches • Usually specific to the learning machines • Measures the "usefulness" of feature subsets • Uses the supervised learning algorithm • Optimize the performance of a learning algorithm	**Benefits** • Better classifier interaction • Computationally inexpensive than wrappers • Dependencies between features can be captured effectively • Better usage of available data and provides faster solutions • Less prone to over-fitting than wrapper techniques **Limitations** • Specific to the learning machine • Poor Generality • Selection of relevant features give due consideration to the classifier • Computationally costlier than the filter methods	Using cross – validation

Table 2
Categorization of feature selection algorithms based on search strategy, evaluation criteria and data mining tasks.

		Search strategies			Data mining tasks		Assessment
	Feature selection type	Complete	Sequential	Random	Classification	Clustering	
Evaluation Criteria	Filter method	FOCUS	Relief		**Y**	–	Using statistical tests (*t*-test, *F*-test)
		MIFES1	ReliefF	LVI			
		BFF	ReliefS	QBB			
		Bobrowski's	SFS	LVF			
		MDLM	FCBF				
		Schlimmer's	CFS				
			Dash's		–	**Y**	
			SBUD				
			Mitra's				
	Wrapper method	AMB&B	WSBG	RGSS	**Y**	–	Using cross – validation
		FSBC	WSBG	LVW			
		BS	PQSS	RMHC-PF			
		FSLC	RC	GA			
			SS	RVE			
			SBS-SLASH				
			AICC		–	**Y**	
			FSSEM				
			ELSA				
			BBHFS		**Y**	–	
	Hybrid method		Xing's				Using cross – validation
			Dash-Liu's		–	**Y**	

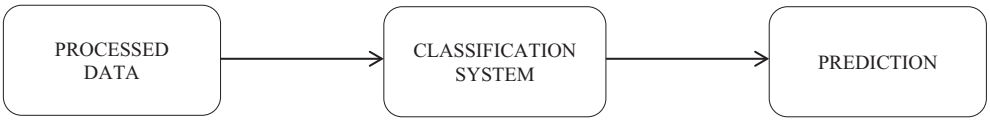

Fig. 2. Classification process.

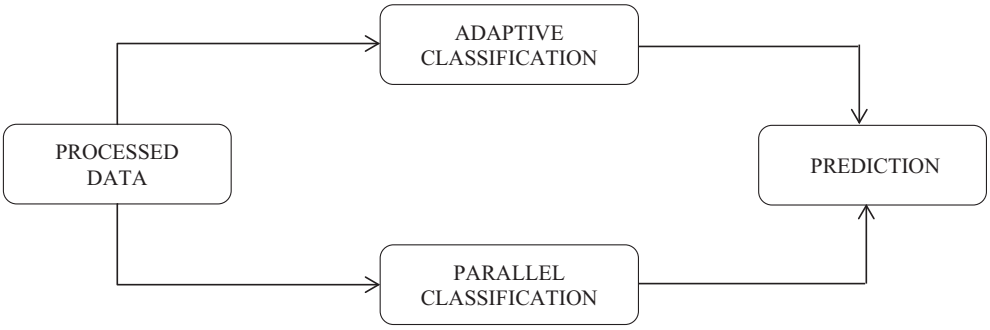

Fig. 3. Adaptive classification process and parallel classification process.

results but multilayer perceptron neural network did not performed well and the classification accuracy reduced drastically to 85.4%. With these results, the authors concluded that the 'Multiclass SVM' and 'RNN' can be used for the diagnosis of the erythemato-squamous diseases.

A hybrid intelligible model was proposed in Ref. [35] using support vector machines for the prediction of diabetes. The key aspect was the proposed rules extracted from the SVM. The proposed work also helped in identifying significant risk factors for diabetes. High prediction accuracy was obtained which would greatly help in medical diagnosis.

ASVM (Adaptive Support Vector Machines) algorithm was proposed by the authors in Ref. [36] who aimed at adapting classifiers to different distributions of data. The proposed algorithm (ASVM) was able to increase efficiency of classifiers by adding the feature of adaptvity to standard SVM algorithm. The key point was that it made it possible to adapt classifiers to any kind of dataset. This was done with the inclusion of 'delta function' in standard SVM.

Table 3
List of popular chronic disease datasets with their descriptions.

S. no	Chronic disease dataset	Type of attributes	Number of instances	Number of attributes
1	Pima Indians Diabetes Dataset	Integer, Real	768	8
2	Chronic Kidney Disease Dataset	Real	400	25
3	Statlog (Heart) Data Set	Categorical, Real	270	13
4	Breast Cancer Wisconsin (Diagnostic) Data Set	Real	569	32
5	Arrhythmia Data Set	Categorical, Integer, Real	452	279
6	Hepatitis Data Set	Categorical, Integer, Real	155	19
7	**Lung Cancer Data Set**	Integer	32	56
8	**Parkinson's Data Set**	Real	197	23

Table 4
Strengths and weaknesses of various classification methods for medical diagnosis.

Classifier	Merits	Demerits
Support vector machines	• Provides flexibility with the concept of kernels • Less over fitting • Robust to noise	• Lack of transparency of results • Computationally expensive
k-nearest neighbour algorithm	• Easy to understand and implement • Robust to noisy training data	• Problem of memory limitation • Runs slowly • Supervised learning lazy algorithm
Naïve Bayes classifier	• Low cost of computation • Works well on numeric as well as nominal data	• Requires a very huge number of records for obtaining good results
Neural networks	• Provides high accuracy with prediction • Shows arbitrarily complex relationship between input and output	• Take long training time for computation • Does not work well when with large number of input features
Bayesian networks	• Provides extensive support with missing data	• High computational cost is involved during computation • Requires initial knowledge of large number of probabilities
C4.5 classifier	• Works well with both categorical and continuous attributes • Takes the less memory to large program execution	• Problem of Over fitting • Problem of insignificant branches

With the results obtained, it was found that the performance of ASVM was better than the ensemble approach and other adaptation techniques.

An automated diagnosis of coronary artery disease (CAD) was done through a decision support system based on fuzzy rules by Tsipouras et al. [37]. The methodology consisted of four phases. In the first phase, a decision tree was constructed from the training dataset using C4.5 algorithm. Rules were extracted from the tree to generate a crisp model. Then a fuzzy model was developed from crisp rules. It was found from results that when the fuzzy model is optimized, the classification accuracy got significantly enhanced from 5% to 35%. Moreover, the proposed approach also gave comparable results than the classification results of ANN and ANFIS algorithm. The approach can help in assisting doctors to take effective decisions related to presence or absence of CAD disease.

Table 3 presents a list of popular chronic diseases datasets with their descriptions which are commonly used in many studies for the prediction of chronic diseases [45].

Table 4 depicts merits and demerits of classifiers used for chronic disease prediction.

Table 5 presents a summarized methodology of related works done by different researchers and is more specifically concerned with the utility driven from different papers.

4. Parallel classification systems for chronic disease prediction

Healthcare industry faces many challenges in processing massive health records. Analysis of substantial amount of medical data brings complexities due to the unstructured nature of data. The health care system needs to evolve and innovate continuously to solve the problems associated with the management of huge amount of data. Existing structure of the current health care system can be improved with the effective use of big data analytics [41]. Big data analytics can be done by applying various tools and technologies like Hadoop, STORM, Map Reduce programming etc.

Parallel Classification Systems holds great potential for enhancing predictive accuracy of diagnostic systems for chronic diseases. Most of the authors in the recent literature have used hadoop and map-reduce programming for chronic disease prediction. Signifi-

Table 5
Summarized methodologies and the utility driven from different papers on Pima Indians Diabetes dataset.

Reference	Classifier used	Methodology used	Merits	Demerits
Balakrishnan et al. [12]	Naive Bayes	SVM Ranking with Backward Search Technique	The proposed technique enhanced the predictive accuracy of Naïve Bayes classifier and would assist healthcare professions for the diagnosis of type 2 diabetes	The proposed method increased the classification accuracy by only 1.88%. The system can be made more precise with the application of other feature reduction algorithms
Patil et al. [45]	C4.5 Algorithm	Building a predictive model using hybrid approach	The hybrid approach gave the classification accuracy of 92.38% which can very helpful for the physicians for effective decision – making related to diabetes	The system does not apply any feature technique to remove redundant features. The application of feature selection could have increased the classification accuracy
Kayaer et al. [46]	General Regression Neural Networks	Three different neural network structures, were applied to the Pima Indians Diabetes to find the best approach among them	The General Regression Neural Networks (GRNN) worked best with 80.21% accuracy	The proposed system is limited to only neural networks for finding the best method for classifying diabetic patients
Gürbüz and Kılıç [32]	Support Vector Machines	Adding the feature of adaptively to standard SVM	The approach gave promising results by giving 100% correct classification rates by determining the Most appropriate bias value for classification for both diabetes and breast cancer	The system can be improved by finding the bias value at lower cost in more effective way
Karegowda et al. [47]	Genetic Algorithm and Back Propagation network (BPN)	Hybrid model with the integration of Genetic Algorithm and Back Propagation network (BPN) to diagnose diabetes mellitus	The results showed that the proposed approach has outperformed the BPN approach without GA optimization, giving high classification accuracy	The results are based on classification accuracy. The method can also consider other performance metrics like specificity, recall, precision, F-measure etc
Lekkas and Mikhailov [48]	Fuzzy Classification	Method named as 'eClass' for evolving fuzzy rule-based systems to process data in online mode	The proposed method significantly increased the accuracy and specificity on both Pima Indians diabetes dataset and on dermatological disease datasets	The complexity of the method is a square of the number of input dimensions which needs to be significantly reduced

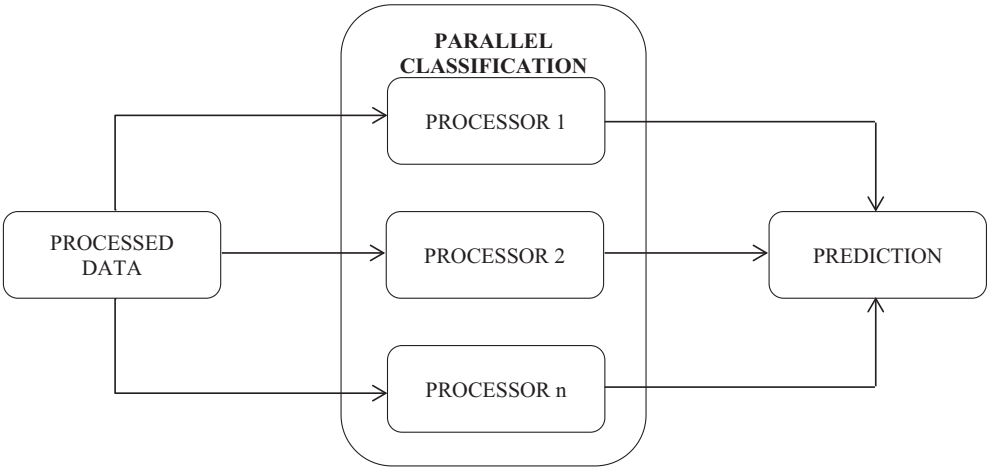

Fig. 4. Parallel classification process.

cant clinical decisions related to chronic disease can be taken by employing parallel classification systems.

Fig. 4 depicts how parallel classification is applied to processed data to get predictive results.

Parallel Support Vector Machine was applied by the authors for the diagnosis of diabetes in Ref. [38]. The usage of parallel SVM facilitated for the distribution of datasets on different machines, thereby, decreasing the computational complexity, power consumption and memory utilization for each machine. With the results obtained, it was found that the time taken by the proposed algorithm on a single machine was 1/3 of time taken by simple SVM. Moreover, the accuracy of the classifier was also comparable with the simple SVM functionality. The designed method was found to very reliable and could be utilized for unbalanced and voluminous databases. In future, this method can be applied on multiple machines to get better results and to enhance classification accuracy.

A parallel classification system based on feed-forward neural networks was designed for the diagnosis of Parkinson Disease in Ref. [39]. Rule-based decision system based on a voting scheme was used to evaluate the output of each neural network. The experimental results showed that this combination of parallel networks with the rule-based decision system significantly improved the robustness of the diagnostic system. However, it was found that as the number of parallel units increased to a great extent, the training time and complexity of the algorithm also increased. Also, after the addition of certain number of parallel units, no significant rise was seen in the predictive accuracy. The designed system performed best with the combination of nine parallel networks as it improved the prediction rate by 8.4% as compared to the implementation on a single network.

PARAMO (PARAllel predictive MOdeling platform), proposed by Ng et al. [40] is a scalable modeling platform employed for prediction to quicken the process of healthcare data analytics domain. The model was built for parallel execution of independent tasks with the inclusion of hadoop map-reduce programming. This model first constructed an efficient graph to demonstrate the dependency between different tasks then applied topological sort to do proper task scheduling and then finally did parallel execution of these tasks. The computational performance of the model was analyzed for three different healthcare EHR data sets and the results depicted that PARAMO model performed very well with significant gains in computational efficiency. This parallel platform was very scalable and can be used to quicken and simplify the use of healthcare analytics data.

A predictive algorithm using hadoop technology was proposed to diagnose the presence of non-communicable disease, Diabetic Mellitus (DM) at an initial stage [41]. The analysis using hadoop's implementation made it easier to understand the diabetic complications and the treatment that can be provided to a diabetic patient. The results concluded that the proposed system can assist healthcare providers to meet the expectations of the patient's needs and to make effective clinical decisions based on the diagnostic results. Moreover with the proposed approach, the risk level of a patient's health condition can also be identified.

Parallel classification was utilized to propose an effective method for rule generation in Ref. [42]. It was based on a distributed approach which was completed in three stages. Firstly, data preprocessing was done in which the voluminous dataset was divided into small data chunks which were simultaneously executed on different machines. Secondly, each data chunk was given to different processors for execution. Next, each processor applied parallel classification to generate "if-then-else" rules. For rule generation, J-48 and ID3 classification algorithms were used. Finally, the rules generated on each machine were merged into a single file. Additionally, the conflicts were resolved in the last step. Compared to the traditional classification approaches, the proposed approach was found to very efficient with huge date sets and was able to generate both specific and generic rules. In future, the overall performance of the proposed approach can be improved by developing some technique to get optimal rules in the final rule set.

5. Comparisons of traditional classification systems, adaptive classification systems and parallel classification systems

This study presents the utilization of different types of classification systems for chronic disease prediction. With the comparative study of various classification systems on disease datasets, it is found that most of the traditional systems for chronic disease prediction are unable to develop accurate diagnostic systems and are more focused on manual processing of tasks. As manual tasks may result in human error, it results in less accuracy and decreases the response time. Adaptive systems, on the other hand, enhance the success rate and can assist doctors and medical professionals to take effective clinical decisions related to the prediction of chronic disease. The current health care systems can be improved with the effective use of parallel classification systems as they facilitate the parallel executes the same implementation on multiple systems. Parallel Classification Systems also holds great potential for enhancing predictive accuracy of diagnostic systems for chronic diseases.

Table 6 presents a summarized comparison of different classification Systems on the diabetes dataset

Table 6
Comparison of various classification systems on diabetes dataset.

Type of system	Reference	Dataset	Methodology	Results
Traditional classification system	Patil et al. [33]	Pima Indians diabetes dataset	Building a predictive model using hybrid approach	The hybrid approach gave the classification accuracy of 92.38% which can very helpful for the physicians for effective decision – making related to diabetes
	Purushottam et al. [43]	Pima Indians Diabetes Database	An effective prediction system for diabetes mellitus using decision trees was proposed for predicting risk level of diabetic patients	The proposed system was 81.27% accurate with the usage of C4.5 algorithm
Adaptive classification system	Gürbüz, Kılıç [32]	Pima Indians diabetes dataset and Wisconsin breast cancer dataset	Diagnosis of diabetes and breast cancer was done by using adaptive support vector machines. The purpose was to design a fast, automatic and adaptive diagnostic system with the help of adaptive SVM	The designed system significantly increased classification accuracy to 100% for both diseases. The proposed adaptive system can be used by medical professionals and doctors for the diagnosis of diabetes and breast cancer
	Alby and Shivakumar [49]	Pima Indians diabetes dataset	Prediction model using adaptive neuro-fuzzy interface system for the diagnosis of or type 2 diabetes	The proposed approach with the aggregation of combination of ANFIS and GA gave better results compared to GRNN approach
Parallel classification system	Shrivastava et al. [38]	Diabetes dataset from S. S. Medical College, Rewa, Master Chart (Study Group)	Parallel Support Vector Machines were utilized for the prediction of diabetes. The designed method could be used with unbalanced and voluminous databases	The proposed algorithm using parallel SVM functionality took 1/3 of the time taken by simple SVM and significantly reduced the computational complexity, power consumption and memory utilization for each machine

6. Performance metrics used in medical diagnosis systems to measure the performance of the classification methods

There exists different metrics in medical diagnosis systems to measure the performance of the classification methods like sensitivity, precision, F-measure, accuracy and specificity, recall. These performance metrics are generally used to analyze the performance of different models.

(a) **Accuracy:** It is defined as the number of all correct predictions made divided by the total number of predictions made. This is defined as ratio of appropriately classified data to overall classified data.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN)$$

In above equation

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

(b) **Sensitivity (Recall or True positive rate)**: Recall is how many relevant items are selected. It is a ratio of true positive to the sum of true positive and false negative. In medical diagnosis, test sensitivity (Recall) is the ability of a test to correctly identify those with the disease (true positive rate). If the test is highly Recall and the test result is negative you can be nearly certain that they don't have disease.

$$Recall = true\ positives/(true\ positive + false\ negative)$$

(c) **Specificity (True negative rate):** Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR).

$$Specificity = true\ negatives/(true\ negatives + false\ positives)$$

(d) **Precision:** Precision is how many selected items are relevant. It is a ratio of true positive to the sum of true positive and false positive. Test specificity (Precision) is the test's ability to correctly recognize those that do not have a disease (true negative rate). If the test output for an extremely precise test is positive user can be nearly certain that they actually have the disease.
$Precision = true\ positive/(true\ positive + false\ positives)$

(e) **False Positive Rate:** False positive rate is defined as the number of incorrect negative predictions divided by the total number of negatives. It is defined as:
$FPR = False\ positive/(true\ negative + false\ positive)$

(f) **F-measure:** The F-measure of the system is defined as the weighted harmonic mean of its precision and recall, that is
$F = 1/(\alpha 1/P + (1 - \alpha)1/R)$

where the weight $\alpha \in [0, 1]$.

7. Conclusion and future perspectives

World's health is badly affected by the chronic diseases which is spreading and increasing day by day. The lack or delay in proper treatment can also lead to the death of patients. So, chronic disease prediction is a vital task in medical field. This article presents a survey on various feature selection and classification techniques which can be very helpful for severity analysis for quick disease diagnosis. Several reliable and efficient feature identification approaches have been developed in the literature according to different principles. Although feature selection is a well-developed field, researchers are focusing on designing novel methods to improve efficiency of the learning machines. This study shows that there is a need to make healthcare professionals aware of reliable

feature selection and classification techniques that can be successfully applied on medical databases for the early detection of diseases.

This paper presents a review on the current feature selection approaches and classification systems for effective disease prediction. The performance metrics used in medical diagnostic systems to measure the performance of the classification methods are also discussed. In addition, this study compares the merits and demerits of various feature selection methods and classifiers separately. Also, this paper summarizes the utility driven from the work done by previous researchers on popular chronic disease datasets. Moreover, this review presents the categorization of feature selection algorithms based on search strategy, evaluation criteria and data mining tasks. A plenitude of variable selection and classification methods have been proposed by various researchers for early diagnosis of patients, yet this study implies that there are still many opportunities for developing new approaches with higher response and increased success rate than previous studies.

This review of the feature selection methods depict that a particular feature selection algorithm plays a vital role for accurate classification of diseases. It is observed from the survey that the filter method is computationally more efficient and provides better generality than other methods. Wrapper and embedded approach should be used when there is a need to find optimal feature subset tailored to a particular learning algorithm. This study reveals that the current advancement in research is found in hybrid approaches that can be applied on disease datasets to remove redundant, noisy and insignificant features. Hybrid approach takes advantage by aggregating the merits of two or more techniques. Unfortunately, there are not ample empirical works that make use of hybrid approaches for disease prediction. Hence, further applications and developments of such approaches are required to be applied on disease datasets.

Among classifier systems, much of the existing work has been done on chronic disease prediction using support vector machines, naive bayes, decision trees and ANNs. This paper put focus on the use of Adaptive Classification Systems and Parallel Classification Systems as they increase the success rate and decrease the decision-making time for the diagnosis of chronic diseases. Future research should focus on designing new hybrid classification methods to improve the classifier accuracy and optimize the computational efficiency of results.

In all, a large and fruitful effort has been performed on various classifier systems during the last years. Previous studies have manly focused on the use of traditional classification systems for medical diagnosis. This paper contributes by reviewing research based on the current and relevant adaptive classification systems and parallel classification systems for chronic disease prediction as these systems significantly enhances the performance of models by providing high classification and prediction accuracy. These systems would assist doctors, physicians and healthcare professionals in effective decision making for disease diagnosis.

References

[1] Shardlow M. An analysis of feature selection techniques. The University of Manchester; 2016.
[2] Dash M, Liu H. Feature selection for classification. Intell Data Anal 1997;1(3):131–56.
[3] Tang J, Alelyani S, Liu H. Feature selection for classification: a review. Data Classif: Algor Appl 2014;37.
[4] Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011.
[5] Tan PN. Introduction to data mining. Pearson Education India; 2006.
[6] Dunham MH. Data mining: introductory and advanced topics. Pearson Education India; 2006.
[7] Muni Kumar N, Manjula R. Role of Big data analytics in rural health care – a step towards svasth bharath; 2014.
[8] Hussein AS, Omar WM, Li X, Ati M. Efficient chronic disease diagnosis prediction and recommendation system. In: Biomedical engineering and sciences (IECBES), 2012 IEEE EMBS conference on. IEEE; 2012. p. 209–14.
[9] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3(Mar):1157–82.
[10] Nalband S, Sundar A, Prince AA, Agarwal A. Feature selection and classification methodology for the detection of knee-joint disorders. Comput Methods Programs Biomed 2016;127:94–104.
[11] Ang JC, Mirzal A, Haron H, Hamed H. Supervised, unsupervised and semi-supervised feature selection: a review on gene selection. IEEE/ACM Trans Comput Biol Bioinform 2015;PP(99).
[12] Balakrishnan S, Narayanaswamy R, Savarimuthu N, Samikannu R. SVM ranking with backward search for feature selection in type II diabetes databases. In: Systems, man and cybernetics, 2008. SMC 2008. IEEE international conference on. IEEE; 2008. p. 2628–33.
[13] Nagpal A, Gaur D. ModifiedFAST: a new optimal feature subset selection algorithm. J Inform Commun Convergence Eng 2015;13(2):113–22.
[14] Battiti R. Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Networks 1994;5(4):537–50.
[15] Kwak N, Choi CH. Input feature selection for classification problems. IEEE Trans Neural Networks 2002;13(1):143–59.
[16] Huang J, Cai Y, Xu X. A filter approach to feature selection based on mutual information. In: 2006 5th IEEE international conference on cognitive informatics, vol. 1. IEEE; 2006. p. 84–9.
[17] Verma L, Srivastava S, Negi PC. A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. J Med Syst 2016;40(7):1–7.
[18] Yu L, Liu H. Feature selection for high-dimensional data: a fast correlation-based filter solution. In: ICML, vol. 3; 2003. p. 856–63.
[19] Kumar V, Minz S. Feature selection. SmartCR 2014;4(3):211–29.
[20] Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97(1):273–324.
[21] Maldonado S, Weber R. A wrapper method for feature selection using support vector machines. Inf Sci 2009;179(13):2208–17.
[22] Shahana AH, Preeja V. Survey on feature subset selection for high dimensional data. In: Circuit, power and computing technologies (ICCPCT), 2016 international conference on. IEEE; 2016. p. 1–4.
[23] Wikipedia, Feature selection; 2016. <https://en.wikipedia.org/wiki/Feature_selection> [accessed 24th October, 2016].
[24] Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinform; 2015.
[25] Maldonado S, Weber R, Famili F. Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines. Inf Sci 2014;286:228–46.
[26] Xiao Z, Dellandrea E, Dou W, Chen L. ESFS: a new embedded feature selection method based on SFS. Rapports de recherché; 2008.
[27] Chhikara RR, Sharma P, Singh L. A hybrid feature selection approach based on improved PSO and filter approaches for image steganalysis. Int J Mach Learn Cybern 1–12.
[28] Das S. Filters, wrappers and a boosting-based hybrid for feature selection. In: ICML, vol. 1; 2001. p. 74–81.
[29] Huang J, Cai Y, Xu X. A hybrid genetic algorithm for feature selection wrapper based on mutual information. Pattern Recogn Lett 2007;28(13):1825–44.
[30] Peng Y, Wu Z, Jiang J. A novel feature selection approach for biomedical data classification. J Biomed Inform 2010;43(1):15–23.
[31] Pujari AK. Data mining techniques. Universities press; 2001.
[32] Gürbüz E, Kılıç E. A new adaptive support vector machine for diagnosis of diseases. Expert Syst 2014;31(5):389–97.
[33] Patil BM, Joshi RC, Toshniwal D. Hybrid prediction model for Type-2 diabetic patients. Expert Syst Appl 2010;37(12):8102–8.
[34] Übeyli ED. Multiclass support vector machines for diagnosis of erythemato-squamous diseases. Expert Syst Appl 2008;35(4):1733–40.
[35] Barakat N, Bradley AP, Barakat MNH. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE Trans Inf Technol Biomed 2010;14(4):1114–20.
[36] Yang J, Yan R, Hauptmann AG. Cross-domain video concept detection using adaptive svms. In: Proceedings of the 15th ACM international conference on multimedia. ACM; 2007. p. 188–97.
[37] Tsipouras MG, Exarchos TP, Fotiadis DI, Kotsia AP, Vakalis KV, Naka KK, Michalis LK. Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. IEEE Trans Inf Technol Biomed 2008;12(4):447–58.
[38] Shrivastava NK, Saurabh P, Verma B. An efficient approach parallel support vector machine for classification of diabetes dataset. J Comput Appl 2011;36(6):19–24.
[39] Åström F, Koker R. A parallel neural network approach to prediction of Parkinson's Disease. Expert Syst Appl 2011;38(10):12470–4.
[40] Ng K, Ghoting A, Steinhubl SR, Stewart WF, Malin B, Sun J. PARAMO: a PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. J Biomed Inform 2014;48:160–70.
[41] Eswari T, Sampath P, Lavanya S. Predictive methodology for diabetic data analysis in big data. Proc Comput Sci 2015;50:203–8.
[42] Ghaffar T, Shahzad W, Baig AR. Parallel rule generation for making an efficient classification system. In: 2012 International conference on information science and applications. IEEE; 2012. p. 1–4.

[43] Purushottam, Saxena Kanak, Sharma Richa. Diabetes mellitus prediction system evaluation using C4.5 rules and partial tree. 2015 4th International conference on reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions).
[44] Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 2005;17(4):491–502.
[45] UCI repository. <https://archive.ics.uci.edu/ml/datasets.html> [accessed 5th August 2017].
[46] Kayaer K, Yıldırım T. Medical diagnosis on Pima Indian diabetes using general regression neural networks. In: Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP); 2003. p. 181–4.
[47] Karegowda AG, Manjunath AS, Jayaram MA. Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes. Int J Soft Comput 2011;2(2):15–23.
[48] Lekkas S, Mikhailov L. Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases. Artif Intell Med 2010;50(2):117–26.
[49] Alby S, Shivakumar BL. A prediction model for type 2 diabetes using adaptive neuro-fuzzy interface system. Biomed Res 2017:1.