International Journal on Advanced Science Engineering Information Technology

Vol.9 (2019) No. 6
ISSN: 2088-5334

Feature Selection Method using Genetic Algorithm for Medical Dataset

Neesha Jothi[#1], Wahidah Husain[#2], Nur'Aini Abdul Rashid[#3], Sharifah Mashita Syed-Mohamad[#4]

[#]*School of Computer Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia*
E-mail:[1]nj14_com042@student.usm.my; [2]wahidah@usm.my; [3]nurainipng@gmail.com; [4]mashita@usm.my

***Abstract*— There is a massive amount of high dimensional data that is pervasive in the healthcare domain. Interpreting these data continues as a challenging problem and it is an active research area due to their nature of high dimensional and low sample size. These problems produce a significant challenge to the existing classification methods in achieving high accuracy. Therefore, a compelling feature selection method is important in this case to improve the correctly classify different diseases and consequently lead to help medical practitioners. The methodology for this paper is adapted from KDD method. In this work, a wrapper-based feature selection using the Genetic Algorithm (GA) is proposed and the classifier is based on Support Vector Machine (SVM). The proposed algorithms was tested on five medical datasets naming the Breast Cancer, Parkinson's, Heart Disease, Statlog (Heart), and Hepatitis. The results obtained from this work, which apply GA as feature selection yielded competitive results on most of the datasets. The accuracies of the said datasets are as follows: Breast Cancer - 72.71%, Parkinson's – 88.36%, Heart Disease – 86.73%, Statlog (Heart) – 85.48 %, and Hepatitis – 76.95%. This prediction method with GA as feature selection will help medical practitioners to make better diagnose with patient's disease.**

***Keywords*— data mining; data mining in healthcare; medical dataset; feature selection; genetic algorithm.**

I. INTRODUCTION

Hospitals nowadays are well equipped with extensive data collection tools that proportionately allocate reasonable means to collect and store the data in the hospital information systems [1]. Large amounts of data that are being accumulated in medical databases require specialized tools for storing, accessing and analyzing the data to make use of the data effectively [1]. There are various types of medical data such as narrative, textual, numerical measurements, recorded signals, and pictures. Lately, it has become laborious to extract useful information for decision support due to the growth in these data sizes [1].

The knowledge discovery in databases (KDD) method provides an exciting approach to probe such data-driven problems. The KDD methodology refers to the comprehensive process of discovering helpful knowledge where the process continues to progress into various research fields, including high-performance computing, data visualization, knowledge acquisition for expert systems, artificial intelligence, statistics, databases, pattern recognition, and machine learning [2].

The KDD methodology involves using the database with preprocessing of data, sub-sampling of data, and transformations of data to apply data mining methods to identify patterns. The KDD method is illustrated in Figure 1. This methodology denotes to the overall process of uncovering useful knowledge from data.

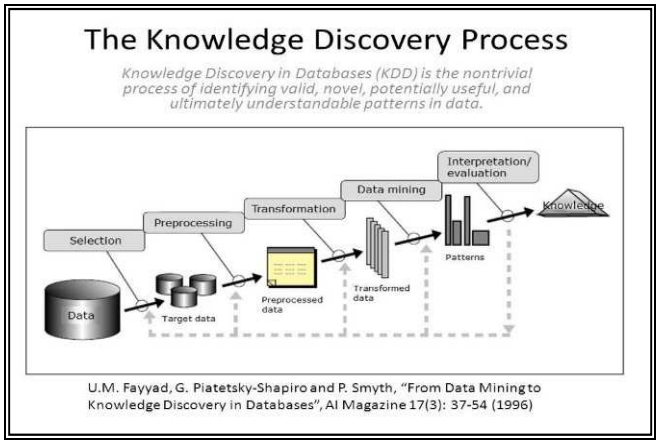

Fig. 1. The KDD Process [2]

Data mining is known as an application of special algorithms to extract patterns from data using the KDD methodology. The additional steps in the KDD methodology include data preparation, data selection, and data cleaning. The inclusion of suitable preceding knowledge alongside an appropriate rendition of the results from the mining steps is

important to fortify the relevant knowledge acquired from the data used.

Since data mining is an important subset of the KDD methodology, it is an iterative task to search for current, invaluable, and non-trivial information from a large capacity of data [2]. The best solution is attained by assessing the knowledge of human experts in outlining the problems and goals that can provide the search capacity. There are various data mining approaches that have been successfully applied in a few main areas of study such as healthcare, medicine, biomedicine, bioinformatics and education [3]–[7]. The competitive results obtained from the said studies above are used to actuate the current research in applying data mining approaches. Based on the outlined problems and their motivations, this work aims to adapt the Genetic Algorithm (GA) into the medical datasets where features are selected and applied into Support Vector Machine as a classifier to predict diseases namely Breast Cancer, Parkinson's, Heart Disease, Statlog (Heart) and Hepatitis.

II. Materials and Methods

A. Nature-Inspired Metaheuristic Algorithms for Feature Selection

Metaheuristic is known as a repetition generation process that accompanies a subordinate heuristic by integrating different concepts of exploring and exploiting in a search space intelligently [8]. The learning strategies are used to structure information to perceive efficiently near-optimal solutions [8]. The metaheuristic algorithms are said to be highly successful and efficient in resolving complex and large problem sets [8]. These algorithms can be mainly categorized into two main classes. The first category is a single-based metaheuristic algorithm, which is the local search method. The second category is the population-based metaheuristic algorithm, which represents global search methods. The single-based algorithms begin with a solution and try to improvise it until the stop condition is attained. This algorithm broadly exploits the current region of the said problem search space. This search method often results in a local optima problem.

Meanwhile, the population-based metaheuristic algorithms perform a search based on global optimization and stochastic methods. The population-based algorithms are inspired by diverse aspects of biological processes that revolve around the living creatures or from the social interactions among animals in the real world [8]. These algorithms are more efficient in their exploration ability of the search space and usually acquire acceptable solutions [8]. Currently, there is numerous work being done using feature selection based on metaheuristic algorithms. The related work using these algorithms are discussed below.

1) Genetic Algorithm (GA): Salem et. al. [9] used GA to classify human cancer diseases. The dataset used in this study was extracted from seven different Cancer Gene Expression Datasets. The results obtained from this work was competitively better with feature selection. In another work, Paul et. al. [10] proposed a GA based Fuzzy Decision Support System for diagnosing heart disease. The proposed algorithm was tested on the heart disease datasets available in the UCI Machine Learning Repository [11]. The proposed system obtained almost 80% accuracy based on the datasets. The proposed algorithms from these studies proved that the classification accuracy is better when feature selection is applied. However, the random convergence on GA for this dataset based on the fitness function remains as a disadvantage. GA is time-consuming whenever large features are involved.

2) Particle Swarm Optimization (PSO): Chatterjee et. al. [12] used PSO to select feature for dengue fever classification. Artificial Neural Network (ANN) was used as a classifier in this work. The dataset used for this work included the acute dengue patients data from Gene Expression Omnibus. The proposed algorithm obtained an accuracy of 90.91% with PSO as feature selection. In another work by Shahsavari et. al. [13], PSO was used to select the features while Extreme Learning Machine was used for the classification of Parkinson's disease. The proposed method obtained an accuracy of 88.72%. Based on the results, it was concluded that PSO yielded good results since the size of the features in these two datasets are really large. The PSO has also proved computational efficiency in the said work.

3) Ant Bee Colony (ABC): Shunmugapriya & Kanmani [14] designed a hybrid algorithm based on the Ant and Bee Colony (AB-ABC) for feature selection. The proposed algorithm was tested on 13 different datasets from the UCI Machine Learning Repository [11]. The classifiers used in this work are Decision Tree and J48. The results obtained were significantly better with feature selection. In addition, Subanya & Rajalaxmi [15] performed feature selection using ABC for cardiovascular disease diagnosis. The classifier used in this work is Naïve Bayes. The dataset used in this study was obtained from the UCI Machine Learning Repository [11]. The accuracy obtained was 86.4%. The prediction accuracies were improved in these works with the implementation of the ABC algorithm as feature selection. The ABC worked well on these datasets due to its ability to explore local solutions well.

B. Data Mining Methods for Classification

According to Fayyad et.al. [2], data mining was employed to detect patterns and to extract hidden information from large databases. There are two main data mining models, which are the predictive model and the descriptive model [16]. The predictive model is often applied to supervised learning functions and to predict the undisclosed variables of interest [16]. While the descriptive model is often applied to the unsupervised learning functions in identifying patterns to describe the data that can be explicated by human [16]. The implementation model of data mining is made through a task. The task used for the predictive models is classification [17], regression [18] and categorization [19].

Meanwhile, the task for a descriptive model is often implemented using the clustering [20], association rules [21], correlation analysis [22] and anomaly detection [23]. When the data mining model and the task are defined, the appropriate data mining method will be used to build the model based on the discipline of study. Many data mining methods are commonly used for classification. The related work using data mining methods is discussed accordingly.

1) K-Nearest Neighbour (KNN): Khateeb and Usman [24] used the KNN to predict heart disease. The dataset used in this study was from the UCI Machine Learning Repository [11]. The classification accuracy obtained was 79.20% based on a 10-fold cross-validation. Moreover, in another work, Hashi. et.al. [25] used KNN to predict diabetes. In this work, the Pima Indians Diabetes Database of the National Institute of Diabetes was used. The KNN algorithm achieved 76.96% of accuracy. On the other hand, Enriko et. al. [26] presented their work on heart disease prediction using KNN. The dataset used in this study was also from the UCI Machine Learning Repository [11]. KNN attained an accuracy of 81.85%. Based on these papers, it was evident that the KNN obtained competitive results on various datasets used. The KNN was found to be an algorithm which constantly evolved based on the learning instances. KNN is capable of determining the neighbours directly from the training models.

2) Support Vector Machine (SVM): Rustam et. al. [27] used the SVM algorithm to classify the imbalanced cerebral infarction. The authors tested them on the imbalanced cerebral infarction dataset obtained from the Department of Radiology at Dr Cipto Mangunkusumo Hospital. An accuracy of 87% was achieved when SVM was implemented on the said dataset. In another work [28], SVM was applied SVM for predicting thyroid disease classification using data retrieved from the UCI Machine Learning Repository [11]. SVM obtained an accuracy rate of 94.55% on the said dataset. Apart from the two disease prediction models stated above, SVM has also been used to predict kidney disease [29]. The dataset used on this work is collected from several medical labs, canters and hospitals. The SVM scored 76.32% in terms of classification accuracy. The results obtained from these studies were found to be acceptable since SVM was designed to minimize structural risk and to find the best hyperplane to classify data from the defined classes.

3) Naïve Bayes: Naïve Bayes classifier was used to detect cardiovascular disease risk level for adults [30]. The dataset employed in this study was collected from cardiac risk assessment. The Naïve Bayes achieved an accuracy of 85.90%, 84.37% sensitivity and 86.19% specificity. In another study, Naïve Bayes method was also used to predict heart disease [31] using the dataset obtained from the UCI Machine Learning Repository [11]. The Naïve Bayes method obtained an accuracy of 76.67%. Meanwhile, Baitharu & Pani [32] have also applied the data mining method for making a healthcare decision support system using liver disorder dataset obtained from the UCI Machine Learning Repository [11] was used in this study. The accuracy obtained from this study was only 55.36%. The results obtained from these studies were mixed. These mixed results could be due the Naïve Bayes classifier's inability to modify the dependencies among the variables. It is not a great method to be implemented as a classifier especially when there is no prior pre-processing before classification.

4) Random Forest: This classifier was used in predicting oesophageal cancer [33]. The authors have used the dataset from oesophageal cancer patients. The Random Forest algorithm obtained an accuracy of 82.3%. In another work conducted by Kumar [34], chronic kidney disease was predicted using Random Forest. The dataset used in this study was retrieved from the UCI Machine Learning Repository [11]. The Random Forest algorithm performed better than the other methods. On the other hand, Random Forest was also used to predict generalized anxiety disorder among women [35]. This method yielded an accuracy rate of 92.85%. The Random Forest methods worked well with these studies because it is flexible and tends to produce high accuracy with large datasets in place.

C. Methodology Development

The research methodology for this work is adapted from the KDD methodology and is illustrated in Figure 2. The proposed test method empirically used five medical diagnostic datasets obtained from the UCI machine learning data repository [36]. The method involves a data pre-processing phase where the collected data is cleaned and transformed suitable for classification prediction. Following this phase, the feature selection was performed. The data from the pre-processing stage (phase two) was carried over to phase three for classification prediction. The prediction algorithm based on the SVM was trained and tested on the data for classification prediction. The last includes the evaluation phase where the enhanced prediction algorithm using classification performance criteria.

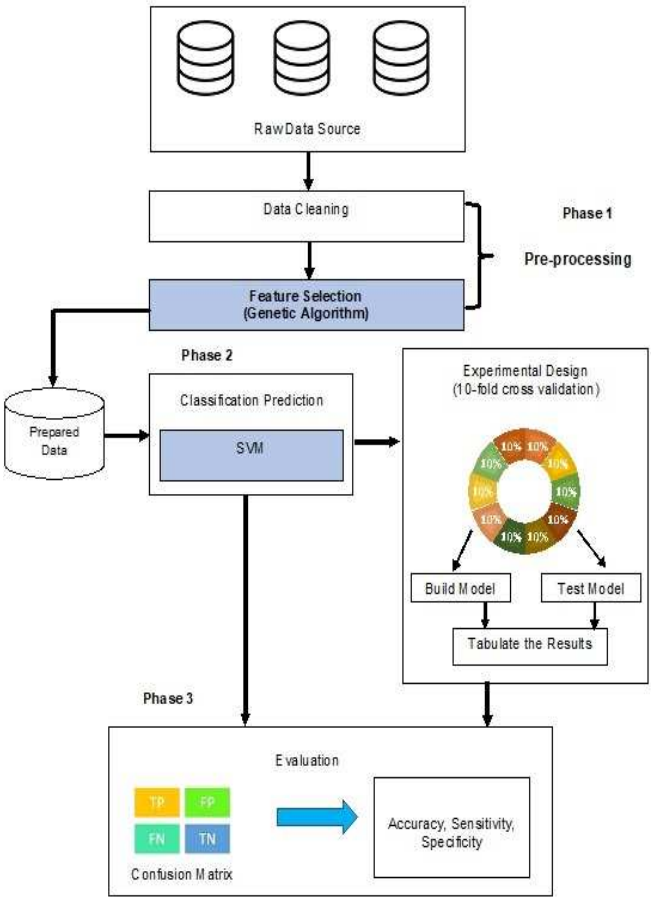

Fig. 2. The proposed methodology

1) Phase 1 (Pre-processing): The dataset cleaning is an important process used to regulate inaccurate, incomplete and noisy data. The cleaned dataset will then be improved in terms of quality through correctly detect the errors and

omissions. In general, dataset cleaning reduces errors and improves the overall data quality. Dataset cleaning is the first step of this methodology [37]. The proposed algorithm was be tested on seven datasets obtained from the UCI machine learning repository [36]. These datasets contain a distinct number of features belonging to the medical diagnosis domain. The attributes of the features are usually diagnosis attribute information namely radius, texture, and smoothness of a malignant cell. The selected data sets provide the maximum variation based on the *number of samples (S)*, the *number of classes (C)* and the *number of features (F)* where variation is widely adopted in related works. However, most of these datasets suffer from missing values. In order to rectify the issue, the missing values were replaced with the mean of the numeric attributes and the mode for the nominal attributes. Based on Table I, the selected datasets was used to prove the performance of feature selection using GA on binary classification. The selected datasets provide an overall insight into the performance of the proposed algorithm in both extreme cases, namely when $S > F$ and when $F > S$. Once the missing values in the dataset were replaced with the mean for numeric attributes and the mode for nominal once, the dataset was used for the feature selection process.

TABLE I
DATASET USED

No	Name	Samples	Features	Classes
1	Breast Cancer	286	9	2
2	Parkinson's	756	754	2
3	Heart Disease	303	75	2
4	Statlog (Heart)	270	13	2
5	Hepatitis	155	19	2

2) Phase 2 (Pre-processing): In this phase, the features selected from phase 1 were passed through SVM to evaluate the classification accuracy by evaluating every feature subset which was generated from the GA algorithm. The SVM classifier was used in the wrapper method because it works well on high dimensional data [38]. In this work, the Radial Basis Function (RBF) kernel was used in SVM classification method because it is very widely used kernel for classification application, and it has a smaller number of hyperparameters. There are many studies that report the effectiveness of RBF kernel over other kernels [39]–[41].

3) Phase 3 (Evaluation): In the evaluation phase, the proposed algorithm was evaluated using several measures to assess the effectiveness and the correctness of the proposed method. The confusion matrix is widely used to evaluate the performance as it encompasses the number of correct and incorrect predictions made by the classification model compared to the actual outcomes in the data [42]. This matrix includes the accuracy, sensitivity, and specificity. Accuracy is the proportion of the total number of predictions that are evaluated as being correct. Sensitivity is the proportions of True Positive (TP) which are correctly identified by the classifier. Meanwhile, specificity is the proportions of True Negative (TN) which are correctly identified by the classifier. Due to this combination of measures, it is possible to have a balanced view of the performance instead of a single numerical value of the method that was being used. These metrics aid in the performance evaluation of the data mining methods. The other measure which was evaluated for effectiveness is classification accuracy, sensitivity, specificity, positive predictive value and negative predictive value [43], [44].

III. RESULTS AND DISCUSSION

Table II demonstrates the performance of the prediction model without feature selection. While Table III summarises the performance of the prediction model with feature selection. Figures 3 and 4 illustrate the classification accuracy, sensitivity, and specificity. The results also indicated an improvement in the classification accuracy of the classifier on the medical dataset because of the decreasing number of features. When the number of features reduces, the performance affects the specificity. The classification accuracy was above 80% for Parkinson's, Heart Disease and Statlog (Heart) after feature selection through GA but not on the Breast Cancer and Hepatitis dataset. Both the datasets had the lowest number of original features compared to the rest of the dataset. The classification accuracy was relatively higher for the Parkinson's, Heart Disease and Statlog (Heart) dataset. It was evident that the overall classification accuracy improved with a reduced number of feature subsets. The assumption was made because the classification accuracy obtained before the application of feature selection was lesser than 60%. The lesser than 60% of classification accuracy is achieved on all the datasets regarding the original number of features on the datasets.

TABLE II
PERFORMANCE OF PREDICTION MODEL WITHOUT FEATURE SELECTION

Dataset	Accuracy (%)	Sensitivity (%)	Specificity (%)
Breast Cancer	55.97	53.50	60.75
Parkinson's	58.17	50.81	47.15
Heart Disease	56.02	43.22	53.16
Statlog (Heart)	52.14	59.44	67.02
Hepatitis	57.80	60.13	69.59

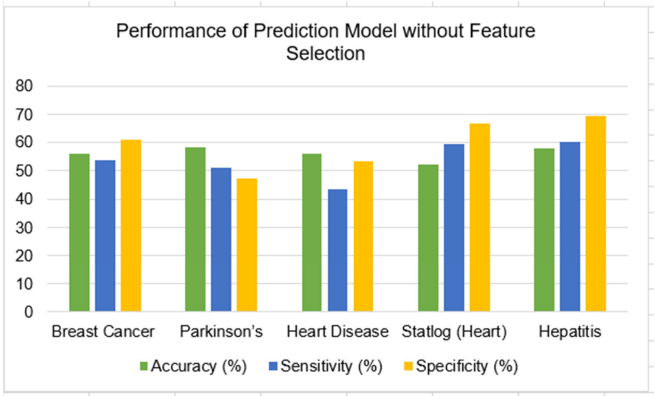

Fig. 3. The performance of the prediction model without feature selection.

TABLE III
PERFORMANCE OF PREDICTION MODEL WITH FEATURE SELECTION

Dataset	Accuracy (%)	Sensitivity (%)	Specificity (%)
Breast Cancer	72.71	85.71	71.42
Parkinson's	88.36	88.70	67.31
Heart Disease	86.73	81.81	59.51
Statlog (Heart)	85.48	82.05	67.51
Hepatitis	76.95	89.68	63.12

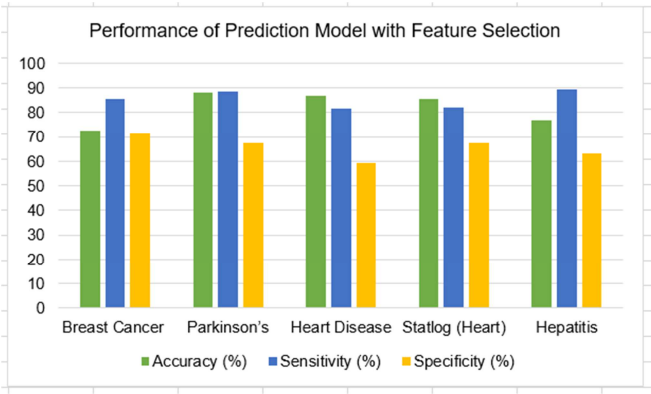

Fig. 4. The performance of the prediction model with GA as feature selection

IV. Conclusions

The proposed approach using medical datasets in this study indicated significant improvements in the classification accuracy with the feature selection. The reduction in features improved the performance in terms of classification accuracy with the correct feature selection. The specificity recorded for both with and without the feature selection approach was relatively within the range of 50% – 60%. The sensitivity increased with the reduction of features. With this proposed prediction method with GA as feature selection and SVM as the classifier, medical practitioners would be able to diagnose patient's diseases more accurately. Despite having good prediction accuracy with the feature selection algorithms, the computation time taken with GA was high and there is evidence of the said algorithm being trapped in local optima. Therefore, our future directions are to enhance the predictions using hybrid feature selection algorithms resolving the issues addressed

Acknowledgment

This work is supported by the Research University Grant (RUI)(1011/PKOMP/8014076) by Universiti Sains Malaysia (USM).

References

[1] N. Lavrač, "Selected techniques for data mining in medicine," *Artif. Intell. Med.*, vol. 16, no. 1, pp. 3–23, 1999.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, pp. 37–54, 1996.

[3] I. Yoo *et al.*, "Data mining in healthcare and biomedicine: A survey of the literature," *J. Med. Syst.*, vol. 36, no. 4, pp. 2431–2448, 2012.

[4] A. Sheikhtaheri, F. Sadoughi, and Z. Hashemi Dehaghi, "Developing and using expert systems and neural networks in medicine: A review on benefits and challenges," *J. Med. Syst.*, vol. 38, no. 9, 2014.

[5] J.-J. J. Yang *et al.*, "Emerging information technologies for enhanced healthcare," *Comput. Ind.*, vol. 69, no. 0, pp. 3–11, 2015.

[6] B. Liao *et al.*, "for High-Throughput Data Analysis," vol. 12, no. 6, pp. 1374–1384, 2015.

[7] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, no. 4 PART 1, pp. 1432–1462, 2014.

[8] I. H. Osman and J. P. Kelly, "Meta-Heuristics: An Overview," in *Meta-Heuristics*, Boston, MA: Springer US, 1996, pp. 1–21.

[9] H. Salem, G. Attiya, and N. El-Fishawy, "Classification of human cancer diseases by gene expression profiles," *Appl. Soft Comput. J.*, vol. 50, pp. 124–134, 2017.

[10] A. K. Paul, P. C. Shill, M. R. I. Rabin, and M. A. H. Akhand, "Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease," *2016 5th Int. Conf. Informatics, Electron. Vision, ICIEV 2016*, pp. 145–150, 2016.

[11] C. Dua, Dheeru and Graff, "{UCI} Machine Learning Repository." University of California, Irvine, School of Information and Computer Sciences, 2017.

[12] S. Chatterjee, S. Hore, and N. Dey, "Dengue Fever Classification Using Gene Expression Data: A PSO Based Artificial Neural Network Approach," vol. 515, pp. 331–341, 2017.

[13] M. K. Shahsavari, H. Rashidi, and H. R. Bakhsh, "Efficient classification of Parkinson's disease using extreme learning machine and hybrid particle swarm optimization," *2016 4th Int. Conf. Control. Instrumentation, Autom. ICCIA 2016*, no. January, pp. 148–154, 2016.

[14] P. Shunmugapriya and S. Kanmani, "A hybrid algorithm using ant and bee colony optimization for feature selection and classification (AC-ABC Hybrid)," *Swarm Evol. Comput.*, vol. 36, pp. 27–36, 2017.

[15] B. Subanya and R. R. Rajalaxmi, "Feature selection using artificial bee colony for cardiovascular disease classification," *2014 Int. Conf. Electron. Commun. Syst. ICECS 2014*, pp. 1–6, 2014.

[16] M. Kantardzic, *Data mining: concepts, methods and algorithms*. Wiley-IEEE Press, 2003.

[17] D. Prilutsky, B. Rogachev, R. S. Marks, L. Lobel, and M. Last, "Classification of infectious diseases based on chemiluminescent signatures of phagocytes in whole blood," *Artif. Intell. Med.*, vol. 52, no. 3, pp. 153–163, 2011.

[18] B. Samanta *et al.*, "Prediction of periventricular leukomalacia. Part I: Selection of hemodynamic features using logistic regression and decision tree algorithms," *Artif. Intell. Med.*, vol. 46, no. 3, pp. 201–215, 2009.

[19] T. M. Lehmann *et al.*, "Automatic categorization of medical images for content-based retrieval and data mining.," *Comput. Med. Imaging Graph.*, vol. 29, no. 2–3, pp. 143–55, 2005.

[20] R. Liu, Y. Chen, L. Jiao, and Y. Li, "A particle swarm optimization based simultaneous learning framework for clustering and classification," *Pattern Recognit.*, vol. 47, no. 6, pp. 2143–2152, 2014.

[21] T. Hong, K. Lin, and S. Wang, "Fuzzy Data Mining for Interesting Generalized Association Rules," *Fuzzy Sets Syst.*, vol. 138, pp. 255–269, 2003.

[22] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical Correlation Analysis: An Overview with Application to Learning Methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[23] B. Liu, Y. Xiao, L. Cao, Z. Hao, and F. Deng, "SVDD-based outlier detection on uncertain data," *Knowl. Inf. Syst.*, vol. 34, no. 3, pp. 597–618, 2013.

[24] N. Khateeb and M. Usman, "Efficient Heart Disease Prediction System using K-Nearest Neighbor Classification Technique," pp. 21–26, 2018.

[25] E. K. Hashi, M. S. Uz Zaman, and M. R. Hasan, "An expert clinical decision support system to predict disease using classification techniques," *ECCE 2017 - Int. Conf. Electr. Comput. Commun. Eng.*, pp. 396–400, 2017.

[26] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient's health parameters," *J. Telecommun. Electron. Comput. Eng.*, vol. 8, no. 12, pp. 59–65, 2016.

[27] Z. Rustam, D. A. Utami, R. Hidayat, J. Pandelaki, and W. A. Nugroho, "Hybrid Preprocessing Method for Support Vector Machine for Classification of Imbalanced Cerebral Infarction Datasets," vol. 9, no. 2, pp. 685–691, 2019.

[28] K. Shankar, S. K. Lakshmanaprabu, D. Gupta, A. Maseleno, and V. H. C. de Albuquerque, "Optimal feature-based multi-kernel SVM approach for thyroid disease classification," *J. Supercomput.*, pp. 1–16, 2018.

[29] S. Vijayarani and S. Dhayanand, "Kidney Disease Prediction Using SVM and ANN Algorithms," *Int. J. Comput. Bus. Res. ISSN (Online*, vol. 6, no. 2, pp. 2229–6166, 2015.

[30] J. Thomas and R. T. Princy, "Human heart disease prediction system using data mining techniques," *Proc. IEEE Int. Conf. Circuit, Power Comput. Technol. ICCPCT 2016*, pp. 1–5, 2016.

[31] M. Sultana, A. Haider, and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," *2016 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol. iCEEiCT 2016*, 2017.

[32] T. R. Baitharu and S. K. Pani, "Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset," *Procedia Comput. Sci.*, vol. 85, no. Cms, pp. 862–870, 2016.

[33] D. Paul, R. Su, M. Romain, V. Sébastien, V. Pierre, and G. Isabelle,

"Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier," *Comput. Med. Imaging Graph.*, vol. 60, pp. 42–49, 2017.

[34] M. Kumar and M. Kumar, "International Journal of Computer Science and Mobile Computing Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm," *Int. J. Comput. Sci. Mob. Comput.*, vol. 5, no. 2, pp. 24–33, 2016.

[35] W. Husain, L. K. L. K. Xin, N. Abdul Rashid, N. Jothi, N. A. Rashid, and N. Jothi, "Predicting Generalized Anxiety Disorder Among Women Using Random Forest Approach," in *2016 3rd International Conference On Computer And Information Sciences (ICCOINS)*, 2016, pp. 42–47.

[36] M. Lichman, K. Bache, and M. Lichman, "UCI machine learning repository," 2013. .

[37] T. Santhanam and M. S. Padmavathi, "Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis," *Procedia Comput. Sci.*, vol. 47, pp. 76–83, 2015.

[38] S. Maldonado, R. Weber, and F. Famili, "Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines," *Inf. Sci. (Ny).*, vol. 286, pp. 228–246, 2014.

[39] J. Kamruzzaman, S. Lim, I. Gondal, and R. Begg, "Gene selection and classification of human lymphoma from microarray data," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3745 LNBI, pp. 379–390, 2005.

[40] W. S. Noble, "Support vector machine applications in computational biology," *Kernel Methods Comput. Biol.*, 2004.

[41] D. Zhang, W. Zuo, D. Zhang, and H. Zhang, "Time series classification using support vector machine with Gaussian elastic metric kernel," in *Proceedings - International Conference on Pattern Recognition*, 2010.

[42] R. Caruana and a. Niculescu-Mizil, "Data mining in metric space: an empirical analysis of supervised learning performance criteria," *Proc. tenth ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pp. 69–78, 2004.

[43] C.-L. Huang, H.-C. Liao, and M.-C. Chen, "Prediction model building and feature selection with support vector machines in breast cancer diagnosis," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 578–587, 2008.

[44] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Syst. Appl.*, vol. 36, no. 2 PART 2, pp. 3240–3247, 2009.