

Received May 8, 2021, accepted May 25, 2021, date of publication June 11, 2021, date of current version June 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3088613

Review of Feature Selection, Dimensionality Reduction and Classification for Chronic Disease Diagnosis

AFNAN M. ALHASSAN^{1,2} AND WAN MOHD NAZMEE WAN ZAINON¹

¹School of Computer Science, Universiti Sains Malaysia, George Town 11800, Malaysia

²College of Computing and Information Technology, Shaqra University, Shaqra 11961, Saudi Arabia

Corresponding authors: Afnan M. Alhassan (aalhassan@su.edu.sa) and Wan Mohd Nazmee Wan Zainon (nazmee@usm.my)

ABSTRACT The early diagnosis of chronic diseases plays a vital role in the field of healthcare communities and biomedical, where it is necessary for detecting the disease at an initial phase to reduce the death rate. This paper investigates the use of feature selection, dimensionality reduction and classification techniques to predict and diagnose chronic disease. The appropriate selection of attributes plays a crucial role in improving the classification accuracy of the diagnosis systems. Additionally, dimensionality reduction techniques effectively improve the overall performance of the machine learning algorithms. On chronic disease databases, the classification techniques deliver efficient predictive results by developing intelligent, adaptive and automated system. Parallel and adaptive classification techniques are also analyzed in chronic disease diagnosis which is used to stimulate the classification procedure and to improve the computational cost and time. This survey article represents the overview of feature selection, dimensionality reduction and classification techniques and their inherent benefits and drawbacks.


INDEX TERMS Adaptive classification, chronic disease, dimensionality reduction, feature selection, parallel classification.

I. INTRODUCTION

In recent decades, chronic disease is the biggest threats to human life, which is essential to diagnose and predict chronic disease prior to reducing the mortality rate. The leading chronic disease includes Parkinson's, heart disease, lung cancer, Hepatitis, breast cancer, chronic kidney disease, etc. In the medical field, maintaining clinical databases is a crucial task that consists of several features and diagnostics related to chronic disease [1]. The data stored in the medical databases consist of redundant data and missing values, so it is necessary to diminish the data before employing data mining algorithms. If the data is consistent and free from noise, chronic disease diagnosis becomes easier and quicker. Feature selection and dimensionality reduction are effective data pre-processing techniques to reduce the data dimension [2]. In the field of healthcare communities, it is important to find the risk factors related to chronic disease. The relevant feature diagnosis helps in removing the redundant attributes and irrelevant information from the chronic disease databases, which

gives a good and quick predictive result. In data mining, the diagnosis and classification techniques utilize training data for developing a model and then the corresponding model is employed on testing data to attain better predictive results. Many classification techniques are employed on the disease dataset for the early diagnosis of chronic disease [3]. The healthcare data includes information related to pharmacy, doctor prescription, clinical and diagnostic test reports of an individual, posts on social media etc. So, it is essential to develop a novel classifier in chronic disease diagnosis that simplifies and expedites the diagnosis process. The chronic disease diagnosis system is utilized as a tool to control the disease that helps the medical profession and clinician to deliver 24/7 healthcare services and monitors the patients' health effectively.

This survey paper is prepared as follows. In Section 2, a short explanation of databases used in chronic disease diagnosis is presented. In Section 3, the description of feature selection and dimensionality reduction for chronic disease diagnosis is mentioned. A tabular study is given for various feature selection techniques that include their characteristics, benefits and drawbacks. A review on traditional, parallel,

The associate editor coordinating the review of this manuscript and approving it for publication was Yin Zhang .

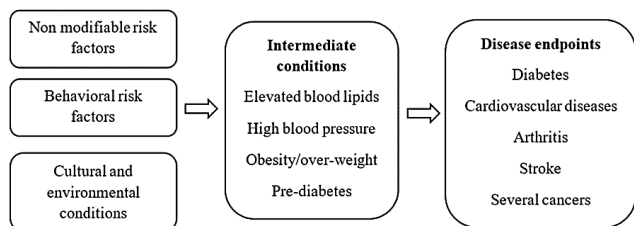


FIGURE 1. Graphical representation of risk behaviours responsible for chronic disease.

TABLE 1. List of popular chronic disease databases.

Chronic disease databases	Attribute type	Number of instances	Number of attributes
Parkinson’s database [5]	Real	197	23
Lung cancer database [6]	Integer	32	56
Hepatitis database [7]	Integer, real, categorical	155	19
Cleveland heart database [8]	Integer, real, categorical	303	75
Mammographic mass database [9]	Integer	961	6
Breast cancer Wisconsin database [10]	Real	569	32
Pima Indians diabetes database [11]	Integer, real	768	8
Arrhythmia database [12]	Integer, real, categorical	452	279
Statlog heart database [13]	Real, categorical	270	13
Chronic kidney disease dataset [14]	Real	400	25

adaptive classification techniques for chronic disease diagnosis is given in Section 4. In section 5, the performance measures used in chronic disease diagnosis is briefly explained. The conclusion of the survey paper is given in Section 6.

II. DATABASE AVAILABLE IN CHRONIC DISEASE DIAGNOSIS

In recent decades, chronic disease is a long-lasting illness that has a huge impact on people health. The most frequent chronic disease are hyperlipidemic arthritis, coronary artery diseases, colon cancer, asthma, heart disease, haemophilia, chronic kidney disease, chronic respiratory disease, etc. [4]. The risk behaviours responsible for chronic disease are; hypertension (raised blood pressure), tobacco use, raised cholesterol, unhealthy diet, physical inactivity, and harmful use of alcohol. The risk behaviours responsible for chronic disease is graphically denoted in figure 1.

In the field of healthcare communities and biomedical, the accurate diagnosis of chronic disease significantly reduces the mortality rate. In recent decades, several databases are available for chronic disease diagnosis in that a few popular databases are denoted in table 1. The selection of the database should be appropriate because the intent operation (feature selection or dimensionality reduction and classification) depends on the selected database.

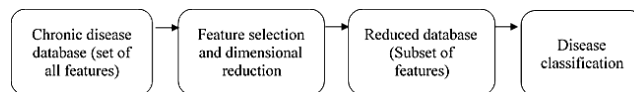


FIGURE 2. Feature selection and dimensional reduction in data mining.

III. FEATURE SELECTION AND DIMENSIONALITY REDUCTION TECHNIQUES IN CHRONIC DISEASE DIAGNOSIS

In data mining, feature selection and dimensionality reduction are the most commonly used pre-processing techniques that minimize the data by reducing the irrelevant attributes in the databases. It also facilitates better data visualization, improves data understandability, and also minimizes the training time of classification techniques in chronic disease diagnosis. However, chronic disease diagnosis occurs in many circumstances such as thalassemia, diabetes, heart disease, high blood pressure, strokes, etc. [2]. In data mining, the feature selection techniques are broadly classified into four types such as embedded techniques, wrapper techniques, filtering techniques, and hybrid techniques [15]. The feature selection techniques remove the repeated and unrelated features from the original database to improve the classification accuracy. The feature selection process in data mining is graphically represented in figure 2.

A. FILTERING TECHNIQUES

The filtering techniques independently select the features from the databases. Filtering is one of the oldest procedure which classifies the attributes based on the certain evaluation criteria since it does not rely on the classification techniques. The filtering techniques are better in eliminating the redundant, constant, duplicate, irrelevant and correlated features. Though, the selected features are used in any machine learning techniques, where it is computationally inexpensive. Presently, there are two types of filtering techniques are available such as univariate and multivariate. The univariate filtering techniques rank the individual features based on certain criteria and treat each feature independently and individually in the feature space. Finally, the highest-ranking feature is selected according to criteria. The univariate techniques select the redundant variables, where it does not consider the relationship between features that is a major concern in univariate techniques.

On the other hand, the multivariate techniques evaluate the whole feature space and able to handle redundant, duplicated, and correlated features. Battiti [16] implemented a feature selection technique based on the concept of mutual information. In this study, the feature selection technique extracts the features, which have maximum mutual information. Additionally, Bannasar *et al.* [17] developed two new non-linear feature selection techniques named as Joint Mutual Information Maximization (JMIM) and Normalized JMIM (NJMIM), where these two techniques utilize mutual information and “a maximum of the minimum criterion” to extract the features from the UCI repository databases.

Chormunge and Jena [18] implemented a novel correlation and clustering-based feature selection technique to reduce the dimensionality problems in data mining tasks. Initially, k means clustering algorithm is utilized for eliminating the irrelevant features and then the non-redundant features are selected by correlation measure for every cluster. Next, naïve Bayes is applied for classification on microarray and text databases. Cigdem *et al.* [19] utilized a correlation-based feature selection technique to rank the features, and the top-ranked features are selected by using the fisher criterion. Then, different classification techniques are used for classifying bipolar disorders on 3D magnetic resonance imaging.

B. WRAPPER TECHNIQUES

The wrapper techniques extract the relevant features based on the performance of the classifier. It solves the real optimization problems significantly, but computationally expensive compared to the filtering techniques. Wrapper techniques work based on greedy search algorithms, where all the combination of features are evaluated and select the combination of features that delivers a better result for machine or deep learning algorithms. The wrapper techniques include two major advantages; (i) effectively finds the optimal feature subsets and (ii) detects the interaction between the variables. Generally, the wrapper techniques result in better predictive accuracy related to the filtering techniques. For feature selection, the wrapper techniques are divided into three categories such as exhaustive feature selection, step forward and step backward feature selection. Lee *et al.* [20] used a wrapper based feature selection technique to effectively deal with the high volume and multi-dimensionality data generated from the medical health care systems. In this paper, a new bagging C4.5 algorithm based wrapper feature selection is developed for clinical decision making in the healthcare and medical fields. Jadhav *et al.* [21] implemented a novel feature selection technique named as information gain directed feature selection technique, which ranks the features based on the information gain and selects the top features utilizing genetic algorithm. Apolloni *et al.* [22] introduced a wrapper feature selection technique based on binary differential evolution algorithm to diminish the dimension of microarray data.

Sawhney *et al.* [23] combined a penalty function with the existing fitness function of the binary firefly algorithm for reducing the feature sets. Selected optimal feature subset effectively increases the classification accuracy of random forest classifier for diagnosing cervical, breast, cervical, liver cancer and hepatocellular carcinoma. Mafarja and Mirjalili [24] utilized a whale optimization algorithm to minimize the dimension of the input features. The experimental outcome showed that the developed algorithm delivers better results compared to other existing techniques like Particle Swarm Optimization (PSO) [25], [26], and genetic algorithm [27]. Additionally, Shen *et al.* [28] and Balasubramanian and Marichamy [29] used a fruit fly optimization algorithm to minimize the dimension of the features. Empirical results show that the developed algorithm obtained more

appropriate model parameters, which generates high classification accuracy. Fruit fly optimization algorithm is regarded as an effective clinical tool for medical decision-making.

C. EMBEDDED TECHNIQUES

The feature selection is integrated as a part of a learning algorithm in embedded techniques, where it combines the quality of both wrapper and filtering techniques. The common examples of embedded techniques are lasso and ridge regression, which has inbuilt penalization functions for reducing the over-fitting problem. Lasso regression technique performs L1 regularization that adds penalty equivalent to the magnitude of coefficients. Whereas, ridge regression technique performs L2 regularization that adds penalty equivalent for squaring the magnitude of coefficients. Liu *et al.* [30] implemented an effective neighbourhood embedding technique for unsupervised feature selection. Initially, a locally linear embedding algorithm is used to obtain the feature weight matrix. Next, the L1 normalization technique is employed for suppressing the impact of noises and outliers in the datasets. The extensive experiment shows that the developed technique attained better performance on benchmark datasets compared to the existing unsupervised feature selection techniques.

Tao *et al.* [31] introduced a new multi-source adaptation embedding technique for feature selection by exploiting the correlation information using L2 normalization, trace norm and l-norm regularizations. Also, the developed adaptation embedding technique is robust to outliers or noises that existed in domains and preserves the original geometrical structure information by applying a sparse regression method and graph embedding via L1 and L2 norm minimization. Wang and Zhu [32] used sparsity preserving and neighbourhood embedding feature selection techniques to handle the large volume of data. The developed feature selection techniques were investigated on eight publicly available datasets from the machine learning repository. The extensive experiment shows that the developed techniques achieved better performance in feature selection related to existing techniques.

D. HYBRID TECHNIQUES

In recent decades, hybrid techniques are extensively utilized by researchers for feature selection. Hybrid techniques combine two or more feature selection techniques to achieve optimal results. Usually, hybrid techniques achieve better computational efficiency compared to filtering techniques and high classification accuracy compared to the wrapper and embedded techniques. Baliarsingh *et al.* [33] combined both emperor penguin optimizer and social engineering optimizer to select the relevant attributes of lung cancer, ovarian cancer, colon tumour, and leukaemia cancer. Next, a Support Vector Machine (SVM) classifier is applied to classify the relevant genes. Further, AlMuhaideb and Menai [34] combined Artificial Bee Colony (ABC) and Ant Colony Optimizer (ACO) to optimize the medical data. Experimental results on real-time and benchmark databases show the efficacy of the hybrid

technique. With the help of the hybrid technique, classification models obtained better predictive accuracy.

Jayaraman and Sultana [35] has combined PSO and Gravitational Cuckoo Search Algorithm (GCSA) to manage the features which present in the heart disease classification system. At first, the data is collected from heart disease database UCI repository. The collected data is the high dimension which is difficult to process and it reduces the efficiency of the heart disease diagnosis system. So, the dimension of the data is reduced by the behaviour of PSO and GCSA. The selected features are fed to associative memory classifier for data classification. Khourdifi and Bahaj [36] combined PSO and ACO algorithms to enhance the quality of heart disease classification. The selected features are fed to different classifiers for data classification. The extensive experiment shows that the hybrid optimization technique significantly improves the diagnostic accuracy of medical databases.

Bharti and Mittal [37] developed a hybrid feature selection based feature fusion system for generating optimal feature subsets to classify liver ultrasound images into 4 classes; cirrhosis, normal, hepatocellular carcinomas, and chronic. The hybrid feature selection eliminates duplicate and irrelevant feature subsets that significantly enhances the performance of classification. Jain and Singh [38] introduced a novel adaptive classification system to diagnose chronic disease. Initially, reliefF and Principal Component Analysis (PCA) are utilized for feature optimization and SVM classifier is applied for data classification. In this study, an effective parameter optimization technique is employed in SVM classifier to achieve higher classification accuracy. For evaluating the system performance, well-known disease databases (ovarian cancer, prostate cancer, leukaemia cancer, lung cancer, colon dataset, and heart disease) are used for medical diagnosis. Table 2 states the advantage and disadvantage of feature selection techniques used in chronic disease diagnosis.

E. DIMENSIONAL REDUCTION TECHNIQUES

The dimensional reduction techniques convert the higher dimension data space into lower dimension data space. While working with higher dimension data space, raw data are often sparse and it leads to “curse of dimensionality” concern and computationally intractable. The dimensional reduction techniques are commonly applied in the medical field, bioinformatics, etc., and it is majorly categorized into two types such as linear and non-linear techniques. The most common techniques applied in chronic disease diagnosis are PCA, Linear Discriminant Analysis (LDA), Generalized Discriminant Analysis (GDA), etc. Muhammad *et al.* [39] used the PCA technique for variable based classification of diabetes mellitus. Diagnosis of chronic disease is a challenging task in the field of health care, where several data mining techniques are employed for decision making. Banu [40] applied LDA technique to classify the hypothyroid disease. Shahbazi and Asl [41] utilized GDA as a feature selection technique to minimize the number of features and overlap of the samples in the feature space. Additionally, K-Nearest Neighbor (KNN)

TABLE 2. Advantage and disadvantage of feature selection techniques utilized in chronic disease diagnosis.

Feature selection techniques	Type	Advantage	Disadvantage
Chi-square	Filter	Computationally effective	Chi-square distance does not attain better results in the non-linear database.
Correlation	Filter	It is a multivariate filtering technique that ranks the features based on the correlation heuristic evaluation function.	Incapable of finding strong interactions.
ReliefF	Filter	Effectively deals with the noisy and incomplete data, and also handles the multiclass concerns.	In a nonlinear database, reliefF fails to remove the redundant features.
Fisher score [19]	Filter	Measures the relevance of feature sub-sets effectively.	Does not consider the dependency of one feature over other features.
Markov blanket filter	Filter	Robust against overfitting problems related to other techniques.	Ignores the interaction among classifiers.
Genetic algorithm [21] [27]	Wrapper	Effectively captures the feature interaction and redundancy.	Increases the system complexity, requires more central processing unit time and consumes more memory to run.
PSO [25-26]	Wrapper		Consumes more time to converge.
Backward elimination technique	Wrapper	Effectively detects the dependencies among feature subsets.	Classifiers need to run many times to assess the quality of features.
Sequential forward selection	Wrapper	Simple and avoids overfitting problem effectively.	The selected subsets are specific to the classifiers under consideration.
Stepwise regression technique	Wrapper	Has better classifier interaction and identifies the most relevant features.	Needs more computational resources and also prone to overfitting problems on small chronic databases.
Weighted naïve Bayes	Embedded	Less prone to overfitting problems compared to the wrapper techniques.	Selection of relevant features completely depends on the consideration of the classifier.
Artificial neural network	Embedded	Computationally inexpensive and has better classifier interaction.	Poor generality
Sequential forward selection	Embedded	Better use of chronic databases and delivers a fast solution.	Computationally expensive related to filtering techniques.

is applied to analyze the performance of feature set in heart disease classification.

IV. CLASSIFICATION TECHNIQUES USED IN CHRONIC DISEASE DIAGNOSIS

In data mining, the classification includes unsupervised and supervised techniques that work based on artificial intelligence, mathematics, probability distributions, and

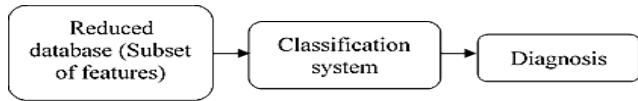


FIGURE 3. Classification process in data mining.

TABLE 3. Merits and demerits of classification techniques utilized in chronic disease diagnosis.

Classifier	Merits	Demerits
SVM	Robust to noise, less overfitting and delivers versatility with kernel concept.	Computationally expensive and lack of transparency of results.
C4.5	Consumes less memory space for large datasets and works well in both continuous and categorical attributes.	Problems like insignificant branches and overfitting.
KNN	Easy to understand the structure and unstructured data, and robust to noise.	Runs slowly and has memory limitation.
Bayesian networks	Extensively supports the missing medical data.	Initial knowledge is required for a large number of probabilities, and highly expensive.
Neural Networks	Gives arbitrarily complex relationship between the input and predicted value, and delivers high diagnosis accuracy.	Consumes more training time if the data is not fitted well.
Naïve Bayes	Computationally low cost and works well on both nominal and numerical data.	Need an enormous amount of medical data to obtain better diagnosis results.

statistics [3]. Usually, classification techniques are utilized to represent the descriptive analysis of data items and to predict the group membership for data items for effective decision making.

In data mining, several researchers used dissimilar classification techniques for chronic disease diagnosis to obtain better diagnosis accuracy, and diagnostic results. The classifiers like SVM [42], [43], random forest [44], KNN [45], Adaboost [46] etc. are utilized for diagnosis and prognosis of chronic disease. Figure 3 indicates the classification process applied to pre-processed medical data to achieve predictive results.

In most scientific applications, automated and intelligible systems are needed for better diagnosis and diagnosis of chronic disease such as lung cancer, Hepatitis, chronic kidney diseases, Parkinson, etc. Whereas, most of the conventional systems are ineffective that reduces the success rate and increases the computational time or decision making time. So, an adaptive classification technique is needed for chronic disease diagnosis that predicts the diseases precisely. In addition, parallel classification techniques are also used for enhancing the diagnosis results. Table 3 states the merits and demerits of classification techniques used in chronic disease diagnosis.

Mohapatra *et al.* [47] used cuckoo search and Extreme Machine Learning (EML) technique for classifying four benchmark datasets; hepatitis, diabetes, Bupa and breast cancer. Experimental result demonstrates the efficacy of the developed model in terms of sensitivity, f-score, specificity, overall accuracy, Gmean, confusion matrix, and normalization value. Polat [48] used attribute weighted techniques for

data pre-processing and classification of heart, Parkinson, Pima Indians and thoracic surgery medical databases. In order to reduce the variance value within the class, three clustering algorithms (mean shifted, k-means and fuzzy C means) are employed in this study. After the attribute weighting process, SVM, KNN, random forest and LDA are utilized for classifying the imbalanced medical databases. The extensive experiment shows that the developed attribute weighting techniques achieved higher classification accuracy compared to random sub-sampling techniques. Seera and Lim [49] developed a hybrid intelligent system for medical data classification on the basis of random forest, regression tree and fuzzy min-max neural network. The performance of the developed intelligent system is analyzed on liver disorders, Pima Indians diabetes, and breast cancer Wisconsin databases.

Jaganathan and Kuppuchamy [50] used feature selection techniques like mean selection, neural network and half selection to diminish the amount of redundant, irrelevant and unnecessary features in Cleveland heart disease, statlog, Wisconsin breast cancer, hepatitis, and Pima Indians diabetes datasets. Next, the selected optimal features are fed to radial basis function for classifying the medical data. Nalband *et al.* [51] used a genetic algorithm and apriori algorithm to select the significant features from the extracted feature vectors. Then, random forest and Least Square SVM (LS-SVM) classification techniques are applied to precisely distinguish the abnormal and normal vibroarthographic signals. Additionally, Cheruku *et al.* [52] introduced a novel hybrid decision support system on the basis of bat optimization algorithm and rough set theory. The hybrid system effectively reduces the redundant features by generating fuzzy rules. Then, the selected features are fed to the fuzzy rule-based classification technique to classify the diseases. In this literature, the developed system performance is investigated on Wisconsin breast cancer, Pima Indians diabetes, iris and Cleveland heart disease datasets in light of g-measure, accuracy, sensitivity and specificity.

A. PARALLEL CLASSIFICATION TECHNIQUES USED IN CHRONIC DISEASE DIAGNOSIS

The traditional classification techniques are ineffective to process massive or unstructured medical data. So, the structure of the present health care system is improved by utilizing data analytics. Currently, big data analytics is done by employing several technologies and tools like Hadoop, map-reduce programming, etc. Whereas, the parallel classification techniques have great potential for improving the predictive accuracy of diagnostic systems. The parallel classification techniques are effective in clinical decisions for chronic disease. Figure 4 indicates the parallel classification techniques applied to the pre-processed data to achieve a better predictive result by Åström and Koker [5].

In this literature, two neural networks are used to reduce the possibility of error decision. For the final decision, the output of each neural network is analyzed using a rule-based system. The developed parallel network effectively increases the

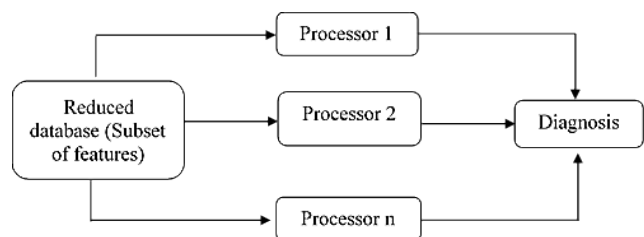


FIGURE 4. Parallel classification process in data mining.

robustness of chronic disease diagnosis. The simulation result shows that the parallel network showed 8.4% improvement in Parkinson disease diagnosis compared to a single unique network. Additionally, Shrivastava *et al.* [53] developed a parallel SVM technique to predict the diabetes chances in human on a survey database. This research paper predicts the future possibilities of diabetes for a person. The survey dataset is high dimensional in nature, so the conventional SVM is ineffective. In order to handle the large number of parameters, parallel SVM concept is introduced in this literature. The parallel SVM distributes the parameters in different machines that reduce the processing power, computational complexity and memory space. The simulation result shows that the parallel SVM reduces 1/3 of computational time compared to the conventional SVM technique.

B. ADAPTIVE CLASSIFICATION TECHNIQUES USED IN CHRONIC DISEASE DIAGNOSIS

In recent times, classification techniques are used in the advanced version to generate categories for the attributes in medical datasets for better classification. The advanced version of the adaptive classifiers is a combination of both clustering and classification techniques. The adaptive systems significantly improve the success rate and assist the medical professionals and doctors to take an effective clinical decision in chronic disease diagnosis.

Jain and Singh [38] developed an adaptive SVM classification technique to diagnose chronic disease. In this literature article, hybrid PCA and reliefF techniques are used for parameter optimization in SVM to achieve high classification accuracy. In order to investigate the developed system performance, nine chronic disease datasets are used for medical diagnosis. The extensive experiment shows that the developed system drastically diminishes the dimension of the databases, enhances the effectiveness of the classifier and decreases the computational time and cost. Yu *et al.* [54] introduced a Hybrid Adaptive Ensemble Learning (HAEL) system for microarray data classification. The developed system performs well on KEEL and real-time microarray databases. Dennis and Muthukrishnan [55] utilized the adaptive genetic fuzzy system to optimize member functions and rules for the medical data classification process. In addition Chandra and Kaur [56] used an adaptive KNN classification technique for liver cirrhosis and lymph node diagnosis. Alhassan and Zainon [57] presented a new approach which Taylor Bird Swarm Algorithm Based on Deep Belief Net-

TABLE 4. Comparison of classification techniques on different datasets.

Classification techniques	Dataset	Performance measure
Improved cuckoo search and EML algorithm [47]	Breast cancer, Bupa, Hepatitis and Diabetes datasets	Classification accuracy of 100%, 94.52%, 100%, and 92.41%
Fuzzy rule-based classification technique [52]	Wisconsin breast cancer, Pima Indians diabetes, iris and Cleveland heart disease datasets	Classification accuracy of 98.23%, 85.33%, 99.11% and 80.66%
Taylor bird swarm algorithm based on deep belief network [57]	Cleveland heart disease dataset	Classification accuracy of 93.40%

work for Heart Disease Diagnosis an approach to classify medical data for medical decision making. The developed technique achieved 93.4% of classification accuracy. The experiment shows that the proposed Taylor-BSA-DBN drastically diminishes the dimension of the databases, enhances the effectiveness of the classifier and decreases the computational time and cost comparing with SVM, and NB, and DBN techniques. The developed technique achieved 90% of classification accuracy on real-time medical databases, which is better compared to the conventional KNN technique. By comparing traditional, parallel and adaptive classification techniques, parallel classification techniques have great potential to enhance the predictive accuracy of diagnostic systems for chronic disease. The comparison of classification techniques on a different dataset is denoted in table 4.

V. PERFORMANCE MEASURES USED IN CHRONIC DISEASE CLASSIFICATION

In a medical diagnosis system, the performance of classification is analyzed by using different performance measures; precision, recall, f-measure, accuracy, specificity, Fowlkes Mallows Index (FMI), etc. Usually, these performance measures are used to analyze the performance of dissimilar classification models.

Precision: It is determined as the ratio of true positive to the sum of false positive and true positive. Among the total predicted positive observation, precision is the ratio of correctly predicted positive observation. The mathematical expression of precision is defined in equation (1).

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall: It is defined as the ratio of true positive to the sum of false negative and true positive, and it is mathematically indicated in equation (2).

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

F-measure: It is defined as the weighted harmonic mean of precision and recall, as mentioned in equation (3).

$$F - measure = \frac{2TP}{2TP + FP + FN} \tag{3}$$

Accuracy: It is a most intuitive performance measure used in chronic disease diagnosis, where it is the ratio of correctly predicted observation among the total observations. The mathematical expression of accuracy is defined in equation (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Specificity: It is defined as the proportion of negatives that are correctly identified. Specificity is also called a true negative rate that is mathematically represented in equation (5).

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

FMI: It is defined as the geometric mean between the recall and precision. FMI is mathematically denoted in equation (6).

$$FMI = \sqrt{\frac{TP}{TP + FP} \times \frac{TP}{TP + FN}} \quad (6)$$

where TP is indicated as true positive, TN is denoted as true negative, FP is represented as false positive and FN is indicated as false negative.

VI. CONCLUSION

This paper reviews the present feature selection, dimensionality reduction and classifiers for effective chronic disease diagnosis. In the medical diagnostic system, the performance measures used to estimate the performance of the classifier are also surveyed. Additionally, this article compares the benefits and drawbacks of several feature selection, dimensionality reduction and classification techniques. Moreover, the feature selection techniques are categorized in this review based on data mining tasks, search strategy, and evaluation criteria. The survey on feature selection techniques depict that the filtering techniques are very efficient and deliver better performance in identifying optimal feature subsets than the wrapper and embedded techniques. This survey article reveals that the current elevation in feature selection is hybrid techniques that are employed on chronic disease databases for eliminating noisy, redundant and insignificant features. On the other hand, numerous classification techniques are used in chronic disease diagnosis such as SVM, naïve Bayes, decision tree, neural network, random forest, etc. Related to the conventional classifiers, the adaptive and parallel classification techniques have a higher success rate and low computation time for diagnosing the chronic disease. Most of the existing studies majorly focused only on traditional classification techniques for medical diagnosis. This paper reviews the adaptive and parallel classification techniques, where it significantly improves the performance of chronic disease diagnosis by providing high diagnosis and classification accuracy. These classification techniques help healthcare professional, doctor, physician, and clinician in effective decision making for chronic disease diagnosis.

VII. RESEARCH GAP

This review paper suggested that future research in chronic disease detection should concentrate on developing new hybrid classification techniques for enhancing classification accuracy and for optimizing the computational effectiveness of the results. In recent decades, fruitful and large efforts have been performed on several classification systems. Existing works majorly focused on the usage of conventional classification systems for medical diagnosis.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

- [1] X. Zhang, H. Zhao, S. Zhang, and R. Li, "A novel deep neural network model for multi-label chronic disease prediction," *Frontiers Genet.*, vol. 10, p. 351, Apr. 2019.
- [2] K. Rajeswari, V. Vaithyanathan, and S. V. Pede, "Feature selection for classification in medical data mining," *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 2, no. 2, pp. 492–497, 2013.
- [3] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009.
- [4] T. Tchkonina and J. L. Kirkland, "Aging, cell senescence, and chronic disease: Emerging therapeutic strategies," *Jama*, vol. 320, no. 13, pp. 1319–1320, 2018.
- [5] F. Åström and R. Koker, "A parallel neural network approach to diagnosis of Parkinson's Disease," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 12470–12474, 2011.
- [6] J. Li and L. Wong, "Using rules to analyse bio-medical data: A comparison between $C_{4.5}$ and PCL," in *Proc. Int. Conf. Web-Age Inf. Manage.* Berlin, Germany: Springer, 2003, pp. 254–265.
- [7] Z. Zhou, Y. Jiang, and S. Chen, "Extracting symbolic rules from trained neural network ensembles," *AI Commun.*, vol. 16, no. 1, pp. 3–15, May 2003.
- [8] B. A. Tama, S. Im, and S. Lee, "Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble," *BioMed Res. Int.*, vol. 2020, pp. 1–10, Apr. 2020.
- [9] M. Elter, R. Schulz-Wendtland, and T. Wittenberg, "The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process," *Med. Phys.*, vol. 34, no. 11, pp. 4164–4172, Oct. 2007.
- [10] V. J. Kadam, S. M. Jadhav, and K. Vijayakumar, "Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression," *J. Med. Syst.*, vol. 43, no. 8, p. 263, Aug. 2019.
- [11] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Syst. Appl.*, vol. 35, nos. 1–2, pp. 82–89, Jul. 2008.
- [12] S. Mishra, H. K. Tripathy, P. K. Mallick, A. K. Bhoi, and P. Barsocchi, "EAGA-MLP—An enhanced and adaptive hybrid classification model for diabetes diagnosis," *Sensors*, vol. 20, no. 14, p. 4036, Jul. 2020.
- [13] J. Bohacik and M. Zabovsky, "Naive Bayes for statlog heart database with consideration of data specifics," in *Proc. IEEE 14th Int. Sci. Conf. Informat.*, Nov. 2017, pp. 35–39.
- [14] M. Elhoseny, K. Shankar, and J. Uthayakumar, "Intelligent diagnostic diagnosis and classification system for chronic kidney disease," *Sci. Rep.*, vol. 9, no. 1, pp. 1–14, 2019.
- [15] K. Sutha and J. J. Tamilselvi, "A review of feature selection algorithms for data mining techniques," *Int. J. Comput. Sci. Eng.*, vol. 7, no. 6, p. 63, 2015.
- [16] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.
- [17] M. Bannasar, Y. Hicks, and R. Setchi, "Feature selection using joint mutual information maximisation," *Expert Syst. Appl.*, vol. 42, no. 22, pp. 8520–8532, Dec. 2015.
- [18] S. Chormunge and S. Jena, "Correlation based feature selection with clustering for high dimensional data," *J. Electr. Syst. Inf. Technol.*, vol. 5, no. 3, pp. 542–549, Dec. 2018.

- [19] O. Cigdem, A. Sulucay, A. Yilmaz, K. Oguz, H. Demirel, O. Kitis, C. Eker, A. S. Gonul, and D. Unay, "Diagnosis of bipolar disease using correlation-based feature selection with different classification methods," in *Proc. Med. Technol. Congr. (TIPTEKNO)*, Oct. 2019, pp. 1–4.
- [20] S.-J. Lee, Z. Xu, T. Li, and Y. Yang, "A novel bagging C_{4.5} algorithm based on wrapper feature selection for supporting wise clinical decision making," *J. Biomed. Informat.*, vol. 78, pp. 144–155, Feb. 2018.
- [21] S. Jadhav, H. He, and K. Jenkins, "Information gain directed genetic algorithm wrapper feature selection for credit rating," *Appl. Soft Comput.*, vol. 69, pp. 541–553, Aug. 2018.
- [22] J. Apolloni, G. Leguizamón, and E. Alba, "Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments," *Appl. Soft Comput.*, vol. 38, pp. 922–932, Jan. 2016.
- [23] R. Sawhney, P. Mathur, and R. Shankar, "A firefly algorithm based wrapper-penalty feature selection method for cancer diagnosis," in *Proc. Int. Conf. Comput. Sci. Appl. Cham, Switzerland: Springer*, 2018, pp. 438–449.
- [24] M. Mafarja and S. Mirjalili, "Whale optimization approaches for wrapper feature selection," *Appl. Soft Comput.*, vol. 62, pp. 441–453, Jan. 2018.
- [25] N. Zemmam, N. Azizi, M. Sellami, S. Cheriguene, A. Ziani, M. AlDwairi, and N. Dendani, "Particle swarm optimization based swarm intelligence for active learning improvement: Application on medical data classification," *Cognitive Computation*, vol. 12, no. 5, pp. 991–1010, 2020.
- [26] K.-C. Lin and Y.-H. Hsieh, "Classification of medical datasets using SVMs with hybrid evolutionary algorithms based on endocrine-based particle swarm optimization and artificial bee colony algorithms," *J. Med. Syst.*, vol. 39, no. 10, p. 119, Oct. 2015.
- [27] P. H. Babu and E. S. Gopi, "Medical data classifications using genetic algorithm based generalized kernel linear discriminant analysis," *Procedia Comput. Sci.*, vol. 57, pp. 868–875, 2015, doi: [10.1016/j.procs.2015.07.498](https://doi.org/10.1016/j.procs.2015.07.498).
- [28] L. Shen, H. Chen, Z. Yu, W. Kang, B. Zhang, H. Li, B. Yang, and D. Liu, "Evolving support vector machines using fruit fly optimization for medical data classification," *Knowl.-Based Syst.*, vol. 96, pp. 61–75, Mar. 2016.
- [29] S. Balasubramanian and P. Marichamy, "An efficient medical data classification using oppositional fruit fly optimization and modified kernel ridge regression algorithm," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 3, pp. 3889–3899, 2021.
- [30] Y. Liu, D. Ye, W. Li, H. Wang, and Y. Gao, "Robust neighborhood embedding for unsupervised feature selection," *Knowl.-Based Syst.*, vol. 193, Apr. 2020, Art. no. 105462.
- [31] J. Tao, D. Zhou, and B. Zhu, "Multi-source adaptation embedding with feature selection by exploiting correlation information," *Knowl.-Based Syst.*, vol. 143, pp. 208–224, Mar. 2018.
- [32] S. Wang and W. Zhu, "Sparse graph embedding unsupervised feature selection," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 48, no. 3, pp. 329–341, Mar. 2018.
- [33] S. K. Baliarsingh, W. Ding, S. Vipsita, and S. Bakshi, "A memetic algorithm using emperor penguin and social engineering optimization for medical data classification," *Appl. Soft Comput.*, vol. 85, Dec. 2019, Art. no. 105773.
- [34] S. AlMuhaideb and M. E. B. Menai, "HColonies: A new hybrid meta-heuristic for medical data classification," *Int. J. Speech Technol.*, vol. 41, no. 1, pp. 282–298, Jul. 2014.
- [35] V. Jayaraman and H. P. Sultana, "Artificial gravitational cuckoo search algorithm along with particle bee optimized associative memory neural network for feature selection in heart disease classification," *J. Ambient Intell. Humanized Comput.*, pp. 1–10, 2019, doi: [10.1007/s12652-019-01193-6](https://doi.org/10.1007/s12652-019-01193-6).
- [36] Y. Khourdifi and M. Bahaj, "Heart disease diagnosis and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," *Int. J. Intell. Eng. Syst.*, vol. 12, no. 1, pp. 242–252, 2019.
- [37] P. Bharti and D. Mittal, "Hybrid feature selection-based feature fusion for liver disease classification on ultrasound images," in *Advances in Computational Techniques for Biomedical Image Analysis*. New York, NY, USA: Academic, 2020, pp. 145–164.
- [38] D. Jain and V. Singh, "A two-phase hybrid approach using feature selection and adaptive SVM for chronic disease classification," *Int. J. Comput. Appl.*, pp. 1–13, 2019, doi: [10.1080/1206212X.2019.1577534](https://doi.org/10.1080/1206212X.2019.1577534).
- [39] M. U. Muhammad, R. Jiadong, N. S. Muhammad, M. Hussain, and I. Muhammad, "Principal component analysis of categorized polytomous variable-based classification of diabetes and other chronic diseases," *Int. J. Environ. Res. Public Health*, vol. 16, no. 19, p. 3593, Sep. 2019.
- [40] G. Rasitha, "Predicting thyroid disease using linear discriminant analysis (LDA) data mining technique," *Commun. Appl. Electron.*, vol. 4, no. 1, pp. 4–6, Jan. 2016.
- [41] F. Shahbazi and B. M. Asl, "Generalized discriminant analysis for congestive heart failure risk assessment based on long-term heart rate variability," *Comput. Methods Programs Biomed.*, vol. 122, no. 2, pp. 191–198, Nov. 2015.
- [42] E. Tuba, I. Strumberger, T. Bezdan, N. Bacanin, and M. Tuba, "Classification and feature selection method for medical datasets by brain storm optimization algorithm and support vector machine," *Procedia Comput. Sci.*, vol. 162, pp. 307–315, 2019, doi: [10.1016/j.procs.2019.11.289](https://doi.org/10.1016/j.procs.2019.11.289).
- [43] M. D. de Lima, J. de Oliveira Roque e Lima, and R. M. Barbosa, "Medical data set classification using a new feature selection algorithm combined with twin-bounded support vector machine," *Med. Biol. Eng. Comput.*, vol. 58, no. 3, pp. 519–528, Mar. 2020.
- [44] M. Z. Alam, M. S. Rahman, and M. S. Rahman, "A random forest based predictor for medical data classification using feature ranking," *Inform. Med. Unlocked*, vol. 15, 2019, Art. no. 100180, doi: [10.1016/j.imu.2019.100180](https://doi.org/10.1016/j.imu.2019.100180).
- [45] R. K. Bania and A. Halder, "R-ensembl: A greedy rough set based ensemble attribute selection algorithm with kNN imputation for classification of medical data," *Comput. Methods Programs Biomed.*, vol. 184, Feb. 2020, Art. no. 105122.
- [46] K. A. Abuhasel, A. M. Ilyyasu, and C. Faticah, "A combined AdaBoost and NEWFM technique for medical data classification," in *Information Science and Applications*. Berlin, Germany: Springer, pp. 801–809, 2015.
- [47] P. Mohapatra, S. Chakravarty, and P. K. Dash, "An improved cuckoo search based extreme learning machine for medical data classification," *Swarm Evol. Comput.*, vol. 24, pp. 25–49, Oct. 2015.
- [48] K. Polat, "Similarity-based attribute weighting methods via clustering algorithms in the classification of imbalanced medical datasets," *Neural Comput. Appl.*, vol. 30, no. 3, pp. 987–1013, Aug. 2018.
- [49] M. Seera and C. P. Lim, "A hybrid intelligent system for medical data classification," *Expert Syst. Appl.*, vol. 41, no. 5, pp. 2239–2249, Apr. 2014.
- [50] P. Jaganathan and R. Kuppuchamy, "A threshold fuzzy entropy based feature selection for medical database classification," *Comput. Biol. Med.*, vol. 43, no. 12, pp. 2222–2229, Dec. 2013.
- [51] S. Nalband, A. Sundar, A. A. Prince, and A. Agarwal, "Feature selection and classification methodology for the detection of knee-joint disorders," *Comput. Methods Programs Biomed.*, vol. 127, pp. 94–104, Apr. 2016.
- [52] R. Cheruku, D. R. Edla, V. Kupplli, and R. Dharavath, "RST-BatMiner: A fuzzy rule miner integrating rough set feature selection and bat optimization for detection of diabetes disease," *Appl. Soft Comput.*, vol. 67, pp. 764–780, Jun. 2018.
- [53] N. K. Shrivastava, P. Saurabh, and B. Verma, "An efficient approach parallel support vector machine for classification of diabetes dataset," *Int. J. Comput. Appl.*, vol. 36, no. 6, pp. 19–24, 2011.
- [54] Z. Yu, L. Li, J. Liu, and G. Han, "Hybrid adaptive classifier ensemble," *IEEE Trans. Cybern.*, vol. 45, no. 2, pp. 177–190, Feb. 2015.
- [55] B. Dennis and S. Muthukrishnan, "AGFS: Adaptive genetic fuzzy system for medical data classification," *Appl. Soft Comput.*, vol. 25, pp. 242–252, Dec. 2014.
- [56] S. Chandra and M. Kaur, "Enhancement of classification accuracy of our adaptive classifier using image processing techniques in the field of medical data mining," in *Proc. Int. Conf. Green Comput. Internet Things (ICGCIoT)*, Oct. 2015, pp. 948–954.
- [57] A. M. Alhassan and W. M. N. W. Zainon, "Taylor bird swarm algorithm based on deep belief network for heart disease diagnosis," *Appl. Sci.*, vol. 10, no. 18, p. 6626, 2020.

AFNAN M. ALHASSAN received the bachelor's degree from Shaqra University, in 2011, and the M.Sc. degree in computer science from North Carolina Agricultural and Technical State University, in 2017. She is currently pursuing the Ph.D. degree with the School of Computer Science, University of Science Malaysia. Her research interests include data mining, knowledge engineering, image processing, medical image analysis, and machine learning.



WAN MOHD NAZMEE WAN ZAINON received the Ph.D. degree in computer science from Universiti Sains Malaysia. He is currently a Senior Lecturer with the School of Computer Sciences, Universiti Sains Malaysia. His research interests include the intersection of visual computing and software engineering with a focus on software reuse, requirement engineering practices, data mining, and big data analytic.