VOL. 10, NO. 15, AUGUST 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

REVIEW OF DATA MINING APPROACHES FOR EXTRACTION AND CLASSIFICATION OF CLINICAL DATA IN DIAGNOSIS OF CORONARY ARTERY DISEASE

Noreen Kausar[1], Azween Abdullah[2], Brahim Belhaouari Samir[3], Sellapan Palaniappan[4] and Bandar Saeed Alghamdi[5]
[1,4]Malaysia University of Science and Technology, Selangor, Malaysia
[2]Taylors University, Selangor, Malaysia
[3]Department of Computer Science, Innopolis University, Kazan, Russia
[5]King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia
E-mail: noreenkausar88@yahoo.com

ABSTRACT

Coronary artery disease (CAD) has been ranked as the top cause of death by world health organization in many countries especially Asia. In Malaysia, 22.18% of total deaths are caused by CAD. In this paper, our focus is to review possible types of data mining algorithms applied for processing of clinical attributes as well as their classification to identify normal and CAD patients in minimal time with optimized accuracy. Various combinations of these techniques and variation have adverse effects as well as increased performance, which will be covered in this paper. Data selection for designing a detection system also varies the system performance and it can be dealt with using standard data sets with relevant feature to ease detection of abnormalities with maximum detection rate.

Keywords: Coronary Artery Disease (CAD), University of California, Irvine (UCI), supervised learning, feature processing, classification, Cleveland dataset.

1. INTRODUCTION

Heart disease is one of the leading causes of deaths around the globe. Coronary Artery Disease (CAD) causes the coronary arteries to contract and slowly solidify, which leads to chronic type of heart attacks [1]. In recent years, computerized diagnostic systems are being designed on ensemble approaches of data mining to enhance the ability of a cardiologist in detecting CAD patients with higher accuracy rate [2]. Selection of rational clinical attributes among the raw data attributes has been truly a complicated yet critical dilemma of biomedical engineering [3].

The idea of computer aided diagnosis of CAD is to generate output on the basis of patient's clinical records. These systems are trained, cross validated and tested well before they can be used in real-time environment. Once an optimized system is developed, they are ready to assist medical staff as per their requirements. But for this purpose, suitable CAD detection approach needs to be applied. An implementing CAD detection system with enhanced performance and lower false rate is the need of the hour.

Timely diagnosis of heart disease is imperative necessity of concerned patients from a cardiac physician in an ideal situation. Symptoms of patients vary with disease indications and demand specific clinical or laboratory procedures for cure. Treatments from physicians and clinical infrastructure all have significant effects on the specific disease diagnosis. Issues in traditional diagnosis procedure are as below in Table 1.

Table-1. Issues in traditional diagnosis procedure.

Patients Profile
- Different symptoms and indications - Difference in clinical findings and laboratory examination - Medical history
Physicians
- Personal experience - Inefficient treatments -Demographic constraints
Treatment Issues
- Clinical constraints - Lack of structural architecture for effective diagnosis - Insufficient technology usage

The organization of this paper is as follows: Section 2 focuses demographic statistics of CAD, whereas Section 3 covers possible datasets used for a CAD system. Section 4 provides various categories of data mining methods and performance comparison of different algorithms. Section 5 discusses limitations of earlier approaches and Section 6 concludes with proposition for future development.

2. DEMOGRAPHIC STATISTICS OF CORONARY DISEASE FOR MALAYSIA

The Coronary Heart Disease (CHD) age adjusted death rate is around 138.75 per 100,000 in Malaysia which is ranked 57 in the world. There are 22,701 deaths in Malaysia due to CHD which means 22.18% of total deaths in the country [4]. Figure 1 and 2 provides the statistics of deaths within the Malaysia.

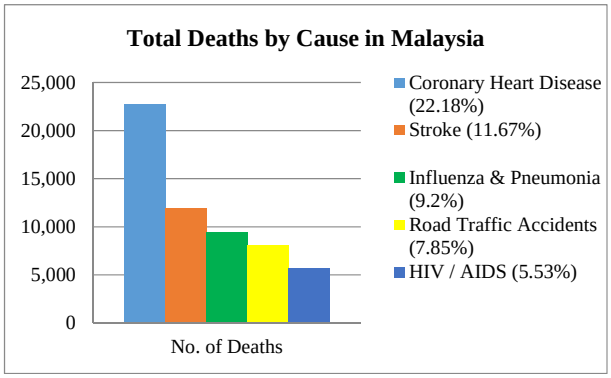

Figure-1. Main five causes of deaths in Malaysia, latest WHO (2011).

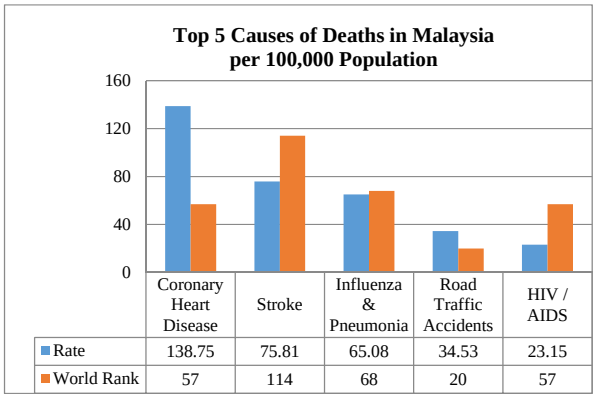

Figure-2. Age standardized death rate. Top 5 deaths causes per 100,000 population.

These details presented in Figure 1 and 2 are from the most recent data of these primary sources: WHO, World Bank, United Nations Educational, Scientific and Cultural Organization (UNESCO), Central Intelligence Agency (CIA) and individual country databases for global health and causes of death [4].

3. EXPERIMENTAL DATASETS FOR DIAGNOSIS OF CAD USING DATA MINING TECHNIQUE

To design, efficient detection system for CAD, the performance of proposed methodologies mainly depends upon the dataset selected. If the dataset does not have enough clinical attributes relevant for disease, it's hard to difficult possible abnormalities using selected data mining approaches. To ease this selection process for CAD data, University of California, Irvine (UCI) has provided publicly a dataset which is frequently used for cardiology research to formulate diagnosis system for detection of cardiac abnormality.

Overall, UCI Heart Dataset has 76 attributes covering all the possible aspects which can help in predicting the presence of heart disease. All these attributes demands different procedures to extract corresponding values for patient records. From these attributes, only thirteen attributes are mainly referred for patient's diagnosis by various researchers. These numeric records are identified with a class label indicating presence or absence of heart abnormality.

UCI heart dataset

The UCI heart dataset has four main databases from various hospitals such as Cleveland, Hungarian, Switzerland and Long Beach. The collected raw data include records represented by zero (0) which have absence of heart disease while patient's records are categorized by 1, 2, 3 and 4. The number of records distributed in different classes of these databases is shown in Table 2.

Table-2. Class distribution.

UCI Heart Databases	UCI Heart Databases					
	0	**1**	**2**	**3**	**4**	**Total**
Cleveland	164	55	36	35	13	303
Hungarian	188	37	26	28	15	294
Switzerland	8	48	32	30	5	123
Long Beach VA	51	56	41	42	10	200

The above mentioned databases of UCI Heart have many missing records. The one which is considered mostly for classification of heart disease is Cleveland with few redundant data samples which were further modified and eliminated by preparing another dataset called Statlog. dataset that does not have any redundant or missing data records. These two are normally preferred datasets from UCI Repository datasets for detecting heart anomalies [5]. Both have same feature types based on clinical attributes, but slight differences in number of patient records. These datasets facilitate to testify the system with earlier applied researches and to analyze the system's generalization ability with varying dimensions of medical data.

There are 13 attributes of Cleveland and Statlog datasets which are presented with their respective input options required for detecting heart patients in Table 3. Overall,

- There are six real valued features.
- There are three binary valued features.
- There are three nominal and one ordered valued features.

Table-3. Features of UCI heart dataset.

Feature No	Name	Description
1	age	age in years
2	sex	sex (1 = male; 0 = female)
3	cp	chest pain type - Value 1: typical angina - Value 2: atypical angina - Value 3: non-anginal pain - Value 4: asymptomatic

Feature No	Name	Description
4	trestbps	resting blood pressure (in mm Hg on admission to the hospital)
5	chol	serum cholesterol in mg/dl
6	fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7	restecg	resting electrocardiographic results - Value 0: normal - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8	thalach	maximum heart rate achieved
9	exang	exercise induced angina (1 = yes; 0 = no)
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	the slope of the peak exercise ST segment
12	ca	number of major vessels (0-3) colored by fluoroscopy
13	thal	3 = normal; 6 = fixed defect; 7 = reversible defect
14	num	diagnosis of heart disease (angiographic disease status) - Value 0: < 50% diameter narrowing - Value 1: > 50% diameter narrowing (in any major vessel: attributes 59 through 68 are vessels)

Other CAD datasets

In various data mining approaches for designing CAD detection system, researchers have used varying datasets with different patient records but, with almost equivalent clinical attributes as UCI repository. These customized datasets are mostly generated by hospitals and clinics for their patients to produce specific results for their circumstances.

- Setiawan et al. selected Ipoh Hospital dataset in comparison with UCI Heart Dataset to develop system using Fuzzy decision support system and rough set theory [6,7].
- Ghumbre et al. utilized Indian heart dataset with 19 features relatively same as Cleveland for designing CAD system based on Support Vector Machines (SVM) in a non-linear technique using Radial Basis Function (RBF) kernel [8].
- Alizadehsani el al. collected clinical data from Tehran's hospital with 54 features and 303 records as much content as Cleveland for a comparative study [9,10]. The data mining techniques used were Naïve Bayes, C4.5, Ada boost and Sequential Minimal Optimization (SMO).
- Akil Jabbar et al. performed his experiment on the data from Andhra Pradesh hospital using Genetic Algorithm, Gini index and Z-Statistics [11,12].

4. DATA MINING TECHNIQUES IN DESIGNING CAD DAIGNOSTIC SYSTEM

Data mining techniques are significantly used to implement the efficient detection system on the basis of their reliability and consistency. Commonly used classifiers are Neural Networks, Support Vector Machines, Decision Tress, Naïve Bayes etc. There are also many popular dimensionality reduction and feature selection techniques which are mathematically based on data mining algorithms.

Data mining approaches for classification

Many recent detection developments have been utilized, single as well hybrid data mining approaches to achieve possible performance with maximized detection rate and decreased false alarms. The classification techniques can be categorized as supervised or unsupervised learning on the basis of the input data samples [13].

- **Supervised learning**: It works on the labeled data patterns which facilitate the training unit to learn the patterns and their concerning labels. This procedure trains the classifier with the labels behavior for both classes which establishes prediction ability for the testing unit to detect unseen or new patterns provided without labels. There are many classifiers which adopt supervised learning such as Support Vector Machines, Neural Network etc.
- **Unsupervised learning**: It works on the data patterns which only include the attribute details and no labels. The behavior of these patterns is learned by the training unit and their distinguishing structure are analyzed and internal grouping is formed from the provided training set. For unsupervised learning, there is no fix number or size of the groups and solely depends on the data provided. Once the groups are formed, patterns for testing are provided which are classified on the basis of these formed groups. k-Nearest Neighbor and K-means clustering are

commonly used unsupervised learning techniques.

Some classification methods applied as single or hybrid approaches for CAD detection system to increase system's sensitivity and specificity measures are as below:

- Support vector machines (SVM)
- K-means algorithm
- Fuzzy systems
- Decision trees (DT)
- Naïve bayes (NB)
- Neural network (NN)
- k-nearest neighbor (k-NN)
- Multilayer perceptron (MLP)
- Rough set theory

Data mining approaches for feature processing

Feature processing facilitates the classifier to process relevant features transformed and extracted using appropriate methods for better detection rate. Feature selection reduces the dimension and accelerates learning ability and prediction rate of the classifier [14]. Filter and wrapper are two common methods of feature selection which are briefly described as following:

- **Filter method**: It differentiates between relevant and irrelevant features by analyzing their dependency on the classification process which reduces classifier overhead in processing redundant features.
- **Wrapper method**: It determines a certain threshold value for the features to qualify for the classification process. The features with a score value above the threshold are selected and the remaining features are discarded as they are not significant enough for enhancing classifier performance.

Some feature processing algorithms which includes feature extraction, feature transformation and selection have been used in combination with classifiers to eliminate irrelevant features and enhanced the accuracy of selected data mining approaches for CAD detection system are as below:

- Genetic algorithm (GA)
- Fisher score (FS)
- Cuckoo search
- Brute force
- Genetic search (GS)
- Kernel principal component analysis (KPCA)
- Principal component analysis (PCA)
- Linear discriminant analysis (LDA)
- Chi square
- Fuzzy weights
- Gini index
- Z- statistics

Data mining approaches for performance enhancement and optimization

Apart from designing CAD systems using predefined classification and feature processing techniques, various optimization algorithms can be applied to enhance the overall system performance. Some commonly used methods for CAD datasets are as below:

- Boosting
- Bagging
- Voting
- Filtering
- Entropy

Table 4 shows the performance comparison of earlier data mining approaches applied as single, hybrid or ensemble for enhancing the system results. Different variations have also been used along with various algorithms to maximize the achieved results for the system. Among these approaches, the results are higher for the standard dataset as compared to other collected datasets by the researchers for the study.

Table-4. Performance analysis of earlier data mining approaches for CAD detection system.

Authors	Dataset	Classification / Feature Processing Techniques	Results
Atul Kumar Pandey et al. [15]	UCI heart dataset	K-means Clustering, PART (Projective Adaptive Resonance Theory	Accuracy: 93.06%
Xiaoyong Liu and Hui Fu [16]	UCI heart dataset	Cuckoo Search, Particle swarm optimization (PSO), SVM	Train Accuracy:100% Test Accuracy: 85%
Shruti Ratnakar et.al. [17]	Cleveland dataset	Genetic Algorithm, Decision Trees (DT), NB	Accuracy for 6 features: Decision Trees: 99.2 Naïve Bayes: 96.5
M.A.Jabbar et al. [18]	Statlog dataset	KNN, Symmetrical uncertainty (SU), Artificial Neural Network (ANN), PCA,GA	Accuracies: KNN with SU: 100% by reducing 7 features ANN with GA: 99.62% ANN with PCA: 98.14%

Authors	Dataset	Classification / Feature Processing Techniques	Results
M.A.Jabbar et al. [12]	Statlog heart	NN, PCA, Genetic search, Chi Square	Accuracies: No subset used:97.4% Subset used: PCA: 98.14% Chi Square: 97.7% GA:99.62%
M.Akhil Jabbar et.al. [11]	Heart data collected from Andhra Pradesh Hospitals	Genetic algorithm, Gini Index, Z-Statistics	Accuracy= 98% This associative rules based prediction was better than doctor's diagnosis.
Nidhi Bhatla and Kiran Jyoti [19]	909 patterns from Cleveland dataset	NN, DT, Naïve Bayes, Genetic algorithm	Accuracy for 6 and 15 features: Naïve bayes : 96.5% & 86.53% DT: 99.2% & 89% NN : 85.53% for 15 features.
R. Alizadehsani et.al. [9]	Tehran's hospital dataset for heart with 303 records	NB, C4.5, Adaboost, SMO	Accuracy for 16 features: NB : 74.89 C4.5: 78.23 Adaboost: 76.86% SMO: 82.16
M.Shouman et al. [20]	UCI heart dataset	KNN, Voting	Without Voting: Accuracy: 97.4% Specificity: 99% Sensitivity: 93.8% Optimal value of k=5 KNN with Voting: Accuracy: 95.7%
S.Ghumbre et al. [8]	Indian heart dataset with 19 features	SVM, Radial basis function (RBF)	Accuracies: SVM: 85.05 using 5 and 10 fold cross validation RBF 82.71% for 5 and 82.24% for 10 fold
M. Anbarasi et.al. [21]	UCI heart dataset 909 records	Genetic algorithm, DT, NB , Clustering	Accuracies for 6 features: NB : 96.5% DT: 99.2% Clustering: 88.3%
Tu et.al. [22]	UCI heart dataset	Bagging algorithm, Decision trees	Accuracy 78.91% with DT and 81.41% with bagging
N.A.Setiawan et al. [6]	UCI heart Dataset, Ipoh Hospital data	Fuzzy decision support system, Rough set theory	83% Accuracy with UCI dataset 87% with FDSS for Ipoh dataset.

4. DISCUSSION ON APPLIED DATA MINING APPROACHES AND THEIR LIMITATIONS

Earlier developed systems have their pros and cons depending upon the combination of techniques implemented. The focus of these systems has always been to achieve a maximum detection rate with lesser false alarms in a limited time.

In Table 4, the performance of different approaches for CAD system has varied with same UCI dataset because of the features selected and classifier applied. Among the feature extraction algorithms, Genetic algorithm has outperformed with classifiers such as Decision trees, Naïve bayes and Neural networks with accuracy between 96% to 100%. Different clinical datasets used apart from the UCI dataset as described in Section 3 have not been able to achieve as much performance except when Genetic algorithm is used in combination with methods such as Gini index and Z-statistics. For k-NN, voting technique have bit lessen the accuracy rate from 97.4% to 95.7%. Selection of approximate 6 features from 13 using appropriate algorithms has much contribution for enhancing the evaluation measures.

The identification of existing challenges helps to propose improved and enhanced CAD detection system. To overcome these issues, concerning solutions should be implemented to contribute earlier research efforts. Table 5 provides existing issues and possible solutions for designing of CAD detection system as follows:

Table-5. Existing issues and concerning solutions for designing CAD system.

Issues of Existing Systems	Suggested Solutions for Future Systems
• Low detection capability • High resources occupation • Excessive human participation • Time consuming statistical analysis • High false alarms • Global and local minima • Complex classifier architecture	• Performance efficiency • Sensitive features selection • Supervised data mining • Time efficient criterion • Less false alarms • Maximum margin classification • Parameter optimization

To overcome these complications, various developments have been done to digitize and automate the diagnosis process for clinical data including signals and images. An ideal heart diagnostic system should 100% accurate to detect normal and cardiac patients. Such systems involve a classification mechanism which distinguishes data records, extracted and carefully selected by a processing algorithm. Appropriate techniques and methods are required to be applied to design the detection system.

The comparison of the approaches is based on the performance factors like accuracy, sensitivity and specificity. Accuracy is the measure of detecting normal and abnormal (CAD) patterns correctly whereas sensitivity is the measure of detecting abnormal (heart disease) patterns accurately and specificity is the measure of detecting normal patterns accurately [23]. The equations for calculating accuracy, sensitivity and specificity are as given below.

$$\text{Accuracy} = \frac{TP + TN}{TN + TP + FN + FP} \%$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \%$$

$$\text{Specificity} = \frac{TN}{TN + FP} \%$$

where

- True positive (TP) means an abnormal (CAD) is detected as an abnormal (CAD).
- True negative (TN) means a normal is detected as normal.
- False positive (FP) means an abnormal (CAD) is detected as normal.
- False negative (FN) means a normal is detected as abnormal (CAD).

CONCLUSIONS AND FUTURE WORK

In this paper, we reviewed and compared various data mining approaches applied as single and hybrid for processing and classification to detect cardiac abnormalities from clinical data. The utmost requirement is a relevant dataset with appropriate clinical and laboratory data which can then be processed or extracted as per need to eliminate redundancy and record duplication. Classifiers can be ensemble with these processing methods to enhance their sensitivity and specificity while minimizing the overhead to have simplified architecture. The imperative aspect in designing a detection system is to optimize and adjust selected methods so that they can perform their best instead of complicating the selected algorithm by an unnecessary procedure with no increase in the system accuracy. The best approach should have optimal architecture and tuned as per the required specification to get maximize detection rate and negligible false alarms.

REFERENCES

[1] AHA. (2013). American Heart Association. Retrieved January, 2013, from http://www.americanheart.org

[2] Das, R., Turkoglu, I., & Sengur, A. (2009). Effective Diagnosis of Heart Disease through Neural Networks Ensembles. Expert Systems with Applications, 36(4), 7675-7680.

[3] Can, M. (2013). Diagnosis of Cardiovascular Diseases by Boosted Neural Networks. Southeast Europe Journal of Soft Computing, 2(1), 91.

[4] World Health Organization. Retrieved 15 January, 2015, from http://www.worldlifeexpectancy.com

[5] Detrano. (2013). UCI. Retrieved 27 July, 2013, from http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[6] Setiawan, N. A., Venkatachalam, P. A., & Fadzil, M. H. (2009). Diagnosis of Coronary Artery Disease Using Artificial Intelligence Based Decision Support System. Proceedings of the International Conference on Man-Machine Systems (ICoMMS), Batu Ferringhi, Penang.

[7] Setiawan, N. A., Venkatachalam, P. A., & Fadzil, M. H. (2009). Rule Selection for Coronary Artery Disease Diagnosis Based on Rough Set. International Journal of Recent Trends in Engineering, 2(5).

[8] Ghumbre, S., Patil, C., & Ghatol, A. (2011). Heart Disease Diagnosis using Support Vector Machine. International Conference on Computer Science and Information Technology (ICCSIT'2011). Pattaya.

[9] Alizadehsani, R., Habibi, J., Alizadehsani, R., Sani, Z. A., Mashayekhi, H., Boghrati, R., et al. (2012). Diagnosis of Coronary Artery Disease Using Data mining based on Lab Data and Echo Features. Journal of Medical and Bioengineering, 1(1).

[10] Alizadehsani, R., Habibi, J., Hosseini, J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., et al. (2013). A Data Mining Approach for Diagnosis of

Coronary Artery Disease. Computer Methods and Programs in Biomedicine, 111(1), 52-61.

[11] Jabbar, M. A., Chandra, P., & Deekshatulu, B. L. (2012). Heart Disease Prediction System using Associative Classification and Genetic Algorithm. International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT. India.

[12] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection. Global Journal of Computer Science and Technology Neural & Artificial Intelligence 13(3).

[13] Medhekar, D. S., Bote, M. P., & Deshmukh, S. D. (2013). Heart Disease Prediction System using Naive Bayes. International Journal of Enhanced Research In Science Technology & Engineering, 2(3).

[14] Kuttikrishnan, M., & Dhanabalachandran, M. (2010). A Novel Approach for Cardiac Disease Prediction and Classification Using Intelligent Agents. International Journal of Computer Science and Information Security, 8(5).

[15] Pandey, A. K., Pandey, P., & Jaiswal, K. L. (2014). Classification Model for the Heart Disease Diagnosis. Global Journal of Medical Research, 14(1).

[16] Lui, X., & Fu, H. (2014). PSO-Based Support Vector Machine with Cuckoo Search Technique for Clinical Disease Diagnosis. The Scientific World Journal, 2014, 7.

[17] Ratnakar, S., Rajeswari, K., & Jacob, R. (2013). Prediction of Heart Disease Using Genetic Algorithm for Selection of Optimal Reduced Set of Attributes. International Journal of Advanced Computational Engineering and Networking, 1(2).

[18] Jabbar, M. A., Deekshatulu, D. L., & Chandra, P. (2013). Heart Disease Classification using Nearest Neighbor Classifier with Feature Subset Selection. Annals. Computer Science Series, 11(1).

[19] Bhatla, N., & Jyoti, K. (2012). An Analysis of Heart Disease Prediction using Different Data Mining Techniques International Journal of Engineering Research & Technology 1(8).

[20] Shouman, M., Turner, T., & Stocker, R. (2012). Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients International Journal of Information and Education Technology, 2(3), 220-223.

[21] Anbarasi, M., Anupriya, E., & Iyengar, N. (2010). Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. International Journal of Engineering Science and Technology (IJESI), 2(10).

[22] My Chau, T., Dongil, S., & Dongkyoo, S. (2009, 17-19 Oct. 2009). Effective Diagnosis of Heart Disease through Bagging Approach. Paper presented at the 2nd International Conference on Biomedical Engineering and Informatics, 2009. BMEI '09.

[23] Kausar, N., Belhaouari Samir, B., Sulaiman, S. B., Ahmad, I., & Hussain, M. (2012). An Approach Towards Intrusion Detection Using PCA Feature Subsets and SVM. 2012 International Conference on Computer & Information Science (ICCIS) (Vol. 2, pp. 569-574).