APPLYING DATA MINING TECHNIQUES TO CHRONIC DISEASE IDENTIFICATION AT THE WORK PLACE

[1]C. Kalaiselvi
Head & Associate professor,
Department of Computer Applications
Tiruppur Kumaran College for women, Tirupur, India.

Abstract : Heterogeneous, chronic diseases such as diabetes, heart diseases and cancer, tuberculosis are commonly occurred and increased now a day. Most of the people don't know the symptoms of these diseases and its chronic complications. There is a chance for everyone getting chronic diseases when people work in a polluted environment. The effects are uncontrolled over time and may lead to serious damage to human body and life. India being a country of huge population is living in working environment. In general, diagnosing diseases at the right time is important in healthcare sector for decision making in correct manner. This study aims to develop enhanced disease identification technique for achieving higher efficiency and to improve advancement in disease identification using data mining techniques.

IndexTerms **– Disease, Chronic, Data mining, Identification, Healthcare.**

1. Introduction

Data mining refers to a set of techniques that use the process of extracting previously unknown but potentially useful data from large repositories of past data. It is also an extraction of hidden predictive information from large amount of data. Data mining has been growing rapidly due to its huge applicability, achievements and sophisticated features. The innovative achievements in various fields and technology have imposed new challenges to data mining techniques. Medical data mining is the analysis of observational data sets to find patterns or models that are both understandable and useful to the medical practitioners and users. Different mining approaches like clustering, classification, association rule learning, regression and outlier detection are frequently used to analyse medical data to gain disease identification and related knowledge **(**Chung et al., 1998**)**. The extracting process of information from a huge data set is denoted by knowledge discovery in data mining. The steps in the KDD process such as data selection, data preparation, cleaning and incorporation of appropriate prior knowledge, and proper interpretation of the results ensure that the extracted patterns correspond to useful knowledge (Fayyad et al., 1996). The KDD process is an interactive, iterative, user driven and involves numerous steps summarized as (Fayyad et al., 1996). The objective of studying and applying data mining techniques available for chronic disease identification at the workplace and to analyse the factors and attributes that influences chronic diseases in which more contribution is in disease identification at the workplace based on the environment they are working.

2. Literature review

Over the past few decades data mining has become a topic of interest to researches and has been applied to many bio-medical applications. Medical data mining is one of the great potential areas for exploring the hidden patterns in the data sets of the medical domain whereas these patterns are utilized for medical treatment.

Here we discuss the brief literature survey on data mining techniques used in disease prediction and diagnosis of chronic diseases using data mining techniques. The goal is to provide the overview of research work so far carried out in this domain, and data mining algorithms represented to solve clinical diagnostic problems. The various methods, techniques, algorithms developed in this area concentrate on the data mining algorithms such as back propagation neural networks, SVM, decision trees, fuzzy systems and etc. which are useful to the healthcare sector.

However, the available raw medical data are distributed widely, and in heterogeneous by nature, and in huge voluminous. These huge data can collected in an organized form and these collected data reports can be then integrated to form a medical information system (Soni et al., 2011).

Data mining technology offers a research oriented approach to find novel and hidden patterns in the data. The healthcare industry collects a huge amount (Parvathi.I and S.Rautaray, 2014) of data which is not properly mined and not used optimally and discovery of hidden patterns and connections often goes unexploited (Kumar et al., 2011). Machine learning techniques are likewise proposed to identify clinical disease diagnosis and prediction.

Data mining is a growing area of research that intersects with many disciplines (Liao.S et al., 2012) such as visualization, databases, Artificial Intelligence (AI), statistics, parallel and high-performance computing. The goal of data

mining is to turn out data into information that are facts, numbers, or text which can be processed by a computer into knowledge. Data mining has become the most effective technique for identification and prediction. It also provides exciting, challenging research and application areas not only for computational sciences but also for biomedical sciences. It enables the researchers to meet the challenge of mining vast amount of data to discover real knowledge. Nowadays, the reliability of health care data is increasing and therefore, this section of research work aims to understand about data mining techniques and its importance in medical systems. In order to achieve the goal of this proposed work, a literature study was made on various data mining techniques used in disease prediction and medical applications which are discussed below.

Prather, J. C et al., (1997) tells that data mining is useful in medical applications such as medical tests, prediction of surgical procedures, and finding the relationships between pathological data and clinical data. W. Lord and D. Wiggins, (2006) reveals that proper usage of knowledge discovery in data mining in the database (KDD) for the growing databases are very important. KDD attempts to gather knowledge by identifying relations from the data sets to help in disease predictions and also, he tells that the medical information databases may contain data such as patient records, physician's diagnosis reports and monitoring information where the data has been useful in many medical decision support systems.

Elmaghraby, A. S et al., (2006) reveals that reliance of health care data is increasing rigorously. Medical researchers, physicians, and health care providers face problems in using these stored data efficiently and effectively when more medical information systems with large databases are used.

W. Lord and D. Wiggins,(2006) summarizes medical decision support systems are systems that helps in the decision making process in the medical domains such Clinical Decision Support Systems (CDSS), medical imaging, and Bioinformatics. The contributions of these systems are to reduce medical errors and costs, earlier disease detection, and to produce preventive medicine. The advantages of using computerized CDSS are the decision support systems can help to manage overloaded data, turn them into knowledge, reduce the complexity of the work, save time, and variety of practices. For better data analysis and decision support, the data mining and decision support systems can be integrated Pur et al., (2005).

Huang et al., (2004) tells that data mining has been used in the medical domain to improve the decision making such as diagnostic and prognostic problems in oncology, liver pathology, neuropsychology, and gynecology.

Elmaghraby, A. S et al., (2006) investigated data mining techniques that helped the physicians and practitioners to improve their health services for the patients by detecting irregularities and unexpected patterns from the data. The task of identifying an association between risk factors and outcomes in the medical area of data mining is difficult to work even for experienced biomedical researchers or health care managers.

Cheng et al., (2010) reveals that the usage of data mining tools with advanced algorithms is popular for pattern discovery in biological data. The biological problems include protein interactions, sequence and gene expression data analysis, and drug discovery, discovering homologous sequences or structures, gene finding and gene mapping, and sequence alignment.

Strumbelj et al., (2010) reveals that machine learning was not fully accepted in the medical community because medical practitioners feel that their work is more complicated using such tools. For example, different models used in healthcare applications have different explanations especially for model-specific methods and which one to choose and use is a big issue. Therefore, the most important things that must be considered when developing an application for medical practitioners are simplicity and the way of explaining the decisions. Another challenge is that the systems must be able to present discovered knowledge in an easy and fast manner Elmaghraby, A. S et al., (2006).

Kumar.D et al., (2011) proposed the ID3 algorithm, decision trees, C4.5 algorithm to classify diabetes, hepatitis and heart diseases. B.M. Patil et al., (2010) proposed the Hybrid Prediction Model (HPM) which uses the Simple K-Means clustering algorithm to validate correctly classified instances and incorrectly classified instances from the class label of given data. C4.5 algorithm is used to create a final classifier model by utilizing the k-fold cross-validation method. The accuracies achieved by the previous method fall within the series of 59.4–84.05% and also the proposed HPM obtained a classification accuracy of 92.38%.

Kumar D et al., (2011) proposed the utilization of the ID3 algorithm, decision trees C4.5 algorithm and CART algorithm to classify hepatitis and heart diseases and compare the effectiveness, correction rate using measures. This part of the medical analysis is done by learning examples through the gathered information of diabetes, hepatitis and heart diseases and also to create clever restorative choice emotionally supportive networks to help the physicians.

ANN is one of the fast-developing and promising techniques used in real-world problems. the accuracy of the processes can be improved using pre-processing techniques to update and change the medical data since most of the data are incomplete. ANN gives classification efficiency experimentally of about 99%.

Vacante et al., (2011) proposed Ant Colony Optimization (ACO) is successful in rule-based classification. A new algorithm is presented in this work for extracting the If-then rule for diagnosing diabetes. They proposed a method called FADD, which is based on an online fixing procedure and it is evaluated using eight biomedical datasets and five versions of the random forests algorithm (40 cases). The method worked out correctly for the maximum number of trees almost 90 % of test cases. Efficiency is less while comparing. Need more attributes FathiGanji, M and M.S Abadeh, (2010).

Ilayaraja.M and T.Meyyappan, (2013) reveal by applying data mining techniques in the Healthcare domain improves the Quality of Service (QoS) by discovering potentially useful trends required for medical diagnosis. Brameier and Banzhaf, (2001) explored and analyzed the two programming models such as neural networks, and linear genetic programming for medical data mining.

SVM is one of the most widely used machine learning techniques. SVM works out by finding linear plane methods for classification model data which are used to implement classification by developing a new model. Barakat et al.,(2010) concentrates more on the classification of the system with reliability. The reliability of the process increases when there is an increase in efficiency and support for the best prediction. for this SVM gives a most promising tool for the prediction of

diabetes where ruleset which is comprehensible one was generated with a prediction accuracy of 94%, the sensitivity of 93%, and specificity of 94%.

3. Research Methodology

Population and Sample

Totally 1000 samples were taken and after pre-processing 853 samples were taken for final analysis. The samples stored in a database can be mined from the total population and samples can be partitioned into several categories based on the disease identification using data mining techniques. The collection of data profiles of each employee's medical data plays a key role in this research.

Theoretical framework and Techniques

Machine learning is a technique to deal with the design and development of algorithms that enhance the systems to work well based on empirical data from large databases. These data can be used as attributes in training and testing.

Support Vector Machine (Svm)

The SVM algorithm is a powerful supervised learning algorithm Furey et al (2000). SVM used a hyperplane that splits two classes of training data. Support Vector Machine is a supervised learning method for classification and regression. The implementation of SVM in classification proves the accuracy level to be improved. The various prediction process based on SVM has been discussed below.

SVM, when compared with the perceptron method and SVM gives superior performance, where SVM yields nearly perfect classification performance. In general, machine-learning methods yield better classification performance than that of the clustering methods, since the main difference between cancer and normal data in the data domain can be revealed in the machine learning methods.

K-Nearest Neighbour

K-means clustering is one of the most famous algorithms for any clustering structure which involves large dataset. It is the simplest and most popular algorithm which is easy to implement. The most commonly used distance measure in K-means clustering is the Euclidean distance metric. This algorithm tends to provide clusters in uniform size even though the inputs are of variable size.

Particle Swarm Optimization (PSO)

The particle swarm optimization algorithm is a population algorithm that creates initial particles and assigns initial velocities for it and also it evaluates the objective function at its every location of the particle. It determines the best lowest function value and location. It chooses the new velocities based on the current velocity, the best location of the particle and its neighbors.

Artificial Neural Networks (ANN)

A soft computing technique has taken a white space in the recent technology world artificial neural network is one of the most important techniques of soft computing. It provides a structured system that consists of simple calculation units with the parallel work process. Many supervised and unsupervised classifiers progress their processing capability very effectively to attain the desired optimal solution. One such algorithm is backpropagation propagates the output towards the proceeding layers and influences the output to the target output. Few neural networks based model for computerized medical diagnosis in disease classification has been discussed below.

An artificial neural network (ANN) learning algorithm is generally known as the neural network (NN). It is a learning algorithm that is stimulated by the structure and functional characteristics of biological neural networks. Computations are controlled in terms of an interrelated group of artificial neurons, processing information with a connectionist approach to computation. Recent neural networks are nonlinear statistical data modeling tools. They are normally used to model the complex associations between the inputs and outputs, to identify patterns in data or to capture the statistical structure in an unidentified joint probability distribution between the observed variables. They are generally presented as systems of interrelated neurons that could compute the values from inputs by supplying information through the network.

The neural network is one of the most widely used machine learning methods for classification and prediction. The basic elements in the construction of NN are the input (units), fed with information from the environment, the "shadow" units within the network (hidden neurons), controlling the actions in the network, and the output (units), which synthesizes with the network response. All these neurons must be interconnected in order that the network becomes fully functional. NN has a feed-forward type structure when the signal moves from input to output, passing through all the network's

hidden units, so that the outputs of neurons are connected to the next layer and not to previous ones. These networks have the property that the outputs can be expressed as a deterministic function of inputs. A set of values entering the network is transmitted through the activation functions over the network to its output, by the so-called forward propagation. Such a network has a stable operating mode. Artificial Neural Network (ANN) combines learning from the neural network and of evolutionary algorithms. The proposed ANN involves dimension reduction with information gain and classification of colon tumor samples based on microarray gene expressions.

Data and Methodology

As the main objective is to identify and apply the data mining techniques to the health sector, this study is based on the primary data generated through direct field survey and questionnaire to the employees at various institutions. At first, a pilot survey was made to identify the areas in which this research has been conducted. After the data collection each record of data is pre-processed for data cleaning to scrutinize the missing data, noisy data and irrelevant data. After pre-processing, the pre-processed data is integrated by the data integration process to store the data in a common format. From this data, the relevant data for this research was selected. For this decision tree approach was carried out and they are clustered to form a new group. Decision Tree is used as a protection model to visualize the tree pattern and to perform prediction of chronic diseases. Then data transformation was carried out and potential data mining techniques such as classification methods were used to retrieve patterns.

The steps in data mining process have been performed by using a data mining tool named MATLAB. Then interesting patterns were evaluated based on measures. The evaluation and performance measurement were done based on accuracy, specificity and sensitivity. Among the result found in this study shows that some factors can take into account to find the useful patterns. The following pictures 3.1, 3.2 illustrates the framework and overview of the research.

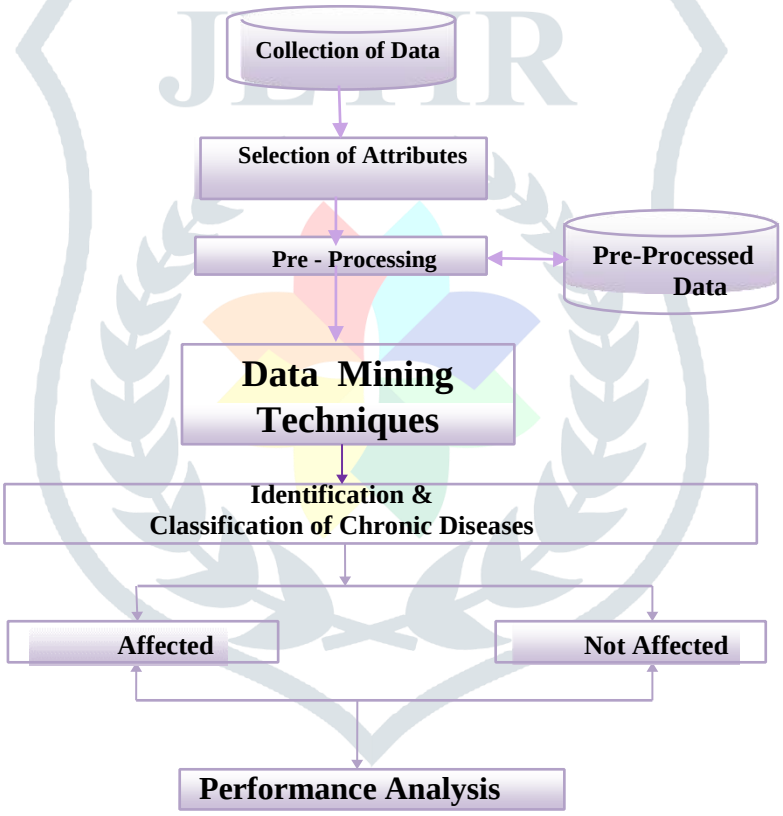

Figure 3.1 Framework of the Proposed System

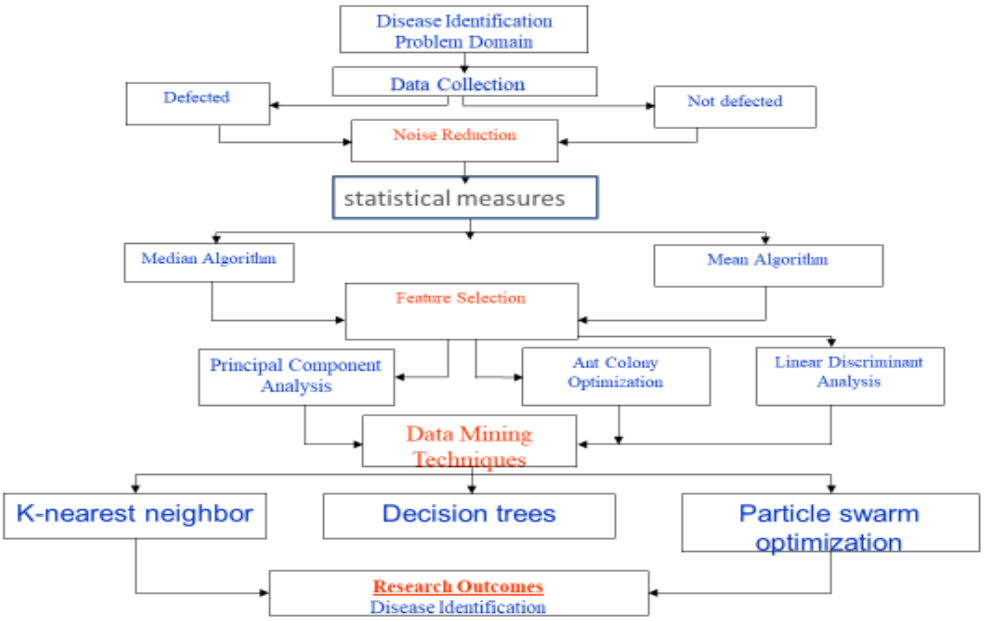

Figure 3.2 OVERVIEW OF PROPOSED SYSTEM

This algorithm when executed results the data set with selecting the 10 prime attributes (Age, BMI, BP, HBA1C, LDL_HDL_ratio, Physical_activity, WBC, ESR, Family Heredity, Habits (Alcoholic/smoking/tobacco) depending on their inference of values in the dataset. It is concluded that if number of iterations increases then the best set of attributes can be obtained. Figure 3.3 shows variations of attributes with respect to number of iterations.

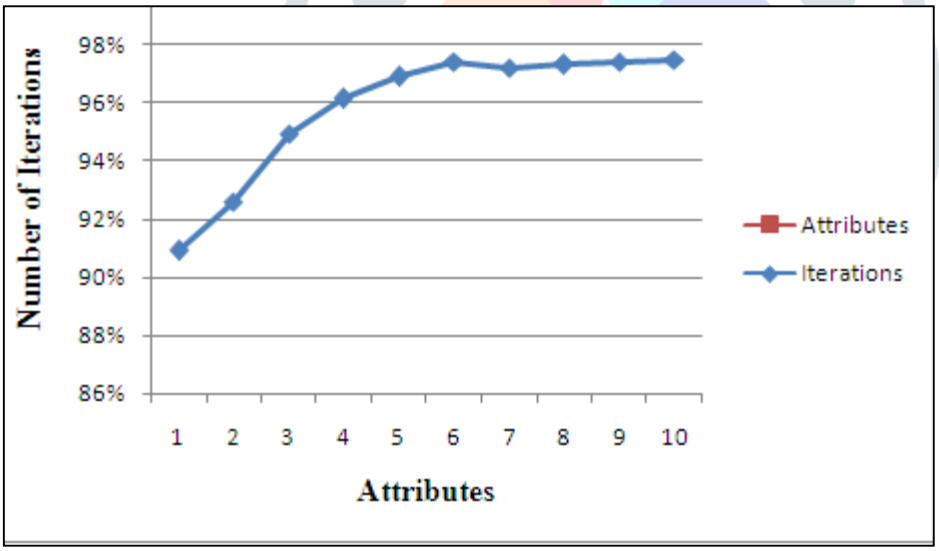

Figure 3.3 Variations of Attributes with Respect to Number of Iterations

The test performance of the classifiers can be determined by the computation of total classification accuracy and RMSE. Sensitivity measures are the actual true positives that are actually measured and the specificity is used to measure the classifiers ability to predict the negative results.

Sensitivity Vs Specificity comparison

TP: Number of True positives

FP: Number of False positives

TN: Number of True Negatives

FN: Number of False Negative

True Positive rate:

$$sensitivity = \frac{TP}{TP + FN}$$

False Positive Rate:

$$FalseAlarm = \frac{FP}{TN + FP}$$

$$specificity = \frac{TN}{TN + FP} = 1 - FalseAlarm$$

The total classification accuracy is defined as ratio of number of correct decisions and total number of cases.

$$Accuracy = (TP + TN)/n$$

Method	Accuracy
K-Nearest Neighbor	87.0%
Particle Swarm Optimization	92.3%
K-Nearest Neighbor With Particle Swarm Optimization	96.0%

Table 3.4 Performance Analysis of Methods

The classification Accuracy of the methods is shown in above table 3.4. It is evident that our proposed algorithm performs better classification accuracy is shown in Figure. 3.5.

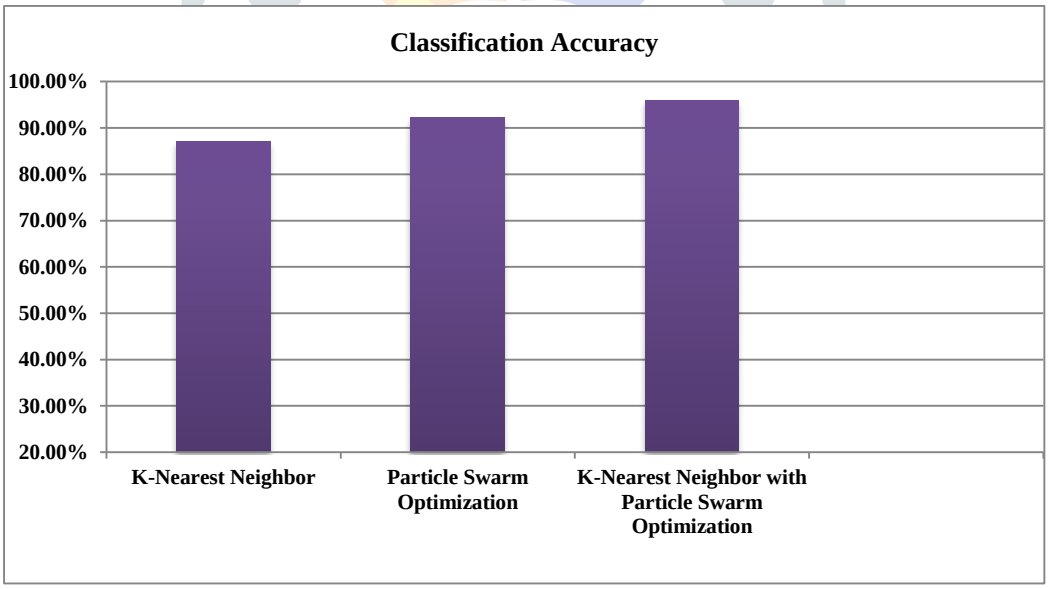

Figure 3.5 Classification Accuracy of Proposed Algorithms

As a conclusion, these results show that the K-Nearest Neighbor with Particle Swarm Optimization approach provide the highest accuracy in the detection of chronic disease categories.

4. Results And Conclusion

Health care costs grow rise and there is an urgency to control the costs, timely analysis of employees healthcare information maintenance and monitorization has become very big issues nowadays. Hospitals, Medical insurance companies require expert analysis of the data and some effective automated mechanisms are needed to support the above said process. Every

organization, institution or companies should maintain their employee's healthcare data or information and it requires great effort and administration is required to maintain and it also requires large memory space.

The steps in data mining process have been performed by using a data mining tool named MATLAB. Then interesting patterns were evaluated based on measures. The performance evaluation and measurement were done based on accuracy, specificity and sensitivity. Among the result found in this study shows that some factors can take into account to find the useful patterns. The results obtained from this study demonstrate potential values of data mining.

Greater accuracy and relevance of the data are used in implementation strategy to predict the chronic diseases. Automatic mechanism is needed to support the processes with use of data mining techniques such as clustering and classification. Disease diagnosis and decision making and automatic document processing is a mere factor in the analysis process. Set of exams frequently done and followed a frequent correlation between exam sets. The evaluation of classification models has been measured using accuracy, F-measures, precision and recall metrics.

The results reveal that the combination of ensemble learning methods of data mining with classification models produce better prediction in comparison with sole classification of medical data.

ACKNOWLEDGMENT

The Author is thankful to UGC (SERO) and Tiruppur Kumaran college for women for providing necessary facilities.

REFERENCES

[1] Abdelaal, M. M. A., Farouq, M. W., Sena, H. A., & Salem, A. B. M. (2010, October). Using data mining for assessing diagnosis of breast cancer. In *Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on* (pp. 11-17). IEEE.

[2] Abdullah, A. S. (2012). A Data Mining Model to Predict and Analyze the Events Related to Coronary Heart Disease using Decision Trees with Particle Swarm Optimization for Feature Selection. *International Journal of Computer Applications*, *55*(8).

[3] Abdullah, U., Ahmad, J., & Ahmed, A. (2008, October). Analysis of effectiveness of apriori algorithm in medical billing data mining. In *Emerging Technologies, 2008. ICET 2008. 4th International Conference on* (pp. 327-331). IEEE.

[4] Abe, H., Yokoi, H., Ohsaki, M., & Yamaguchi, T. (2007, October). Developing an integrated time-series data mining environment for medical data mining. In *Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on* (pp. 127-132). IEEE.

[5] Abushariah, M. A., Alqudah, A. A., Adwan, O. Y., & Yousef, R. M. (2014). Automatic Heart Disease Diagnosis System Based on Artificial Neural Network (ANN) and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) Approaches. *Journal of Software Engineering and Applications*, *7*(12), 1055.

[6] Acharya, K. K., & Patel, R. C. (December 2014), Applications of Fuzzy-Neural and FPGA For Prediction of Various Diseases-A Survey. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering *(An ISO 3297: 2007 Certified Organization)* Vol. 3, Issue 12,

[7] Alkım, E., Gürbüz, E., & Kılıç, E. (2012). A fast and adaptive automated disease diagnosis method with an innovative neural network model. *Neural Networks*, *33*, 88-96.

[8] Al-Qarzaie, S., Al-Odhaibi, S., Al-Saeed, B., & Al-Hagery, M. Using the Data Mining Techniques for Breast Cancer Early Prediction.

[9] Anbarasi, M., Anupriya, E., & Iyengar, N. C. S. N. (2010). Enhanced prediction of heart disease with feature subset selection using genetic algorithm. *International Journal of Engineering Science and Technology*, *2*(10), 5370-5376.

[10] Anooj, P. K. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University-Computer and Information Sciences*, *24*(1), 27-40.

[11] Ansari, A. Q., & Gupta, N. K. (2011, December). Automated diagnosis of coronary heart disease using neuro-fuzzy integrated system. In *2011 World Congress on Information and Communication Technologies*.

[12] Balakrishnan, S., Narayanaswamy, R., Savarimuthu, N., & Samikannu, R. (2008, October). SVM ranking with backward search for feature selection in type II diabetes databases. In *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on* (pp. 2628-2633). IEEE.

[13] Bo, T., & Jonassen, I. (2002). New feature subset selection procedures for classification of expression profiles. *Genome biology*, *3*(4), 1-0017. http://genomebiology.com/2002/3/4/research/0017.1.

[14] Brameier, M., & Banzhaf, W. (2001). A comparison of linear genetic programming and neural networks in medical data mining. *Evolutionary Computation, IEEE Transactions on*, *5*(1), 17-26.

[15] Breault, J. L., Goodall, C. R., & Fos, P. J. (2002). Data mining a diabetic data warehouse. *Artificial Intelligence in Medicine*, *26*(1), 37-54.

[16] Brunie, L., Miquel, M., Pierson, J. M., Tchounikine, A., Dhaenens, C., Melab, N., ... & Morvan, F. (2003, September). Information grids: managing and mining semantic data in a grid infrastructure; open issues and application to geno-medical data. In *Database and Expert Systems Applications, 2003. Proceedings. 14th International Workshop on* (pp. 509-515). IEEE.

[17] Çalişir, D., & Doğantekin, E. (2011). An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. *Expert Systems with Applications*, *38*(7), 8311-8315.

[18] Cao, F., Liang, J., & Jiang, G. (2009). An initialization method for the K-Means algorithm using neighborhood model. *Computers & Mathematics with Applications*, *58*(3), 474-483.

[19] Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*, *47*(10), 44-48.

[20] Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. *Expert systems with applications*, *36*(4), 7675-7680.

[21] Hassan, S. Z., & Verma, B. (2007, October). A hybrid data mining approach for knowledge extraction and classification in medical databases. In *Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference on* (pp. 503-510). IEEE.

[22] Hastings, G., & Ghevondian, N. (1998). A selforganizing estimator for hypoglycemia monitoring in diabetic patients. In *20 th annual international conference of IEEE engineering in medicine and biology society* (Vol. 20, No. 3).

[23] Hathaway, R. J., & Bezdek, J. C. (1995). Optimization of clustering criteria by reformulation. *Fuzzy Systems, IEEE Transactions on*, *3*(2), 241-245.

[24] He, H. T., & Zhang, S. L. (2007, September). A New method for Incremental Updating Frequent patterns mining. In *Innovative Computing, Information and Control, 2007. ICICIC'07. Second International Conference on* (pp. 561-561). IEEE.

[25] Heart disease guide (Web MD), http://www.webmd.boots.com/heart-disease/guide/heart-disease-symptoms-types Heart disease type,

[26] Kalaiselvi. C and Dr. G.M. Nasira (2014) "A new approach for the diagnosis of diabetes and cancer using ANFIS" in world Congress on computing and communication technologies"

[27] Kalaiselvi. C and Dr. G.M. Nasira (2015) "Classification and Prediction of heart disease from diabetes patients using hybrid particle swarm optimization and library support vector machine algorithm" International Journal of Computing Algorithm (IJCOA) .

[28] Kalaiselvi. C and Dr. G.M.Nasira (2014) "A Novel Approach for the Diagnosis of Diabetes and Liver Cancer using ANFIS and Improved KNN " Research Journal of Applied Sciences, Engineering and Technology 8(2): 243-250, 2014

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

[30] Paolo Vigneri, Francesco Frasca, Laura Sciacca, Giuseppe Pandini (2009) "Diabetes and cancer" Endocrine Related cancer,

[31] Parthiban, G., Rajesh, A., & Srivatsa, S. K. (2011). Diagnosis of heart disease for diabetic patients using naive bayes method. *International Journal of Computer Applications*, *24*(3), 7-11.

[32] Parthiban, L., & Subramanian, R. (2007). Intelligent heart disease prediction system using CANFIS and genetic algorithm. *International Journal of Biological and Life Sciences*, *3*(3), 157-160.

[33] Parvathi, I., & Rautaray, S. (2014). Survey on Data Mining Techniques for the Diagnosis of Diseases in Medical Domain. *International Journal of Computer Science & Information Technologies*, *5*(1).

[34] Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010). Hybrid prediction model for Type-2 diabetic patients. *Expert systems with applications*, *37*(12), 8102-8108.

[35] Patil, R. R. (2014). Heart disease prediction system using Naive Bayes and Jelinek-mercer smoothing. *Int J Adv Res Comput Commun Eng*.

[36] Pradhan, M., & Sahu, R. K. (2011). Predict the onset of diabetes disease using Artificial Neural Network (ANN). *International Journal of Computer Science & Emerging Technologies (E-ISSN: 2044-6004)*, *2*(2).

[37] Rajeswari, Vaithiyanathan, V., & Pede, S. V. Feature Selection for Classification in Medical Data Mining.

[38] Ramachandran, P., Girija, N., & Bhuvaneswari, T. (2014). Early Detection and Prevention of Cancer using Data Mining Techniques. *International Journal of Computer Applications*, *97*(13).

[39] Researchgate http://www.researchgate.net

[40] Rukshan Athauda, (2008) Data mining Applications: Promise and challenges Data Mining and Knowledge Discovery in Real Life Applications (Eds. Julio Ponce and AdemKarahoca), I-Tech Education and Publishing Denmark, 202-214.

[41] Sethukkarasi, R., & Kannan, A. (2012). An Intelligent System for Mining Temporal Rules in Clinical Databases using Fuzzy Neural Networks. *European Journal of Scientific Research ISSN*, 386-395.

[42] Srideivanai Nagarajan, R.M. Chandrasekaran, (April 2015) "Design and implementation of expert clinical system for diagnosing diabetes using datamining techniques" IJST Vol.8, 771-776.

[43] Štrumbelj, E., Bosnić, Z., Kononenko, I., Zakotnik, B., & Kuhar, C. G. (2010). Explanation and reliability of prediction models: the case of breast cancer recurrence. *Knowledge and information systems*, *24*(2), 305-324.

[44] Subbalakshmi, G., Ramesh, K., & Rao, M. C. (2011). Decision support in heart disease prediction system using naive bayes. *Indian Journal of Computer Science and Engineering (IJCSE)*, *2*(2), 170-176.

[45] Vijiyarani, S. (2013). An Efficient Classification Tree Technique for Heart Disease Prediction. In *International Conference on Research Trends in Computer Technologies (ICRTCT-2013) Proceedings published in International Journal of Computer Applications (IJCA)(0975–8887)* (pp. 6-9).

[46] World heart federation,Cardiovascular disease risk factors"http://www.world-heart-federation.org /cardiovascular-health/cardiovascular-disease-risk-factors/"

[47] Xu, J., Sun, L., Gao, Y., & Xu, T. (2013). An ensemble feature selection technique for cancer recognition. *Bio-medical materials and engineering*, *24*(1), 1001-1008.

[48] Xu, J., Xu, T., Sun, L., & Ren, J. (2013). An Improved Correlation Measure-based SOM Clustering Algorithm for Gene Selection. *Journal of Software*, *8*(12), 3082-3087.

[49] Xu, J., Xu, T., Sun, L., & Ren, J. (2014). An Efficient Gene Selection Technique based on Fuzzy C-means and Neighborhood Rough Set. *Appl. Math*, *8*(6), 3101-3110.

[50] Yadav, R., Khan, Z., & Saxena, H. (2013). Chemotherapy Prediction of Cancer Patient by using Data Mining Techniques. *International Journal of Computer Applications (0975-8887)*, *76*(10).