I J C T A, 8(5), 2015, pp. 2153-2160

Intelligent Prediction Methods and Techniques Using Disease Diagnosis in Medical Database: A Review

M. Durairaj* and Nandhakumar Ramasamy**

Abstract: In medial field, Prediction is considered to be a systematic process, which helps to reduce the complexity of examining the diseases and their computerized system for the treatment ofthat diseaseparticularly their stages of diseases. To develop a prediction technique based on Artificial intelligence, Data mining approaches will be well suited in this context. Furthermore, medical practitioners and other related people needa particular methodology for immediate rectification.Medical data are analysed and relegated by using different techniques. This review also aims to tabulate various techniques applied in different works for disease diagnosis based on clinical data and the patient's history.Information technology application improves the treatments of Heart diseases, Cancer, Infertility and almost all the field of medical treatments with more effective and reduces the death ratio drastically.Proper pre-processing and cleaning of disease data helps in accurate diagnosis of diseases and to device treatment methodologies.

Index terms: Data mining, preprocessing Algorithms, Prediction Level, Accuracy, Disease diagnosis.

1. INTRODUCTION

Medical data mining plays a vital role for exploring hidden patterns and knowledge from the stored clinical data. These patterns can be utilized for diagnosis ofhealth care issues. However, the available clinical data are broadly distributed, heterogeneous in the environment, and raw in nature. These raw data should be cleaned and prepared in an acceptable format before subject to the data analysis for information retrieval [23]. Data mining is a term used for extracting useful information from unknown patterns in the clinical database. The systematize design for data mining is designed to reduce the noisy part of the data and to extract needed parameters to predicting the required details. Data preprocessing has to be performed in achieving correct information [13].

Clinical diagnosis is an important, yet problematic task that needs to be accomplished accurately and efficiently since the error in diagnosis may lead to fatal [20]. The automation of clinical diagnosis with the support of computers and tools would be useful for the physicians to diagnosis healthcare problems accurately. Lamentably all doctors do not possess expertise in every individual field and moreover there is a lack of resource persons inassuring places [8] [25]. Therefore, an automatic clinical diagnosis system would be extremely useful by taking all of them together. Felicitous computer-predicated data and/or decision support systems can aid in succeeding medical tests at a reduced cost [4].

Efficiency and precision of automated systems' execution need a comparative study of sundry techniques available for the evaluation. Using the support of clinical diagnostic system, the decision making can be done quickly with increased accuracy [6]. The medical databases contain a heterogeneous type of disease data and in this scenario, a particular algorithm cannot be suitable. So, the identification of algorithm which is suitable for particular diseases is difficult task [11] [26]. This paper aims to evaluate the different

* Assistant Professor,

** Research Scholar, School of Computer Science, Engineering and Applications, Bharathidasan University, Tiruchirappalli.

predictive data mining methods proposed in now a days for the diagnosis of different diseases. The organization of this paper is as, the Section 2 describes the methodology and section 3 defines the findings from the papers reviewed.Section 4 explains the results and discussion involved and section 5 comprises the conclusion of the survey.

2. METHODOLOGY

The methodologies discussed in this paper are compiled through the survey of journals and publications in the fields of medicine and information technology.This study has concentrated on more recent publications.

3. DATA MINING APPLICATIONS IN HEALTHCARE

3.1. Heart Disease Prediction

The term Heart disease involves various diseases that affect the heart. Heart disease is the major cause of death in different nations including India. Different data mining methods involved in the Heart disease prediction are discussed here [1] [17] which include Naïve Bayes, KNN, Decision tree, K means, ANN, J48 etc. These data mining algorithms are applied based on the nature of data to find heart disease. Naive Bayes is a machine learning technique and Naive Bayes classifiers is a set of probabilistic classifiers based on Bayes' theorem with strong independence suppositions between the features. Decision Tree is a standard classifier which is simple and easy to implement [2] [27]. Domain knowledge is not required and it can be able to hold high dimensional data. Using Decision Trees processes is the easiest to apply and understand [17].

KNN (K Nearest Neighbors) is a simple algorithm that solves all suitable cases and relegates incipient cases predicated on a kindred attribute. KNN has been utilized in statistical estimation and pattern recognition from the commencement of 1970's as a non-parametric method. J48 is developed in open source Java for the applications of C4.5 algorithm as a data mining tool. C4.5 is analgorithm that creates a decision tree related to a set of labelled input data [5]. In this paper, two methods of prediction process are taken for the study.There are predicting the disease before and after preprocessing the data. The preprocessing can reduce time factor and attributes [9]. The table 1 shows the clear view of the data mining technologies which are applied before preprocessing for the prediction.

Table 1
Prediction level of heart diseases before preprocessing

S. No	*Data MiningTechniques*	*Prediction Level (%)*
1	Navie Bayes	86.53
2	Decision Tree	89
3	NN	85.53
4	J48	77
5	Random Tree	75

The clinical data are preprocessed using Rough Set Theory. The result of knowledge discovery process can be obtained by using decision tree, association rules, decision rules, sequential pattern, etc. [18]. The most inclusive and interpretable erudition extracted is in the form of rules.Some rule induction algorithms and Rough Set Theory can generate in a sizable voluminous number of rules [24]. Lack of interpretability cuts down the advantages of rule based systems. The resulting inan immensely large number of rules is due to noise redundancy observed inthe input and/or training data sets.Rule pruning is the method which reduces the number of rules while maintain the quality of the systems. Rough set theory (RST) is a mathematical and an artificial perspicacious technique developed by ZdzislawPawlak, Warsaw University of Technology,

in the early 1980. RST is especially useful to find the relationships between parameters in data. The revelation of the relationship between parameters in the data is called cognizance revelation or data mining [18] [28]. The result of knowledge discovery is an information which is understandable and meaningful. The table 2 shows the results obtained by applying data mining technologies on the data after preprocessing.

Table 2
Prediction Level of Heart disease after Preprocessing

S. No	*Data MiningTechniques*	*Prediction Level (%)*
1	Navie Bayes	96.5
2	Decision Tree	97.2
3	NN	88.3
4	J48	87.6
5	Random Tree	89.8

The comparisonwith other algorithms such as j48 and simple cart algorithm, the performance of the decision treeis high. In another experiment, Quad-clustering algorithms like Simple K means, D-stream, Global K-means, K-means++ are compared and theperformancesin term of accuracy are shown in table 3.

Table 3
Prediction Level of Heart disease using Clustering Algorithms

S. No	*Clustering Algorithms*	*Prediction Level (%)*
1	Simple K-means	83
2	D-Stream	87
3	Global K-means	90
4	K-means++	93

During the cluster formation, the accuracy of suitably classified instancesincreases, whereasincorrectly classified instances decrease. The cluster analyses show that the performance of the k-means++ algorithm is relatively high.

3.2. Diabetic Disease Prediction

Diabetes mellitus is a group of metabolic diseases categorized by high blood sugar levels that result from defects in insulin secretion, or its action, or both. Diabetes mellitus, commonly referred to as diabetes was first identified as a disease related to “sweet urine,” and excessive muscle loss in the early world [6]. The following algorithms are frequently used in many of the reviews. Nearest Neighbors, Bagging, Multilayer perceptron, K-Nearest Neighbors are set of algorithms compared with each other’s to find the accuracy of classification [3]. J48, Navie Bayes, Decision Tree and Random tree are used for classifications. The comparison resultsof these algorithms with their performances are shown in table 4.

Table 4
Prediction Level of Diabetics disease

S. No	*Data Mining Techniques*	*Prediction Level (%)*
1	Nearest Neighbor	70.53
2	Bagging	76.45
3	Multilayer Perceptron	75.91
4	Random forest	75.78
5	KNN	77.08

The table 4 depictsthe accuracy level, obtained by executing all thetechniques.Each algorithm may vary from others and results show the best suitable technique for classification and prediction.

3.3. Diseases Prediction using Hemogram Blood samples

A Hemogram/Complete Blood Count (CBC) is possibly the commonest laboratory test conducted in any labora-tory. As the name suggests, a hemogram/CBC analyses various blood components like RBC, WBC and platelets. These components are non-specific, but very important indicators of general well being of an individual. Many of the parameters included in a hemogram do not pinpoint towards a specific diagnosis.Although careful analysis of a combination of several parameters included in a hemo-gram may help the clinician/hematologist to arrive at important conclusions. Sample required is EDTA whole blood and the entire range of tests can easily be conducted with just 2/3 ml of blood[3]. Five diseases leukaemia, Inflammatory disease, Bacterial infection or viral infection, HIVdisease and pernicious anaemia are predicted from this hologram blood test sample analysis [12].

The K-means, fuzzy C means algorithms are used and proposed an algorithm called weight based K-means algorithm for analysing the blood sample data for the accuracy of prediction [23]. The table 5 shows the results of the accuracy of the Algorithms.

Table 5
Accuracy of the Blood related disease prediction using Blood test dataset

S. No	*Diseases*	*K Means (%)*	*Fuzzy C Means (%)*	*Weighted K Means (%)*
1	Leukemia	78	75	85
2	Inflammatory Disease	85	80	90
3	Bacterial or Viral Infection	80	74	98
4	HIV Infection	88	88	93
5	Pernicious Anaemia	88	83	94

The Table 5 explains the performance of the algorithmwhich gave a high accuracy rate and clear view of the accuracy range.Based on the table, performance of the Weighted K means clustering algorithm is high. This weighted K Means clustering algorithm is an example of integrated algorithm. Based on a single data set, it is possible to predict multiple diseases such asleukemia, Inflammatory disease, Bacterial or viral infection, HIV infection and pernicious anaemia.

3.4. Breast Cancer Prediction

Breast cancer is a malevolenttumor that starts in the cells of the breast. A malevolenttumor is a group of cancer cells that can grow into adjacent tissues or spread to distant areas of the body [15]. The disease occurs almost entirely in women, but men too can get it.Detection of breast cancer in early stages can be cured easily. This review summarizes the prediction accuracy of breast cancer by applying data mining algorithms[4].Figure 1 comprises the prediction accuracy by employing data mining algorithms.

3.5. InfertilityPrediction

Infertility is one of the major issues in the medical field. The infertilityaffect patientspsychologically and physically [19].There are different treatment procedures such as Intrauterine Insemination (IUI), Intracytoplasmic Sperm Injection (ICSI), In-Vitro Fertilisation (IVF) and Gamete Intrafallopian Transfer (GIFT) [16] available to treat infertility patients.The success rate of treatment is minimal. The data mining algorithms applied to predict the success rateof infertility treatment which can be helpful to counsel a patient to make them fit psychologically for the treatments [21]. The psychological factor affects the success rate of treatments. The

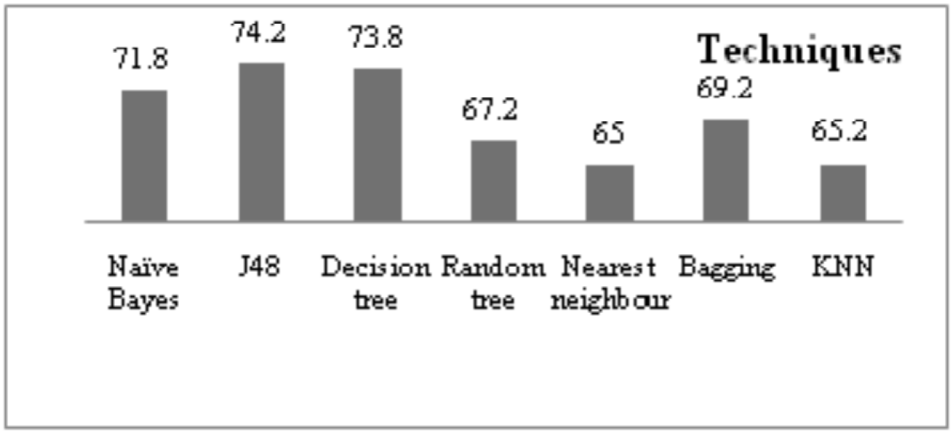

Figure 1: Breast cancer prediction level

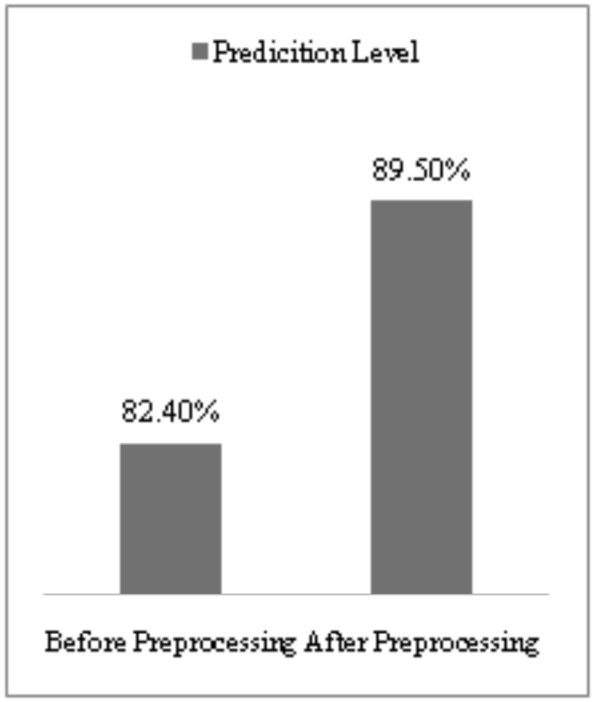

Figure 2: Human Artificial insemination prediction level using data mining algorithms

combination of different in-vitro test results and past history of treatments acquired from the data set plays a vital role for predicting success rate of the suggested treatment [22]. Data mining algorithms such as Rough Set Theory and Artificial Neural Network are employed for identifying influential parameters (test results) by extracting reduct data setwhich can be used fortreatment success rate prediction. Figure 2 shows the prediction accuracy ofsuccess rate of treatment before and after preprocessing infertility test data.

In figure2, the importance of preprocessing and the level of accuracywhich increases due to the data cleaning / preprocessing. Different data mining algorithms and their performance are shown in the figure 2.

4. RESULTS AND DISCUSSION

The data mining techniques applied for the prediction and analysis of diseases, and the results acquired are tabulated in Table 6.

These results and finding are summarized from the recent research works published in journals.Every methodhas their own merits and demerits in processing the data. The data set used in these works isplayinga

Table 6
Preferred Data Mining algorithms for Human Diseases prediction

S. No	*Diseases*	*Methods*	*Result*
1	Heart Disease	Decision Tree	97.2%
2	Diabetic Disease	KNN	77.08%
3	Leukemia	Weighted K Means	85 %
4	Inflammatory Disease	Weighted K Means	90 %
5	Bacterial Or Viral Infection	Weighted K Means	98 %
6	HIV Infection	Weighted K Means	93 %
7	Pernicious Anaemia	Weighted K Means	94 %
8	breast cancer	J 48	74.2%
9	Infertility	Ann	89.50%

role in calculating accuracy. The reviewed results demonstrate that the preprocessed or cleaned data gives more accurate results than the data set without preprocessed. The data mining techniques should be chosen based on clinical data set, disease and the type of information required. In recent times, researchers propose hybrid methodologies for clinical data by combining different data mining techniques in order to increase the accuracy of prediction or classification and to reduce the processing time.

Medical diagnosis is considered as a significantly complicated task that needs domain knowledge and the proper application of information technology to perform accurately and efficiently. The computerization of the same would be exceedingly valuable. Medical decisions are often made based on doctor s perception and practice rather than on the knowledge rich data hidden in the database. This practice leads to undesirable biases, errors and excessive medicinal costs which affects the quality of service provided to patients. Data mining have the probable to produceknowledge-rich surroundings which can help to expressively improve the quality of medical decisions.

5. CONCLUSION

In this paper, we discussed different prediction algorithms of data mining used in the field of medical diagnosis. This work focused on reviewingvariousdata mining algorithms and combining the several target attributes for intelligent and effective medical diagnosisto support medical practitioners.Heart disease, Diabetic, Blood disease, Breast cancer and infertility data processing using different data mining algorithms are compiled and compared. Each data mining algorithms performs its respective role effectively. The comparison results support the development of hybrid data mining technologies for the more accurate processing of clinical data.Hybridizing two or more algorithms may process with increased accuracy on erroneous / redundant data set since it holds the characteristics of individual algorithms involved.The future work of this work is to propose an effect hybrid tool for effectively process the clinical data with an increased accuracy.

References

[1] Jyoti Soni, Ujma Ansari, Dipesh Sharma and SunitaSoni,(2011), "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" International Journal of Computer Applications, (0975–8887), Volume 17– No. 8, pp. 45-48, March 2011.

[2] Hlaudi Daniel Masethe and Mosima Anna Masethe, "Prediction of Heart Disease using Classification Algorithms" Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, 22-24 October, 2014, San Francisco, USA.

[3] S. Vijayaraniand S. Sudha (2015) "An Efficient Clustering Algorithm for Predicting Diseases from Hemogram Blood Test Samples" Indian Journal of Science and Technology, Vol. 8(17), 52123, August 2015.

[4] Alpna Sharma and Seema Sharma, "An Analytical Study on Classification Algorithms for Medical Datasets" International Journal of Electrical Electronics & Computer Science Engineering, Special Issue-TeLMISR 2015, ISSN: 2348-2273, pp. 72-76.

[5] K. Rajalakshmi, S.S. Dhenakaran and N. Roobini, "Comparative Analysis of K-Means Algorithm in Disease Prediction"International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 7, July 2015 pp. 2697-2699.

[6] Aanchal Oswal, Vachana Shetty, Mustafa Badshah, Rohit Pitre and Manali Vashi, "A Survey On Disease Diagonsis Algorithms" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 11, November 2014, pp. 3757-3761.

[7] Sathyabama Balasubramanian and Balaji Subramani, "Symptom's Based Diseases Prediction In Medical System By Using K-Means Algorithm" International Journal of Advances in Computer Science and Technology, ISSN 2320–2602, pp. 123-128.

[8] Monali Dey and Siddharth Swarup Rautaray Monali Dey, "Study and Analysis of Data mining Algorithms for Healthcare Decision Support System" International Journal of Computer Science and Information Technologies, Vol. 5 (1), 2014, pp. 470-477.

[9] D.Nithya and R.Shanmugavadivu, "A Scrutiny Prediction of Quad-Clustering Algorithms of Healthcare Application" International Journal of Advanced Research in Computer Science and Software EngineeringVolume 3, Issue 12, December 2013, pp. 68-71.

[10] P. Radha and B. Srinivasan, "Predicting Diabetes by consequence the various Data Mining Classification Techniques" International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014, pp. 334-339.

[11] Mangesh Metkari and M.A. Pradhan, "Comparative Study of Soft Computing Techniques on Medical Datasets" International Journal of Science and Research Volume 3 Issue 12, December 2014, pp. 761-765.

[12] Tarigoppula V.S Sriram, M. Venkateswara Rao, G V Satya Narayana, DSVGK Kaladhar and T Pandu Ranga Vital," Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms" International Journal of Engineering and Innovative Technology, Volume 3, Issue 3, September 2013, pp. 212-215.

[13] B.S. Ahn, S.S.Cho and C.Y.kim(2000) "The integrated methodology of rough set theory and artificial neural network for business failure prediction" Expert System with applications/Elsevier/locate.P.no 65-74.ICLE

[14] S. Anto and S.Chandramathi, " Supervised Machine Learning Approaches for Medical Data Set Classification-A Review" IJCST, Vol. 2, Issue 4, Oct .-Dec. 2011, pg.no: 234-270.

[15] NN Das and Anjali Saini, "A Study on Association Rule Mining Using ACO Algorithm for Generating Optimized Result Set", International Journal of Computer Science and Mobile Computing, Vol. 2, Issue. 11, November 2013, pp. 123–128.

[16] M. Durairaj and R. Nandhakumar (2013), "Data Mining Application on IVF Data for the Selection of Influential Parameters on Fertility" International Journal Engineering and Advanced Technology, Vol-2, Issue-6, pp. 262–266.

[17] K. Srinivas, G. Raghavendra Rao, A. Govardhan (2012), "Analysis of Attribute Association in Heart Disease Using Data Mining Techniques" International Journal of Engineering Research and Applications (IJERA) pp. 1680-1683.

[18] Lixiang Shen, Francis, E.H. Tay, Liangsheng Qu and Yudi Shen, "Fault Diagnosis using Rough Sets Theory" computer industry (2000) pp. 61-72.

[19] S.J. Kaufmann, J.L. Eastaugh, S. Snowden, S.W. Smye and V. Sharma (1997), "The application of Neural Networks in predicting the outcome of in-vitro fertilization "Human Reproduction vol. 12 no. 7 pp. 1454–1457.

[20] Norsalini Salim., "Medical diagnosis using Neural Networks", Faculty of Information Technology, University Utara Malaysia, Sintok, Kedah, 2004.

[21] M. Durairajand P. Tamilselvan (2013), "Applications of Artificial Neural Network for IVF Data Analysis and Prediction" Journal of Engineering, Computers & Applied Sciences, vol-2, issue-9, pp. 11-15.

[22] Kay Elder, & Brian Dale., 2000, "In-Vitro Fertilization", Second Edition, United Kingdom at the University Press, Cambridge.

[23] Muller H., Freytag J.," Problems, Methods, and Challenges in Comprehensive Data Cleaning" Humboldt-Universitat zu Berlin, Germany.

[24] M.A. Nishara Banu and B Gomathy, "Disease Predicting System Using Data Mining Techniques" International Journal of Technical Research and Applications, Volume 1, Issue 5, Nov-Dec 2013, PP. 41-45.

[25] Sivagowry. S, Durairaj. M, Persia. A "An Empirical Study on applying Data Mining Techniques for the Analysis and Prediction of Heart Disease", International Conference on Information, Communication and Embedded Systems, S.A . Engineering College, Chennai, February 2013.

[26] Durairaj. M, Sivagowry. S, "An Intelligent Hybrid Quick Reduct Particle Swarm Optimization Algorithm for Feature Reduction in Cardiac Disease Prediction", International Journal of Emerging Technologies in Computational and Applied Sciences, Vol 12(2), pp. 163-173, May 2015.

[27] Sivagowry. S, Durairaj. M, "An Intelligent System based on Fuzzy Inference System to prophesy the brutality of Cardio Vascular Disease", An International Journal of Advances in Computer Science, Vol 6(18), November 2015.

[28] Durairaj. M, Sivagowry. S, "A Survey on Particle Swarm Optimization and Rough Set Theory in Feature Selection for Heart Disease Prediction", International Journal of Computer Science and Mobile Computing, Vol 4(3), pp 87-92, March 2015.