IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

A Survey on Prediction of Disease with Data Mining

Niyati I. Patel[1*], Hiren R. Patel[2]

[1,2]Dept. of Computer Engineering & Information Technology, Government Engineering Collage, Modasa, India

[*]*Corresponding Author: patelniyati713@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i2.289293 | Available online at: www.ijcseonline.org

Accepted: 13/Feb/2019, Published: 28/Feb/2019

***Abstract*—** In today's era there is a huge amount of data available with health care; however, the knowledge about the data is rather poor. So there is a need to process that enormous size of medical dataset instead of just storing extract valuable information or useful knowledge. Data mining is the process of extracting hidden knowledge from large volumes of raw data using techniques like statistical analysis, machine learning, clustering, neural networks and genetic algorithms. A logical combination of multiple pre-existing techniques or Hybrid algorithms for data mining to enhance performance and provide better results. Data mining is used to discover hidden patterns and relationships out of data and presenting it in a form that can be easily understood. Data mining plays an important role in disease prediction. Data Mining is used intensively in the medical field to predict diseases such as heart disease, diabetes, breast cancer etc. In this paper, a survey is carried out on several single and hybrid data mining approaches used for disease prediction.

***Keywords*—** Data mining, Data Mining Techniques, Hybrid Approach, Diseases

I. INTRODUCTION

The growth of medical databases is very high with the development of information technology. Extensive medical data is available in electronic form through some of the database management systems. These systems generate a large volume of data on daily basis. Medical data interpretation plays a crucial role in many medical applications. Nowadays researchers have the motivation to mine useful information from these medical databases. Data mining techniques play an important role in finding patterns and extracting knowledge to provide better patient care and effective diagnostic capabilities with the increasing volume of data. This technique can help in data modelling and creating disease analysis tools, in order to produce information that enhances the decision-making process in the healthcare domain [1].

Giudici defines Data mining as "a process of selection, exploration, and modelling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database". Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases [2].

Data mining is the process of extraction of implicit, unique, and potentially useful information from data. Knowledge Discovery from Data (KDD) is the goal of the data mining process [3]. KDD involves processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation, and Data Presentation. Once all these processes are over, we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Healthcare analysis, Production Control, Science Exploration, etc.

A. Data Mining Tasks

There are the number of data mining tasks such as classification, prediction, time-series analysis, association, clustering, summarization etc. All these tasks are either predictive data mining tasks or descriptive data mining tasks [4]. With a model from the available data set, Predictive data mining tasks are helpful in predicting unknown or future values of another data set of interest.

A medical practitioner trying to diagnose a disease based on the medical test results of a patient can be considered as a predictive data mining task [5]. Descriptive data mining tasks usually find data describing patterns and come up with new, significant information from the available data set [6],[7]. For better prediction of result nowadays a combination of the above techniques is the main focus. Hybrid Data Mining Technique is a "logical combination of multiple pre-existing techniques or tasks." Hybridization of clustering and classification method wherein the first state the clustering algorithm K-mean is applying and after that in the second state for each cluster the Decision tree algorithm is applied to generate the tree for batter classified result [8].

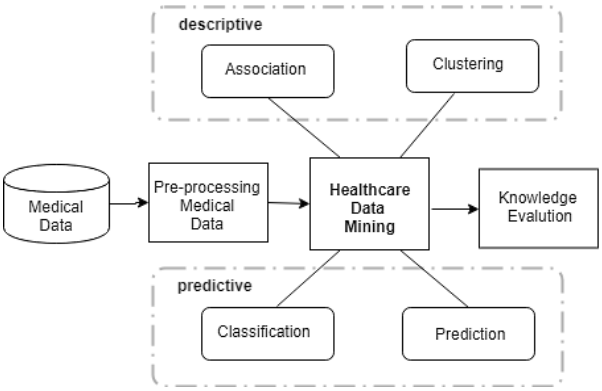

Figure 1. *Healthcare Data mining*

Classification
Classification is a supervised learning approach that derives a model to determine the class of an object based on its attributes. Each record with a set of attributes, a collection of records will be available. One of the attributes will be a class attribute and the goal of the classification task is assigning a class attribute to a new set of records as accurately as possible. The data mining classification techniques, namely Decision Trees, Naive Bayes, and Neural Networks are analyzed on the Heart disease database [9], [22].

Prediction
The possible values of missing or future data predict by prediction task. Prediction involves developing a model based on the available data and this model is used in predicting future values of a new data set of interest [5].

Association
The association or connection among a set of items or attributes discovered by association. Association identifies the relationships between attributes used for analysis. Associations are represented in the form of rules, or implications [6], [10].

Clustering
Clustering is unsupervised classification, no predefined classes are there. Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters [11]. In other words, the natural grouping or structure in a data set is done with clustering. Data mining adds to clustering the complications of very large datasets with many attributes of different types [7]. A variety of algorithms successfully applied to real-life data mining problems for unique computational requirements. K-Means Clustering algorithm is based on this partitioned methods to decompose the data set into a set of k disjoint clusters [8]. Clustering based on density (local cluster criterion), such as density-connected points to discover clusters of arbitrary shape [12].

B. Objectives
The primary challenge in health care is to accurately predict the disease and make intelligent decisions regarding that disease. An enormous amount of information on the medical field can be used in the decision making of the disease. In this data, there is hidden information which is mostly unexplored, so the motivation is to discover this hidden knowledge [1]. A data mining model can accomplish these challenges by utilizing real-world data and provide accurate decision making. Various data mining techniques used to predict the disease and also the effective development of the prediction model using the combination of various data mining techniques to predict disease and performance in prediction. So, it shows that to get reasonable accuracy in the medical field, data mining can be applied to predict or classify the data.

C. Applications
Data mining is helpful in the detection of, gaining deeper knowledge about the disease, making better health care policies, taking decision for preventing disease, minimizing death rates in hospitals, saving money, fraudulent insurance claims detection, diagnosis of various diseases, namely cancer, sugar, etc.

The organization of the paper as follows, Section I contains the introduction of data mining, tasks of data mining, objective and applications of data mining supported to healthcare domain, Section II contains the related work of data mining methods for prediction of disease and the Section III conclusion of the survey.

II. RELATED WORK

Humar Kahramanli and Novruz Allahverdi (2007) [13] presented a hybrid system to classify the data and more reliable diagnosis of the heart and diabetes diseases, system obtained data with use of artificial neural network (ANN) and fuzzy neural network (FNN) on the datasets Pima Indians diabetes and Cleveland heart disease from University of California, Irvine(UCI) repository. The proposed hybrid neural networks algorithm helped achieving high accuracy rate than previous studies.

Liang Sun et al. (2010) [11] proposed an effective novel support vector and K-Means based hybrid algorithm that exploits the advantages of both support vector clustering and K-Means for data clustering. Tasks considered on Iris data, first to identify the outliers and overlapping data points through the support vector approach (SVA). Second to remove the outliers and overlapping data points and then run the K-Means on the rest data points to obtain clustered data set. Here the SVC has an advantage of generating cluster boundaries of an arbitrary shape. Thus, a promising way to solve this problem is to build a support vector description for

every single cluster and then assign the removed data points to the cluster with the smallest distance.

Anchana Khemphila and Veera Boonjing (2011) [14] introduced a classification approach by using Multi-Layer Perceptron (MLP) with a Back-Propagation learning algorithm and a feature selection using information gain with heart disease patients database from Cleveland Clinic Foundation at UCI database collection. Feature selection increases computational efficiency while ANN improving classification accuracy. The multilayer with error correction learning is feed-forward neural networks trained with the standard back-propagation algorithm. The most popular static network in the multilayer gives the approximate performance of optimal statistical classifiers in difficult problems.

Jayaram et al. (2012) [15] described a hybrid model with two tasks for classifying the Pima Indian Diabetic Database (PIDD).In the first clustering done using the K-means algorithm to identify and eliminated incorrectly classify instances. In the second task, a finely tuned classification done using Decision tree C4.5 by taking the correctly clustered instance of the first task. Type II Diabetes Prediction using the cascaded K-means clustering and the rules generated by cascaded C4.5 tree with categorical data is easy to interpret as compared to rules generated with C4.5 alone with continuous data. The cascaded model with categorical data obtained the efficient classification accuracy.
Kavita Agarwal et al. (2013) [16] proposed an intelligent data mining system based on genetic algorithm optimized neural network for the prediction of heart disease based on risk factors categories. Firstly the system determines the number of inputs, layers and hidden neurons of the neural network. Second, it uses the back propagation algorithm to train the networks using the weights optimized by a genetic algorithm.

Nihat Yilmaz et al. (2014) [17] introduced a new modified K-means Algorithm for the clustering based data preparation system for the elimination of noisy and inconsistent data and to achieve greater stability. This modified k means algorithm is using a weight factor which incorporated as a factor to the Euclidean distance measurement that determines the distance between the set centres and the samples. Databases the Statlog (Heart) data set, SPECT Heart data set and Pima Indians Diabetes in the UCI database that was subject to data preparation are classified reliably with SVM.

Neelam Singhal and Mohmmad Ashraf (2015) [18] advanced an old hybrid classifier by combining evolutionary and non-evolutionary algorithms to improve the accuracy, comprehensibility, and timing of the classification. DT used for feature selection because it doesn't suffer from oversensitivity to irrelevant or/and redundant attributes. Simple K-Means used for clustering and then after the dataset classify using a genetic algorithm (GA). This algorithm applied on five datasets Lymph, Heart-statlog, Molecular-Biology, Sonar, and Zoo from UCI Repository.
K.Gomathi and Dr. D. Shanmuga Priyaa (2016) [19] proposed a study with datasets of diabetics and breast cancer in order to evaluate the precise and computationally efficient classifier in data mining and machine learning areas for Medical applications. A decision tree (J48) is a classifier in the tree structure which induce the tree and its rules that will be used to make predictions. The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors, simple to build, no difficult iterative parameter estimation which makes it particularly useful for very large datasets.

C. Kalaiselvi (2016) [20] proposed the Average K-Nearest Neighbor (AKNN) algorithm to make the KNN a faster which improved classification accuracy, efficiency and reduces complexity based on attribute reduction. Diagnosing of heart diseases using the AKNN algorithm, the supersample is created for each class which is the average of every training sample in that particular class. The AKNN searches sample data and find the closest to the input when the test samples are given to the AKNN. The closest neighbor is identified by measuring the distance between the neighbors.

Garima Singh et al. (2017) [4] proposed the system approach uses the Naïve Bayes Classifier. The system web application enables users to induce instant steerage on their cardiopathy through an intelligent system online. The application required numerous details and also permits the user to share their heart-related problems. It then processes user specific details to see for numerous health problem that might be related to that given details.

Kanika Bhatia and Rupali Syal (2017) [21] introduced a hybrid approach with using K*-Means, genetic algorithm (GA), SVM. K*-Means is an optimized hierarchical clustering method used to remove the inconsistency found in the data, and for optimal feature selection, the genetic algorithm used with SVM for the purpose of classification. The application of the proposed hybrid clustering model applied to the Pima Indians Diabetes dataset shows an increase in accuracy, sensitivity and positive predicted value and reduction of computational cost.

Anjana Suresh et al. (2017) [12] presented PCACRA, a clustering approach acquired by an unsupervised process. As the clustering approach DBSCAN is used and for regression multiclass logistic regression. DBSCAN clustering algorithm fragmented entire dataset into disjoint clusters and the resulted clusters which are found to contain fewer instances focused by multiclass logistic regression. This methodology

found to be 80% accurate for the prophecy of the type of cardiac arrhythmia by assembling the use of DBSCAN clustering and multiclass logistic regression algorithms.

P. Manivannan and Dr. P. Isakki (2017) [7] introduced the system for diagnosis of Dengue disease. The dengue viruses occur in 4 serotypes (DENV-1 to DENV-4). Predicting the relationship between the dengue serotypes helps to discover antibiotic for dengue. One R method used for attribute selection and K-means clustering algorithm has been implemented to predict the dengue patients who are affected by dengue depending upon categorization of age group.

Table 1. Comparative Study of Different Techniques

No.	Title	Techniques	Strong Points	Remarks
1.	Design of a hybrid system for the diabetes and heart diseases	artificial neural network (ANN), fuzzy neural network (FNN)	With rules extraction more reliable diagnosis.	Handling the extraction of rules from trained hybrid neural networks.
2.	A Novel Support Vector and K-Means based Hybrid Clustering Algorithm	Support Vector Clustering (SVC), K-Means	Deal with overlapping data points.	Handling of Parameters Selection effectively.
3.	Heart disease Classification using Neural Network and Feature Selection	Multi-Layer Perceptron (MLP), ANN	feature selection increase computational efficiency while improving classification accuracy, decrease the complexity	Handling the discretization of continuous-valued attributes
4.	Rule based classification for diabetic patients using cascaded k-means and decision tree c4.5	K-Means, Decision Tree (c4.5)	rules generated with categorical data	More time
5.	Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors	Genetic Algorithm, Neural Networks	Identified important risk factors.	Initialization of the NN weights, very slow convergence.
6.	A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases	Modified K-means Algorithm, SVM	distance measurement has been modified, the success rate of the algorithm increases	Handle the inconsistency, disruptive data.
7.	Performance Enhancement of Classification Scheme in Data Mining using Hybrid Algorithm	Decision Tree (DT), Genetic Programming (GP)	Improve the accuracy, time and comprehensibility. GP/DT hybrid approach is working better than that of the GP and DT individually.	Need focus on bias.
8.	Multi Disease Prediction using Data Mining Techniques	Naive Bayesian, Decision Tree (J48)	No difficult iterative parameter estimation. Dimensionality reduction.	
9.	Diagnosing of Heart Diseases using Average K-Nearest Neighbor Algorithm of Data mining.	AKNN	To remove the voting drawback of KNN and make KNN faster AKNN is used.	KNN takes more time to train data, It does not work if the data is noisy.
10.	Heart Disease Prediction using Naïve Bayes.	Naïve Bayes	Proposes web application that enables the user to induce instant steerage on their cardiopathy through an intelligent system online	Class conditional independence, loss of accuracy, Probability value zero problems, very strong assumption on the shape of data distribution.
11.	Predictive Analysis Using Hybrid Clustering in Diabetes Diagnosis	K*-Means, GA, SVM	Modification in K-Means technique by including weight factor to provide higher priorities.	The time required to obtain results is quite high. The model uses three algorithms, so the complexity increases.
12.	Prediction of Cardiac Arrhythmia type using Clustering and Regression Approach (P-CA-CRA)	DBSCAN, multiclass logistic regression algorithms	The proposed system is found to be 80% accurate.	Difficulty in work with clusters of different sizes and density.
13.	Dengue Fever Prediction Using K-Means Clustering Algorithm	One R, K-Means	Increasing the proficiency of the output.	Selection of group (dengue serotypes depending upon the age group)

III. CONCLUSION

This survey presents a systematic and healthy review of the applications of data mining methods in the healthcare domain, with a focus on the application and the used techniques can deliver optimized the results to solve the problems in the healthcare domain. In this survey, we presented an overview of the data mining techniques for the diagnosis and prognosis of different diseases. Each technique is unique in its very own way, which may be reasonable for various applications. These applied hybrid data mining techniques have shown promising results in the diagnosis of diseases. By using hybrid data mining techniques in the real data, we can get more accurate results for the system. The advantages of such a system are remarkable in their own way. By using accurate data in the health care system, People can be analysed for diseases quickly and may possible to find cure for that particular disease by detecting the disease at an early stage.

ACKNOWLEDGMENT

I would like to thank my guide, Assistant Professor of Department of Information Technology Government Engineering College, Modasa for providing continual encouragement through a relaxed approach and support and proper guidance for holding us to a higher standard.

REFERENCES

[1] N. Prabakaran, R. Kannadasan, "Prediction of Cardiac Disease Based on Patient's Symptoms", In the Proceedings of Second International Conference on Inventive Communication and Computational Technologies, India, pp. 794 – 799, 2018.

[2] P. Giudici, "Applied Data Mining: Statistical Methods for Business and Industry", Wiley Publisher, New York, 2003.

[3] J. Han, M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kauffmann Publishers, San Francisco, 2001.

[4] G. Singh, K. Bagwe, S. Shanbhag, S. Singh , S. Devi, "*Heart Disease Prediction using Naïve Bayes*", International Research Journal of Engineering and Technology, Vol. 4, No.3, pp.1-3,2017.

[5] F. Babič, J. Olejár, Z. Vantová, J. Paralič, "Predictive and Descriptive Analysis for Heart Disease Diagnosis", In the Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, Czech Republic, pp. 155 - 163, 2017.

[6] Carlos O., "Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction", Ieee Transactions on Information Technology in Biomedicine, Vol. 10, No. 2, pp.334–343, 2006.

[7] P. Manivannan , P. Isakki Devi, "Dengue Fever Prediction Using K-Means Clustering Algorithm", In the Proceedings of IEEE International Conference on Intelligent Techniques In Control, Optimization and Signal Processing, India, pp. 1 – 5, 2017.

[8] A. Ahlawat, B. Suri, "Improving classification in data mining using hybrid algorithm", In the Proceedings of the 2016 1st India International Conference on Information Processing, India, pp. 1-4, 2016.

[9] C. S. Dangare, S. S. Apte, "*Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques*", International Journal of Computer Applications, Vol. 47, No.10, pp.44-48, 2012.

[10] B. Milovic, "*Prediction and Decision Making in Health Care using Data Mining*", International Journal of Public Health Science, Vol. 1, No. 2, pp. 69-78, 2012.

[11] L. Sun, S. Yoshida, Y. Liang, "A novel support vector and K-Means based hybrid clustering algorithm", Proceedings of the 2010 IEEE International Conference on Information and Automation, China, pp.126-130, 2010.

[12] Cp. Prathibhamol, A. Suresh, G. Suresh, "Prediction of Cardiac Arrhythmia type using Clustering and Regression Approach (P-CA-CRA)", In the Proceedings of International Conference on Advances in Computing, Communications and Informatics, India, pp. 51 – 54, 2017.

[13] H. Kahramanli, N. Allahverd, "Design of a hybrid system for the diabetes and heart diseases", Expert Systems with Applications, Vol. 35, No.1-2, pp.82-89, 2008.

[14] A. Khemphila, V. Boonjing, "Heart Disease Classification Using Neural Network and Feature Selection", In the Proceedings of the 2011 21st International Conference on Systems Engineering, USA, pp.406-409, 2011.

[15] A. G. Karegowda, V. Punya, M. A. Jayaram, A. S. Manjunath, " *Rule based classification for diabetic patients using cascaded k-means and decision tree c4.5*", International Journal of Computer Applications, Vol. 45, No.12, pp.45-50, 2012.

[16] S. U. Amin, K. Agarwal, R. Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors", In the Proceedings of the IEEE Conference on Information and Communication Technologies, India, pp. 1227-1231, 2013.

[17] N. Yilmaz, O. Inan, M. S. Uzer, "A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases" , Springer Science + Business Media, New York, pp.38-48, 2014.

[18] N. Singhal, M. Ashraf, "Performance Enhancement of Classification Scheme in Data Mining using Hybrid Algorithm", In the Proceedings of the International Conference on Computing, Communication and Automation, India, pp.138-141, 2015.

[19] K. Gomathi, D. S. Priyaa, "*Multi Disease Prediction using Data Mining Techniques*", International Journal of System and Software Engineering, Vol. 4, No.2, pp.12-14, 2016.

[20] C. Kalaiselvi, "Diagnosis of Heart Disease Using K-Nearest Neighbor Algorithm of Data Mining", In the Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development , India, pp. 3099-3103, 2016.

[21] K. Bhatia, R. Syal, "Predictive Analysis Using Hybrid Clustering in Diabetes Diagnosis", In the Proceedings of the Recent Developments in Control, Automation & Power Engineering, India, pp. 447 – 452, 2017.

[22] S. Palaniappan, R. Awang, "*Intelligent heart disease prediction system using data mining techniques*", IJCSNS International Journal of Computer Science and Network Security, Vol. 8, No.8, pp. 108-115, 2008.

Authors Profile

Ms. Niyati Ishvarbhai patel pursued Bachelor of Technology degree from Dharmsinh Desai University, Nadiad, India in 2017 and currently pursuing Master of Engineering Computer Engineering degree from Department of Computer Engineering & Information Technology, Government Engineering Collage (GEC), Modasa, India in year 2018.

Mr. Hiren R. Patel received the Bachelor of Engineering degree from SCET, Surat, India in 2004 and Master Of Engineering Computer Engineering at GEC Modasa in 2016. He is now with Government Engineering Collage, Modasa as the assistant professor since 2009. Currently he is pursuing PhD from Gujarat Technological University, Ahmedabad. His main research work focuses on Cloud Computing, Web Development, Data mining. He has 12 years of teaching experience.

