# RESEARCHER REVIEWS ON DATA MINING TECHNIQUES FOR MEDICAL DATASET: A SURVEY

**A.Ameer Rashed Khan[1], Dr. S. Shajun Nisha [2], Dr. M. Mohamed Sathik [3]**

[1]Research Scholar PhD.

Email : ameerkhan.a1694@gmail.com

[2]Assistant Professor & Head, Research Supervisor.

[3]Principal

[1,2,3]PG & Research Department of Computer Science, Sadakathullah Appa College, Tirunelveli, India. Affiliation of Manonmaniam Sundaranar University, Abhishekapatti, Tirunelveli, 627012, Tamilnadu, India.

***ABSTRACT - In advanced computer technologies huge amount of raw data are being generated day by day in every field especially health care industry, to make the raw medical data into useful information to predict and diagnosis the diseases data mining techniques have been used. Data mining transforms tremendous amount of raw medical data into valuable information that can be helpful to take quick decision and prediction. This paper aims to provide a quick and easy review and understanding some of the recent research on predicting diseases using data mining techniques from 2009-2018. The comparison demonstrates the accuracy level of each model given by different researchers.***

**KEYWORDS:** Raw Data; Data Mining; Medical Data; Disease Diagnosis; Prediction; Accuracy

## INTRODUCTION

Terabytes of data are being generated day by day. The health care industry generates a massive amount of data every day. Data mining is the way of discovering meaningful patterns, knowledge and information are extract from a datasets it is also known as KDD. Researchers are using data mining algorithm such as Support Vector Machine, Naïve Bayes, Decision tree, Random Forest, Multi Layer Perceptron, Artificial Neural Network, and KNN in the diagnosis of several diseases such as Respiratory Infection Diseases, Liver Disorder, Bone Tumor, Kidney Diseases, Diabetic & Breast Tissue, different types of heart diseases. This Survey mainly focused on Heart Diseases, in 2016 worldwide 56.9 million deaths occurred due to Heart diseases. It is the world's biggest killer according to World Health Organization (WHO). Several techniques have been applied on medical data to improve such diagnosing efficacy, regarding performance parameters such as prediction accuracy, sensitivity and specificity.

## MOTIVATION & JUSTIFICATION

The objective of this study is to provide a literature review on medical diagnosis. This will provide researchers and medical practitioners to establish a better understanding and help them identify a solution to identify the best method for predicting the diseases.

## CLASSIFIFCATION TECHNIQUES

Classification is a supervised learning method it's the initial procedure of data mining supported machine learning functions. It tends to be perform on structured or unstructured data that assigns objects in a collection to target classes or categories according to some constrains. The main goal is to exactly define the target class within the given data set. Many major varieties of classification technique including decision tree, naïve bayes, support vector machine, k-nearest neighbor, neural networks and J48.

## DECISION TREE

It is a diagram representation of possible consequence steps that comprises of nodes connected with lines called edges. Decision is a conclusion reached after single or set of consideration decision tree is a tree based supervised machine learning algorithm that makes a machine to solve the problem each of it node is an attribute of given data sets they are connected by the edges that shows the successful relationship between them.

## NAIVE BAYES

This algorithm is a classification technique base on the famous probability theory called "Bayes Theorem". The strong part is called Naive relationship between the features. Relationship means that how change in one attribute affect or don't affect the other one. It is done by using existing inputs in a relatively fast way. This performance measured by the value of predictive accuracy.

## SUPPORT VECTOR MACHINE (SVM)

This method is used for both classification as well as regression. It categories the given dataset based on their feature. Then it predicts or categories the newly taken sample data. A special character of SVM is, it simultaneously minimize the

empirical classification error and maximize the geometric margin.

**DIFFERENT DATA MINING TECHNIQUES FOR MEDICAL DATA (2009 - 2018)**

| S.N | AUTHOR | YEAR | DATASETS | TECHNIQUES | ACCURACY |
|---|---|---|---|---|---|
| 1. | Sitar-Taut, et al.[1] | 2009 | Different Heart Diseases | Naïve Bayes | 62.03% |
| | | | | Decision Trees | 60.40% |
| 2. | Tu, et al.,[2] | 2009 | Cleveland Heart Diseases | J4.8 Decision Tree | 78.9% |
| | | | | Bagging Algorithm | 81.41% |
| 3. | Das, et al.,[3] | 2009 | Cleveland Heart Diseases | Neural Network Ensembles | 89.01% |
| 4. | Razali et al.,[4] | 2009 | Respiratory Infection Disease | Decision Tree | 94.73% |
| 5. | Minas A Karoliset al., [5] | 2009 | | C4.5 | 82% |
| 6. | Rajkumar, et al.[6] | 2010 | Different Heart Diseases | Naive Bayes | 52.33% |
| | | | | KNN | 45.67% |
| | | | | Decision List | 52% |
| 7. | Srinivas, et al.[7] | 2010 | Different Heart Diseases | Naïve Bayes | 84.14% |
| | | | | Augmented Naïve Bayes Classifier | 80.46% |
| 8. | Kavitha, et al.[8] | 2010 | Different Heart Diseases | Back-Propagation Neural Network | 78.43% |
| | | | | Bayesian Neural Network | 78.43% |
| | | | | Probabilistic Neural Network | 70.59% |
| | | | | Linear SVM | 74.51% |
| | | | | Polynomial SVM | 70.59% |
| | | | | Radial Basis Function Kernel SVM | 60.78% |
| 9. | Anbarasi, et al.[9] | 2010 | Different Heart Diseases | Genetic With Decision Tree | 99.2% |
| | | | | Genetic With Naïve Bayes | 96.5% |
| | | | | Genetic With Classification Via Clustering | 88.3% |
| 10. | K. Srinivas et al.[10] | 2010 | Heart Disease | SVM | 82.5% |
| | | | | C4.5 | 82.5% |
| | | | | Multilayer Perceptron | 89.7% |

| | | | | | |
|---|---|---|---|---|---|
| 11. | Maishowman et al.[11] | 2012 | Benchmark Dataset | SVM with Bagging | 84.1% |
| 12. | AbhishekTaneja et al. [12] | 2013 | Heart disease with selected attributes | J48 Pruned | 95.56 % |
| | | | | J48 Un Pruned | 95.52% |
| | | | | Naive Bayes | 92.42 % |
| | | | | Neural Network | 94.85 % |
| 13. | Syed Umar Amin et al.[13] | 2013 | | NN Backpropagation | 96.2% |
| 14. | ShamsherBahadur et al.[14] | 2013 | Heart Disease | Decision Tree | 99.2% |
| | | | | Naïve Bayes | 96.5% |
| | | | | Classification Clustering | 88.3% |
| 15. | I.S. Jenzi et al.[15] | 2013 | Heart Disease | Naïve Bayes | 80.7% |
| 16. | Hlaudi Daniel et al.[16] | 2014 | Heart Disease | J48 | 99.0741% |
| | | | | REP TREE | 99.222% |
| | | | | Naïve Bayes | 98.148% |
| | | | | CART | 99.0741% |
| 17. | RajendraAcharya et al.[17] | 2015 | Diabetic Subject By Heart Rate Variability Signals | Decision Tree | 92.64% |
| | | | | KNN | 92.02% |
| | | | | Naïve Bayes | 62.58% |
| | | | | SVM | 87.12% |
| 18. | LokanathSarangi et al.[18] | 2015 | | Hybrid GA & NN Technique | 90% |
| 19. | Purusothama et al.[19] | 2015 | Heart Disease | association rule | 55% |
| | | | | K-NN | 58% |
| | | | | Artificial Neural Network | 85% |
| | | | | Naïve Bayes | 86% |
| | | | | Decision Tree | 76% |
| 20. | Marjia et al.[20] | 2016 | Heart disease | KStar | 75% |
| | | | | J48 | 86% |
| | | | | SMO | 89% |
| | | | | Bayes Net | 87% |
| | | | | Multilayer Perceptron | 86% |

| | | | | | |
|---|---|---|---|---|---|
| 21. | KumariDeepika et al.[21] | 2016 | Heart Disease and Diabetic | Naïve Bayes | 93.85% & 65.07 |
| | | | | Decision tree | 92.59% & 87.46 |
| | | | | SVM | 95.2% & 87.32 |
| | | | | Artificial Neural Networks | 94.27% & 76.2 |
| 22. | Acharya et al.[22] | 2016 | Coronary Artery Disease | KNN | 98.17% |
| | | | | Decision Tree | 98.99% |
| 23. | Ashok Kumar Dwivedi [23] | 2016 | Heart Disease | Naïve Bayes | 83% |
| | | | | Classification Tree | 77% |
| | | | | KNN | 80% |
| | | | | Logistic Regression | 85% |
| | | | | SVM | 82% |
| | | | | ANN | 84% |
| 24. | Tapas RanjanBaitharuet al., [24] | 2016 | Liver Disorder | J48 | 68.97% |
| | | | | ZeroR | 57.971% |
| | | | | Multilayer Perceptron | 71.59% |
| | | | | IBK | 62.8986% |
| | | | | Naïve Bayes | 55.3623% |
| | | | | Naïve Bayes | 87.7% |
| 25. | Azam et al.[25] | 2017 | Heart disease | Optimized SVM | 99.2% |
| 26. | Curtis Langlotz et al.[26] | 2017 | Bone Tumor | Naïve Bayesian | 80% |
| 27. | EmranaKabirhashi [27] | 2017 | Different Disease | C4.5 | 90.43% |
| | | | | KNN | 76.96% |
| 28. | HuseyinPolat et al.[28] | 2017 | Kidney Disease | SVM | 98.5% |
| 29. | MeghaShahi et al.[29] | 2017 | Heart disease | SVM | 85% |
| 30. | Syed Muhammad Saqlain Shah et al [30] | 2017 | Heart Disease | SVM | 91.30% |
| 31. | Malarvizhi et al.[31] | 2018 | Diabetic and Breast Tissue | K-means | 85% & 88% |
| | | | | k-means ++ | 90% & 93% |
| | | | | Fuzzy C means | 80% & 88% |
| 32. | Shadman et al.[32] | 2018 | Heart Disease | Naïve Bayes | 86.40% |
| | | | | SVM | 97.53% |
| | | | | Random Forest | 95.76% |
| | | | | Simple Logistic | 95.05% |

TABLE 1. VARIOUS DATA MINING TECHNIQUES USED FOR MEDICAL DATASETS

**K-NEAREST NEIGHBOR (KNN)**

KNN is a data mining process used for classification and regression in machine learning. KNN classifies the given set of data based on their features. It is called as "lazy learner" algorithm because there is no model building. It triggers only when a new sample or query of the data is performed. In KNN there is K- nearest training samples of feature space. It classifies the new testing samples that have nearest or closed features with trained sample.

**NEURAL NETWORKS**

Neural system is a parallel handling system which made with mimicking the natural considering human. It is based on human nerve system. It is a set of nodes that are transformed by the lines known as edges. The framework of neural network that changes its structure dependent on outside or inner data that courses through the system during the learning stage. They can be utilized to show complex connections among information sources and yields or to discover designs in information.

**J48**

It is an extension version of ID3; it is an open source Java execution of the C4.5 method which is very effective and efficient. The highlights of J48 are representing missing qualities, choice trees pruning, nonstop property estimation ranges, induction of rules etc. This method produces the rules from which specific character of that information is created. It depends on divide and conquers approach.

**CONCLUSION**

In this survey paper, we have taken from the year of decade paper to analyze the approaches for diseases diagnosis and prediction using various data mining techniques. It is concluded that although most of the researchers where concentrate on the prediction and diagnosis of heart disease. The performance of different methods varies from one dataset to other datasets they are compared on basis of accuracy metrics. The objective of different research papers is to provide higher accuracy. The hybrid techniques give best accuracy results than other existing techniques by observing various research papers. From this, we come to known that data mining is a milestone in the field of medicine as it plays a major role in effective diagnosis and also for providing a better decision making.

**REFERENCES**

[1] Sitar-Taut, V.A., et al., *Using machine learning algorithms in cardiovascular disease risk evaluation.* Journal of AppliedComputer Science & Mathematics, 2009.

[2] Tu, M.C., D. Shin, and D. Shin, *Effective Diagnosis of Heart Disease through Bagging Approach.* Biomedical Engineering andInformatics, IEEE, 2009.

[3] Das, R., I. Turkoglu, and A. Sengur, Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications, Elsevier, 2009. 36 (2009): p. 7675–7680.

[4] Razali, A.M. and S. Ali, *Generating Treatment Plan in Medicine: A Data Mining Approach.* American Journal of Applied Sciences,2009. 6 (2): 345-351.

[5] Minas A. Karaolis, Joseph A. moutiris, Dementra Hadjipanayi, " Assessment of the risk factors of coronary heart events based on data mining with decision trees", IEEE transactions on information technology in biomedicine,2010.

[6] Rajkumar, A. and G.S. Reena, *Diagnosis Of Heart Disease Using Datamining Algorithm.* Global Journal of Computer Science andTechnology, 2010.Vol. 10 (Issue 10).

[7] Srinivas, K., B.K. Rani, and A. Govrdhan, *Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks.* International Journal on Computer Science and Engineering (IJCSE), 2010. Vol. 02, No. 02: p. 250-255.

[8] Kavitha, K.S., K.V. Ramakrishnan, and M.K. Singh, Modeling and design of evolutionary neural network for heart disease detection. International Journal of Computer Science Issues (IJCSI ), 2010. Vol. 7, Issue 5.

[9] Anbarasi, M., Anupriya, E. And Iyengar.(2010) Enhanced Prediction Of Heart Disease With Feature Subset Selection Using Genetic Algorithm', International Journal Of Engineering Science And Technology, 2(10), Pp.5370- 5376.

[10] K. Srinivas,G. RaghavendraRao ; A. Govardhan, "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques", in IEEE 5th International Conference on Computer Science and Education (ICCSE), 2010.

[11] Mai Shouman, Tim Turner, Rob Stocker, " Using data mining techniques in heart disease diagnosis and treatment", IEEE Japan-Egypt Conference on Electronics, Communications and Computers, 2012.

[12] A. Taneja, "ORIENTAL JOURNAL OF Heart Disease Prediction System Using Data Mining Techniques," 2013.
[13] Syed Umar Amin, KavitaAgarwal, Dr. Rizwan Beg, "Genetic Neural Network based Data mining in prediction of Heart disease using risk factors". Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013).
[14] ShamsherBahadur Patel, Pramod Kumar Yadav and Dr. D.P. Shukla, "Predict the Diagnosis of Heart Disease Patients using classification Mining Techniques", IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS), 2013.
[15] I.S.Jenzi, P.Priyanka, Dr.P.Alli, "A Reliable Classifier Model Using Data Mining Approach for Heart Disease Prediction", International Journal of Advanced Research in Computer Science and Software Engineering, 2013.
[16] Hlaudi Daniel Masethe, Mosima Anna Masethe, "Prediction of Heart Disease using classification Algorithm", Proceedings of the World Congress on Engineering & Computer Science 2014, WCECS.
[17] U. RajendraAcharya, K. S. Vidya, D. N. Ghista, W. J. E. Lim, F. Molinari, and M. Sankaranarayanan, "Computer-aided diagnosis of diabetic subjects by heart rate variability signals using discrete wavelet transform method," Knowledge-Based Syst., vol. 81, pp. 56–64, 2015.
[18] LokanathSarangi, Mihir Narayan Mohanty, SrikantaPattnaik, "An Intelligent Decision Support System for Cardiac Disease Detection", IJCTA, International Press 2015.
[19] G. Purusothama and P. Krishnakumari, "A Survey of Data mining techniques on risk prediction: Heart disea se", Indian Journal of Science and Technology, 2015.
[20] M. Sultana, A. Haider, and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," 2016 3rd Int. Conf. Electr.Eng. Inf. Commun.Technol. iCEEiCT 2016.
[21] K. Deepika, "Predictive Analytics to Prevent and Control Chronic Diseases," pp. 381–386, 2016.
[22] U. R. Acharya et al., "Application of higher-order spectra for the characterization of Coronary artery disease using electrocardiogram signals," Biomed. Signal Process. Control, vol. 31, pp. 31–43, 2016.
[23] Ashok Kumar Dwivedi, "Analysis of computational intelligence techniques for diabetes mellitus prediction," Neural Comput.Appl., vol. 13, no. 3, pp. 1–9, 2017.
[24] T. R. Baitharu and S. K. Pani, "Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset," ProcediaComput.Sci., vol. 85, no.Cms, pp. 862–870, 2016.
[25] A. DavariDolatabadi, S. E. Z. Khadem, and B. M. Asl, "Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM," Comput. Methods Programs Biomed., vol. 138, pp. 117–126, 2017.
[26] B. H. Do, C. Langlotz, and C. F. Beaulieu, "Bone Tumor Diagnosis Using a Naive Bayesian Model of Demographic and Radiographic Features.," J. Digit. Imaging, 2017.
[27] E. K. Hashi, M. S. U. Zaman, and M. R. Hasan, "An expert clinical decision support system to predict disease using classification techniques," 2017 Int. Conf. Electr. Comput.Commun.Eng., pp. 396–400, 2017.
[28] H. Polat, H. DanaeiMehr, and A. Cetin, "Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods," J. Med. Syst., vol. 41, no. 4, 2017.
[29] M. Shahi and R. KaurGurm, "Heart disease prediction system using data mining techniques," Orient. J. Comput. Sci. Technol., vol. 6, no. 4, pp. 457–466, 2017.
[30] S. M. S. Shah, S. Batool, I. Khan, M. U. Ashraf, S. H. Abbas, and S. A. Hussain, "Feature extraction through parallel Probabilistic Principal Component Analysis for heart disease diagnosis," Phys. A Stat. Mech.Its Appl., vol. 482, pp. 796–807, 2017.
[31] Malarvizhi, Dr. S. Ravichandran, "Data Mining's Role in Mining Medical Datasets for Disease Assements – a case Study", International Journal of Pure and Applied Mathematics,2018. Vol.119, No.12.
[32] ShadmanNashif, Md. RakibRaihan, Md. Rasudul Islam, Mohammad Hasan Imam, "Heart Disease Detection by Using Machnine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System", World Journal of Engineering and technology, 2018.