A REVIEW ON DATA MINING TECHNIQUES FOR HEART DISEASE PREDICTION

Taye Oladele Aro [1], Besiru Jibrin Muhammed [2], Olufisayo Babatope Ayoade [1], Idowu Dauda Oladipo [1]

[1] Department of Computer Science, University of Ilorin, Ilorin, Nigeria
[2]Department of Computer Science, Federal University of Kashere, Gombe, Nigeria

Corresponding author: Taye Oladele Aro, taiwo_aro@yahoo.com

ABSTRACT: *Data mining is an important stage in knowledge discovery in database (KDD) of clinical data due to its ability to extract concealed (hidden) patterns from huge amount of datasets to produce more useful and understandable information. One of the most important applications of data mining is the prediction of heart disease. Techniques in data mining can be employed for management, diagnosis and prediction of heart disease in healthcare establishments. This paper discusses review on different approaches in data mining that have been employed by several researchers to predict heart disease.*
KEYWORDS: *Prediction, Data mining, Diagnosis, Healthcare, Knowledge discovery in database.*

1. INTRODUCTION

Heart disease is one of the common diseases among adults, this has recently increased the mortality rate in the world ([SP15, Bin16]). The term disease of heart applies to a number of illnesses that affect circulatory system of the heart ([Cha14a]). Healthcare system is endowed with large quality of data but little or no effort has been applied to this data in solving some critical problems in medical diagnosis of diseases ([AA15]). Among numerous techniques to achieve this task, data mining remains a most significant technique [STS12].
Information given by patients in biomedical diagnosis may include redundant, interrelated symptoms and signs especially when a patient suffers from than one type of disease of the same category. It becomes a serious issue for physicians to diagnose perfectly. Data mining with computational intelligent algorithms can be used to handle prediction in clinical datasets with multiple inputs ([Bin16]).
The prediction of heart disease is an intricate task that requires to be performed accurately and effectively ([MBD13]). Decisions making by doctors are often based on intuition rather than on the knowledge from rich data hidden in the database. This may sometimes result into undesired errors and excessive medical costs which also affect quality of service render to patients during treatment.
The available methodologies and techniques in data mining aid the conversion of huge data into relevant data for intelligent decision making, knowledge discovery and prediction ([RD15]). Data mining methods involve the analyses of data from different perceptives and summarizing it into useful information that can be applied to predict trend analysis ([SS12]). Data mining employs data to systematically or analytically find inadequacies, reduce cost and enhance best practices in healthcare organizations ([KRK16]).
Several techniques in data mining have been used to analyze and deduce unknown relationship that exist among features of clinical data to perform some operations such as prediction, diagnosis, control and treatment of diseases ([KPD13]). The data mining method is an emerging area of great prominence for providing diagnosis and in depth information about medical data ([Cha14b]).
The use of data mining methods for prediction of heart disease in hospitals is seen as positive answers to some perplexing questions, which provide assistance to medical personnel in arriving at intelligent conclusions to improve quality of medical decisions ([LKK13]). Data mining technique may be applied to convey knowledge out of dataset in more appropriate way that can be easily interpreted by people ([AA14]).

2. KNOWLEDGE DISCOVERY IN DATABASE

The significant phase in knowledge discovery in database is data mining, it is a domain in which unique and useful information is obtained from huge amount of dataset ([Bin16]). The KDD procedure represents mathematical approach for determining pattern types in a large database as illustrated in figure 1. This method has been used for the extraction of data associated with numerous diseases from old dataset in order to facilitate easier prognosis of diseases ([AA15]). A number of data mining methods have been applied to medical

prediction of diseases such as heart disease, diabete, cancer and stroke ([Cha14b]).

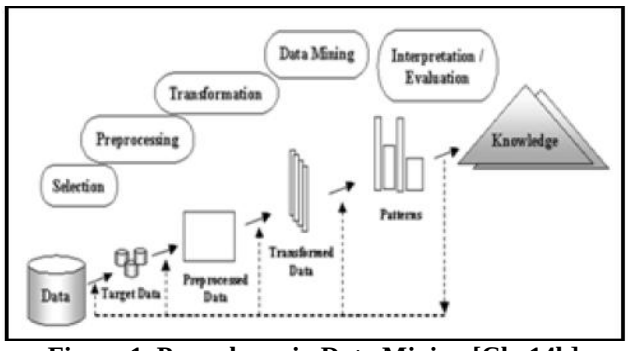

Figure 1. Procedures in Data Mining [Cha14b]

3. HEART DISEASE

Cardiovascular disease (CVD) refers to all the conditions that affect the functioning of the heart, which include coronary heart disease, angina (chest pain) and heart attack ([JCD11]) .

(a) Coronary Heart Disease (Atherosclerotic):- This disease happens when plaque (substances such as cholesterol, calcium and fat) builds up in the wall of arteries ([SPS13]). The condition narrows the arteries, reduces blood flow to heart muscle. If a blood clot is formed, it might eventually stop flow of blood which results into heart attack or stroke.

(b) Angina (Chest pain):- The condition occurs when enough oxygen rich blood is not passing through area of heart muscle ([KK16]) Pain may radiate or move to the arm, neck and back.

(c) Heart Attack:- It happens when blood that flows to the heart muscle is completely blocked. This disorder prevents oxygen rich in blood from reaching heart muscle part. If quick treatment is not taken this can result to serious problems or death.

4. TYPES OF DATA MINING TECHNIQUES

There are many methods in data mining employed for arrangement of data in order to find patterns, these techniques include Association, Classification, Prediction and Clustering ([AA14]).

(a) Classification: - It is an important technique in data mining that identifies relevant features from different classes based on data attribute values ([RD15]). This technique discovers a predictive learning function that categorizes a data into a one of numerous predefined classes ([O+01]). Mathematical methods like linear programming, Decision Trees, Naïve Bayes, Support Vector Machine and Artificial Neural Network are used for classification.

(b) Prediction:- The prediction method fits in the prognostic model level of data mining ([KRK16]). It determines patterns in data that can lead into a reasonable forecast about the future.

(c) Association:- It looks for correlation that exists between different attributes in a dataset ([AA14]). It discovers pattern based on a relationship of a particular item when such items are in transaction ([Bin16]). Asssociation is popularly used technique for prediction of heart disease.

(d) Clustering: - This is unsupervised machine learning technique in which no class labels are given. It locates or analyses pattern present in the vast data chunk ([PV01]).

5. RELATED WORK

Several studies have been proposed for which different techniques in data mining employed by researchers to carryout prediction of heart disease. These studies include:

Sen et al. ([SPS13]) proposed data mining diagnosis for coronary heart disease using Neuro-fuzzy integrated approach two level. The system introduces a layered Neuro-fuzzy based method. Input in terms of patient basis information and medical tests were collected. Critical parameters that are mandatory for occurrence of coronary heart disease were taken at first level and the rest taken at second level. The two level method increases the performance of system in predicting disease chances accurately. UCI heart disease database was used to train Neural Network while Fuzzy rules were applied to predict the chances of coronary heart disease as low, medium or critical.

Automated diagnosis of coronary heart disease using Neuro-fuzzy integrated was conducted by ([AG11]). The system combined computational intelligence fuzzy systems, Neural Network and Evolutionary computing. To show effectiveness of the diagnostic mode, simulation for automated diagnosis system was performed using the realistic causes of heart disease. Experimental results suggested that the hybrid system is good for identification of patients with low or high cardfiac risk.

Dangare and Apte ([DA12]) developed an improved system for prediction of heart disease using classification techniques in data mining. The model used Naïve Bayes, Decision trees and Neural Network for classification purpose. The system analyzed prediction for heart disease with more number of input attributes. Medical terms such as blood, sex, pressure cholesterol like 13 attributes were used to predict the chance of patient having heart disease. Two features which include obesity

and smoking were added. The techniques performance compared based on accuracy.The results gave accuracy of 100% for Neural Network, 99.62% for Decision Trees and 90.74% for Naïve Bayes. The comparative analysis carried out showed that out of three classification models, Neural Network predict heart disease with highest accuracy.

Bindushree ([Bin16]) conducted general review on prediction of cardiovascular heart hazard analysis and evaluation performance using several techniques of data mining. The study reviewed various performances of algorithms, approaches and results for prediction of heart disease by applying data mining. Result methods, evaluation and summary of findings were discussed. The study concluded that data mining techniques can offer a reliable performance for heart disease prediction.

Shouman et al ([STS12]) applied K-Nearest Neighbour (KNN) in diagnosis of heart disease. KNN was used on a benchmark dataset to investigate its efficiency in heart disease diagnosis. The system also considered the investigation of whether integrating voting with KNN would improve its accuracy. Results revealed that applying KNN achieved an accuracy of 97.4% which was higher than any other published works on the benchmark dataset. The experimental results also further showed that the use of voting could not improve the KNN accuracy in the heart disease diagnosis.

Akhil et al.([ADC13]) presented a genetic algorithm and KNN method to improve heart disease classification accuracy. Genetic algorithm was employed to reduce irrelevant, redundant data attributes and select attributes that contributed more towards classification. KNN classifier was trained to categorize heart disease based on dataset as either healthy or sick. The performance of the data mining approach for heart disease prediction was carried out with 6 medical datasets and 1 non-medical dataset. The results obtained reviewed that integrating KNN and genetic algorithm improved the classification accuracy for many datasets.

Chaudhari and Akarte ([Cha14b]) used a Fuzzy system and K-Nearest Neighbour classifier to perform prediction of heart disease. The study integrated voting with KNN to enhance accuracy in the diagnosis phase. The KNN algorithm proved to be more efficient when implemented with fuzzy rules for certain set of disease. Experimental results showed that KNN performed better when quality datasets was used.

Venkatashmi and Shivsankar ([VS14]) came up with heart disease diagnosis using predictive data mining methods. The work focused on the development of heart disease diagnosis and prediction model based on data mining. Several experiments were used to compare the performance of various predictive data mining methdos including Naïve Bayes and Decision Tree Algorithms. A clinical database of 13 attributes from UCI Machine leaving Repository was used as a data source. The experimental results showed that Bayes outperformed when compared to Decision Tree.

Ratnakar et al. ([RRJ13]) proposed a model for the prediction of risk level of heart ailment. The study compared the performance of two modelling methods; Naïve Bayes and Genetic Algorithm. Genetic algorithm was mentioned to reduce set of features or factors that contribute more to heart ailment. A conditional probability method known as Naïve Bayes method was applied on antique heart disease database to produce relationships that exist among the factors. The comparative analysis showed that two modelling data mining methods not accurate for prediction of intensity risk level of heart disease. An intelligent decision support system was developed by combining Apriori, Genetic and Fuzzy methods in order to build optimal prediction system for risk level of heart disease.

Jabbar et al. ([JCD11]) used Association Rule mining based on the sequence number and clustering for heart disease prediction. The database was divided into equal partitions. 14 attributes in the dataset were used and each clustering was considered one at a time for computing frequent item sets. The system reduced memory space usuage. Patterns were obtained from the database with significant weight calculation for prediction purpose. The frequent possessing a value higher than a predefined threshold were selected for valuable prediction of heart attack.

Souza ([Sou15]) developed a predictive system for heart disease using methods of data mining. They employed Neural Network, frequent item set generation and K-means clustering as data mining techniques with Apriori techniques to predict whether a person suffers from heart disease or not. Medical profiles such as age, blood sugar, blood pressure and sex were also considered as parameters to predict the chance of a person getting heart disease. These techniques performance were compared through sensitivity, specificity and accuracy. Result reviewed that Artificial Neural Networks performed better than K-means clustering in all the parameters.

6. SUMMARY / DISCUSSION

The prediction of heart disease is an important process in healthcare system which assists individual to commence earlier treatment before it gets to critical stage. Several studies have been proposed for treatment and prediction of heart disease, but most

commonly proposed algorithms so far engaged data mining classification techniques such as Decision Tree, Artificial Neural Network, Linear Discriminant Analysis, Decision Tree and K-Nearest Neighbour ([AA15]). There is need to identify the performance of each classification method. To handle this work, it is therefore necessary to conduct a comparative performance evaluation on different classification techniques in data mining to reveal the accuracy of each classification techniques. A number of comparative performance analysis of data mining techniques have been done for the prediction of heart disease. The most existing comparative performance evaluation for heart disease prediction with different classification methods only validate using a dataset with just only one data mining software tool ([DA12]). In future work, two or more datasets (Cleveland Heart Disease & Statlog Heart Disease Database) and data mining software tools like Weka, Rapid Miner, Matlab, Orange and Tanagra ([KK16]) should be applied to substantiate and also produce comprehensive comparative performance evaluation for a heart disease predictive system.

In prediction of heart disease, it requires to preprocess and normalize huge data from dataset. An efficient predictive system can be achieved through application of robust preprocessing or feature data reduction techniques on heart disease database such as Random Projection, Principal Component Analysis, Linear Discriminant Analysis. These techniques remove redundant and irrelevant information from the large data prior classification. Therefore, it is important to introduce techniques aforementioned into data mining methods to reduce memory space usuage of data in the database. This will further enable prediction of heart disease to be done in a reasonable time with better accuracy.

7. CONCLUSION

Heart disease remains one of the major leading causes of high mortality in the society. The early prediction is very important in order to reduce the possibility of this disease. Applications of data mining techniques in health care systems have contributed positively to the improvement of predictive system for heart disease diagnosis. This paper has reviewed recent techniques in data mining for heart disease prediction and also suggested future work.

REFERENCES

[AA14] **A. S. Aslam, I. Ashraf** - *Data Mining Algorithms and their applications in Education Data Mining*, Int. J. Adv. Res. Comput. Sci. Mnangement Stud., vol. 2, no. 7, pp. 50–56, 2014.

[AA15] **O. O. Adeyemo, T. O. Adeyeye** - *Comparative Study of ID3 / C4 . 5 Decision tree and Multilayer Perceptron Algorithms for the Prediction of Typhoid Fever*, African J. Comput. ICT, vol. 8, no. 1, pp. 103–112, 2015.

[AG11] **A. Q. Ansari, N. K. Gupta** - *Automated diagnosis of coronary heart disease using neuro-fuzzy integrated system*, World Congr. Inf. Commun. Technol., pp. 1379–1384, 2011.

[ADC13] **M. Akhil, B. L. Deekshatulu, P. Chandra** - *Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm*, Procedia Technol., vol. 10, pp. 85–94, 2013.

[Bin16] **D. C. Bindushree** - *Prediction of Cardiovascular Risk Analysis and Performance Evaluation Using Various Data Mining Techniques: A Review*, Int. J. Enginnering Res., vol. 5013, no. 5, pp. 796–800, 2016.

[Cha14a] **D. Chandna** - *Diagnosis of Heart Disease Using Data Mining Algorithm*, International Comput. Sci. Inf. Technol., vol. 5, no. 2, pp. 1678–1680, 2014.

[Cha14b] **A. A. Chaudhari** - *Fuzzy & Datamining based Disease Prediction Using K-NN Algorithm*, Int. J. Innov. Engineeering Technol., vol. 3, no. 4, pp. 9–14, 2014.

[DA12] **S. S. Dangare, C. S. Apte** - *Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques*, Int. J. Comput. Appl., vol. 47, no. 10, pp. 44–48, 2012.

[JCD11] **M. Jabbar, P. Chandra, B. L. Deekshatulu** - *Cluster based Association rule mining for Heart Attack Prediction*, J. Theor. Appl. Inf. Technol., vol. 32, no. 2, pp. 196–201, 2011.

[KK16] **V. A. Kanimozhi, T. Karthikeyan** - *A Survey on Machine Learning Algorithms in Data Mining for Prediction of Heart Disease*, Int. J. Adv. Res. Comput. Commun. Eng., vol. 5, no. 4, pp. 552–557, 2016.

[KPD13] **M. A. Khaleel, S. K. Pradhan, and G. N. Dash** - *Finding Locally Frequent Diseases Using Modified Apriori Algorithm*, Int. J. Adv. Res. Comput. Coomunication Eng., vol. 2, no. 10, pp. 3792–3797, 2013.

[KRK16] **T. Karthikeyan, B. Ragavan, V. A. Kanimozhi** - *A Study on Data mining Classification Algorithms in Heart Disease Prediction*, Int. J. Adv. Res. Comput. Eng. Technol., vol. 5, no. 4, pp. 1076–1081, 2016.

[LKK13] **K. R. Lakshmi, M. V. Krishna, S. P. Kumar** - *Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability*, Int. J. Sci. Res. Publ., vol. 3, no. 6, pp. 1–10, 2013.

[MBD13] **D. S. Medhekar, M. P. Bote, and S. D. Deshmukh** - *Heart Disease Prediction System using Naive Bayes*, Int. J. Enh. Res. in Sc. Tehn. & Eng., vol. 2, no. 3, pp. 1–5, 2013.

[O+01] **C. Ordonez, E. Omiecinski, L. De Braal, C. A. Santana, N. Ezquerra, J. A. Taboada, D. Cooke, E. Krawczynska, E. V. Garcia** - *Mining constrained association rules to predict heart disease*, Proc. 2001 IEEE Int. Conf. Data Min., pp. 1–7, 2001.

[PV01] **T. Puyalnithi, V. M. Viswanatham** - *Preliminary Cardiac Disease Risk Prediction Based on Medical and Behavioural Data Set Using Supervised Machine Learning Techniques*, Indian J. Sci. Technol., vol. 9, no. 31, pp. 1–5, 2016.

[RD15] **S. Ram, A. Doegar** - *A Comparative Study of Data Mining Techniques for Predicting Disease Using Statlog Heart Disease Database*, Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 5, no. 6, pp. 1202–1210, 2015.

[RRJ13] **S. Ratnakar, K. Rajeswari, R. Jacob** - *Prediction of Heart Disease Using Genetic Algorithm*, Int. J. Adv. Comput. Eng. Neworking, vol. 1, no. 2, pp. 51–55, 2013.

[Sou15] **A. D. Souza** - *Heart Disease Prediction Using Data Mining Techniques*, Int. J. Res. Eng. Sci., vol. 3, no. 3, pp. 74–77, 2015.

[SP15] **S. B. Shinde, A. Priyadarshi** - *Decision Support System on Prediction of Heart Disease Using Data Mining Techniques*, Int. J. Eng. Res. Gen. Sci., vol. 3, no. 2, pp. 1453–1458, 2015.

[SS12] **T. Smitha, V. Sundaram** - *Knowledge Discovery from Real Time Database using Data Mining Technique*, Int. J. Sci. Res. Publ., vol. 2, no. 4, pp. 2–4, 2012.

[SPS13] **A. K. Sen, S. B. Patel, D. P. Shukla** - *A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level*, Int. J. Eng. Comput. Sci., vol. 2, no. 9, pp. 1663–1671, 2013.

[STS12] **M. Shouman, T. Turner, R. Stocker** - *Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients*, Int. J. Inf. Educ. Technol., vol. 2, no. 3, pp. 220–223, 2012.

[VS14] **B. Venkatalakshmi, M. V Shivsankar** - *Heart Disease Diagnosis Using Predictive Data mining*, Int. J. Innov. Res. Sci. Eng. Technol., vol. 3, no. 3, pp. 1873–1877, 2014.