International Conference on Recent Innovations in Science, Agriculture, Engineering and Management
University College of Computer Applications Guru Kashi University, Bathinda, Punjab (India)
20th November 2017, www.conferenceworld.in ISBN: 978-93-86171-80-1

REVIEW ON EARLY PREDICTION OF CARDIOVASCULAR DISEASES USING DATA MINING TECHNIQUES

Ankita R. Bansal[1], Dr. Rajinder Singh[2]

[1] Research Scholar, [2]Assistant Professor

Guru Kashi University, Talwandi Saboo, Bathinda.

ABSTRACT

According to the World Health Organization survey report published on World Health Day year 2016, 31% of all global deaths (17.7 million deaths) are occurring every year, primarily due to heart attacks and strokes which can be easily preventable by adopting healthier life style. It is also found that lack of physical activity, unhealthy dietary habits are a part of current life style which play major role in occurrence of life style disease and its long term morbidity and mortality. In this paper we will discuss about life style diseases, risk of developing coronary heart disease and factors affecting it. With the growing research on heart disease prediction, it has become important to categorize the research outcomes and provides researchers with the overview of the existing heart disease prediction techniques. Authors will discuss different data mining methodologies and technology reviewed in different published research papers in which one or more algorithm of data mining is applied to transform voluminous data to be processed and analyzed which can benefit all parties involved in the health sector such that physicians can identify effective treatments and best practices whereas patients can receive better and more affordable healthcare services delaying the complications and increasing the lead time between cause and effects.

Keywords: *Non communicable Disease, life style disease, cardiovascular diseases, data mining techniques*

I. INTRODUCTION

Life style diseases were defined as a diseases or condition that occur in , affect known an individual over an extensive period of time and for which there are no known causative agent that are transmitted from one affected individual to another. CVD/CHD is a major health concern all around the globe. According to the report published by World Health organization (WHO), non-communicable diseases (NCDs) kill 40 million people each year (70% of all deaths globally). Each year, 15 million people die from a NCD between the ages of 30 -69 years; over 80% of these "premature" deaths occur in low- and middle-income countries. Cardiovascular diseases account for most NCD deaths, or 17.7 million people annually, followed by cancers (8.8 million), respiratory diseases (3.9million), and diabetes (1.6 million). These all four group of diseases account for over 80% of all premature NCD deaths. Tobacco use, physical inactivity, use of alcohol and unhealthy diets all

 ISBN: 978-93-86171-80-1

increase the risk of dying from a NCDs. Detection, screening and treatment of NCDs, as well as palliative care, are key components of the response to NCDs. Mainly two groups of attributes play a cruicial role in the occurrence of life style related health problems are modifiable and non-modifiable risk factors. Modifiable risks factors are the risks factor that can be modify or reduced to some extent which reduce the probability of CVD among individual. Modifiable risk factors are smoking and physical inactivity, alcohol, obesity, hypertension, diabetes, high blood cholesterol and poor diets, stress. While non-modifiable risks are the factors that cannot be altered. It includes age, gender, personality and family history. Both modifiable and non-modifiable risks factors do not cause CVD but have a positive association with acquiring the disease. The presence of multiple risks factors does increase the chance of developing CVD.

Symptoms of a heart attack includes discomfort, pressure, heaviness, or pain in the chest, arm or below the breastbone, discomfort radiating to the back, jaw, throat or arm, fullness, indigestion or choking feeling, sweating, nausea, vomiting or dizziness, extreme weakness, anxiety or shortness of breath, rapid or irregular heartbeats.

Some of the attributes used to predict cardio vascular disease:

1. age: age in years
2. sex: sex (1 = male; 0 = female)
3. cp: chest pain type
    Value 1: typical angina
    Value 2: atypical angina
    Value 3: non-anginal pain
    Value 4: asymptomatic
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholestoral in mg/dl
6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. restecg: resting electrocardiographic results
    Value 0: normal
    Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. thalach: maximum heart rate achieved
9. exang: exercise induced angina (1 = yes; 0 = no)
10. oldpeak = ST depression induced by exercise relative to rest
11. slope: the slope of the peak exercise ST segment
    Value 1: upsloping
    Value 2: flat
    Value 3: downsloping
12. ca: number of major vessels (0-3) colored by flourosopy

 ISBN: 978-93-86171-80-1

13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
14. num: diagnosis of heart disease (angiographic disease status)
    Value 0: < 50% diameter narrowing
    Value 1: > 50% diameter narrowing
15. Smoking, 1 = yes, 0 = No
16. Alcohol, 1 = yes, 0 = No
17. Family History, 1 = yes, 0 = No
18. Past History, 1 = yes, 0 = No

Open source tools used for Data Mining

Weka

Weka is a Java based free and open source software licensed under the GNU GPL and available for use on Linux, Mac OS X and Windows. It comprises a collection of machine learning algorithms for data mining.

RapidMiner

Rapid Miner is available in both FOSS and commercial editions and is a leading predictive analytic platform. Rapid Miner is helping enterprises embed predictive analysis in their business processes with its user friendly, rich library of data science.

Orange

Python users playing around with data sciences might be familiar with Orange. It is a Python library that powers Python scripts with its rich compilation of mining and machine learning algorithms for data pre-processing, classification, modeling, regression, clustering and other miscellaneous functions.

Knime

Knime is one of the leading open source analytic, integration and reporting platforms that comes as free software and as well as a commercial version. Written in Java and built upon Eclipse, its access is through a GUI that provides options to create the data flow and conduct data pre-processing, collection, analysis, modeling and reporting.

DataMelt

DataMelt is a computational platform, offering statistics, numeric and symbolic computations, scientific visualisation, etc. DMelt provides data mining features like linear regression, curve fitting, cluster analysis, neural networks, fuzzy algorithms, analytic calculations and interactive visualisations using 2D/3D plots and histograms.

Apache Mahout

Mahout is primarily a library of machine learning algorithms that can help in clustering, classification and frequent pattern mining. It can be used in a distributed mode that helps easy integration with Hadoop. Mahout is currently being used by some of the giants in the tech industry like Adobe, AOL, Drupal and Twitter.

ELKI

ELKI is open source software written in Java and licensed under AGPLv3. This software focuses especially on

cluster analysis and outlier detection with a compilation of numerous algorithms from both these domains. ELKI's design goals are performance, scalability, completeness, extensibility and a modular design.

MOA

Massive Online Analysis (MOA) is primarily data stream mining software that can handle volumes of real-time data streams at a high speed. MOA is distributed under GNU GPL, and can be used via the command line, GUI or Java API. It is a rich compilation of machine learning algorithms and has proved to be a great choice during the design of real-time applications. Stream mining algorithms typically require faster computations without storing all of the datasets in the memory and have to get the work done within a limited time.

KEEL

KEEL (Knowledge Extraction for Evolutionary Learning) is a Java based open source tool distributed under GPLv3. GUI based interface helps to manage data with different file formats.

Rattle

Rattle, expanded to 'R Analytical Tool To Learn Easily', has been developed using the R statistical programming language. The software can run on Linux, Mac OS and Windows, and features statistics. There is a few other machine learning, NLP and data analytic tools that could aid in mining, likes scikit-learn, NLTK, GraphLab, Neural Designer, Pandas and SPMF, which readers could also explore.

MATLAB

MATLAB (Matrix Laboratory) is a powerful and versatile tool, more than capable of performing data mining, which can be used to examine data, create algorithm, develop models and applications. MATLAB can be used as a standalone tool, rather than in conjunction with other packages.

II. LITERATURE SURVEY

Data mining techniques is used to extract the knowledge and determine interesting and useful patterns. The knowledge gained can be used to improve work efficiency and enhance the quality of decision making process for diagnosis and treatment planning. In medical field, data mining algorithms are use to mine the hidden knowledge in the data set of the medical domain. There are various data mining approaches such as classification, clustering, association rule mining, statistical learning and link mining, logistic regression.

Authors in the year 2014 have used various data mining classification algorithms such as J48, Naïve Bayes, REPTREE, CART and Bayes Naïve for predicting heart disease with 99% accuracy using 12 attributes. The predictive accuracy determined by J48, REPTREE and SIMPLE CART algorithms suggests that parameters used are reliable indicators to predict the likelihood of heart diseases. [1]

Researcher approached by taking a sample of nearly 6000 patients between 30-74 years of age group through prospective study methodology with 12 years of follow up. He found that nearly 10 percent of study population develops the heart problems. He adopted sex-specific prediction equations by using 10 attributes. Out of them few are modifiable and rest non-modifiable. He also tried after adjustment for other factors; blood pressure was a single attribute which contributed 25-30% among male and female CHD events then another attributing factor

 ISBN: 978-93-86171-80-1

is elevated total cholesterol. As per his study result if we are capable to eliminate or reduced attributing factors then likelihood of CHD event will be decrease significantly. [2]

Another researcher added two more attributes i.e. obesity and smoking which also play a crucial role in occurrence of heart d disease. The data mining classification techniques, namely Decision Trees, Naive Bayes, and Neural Networks are analyzed on Heart disease database. The performance of these techniques is compared, based on accuracy. As per results accuracy of Neural Networks, Decision Trees, and Naive Bayes are 100%, 99.62%, and 90.74% respectively. [3]

In the paper published in year 2016 authors has predicted non communicable disease like diabetes, heart disease, urinary system diseases, lung cancer and breast cancer. Authors have predicted heart disease using classification algorithms Naïve Bayes, KNN, J48 decision tree, CART, ANN and SMO giving 85.92%, 100% , 91.85%, 95.92%, 99.25%, 85.55% accuracy respectively.[4]

Authors in 2013 predicted cardio vascular heart diseases taking in to account different number of attributes. Heart disease is a fatal disease by its nature and diagnosis of this disease can cause serious, even life threatening complications such as cardiac arrest and death. The best model selected for predicting heart disease could not exceed a classification accuracy of 95.56% and still much remains to fill the gap of 4.44% misclassified cases. J48 pruned and unpruned, Naïve Bayes and neural network with all attributes and selected attributes are the classification algorithms used. [5]

Authors in 2013 used data mining techniques to predict the diagnosis of heart disease with reduced number of attributes from fourteen to six attributes by using Genetic algorithm. Subsequently three classifiers like Naive Bayes, Classification by Clustering and Decision Tree were used showing 96.5%, 88.3% and 99.2% accuracy respectively. [6]

Published in year 2014, authors used 13 attributes structured clinical database from UCI Machine Learning Repository has been used as a source data. Decision tree and Naive Bayes have been applied and their performance on diagnosis has been compared. Naïve Bayes outperforms when compared to Decision tree. [7]

Author has used three popular data mining algorithms CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision table (DT) extracted from a decision tree or rule-based classifier to develop the prediction models using a large dataset. They also used 10-fold cross validation methods to measure the unbiased estimate showing 83.49%,72.93% and 82.50% accuracy respectively. [8]

This paper explores the utility of various decision tree algorithms in classify and predict the disease. Decision tree has performed well with 99.62% accuracy by using 15 attributes. Moreover, in combination with genetic Algorithm and 6 attributes, Decision tree has shown 99.2% efficiency. [9]

In this work a group of 14 important heart risk attributes is developed from the data set of 76 parameters. The most popular 5 algorithms of J48, Naïve, CART, KNN, and NN are adopted for the prediction. The comparative studies have resulted in finding a new algorithm for higher accuracy. Consequently an algorithm of ArAfPha2016 is introduced. The accuracy of this proposed steps are tested in confusion matrix. The accuracy is found out as 0.9898, which is greater than the values of other methods studied. The above results are discussed

with relation to the heart attack risk assessment. Algorithms such as J48, Naive, CART, KNN, NN had the corresponding accuracy units as 0.9478, 0.7199, 0.8599, 0.8837, and 0.8023 respectively. [10]

Sr. No.	Author	Title	Data Mining Algorithms	Variables	Result Accuracy
1.	Hlaudi Daniel Masethe, Mosima Anna Masethe	Prediction of Heart Disease using Classification Algorithms	• J48, • Naïve Bayes, • REPTREE, • CART, and • Bayes Net	12	99%
2.	Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB.	Prediction of Coronary Heart Disease Using Risk Factor Categories	Framingham CHD prediction equation, NCEP ATP II algorithm	7	99.2%
3	Chaitrali S. Dangare Sulabha S. Apte,	Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques	Decision Trees, Naive Bayes, Neural N/W	13	Neural Networks100%, Decision Trees 99.62%, Naive Bayes 90.74%
4	Dr.B.Srinivasan , K.Pavya	A Study and Analysis of Disease Prediction Techniques in Data Mining for Healthcare	• Naïve bayes • IBK (KNN) • J48 (decision tree) • CART • ANN • SMO	18	Naïve bayes : 85.92 % IBK (KNN) 100% J48 (decision tree) 91.85% CART 95.92% ANN 99.25% SMO 85.55%
5	Abhishek Taneja	Heart Disease Prediction System Using Data Mining Techniques	• J48 unpruned with all attributes • J48 pruned with all attributes	15	• J48 unpruned with all attributes 94.29 • J48 pruned with all attributes 95.41 • J48 unpruned with selected attributes

 ISBN: 978-93-86171-80-1

			• J48 unpruned with selected attributes • J48 pruned with selected attributes • Naive Bayes with all attributes • Naive Bayes with selected attributes • Neural Network with all attributes Neural Network with selected attributes		95.52 • J48 pruned with selected attributes 95.56 % • Naive Bayes with all attributes 91.96 • Naive Bayes with selected attributes 92.42 • Neural N/W with all attributes 93.83 • Neural N/W with selected attributes 94.85
6	Shamsher Bahadur Patel, Pramod Kumar Yadav, yDr. D. P.Shukla	Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques	• Decision tree • Naïve Bayes • Classification Clustering	14	Decision tree 99.2% Naïve Bayes 96.5% Classification Clustering 88.3%
7	B.Venkatalakshmi, M.V Shivsankar	Heart Disease Diagnosis Using Predictive Data mining	Naïve Bayes Decision Tree	13	85.03 84.01
8.	Vikas Chaurasia Saurabh Pal	Early Prediction of Heart Diseases Using Data Mining Techniques	CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision	11	CART -83.49% ID372.93% DT 82.50%

 ISBN: 978-93-86171-80-1

			table (DT)		
9	Nidhi Bhatla, Kiran Jyoti	An Analysis of Heart Disease Prediction using Different Data Mining Techniques	• Naive Bayes • Decision Tree • Neural Network • K-Mean based on MAFIA • K-Mean based on MAFIA with ID3 • K-Mean based on MAFIA with ID3 and C4.5	13	Naive Bayes 94.44 Decision Tree 96.66 Neural Network 99.25 K-Mean based on MAFIA 74% K-Mean based on MAFIA with ID3 85% K-Mean based on MAFIA with ID3 and C4.5 92%
10	Mr.K.Aravintha n Dr.M.Vanitha	A Novel Cluster and Rank Based Method for Prediction of Heart Diseases	• J48 • Naïve • CART • KNN • NN	14	J48 94.78 % Naïve 71.99 % CART 85.99 % KNN 88.37 % NN 80.23 %
11	Illayaraja M. Meyyapan T	Efficient data mining methods to predict the risk of heart diseases through frequent item sets	• Based on symptoms chosen and minimum support value	19	
12	M. Akhil Jabbar	Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm	• KNN + GA	12	100

 ISBN: 978-93-86171-80-1

13	V. Manikantan & S. Latha	Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods	• Maximal Frequent item set Algorithm	13	K-Mean based MAFIA 74% K-Mean base MAFIA with ID3 85% K-Mean based MAFIA with ID3 and C4.5 92%
14	R. Chitra and Dr.V. Seenivasagam	Heart Disease Prediction System Using Supervised Learning Classifier	Cascaded Neural Network Support Vector Machine	13	85% 82%
15	Aditya Methaila, Prince Kansal, Himanshu Arya, Pankaj Kumar	Early Heart Disease Prediction Using Data Mining Techniques	Naive Bayes Decision Tree Classification via clustering	14	Naive Bayes 96.53 % Decision Tree 99.2% Classification via Clustering 88.3%

The authors devised a method to predict the risk level of the patients having heart disease through frequent item sets. Frequent item sets are generated based on the chosen symptoms and minimum support value. The extracted frequent item sets help the medical practitioner to make diagnostic decisions and determine the risk level of patients at an early stage. The proposed method can be applied to any medical dataset to predict the risk factors with risk level of the patients based on chosen factors. An experimental result shows that the developed method identifies the risk level of patients efficiently from frequent item sets. [11]

Researcher used a novel approach which combines KNN and genetic algorithm to improve the classification accuracy of heart. The performance of approach has been tested with 6 medical data sets and 1 non-medical data-set. His research results reveal that by integrating GA with KNN will improve the classification accuracy. [12]

Association rule mining procedures are used to extract item set relations. Item set regularities are used in the rule mining process. The data classification is based on MAFIA algorithms which result in accuracy, the data is evaluated using entropy based cross validations and partition techniques and the results are compared. Using the C4.5 algorithm as the training algorithm, author has showed rank of heart attack with the decision tree. Finally, the heart disease database is clustered using the K-means clustering algorithm, which will remove the data

applicable to heart attack from the database. The results showed that the medicinal prescription and designed prediction system is capable of prophesying the heart attack successfully. [13]

In the classification stage 13 attributes are given as input to the CNN classifier to determine the risk of heart disease. The proposed system will provide an aid for the physicians to diagnosis the disease in a more efficient way. The efficiency of the classifier is tested using the records collected from 270 patients. The results show the CNN classifier can predict the likelihood of patients with heart disease in a more efficient way. [14]

In this paper the focus is on using different algorithms and combinations of several target attributes for effective heart attack prediction using data mining. Decision Tree has outperformed with 99.62% accuracy by using 15 attributes. Also the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction [15]

III. CONCLUSION

The model from the classification will be able to answer more complex queries in the prediction of heart attack diseases. The most popular 5 algorithms of J48, Naïve, CART, KNN, and NN are adopted for the prediction. An algorithm of ArAfPha2016 is introduced, reducing the number of attributes using genetic algorithm, using frequent item set and predicting CVD using different number of attributes giving different accuracy. More positive predictive value of the algorithm, will not only helps in decision making for service provider but also helps the patient to take the appropriate action to reduce likelihood of cardiovascular events and extend the healthy & fruitful life by reducing morbidity and mortality. But always a grey area between prediction and actual occurrence of diseases among individual, community, sub national and national level so more and more refined and more accuracy model to be developed.

REFERENCES

[1.] Hlaudi Daniel Masethe, 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

[2.] Peter W. F. Wilson, Ralph B. D'Agostino, Daniel Levy, Albert M. Belanger, Halit Silbershatz, William B. Kannel, Prediction of Coronary Heart Disease Using Risk Factor Categories. by the American Heart Association. doi: 10.1161/01.CIR.97.18.1837,Circulation. 1998;97:1837-1847.

[3.] Chaitrali S. Dangare Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques, International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012

[4.] Dr.B.Srinivasan [1], K.Pavya [2], A STUDY ON DATA MINING PREDICTION TECHNIQUES IN HEALTHCARE SECTOR ,International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 03 Issue: 03 | Mar-2016 www.irjet.net p-ISSN: 2395-0072

[5.] Abhishek Taneja, Heart Disease Prediction System using Data Mining Techniques, An International Open Free Access, Peer Reviewed Research Journal Published By: Oriental Scientific Publishing Co., India. ISSN: 0974-6471 December 2013, Vol. 6, No. (4): Pgs. 457-466

[6.] Shamsher Bahadur Patel 1, Pramod Kumar Yadav2, Dr. D. P.Shukla3, Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques, IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS) e-ISSN: 2319-2380, p-ISSN: 2319-2372. Volume 4, Issue 2 (Jul. - Aug. 2013), PP 61-64

[7.] B.Venkatalakshmi, M.V Shivsankar, B.Venkatalakshmi, M.V Shivsankar International Journal of Innovative Research in Science, Engineering and Technology Volume 3, Special Issue 3,2014, ISSN (Online) : 2319 - 8753 ISSN (Print) : 2347 – 6710

[8.] Vikas Chaurasia, et al, Early Prediction of Heart Diseases Using Data Mining Techniques, Carib.j.SciTech,2013,Vol.1,208-217

[9.] Nidhi Bhatla, Kiran Jyoti, " An Analysis of Heart Disease Prediction using Different Data Mining Techniques" International Journal of Engineering and Technology Vol.1 issue 8 2012.

[10.] Mr.K.Aravinthan Dr.M.Vanitha, A Novel Cluster and Rank Based Method for Prediction of Heart Diseases, International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682 Vol.5, No.11, November 201

[11.] Ilayaraja M., Meyyapan T, Efficient data mining methods to predict the risk of heart diseases through frequent item sets, 1877-0509 © 2015 The Authors. Published by E,lsevier B.V

[12.] M.Akhil jabbar* B.L Deekshatulua Priti Chandra, Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, 2212-0173 © 2013 The Authors. Published by Elsevier Ltd

[13.] V. Manikantan & S. Latha, Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods, International Journal on Advanced Computer Theory and Engineering (IJACTE ISSN (Print) : 2319 – 2526, Volume-2, Issue-2, 2013

[14.] R. Chitra and Dr.V. Seenivasagam, EARLY HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES pp. 53–59, 2014. © CS & IT-CSCP 2014

[15.] Aditya Methaila, Prince Kansal, Himanshu Arya, Pankaj Kumar, EARLY HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES, pp. 53–59, 2014. © CS & IT-CSCP 2014