Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A Literature Survey on Data Mining Techniques to Predict Lifestyle Diseases

Divya Sharma[1], Anand Sharma[2], Vibhakar Mansotra[3]
[1,3]*Department of Computer Science and IT, University of Jammu, J&K, India,* [2]*UCCA, Guru Kashi University, Talwandi Sabo, Bhatinda, Punjab, India*

Abstract: ***Data Mining is the process of extraction hidden patterns from previously unknown and imaginably useful information from huge amount of data. The diagnosis of Disease is one of the major application where data mining tools are showing successful results. Lifestyle Diseases linked with the way people live their life. Heart Disease and Type II Diabetes are the two complex diseases that has impact on our lifestyle. The diagnosis of these diseases is complex task which requires much experience and knowledge. Type II Diabetes is one of the silent killer disease worldwide where as Heart Disease is the major cause of the death all over the world in the last few years. In 2000, India(31.7 million) topped the world with highest number of people with diabetes and it is also predicted that by 2030 type II diabetes may afflict to 79.4 million individuals in India. About 17.5 million people die each year in India form Heart disease. Many researchers are using different data mining tools to help medical professionals in the diagnosis of lifestyle diseases. Researchers reviewed literature on the prediction and diagnosis of heart disease and Type II diabetes by using data mining techniques and applied on healthcare data of patients. This paper highlights the important role played by data mining tools in analysis of huge volume of healthcare related data in prediction and diagnosis of lifestyle diseases.***
Keywords: ***Lifestyle Disease, Data mining techniques, Data mining Tools, Diseases prediction and diagnosis.***

I. INTRODUCTION

Data mining is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistical methods. In short, Data mining is a process of analyzing data from different perspective and gathering the knowledge from it [1]. Nowadays, data mining is becoming popular in healthcare domain as there is need of efficient analytical methodology for detecting unknown and valuable information in healthcare domain. Data mining is used intensively in the field of medicine to predict and diagnose lifestyle diseases such as heart disease, type II diabetes, stroke and obesity. In this fast moving world people want to live a very luxurious life so they work like a machine in order to earn lot of money and live a comfortable life therefore in this race they forget to take care of themselves, as their food habits change their entire lifestyle change, in this type of lifestyle they are more tensed they have blood pressure, sugar at a very young age and they don't get enough rest for themselves and eat what they get and they even don't bother about the quality of the food ,as a result of all these small negligence it leads to major threat that is the heart disease and sometimes increase in the level of sugar leads to diabetes. So in order to overcome this serious issues data mining techniques are used.

II. DATA MINING TECHNIQUES

Data mining techniques are used to explore, analyze and extract useful medical data using complex algorithms in order to discover the unknown patterns. Researchers have been applying different data mining techniques that are as under:

A. Classification

Classification is a data mining function that assigns items in a collection to target categories or classes. Classifications are discrete and do not imply order. In a general point of view, regression and classification are two types of predictive factors that regression is used for prediction of continuous data and classification is used for prediction of discrete and nominal data.

B. Clustering

Clustering is the process of grouping set of objects in such a way that objects in the same group(called a cluster) are more similar (in some sense or another) to each other to those in other than to those in other groups(clusters).

C. Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a

particular item on other items in the same transaction..

D. Correlation rules

Correlation rules considered as the most important form of discovery and extraction of patterns. This method retrieves all possible patterns of databases. Each algorithm implement on different on different records, and each of them has indicated different functions according to implementation conditions and also data types.

III. DATA MINING CLASSIFICATION TECHNIQUES

The approaches followed by every data mining technique are different. Researchers are using different data mining techniques for the diagnosis of many diseases. Some of the classification techniques are as under:

A. C4.5 algorithm

This algorithm is one of the types of decision tree that was introduced after upgrading the ID3 algorithm. This algorithm can classify the records with noisy and continuous amplitude. When the records are with discrete amplitude, this algorithm operates like ID3 algorithm but when the data amplitude is continuous, it will consider a threshold for all selectable modes and an effective standard is assessed for the threshold and then, the threshold with the highest rate is chosen as the decision index of that node [2 and 3].

B. SVM algorithm

Support Vector Machine(SVM) algorithm is a supervised data mining algorithm in which we plot each data item as a point in n-dimensional space (where n is the no. of features we have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

C. KNN algorithm

K-Nearest Neighbor is an algorithm which is based on similarity with other items. The items which are similar to each other are called neighbors. Once a new item is found, its distance from other items in the model is calculated. This classification partitions the item to the nearest neighbor which is also the most similar one; so places the item in a group that includes the nearest neighbors [4].

D. Neural Network

Artifical Neural Network is inspired from the brain that is considered as a data processing system. In this algorithm, many microprocessors are responsible for data processing and they are acting as an interconnected and parallel network with each other to solve a problem. By using programming science in this network, a data structure is designed that can act as a neuron and this data structure is called *neuron*. By setting a network between the neurons and applying a learning algorithm, the network is trained. In this Neural Network, neurons are divided into two enable (NO or 1) or disable (OFF or 0) modes and each edge (synapses or connections between nodes) has a weight. Edges with positive weight, stimulate or enable the next disable nodes and edges with negative weights, disable or inhibit the next connected nodes (if they are enabled) [14, 15, 16, 17]

E. Naïve Bayes

Naïve Bayes classifier is based on Bayes theorem. This classifier uses conditional independence in which attribute value is independent of the values of other attributes. The Bayes theorem is as follows:

Let X={x1, x2,……,xn} be a set of n attributes.

In Bayesian, X is considered as evidence and H be some hypothesis means, the data of X belongs to specific class C.

We have to determine P (H|X), the probability that the hypothesis H holds given evidence i.e. data sample X. According to Bayes theorem the P (H|X) is expressed as

$$P(H|X) = P(X|H)\, P(H) / P(X)$$

IV. LIFESTYLE DISEASES

Lifestyle disease are the disease that are linked with the way people live their lives. Diseases that impact on our lifestyle are Heart disease, Stroke, Obesity and Type II Diabetes. Two of the lifestyle disease are explained below:

A. Heart Disease

Heart is the most vital part of the human body as the life is dependent on efficient working of heart. If functioning of heart is not proper then it will influence the other body parts. Heart disease is the major cause of causalities in the different countries including

India. Heart disease kills one person in every 34 seconds in the United States. A heart disease is caused due to narrowing or blocking of coronary arteries. This is caused by the deposition of fat on the inner walls of the arteries and also due to build up cholesterol.

There are number of factors which increase the risk of Heart disease..

Family history
Smoking
Cholesterol
Poor Diet
High blood pressure
Obesity
Physical inactivity
Hypertension

1) *Symptoms:* The symptoms include tightness or pain in the chest, back or arms, neck, as well as fatigue, lightheadedness, abnormal heartbeat and anxiety. Women are more likely to have atypical symptoms than men.
a) *Pain area:* area between shoulders blades, arm, chest, jaw, left arm or upper abdomen.
b) *Pain types*: can be crushed, like a clenched first in the chest, radiating from the chest, sudden in the chest, or mild.
c) *Pain circumstances:* may occur during rest
d) *Whole body:* dizziness, fatigue, light-headedness, clammy skin, cold sweat, or sweating.
e) *Gastrointestinal:* heartburn, indigestion, nausea or vomiting.
f) *Chest:* discomfort, fullness or tightness.
g) *Neck:* discomfort or tightness.
h) *Arm:* discomfort or tightness.
2) *Types of Heart diseases:* Heart disease includes all types of disease affecting different components of the heart. Heart means 'cardio'. Therefore, all heart diseases belong to the category of cardiovascular diseases. Some types of Heart diseases are:

Coronary Artery Disease is the most common type of heart disease. In coronary artery disease, the arteries carry blood to the heart muscle which contain cholesterol and fat are lined with plague.

Angina is a pain that occurs when your heart is not getting enough oxygen and nutrients. It is the medical term for chest pain that occurs due to insufficient supply of blood to the heart.

Myocarditis

It is an inflammation of the heart muscle usually caused by viral, fungal, and bacterial infections affecting the heart. It is uncommon disease with few symptoms like join pain, leg swelling or fever that cannot be directly related to the heart.

3) *Heart Failure :* Heart failure happens when the heart isn't pumping enough blood to meet your body's needs.
4) *Arrhythmia:* Sometimes the heart's electrical system does not function normally. It may skip beats or sometimes the heart's electrical signal does not move in the proper sequence. These abnormal rhythms are called arrhythmia.
5) *Cardiomyopathy:* It is the weakening of the heart muscle or a change in the structure of the muscles due to inadequate heart pumping. Hypertension, alcohol consumption, viral infections, and genetic defects are common causes of cardiomyopathy.
6) *Congenital heart disease:* It is also known as congenital heart defect, it refers to the formation of an abnormal heart due to a defect in the structure of the heart or its functioning. It is also a type of congenital disease that children are born with.

B. Diabetes

Diabetes is a disease in which the body could not produce insulin or sometimes could not use the produced insulin properly. Diabetes leads to gathering of glucose particle in the blood instead of going into body cell. The gathering of glucose particle in the blood invites various kinds of instabilities in the body.

The various factors that helps to determine the presence of diabetes are:

Number of times pregnant.
Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
Diastolic blood pressure(mm Hg)
Triceps skin fold thickness(mm)

2-Hour serum insulin(mu U/ml)
Body mass index(weight in kg/(height in m)^2)
Diabetes pedigree function

1) Symptoms : The symptoms of diabetes are:
Polyuria(frequent urination)
Polydipsia(increased thirst)
Polyphagia(increased hunger)
Weight loss

V. LITERATURE REVIEW

Numerous works has been done related to lifestyle disease diagnosis using different data mining techniques. The dataset, algorithms, methods used by the authors and the observed results along with the future work is carried out in finding out efficient methods of medical diagnosis for various lifestyle diseases. Here is a brief discussion about two lifestyle diseases i.e. heart disease and type II diabetes and the work that has been already carried out in past few years.

A. Heart disease diagnosis using classification methods

Hlaudi Daniel Masethe predicts and diagnose heart disease by using different data mining algorithms such as J48, REPTREE, Naïve Bayes, Bayes Net, Simple CART. The author analyze the performance of these algorithms through evaluation criteria such as Kappa Statistics, Mean Absolute Error, Root Mean Squared, Relative Absolute Error and Root Relative Squared Error.
Accuracy of J48, REPTREE, Naïve Bayes, Bayes Net and CART are 99.0741%, 99.0741%, 97.222%, 98.1481% and 99.0741% respectively [5].

B. Heart Attack Prediction System using Clustering

Shantakumar B. Patil applied K-mean clustering algorithm on the pre-processed data. And the recurrent patterns applicable to heart disease are mined with the MAFIA algorithm from the data extraction. The neural network is trained with the selected important patterns for effective prediction of Heart Attack on the basis of computed significant weightage [6].

C. Heart Disease Diagnosis Using Fuzzy Logic Approach

P.K. Anooj has proposed a weighted fuzzy rule based CDSS for the diagnosis of heart disease. It automatically obtains the knowledge from the patient clinical data. The proposed CDSS for risk of heart patients consists of two phases. First is an computerized approach for generation of weighted fuzzy rules and decision tree and the second is creating a fuzzy rule based decision support system [7].

D. Heart Disease Prediction Using Association Rule

V.Manikandan et al. extract the item set relations by using association rule. The data classification was based on MAFIA algorithms which resulted in better accuracy. The data was evaluated using entropy based cross validation and partition techniques and the results were compared. MAFIA (Maximal Frequent Itemset Algorithm) used a dataset with 19 attributes and the goal of the research work was to have highly accurate recall metrics with higher levels of precision [8].

E. Heart Disease Prediction System using Hybrid System

R. Chitra et.al. Present Hybrid Intelligent techniques for the prediction of heart disease. Some Heart disease classification system was reviewed in this study and concluded with justification importance of data mining in heart disease diagnosis and classification. The classification accuracy can be improved by reduction in features [9].

F. Type II Diabetes Prediction Using Hybrid Model

Jayaram et al. develop of a hybrid model for classifying Pima Indian Diabetic Database(PIDD). The model consisted of two stages. In the first stage, the K-means clustering was used to identify and eliminated incorrectly classificated instances. In the second stage a fine tuned classification was done using Decision tree C4.5 by taking the correctly clustered instance of first stage. Experimental results signify that 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 classification accuracy of 93.33% [10].

G. Type II Diabetes Prediction Using Classification

Han et al. used data mining techniques through Rapid Miner for the classification of diabetes data analysis and diabetes prediction model. A Decision tree and ID3 algorithm were used for prediction with 72% and 80% of accuracy respectively [11].

H. Type II Diabetes Prediction Using Rough Sets

Breault applied rough sets on the PIMA for the first time. He first pre-processed the data and discrete it by making intervals of data. He used the equal frequency binning criteria for intervals and then he created reducts by using Johnson reducer algorithm and classified using the batch classifier with the standard/tuned voting method (RSES). The rules were constructed for each of the 10 randomizations of the PIDD training sets from above [12]. The tests sets were classified according to defaults of the naïve Bayes classifier, and the 10 accuracies ranged from 69.6% to 85.5% with a mean of 73.8% and 95% CI of (71.3%, 76.3%)

I. Type II Diabetes Prediction Using Clustering

Vijayalakshmi et al. developed a clustering algorithm that is used for predicting diabetes based on graph b-coloring technique. They implement and perform experiments by comparing their approach with K-NN classification and K-means clustering. The results showed that the clustering based on graph coloring is much better than other clustering approaches in terms of accuracy and purity. The proposed technique presented a real representation of clusters by dominant objects that assures the inter cluster disparity in a partitioning and used to evaluate the quality of clusters [13].

Table i

Different data mining techniques used in the diagnosis of heart disease over different datasets.

Author	Year	Data Mining Tool	Techniques used	Accuracy
Abhishek et al.	2013	WEKA 3.6.4	J48	95.56%
			Naïve Bayes	92.42%
			Neural Network	94.85%
Chaitrail et al.	2012	WEKA 3.6.6	Neural Network	100%
Nidhi et al.	2012	WEKA 3.6.6 TANAGRA .NET	Naïve Bayes	99.52%
			Decision Tree	52.33%
			Neural Network	96.5%
Vikas Chaurasia et al.	2013	WEKA	CART	83.49%
			ID3	72.93%
			Decision Table	82.50%
Hlaudi Daniel Masethe et al.	2014	WEKA	J48	99.074%
			REPTREE	99.74%
			Naïve Bayes	97.22%
			Bayes Net	98.14%
			Simple CART	99.74%
Rashedur et al.	2013	WEKA	Neural Network	79.19%
		TANAGRA	Fuzzy logic	83.85%

Table ii
Different data mining techniques used in the diagnosis of type ii diabetes over different DATASETS

Authors	Years	Data Mining Tools	Techniques Used	Accuracy
Han et al.	2008	Rapid Miner	Decision Tree	72%
			ID3	80%
Karegoneda et al.	2012	WEKA	K-NN	96.68%
Vijayarami et al.	2013	WEKA	C4.5 Decision Tree	91%
Taba Pala	2014	WEKA	SVM	97.21%
			LR	98.60%
			MLP	98.83%
		TANAGRA	SVM	97.98%
			LR	98.65%
			MLP	99.10%
Vinod Sharma	2012	MATLAB	Naïve Bayes	95%

VI. CONCLUSION

The objective of our work is to provide a study of different data mining techniques that can be employed in automated lifestyle diseases prediction systems. Various techniques and data mining classifiers are defined in this work which has emerged in recent years for efficient and effective heart disease and type II diabetes diagnosis. This analysis shows that different technologies are used with different number of attributes. So, different technologies used shown the different accuracy to each other. In some heart disease papers it is shown that neural network given the accuracy of 100 % whereas decision tree and naïve bayes gives 99.0741 % and 99.52% accuracy respectively. And in some diabetes papers it is shown that MLP gives the accuracy of 99.10% , SVM gives 97.98% and Naïve Bayes gives 95% of accuracy. So, different technologies used shown the different accuracy depends upon number of attributes taken and tool used for implementation. The availability of huge of amount and world-wide increasing mortality of lifestyle diseases, researchers are using data mining techniques in the diagnosis of heart disease and type II Diabetes. Although applying data mining techniques to help healthcare professionals in the diagnosis of lifestyle disease is having some success, the use of data mining techniques to identify a suitable treatment for lifestyle disease patients has received less attention.

VII. FUTURE WORK

The paper provides a comparison study of many data mining tools and data mining techniques like Decision tree, SVM, Naïve Bayes, Neural Network over disease prediction system. The experimental results show that many of rules and techniques help in the best prediction of disease. In future these techniques can be implemented on Indian dataset with different parameters.

REFERENCES

[1] Han and Kamber, “Data mining concepts and techniques”, 2nd edition(2010)
[2] Quinlan j r. (1986). Induction of decision trees. machine learning. pp.(4): 81-106.
[3] Quinlan j r. (1994). C4.5: Programs for machine learning. machine learning. pp.(3): 235-240.
[4] Yazdani a, Ebrahimi t, Hoffmann u. (2009), “ Classification of eeg signals using dempster shafer theory and a k-nearest neighbor classifier”,IEEE. in proc of the 4th int embs conf on Neural Engineering, pp. 327-30.
[5] 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
[6] Shantakumar b. patil, “Extraction of significant patterns from heart disease warehouses for heart attack prediction” , international journal of computer science and network security, vol.9, no. 2, february 2009
[7] P. K. Anooj, “Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules”, Journal of computer sciences, vol.24, pp. 27-40,2012
[8] V. manikandan and S. Latha ,“Predicting the analysis of heart disease symptoms using medicinal data mining methods”,International Journal of Advanced Computer theory and Engineering, vol. 2 ,pp.46-51,2013.
[9] R. chitra, “ Review of heart disease prediction system using data mining and hybrid intelligent techniques”, ictact journal on soft computing, july 2013,volume: 03, issue: 04
[10] Jayaram .Kkaregowda, A.G. ., Punya, v., , M.a., Manjunath, a.s., “ Rule based classification for diabetic patients using cascaded k-means and decision tree c4.5.”,International Journal of computer applications. 45–12, (2012)
[11] Han, J., Rodriguze, J.C ., Beheshti, m., “ Diabetes data analysis and prediction model discovery using rapidminer”, Second International Conference on Future

Generation Communication and Networking.96-9 (2008)
[12] Breault, j.l., "Data Mining Diabetic Databases: are rough sets a useful addition?"
[13] Vijayalakshmi, d., Thilagavathi, k., " An approach for prediction of diabetic disease by using b-colouring technique in clustering analysis", in: International Journal of applied mathematical research, 1 (4) pp. 520-530 science publishing corporation www.sciencepubco.com/index.php/ijamr (2012)
[14] Daubechies i.(1990). The wavelet transform, time-frequency localization and signal
analysis. ieee. trans inform theor pp. 36:961–1005.
[15] Demuth h, Beale m, Hagan m. (2009). Neural Network toolbox™ user's guide. the
mathworks, inc, pp. 1-901.
[16] Leng, G., Mcginnity, T.M ., Prasad, g. (2006), " Design for self-organizing fuzzy neural networks based on genetic algorithms",. IEEE. trans. fuzzy syst, vol 14, no. 6, pp. 755–766.
[17] Leung, F.H.F. ., Lam, H.K.., Ling, S.H., et al.(2003)., "Tuning of the structure and parameters of a neural network using an improved genetic algorithm", IEEE. Trans. neural networks , vol 14, no. 1, pp. 79–88.