Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

Methods for the Prediction of Cardio Vascular Diseases in Diabetes patients using Machine Learning Techniques

S.V.Evangelin sonia[1*] and Sindhuja S[2]

[1,2]Department of CSE,Sri Shakthi Institute of Engineering and Technology,Coimbatore, India

ABSTRACT:

Coronary illness is basic in individuals with diabetes. Information from the National Heart Association from 2012 shows 65% of individuals with diabetes will kick the bucket from a type of coronary illness or stroke. As a rule, the danger of coronary illness passing and stroke are twice as high in individuals with diabetes. Early expectation and intercession would hence be of colossal advantage to society. In this paper, we describe effective information learning systems that can be used to predict the cardiovascular diseases in diabetes patients.These strategies are useful to choose the best highlights with the least expenses and briefest occasions and machine learning calculations to accomplish the exactness. These strategies will lessen the outstanding burden and cost for patients just as medicinal services unit.

KEYWORDS— Cardiovascular disease, Diabetes mellitus, Computational Intelligence, Data Mining, Machine Learning algorithms, Deep Learning.

***Corresponding author**

S.V.Evangelin Sonia

Department of Computer Science and Engineering,

Sri Shakthi Institute of Engineering and Technology,

Coimbatore-641005, Tamilnadu, INDIA.

Email:evangelinsonia.vs@gmail.com

Mob:8489827261.

1. INTRODUCTION

The more wellbeing dangers factors an individual has for coronary illness, the higher the odds that they will create coronary illness and even kick the bucket from it. Much the same as any other person, individuals with diabetes have an expanded danger of passing on from coronary illness on the off chance that they have more wellbeing hazard factors. Be that as it may, the likelihood of kicking the bucket from coronary illness is 2 to multiple times higher in an individual with diabetes. In this way, while an individual with one wellbeing hazard factor, for example, hypertension, may have a specific shot of passing on from coronary illness, an individual with diabetes has twofold or even fourfold the danger of kicking the bucket.

It's safe to say there are too many manual processes in medicine. While in training, it is necessary to hand write lab values, diagnoses, and other chart notes on paper. It is always known this was an area in which technology could help improve the workflow and hoped it would also improve patient care. Since then, advancements in electronically medical records have been remarkable, but the information that is provided is not much better than the old paper charts they replaced. If technology is to improve care in the future, then the electronic information provided to doctors needs to be enhanced by the power of analytics and machine learning.Using these types of advanced analytics, one can provide better information to doctors at the point of patient care. Having easy access to the blood pressure and other vital signs when a patient is seen routine and expected. Imagine how much more useful it would be if the patient's risk for stroke, coronary artery disease, and kidney failure are shown based on the last 50 blood pressure readings, lab test results, race, gender, family history, socio-economic status, and latest clinical trial data.

In many countries, Healthcare is something people don't trust on machines and technologies unless it's the only way to treat a disease.Thus ends up visiting any doctor nearby. But with an analytics platform and machine learning running in the background, the human algorithm, the extra layer of a back-up physicianwouldn't be necessary. The analytics engine would have infinitely more data than any one person could ever process. It would have a library of patients, with their diagnosis and tissue type. It would have treatment options available with predictions of how long they would be effective, mortality rates, side effects, and cost. Regardless of all the effort by a human caregiver, an analytics platform could put in infinitely more work behind the scenes and deliver decisive information to the physician in real time.

Modern medicine has evolved tools and techniques which may be used invarious combinations for the assessment of physical health. They include self-assessment of overall health, inquiry into symptoms of ill health and risk factors,inquiry into the use of medical services,

standardized questionnaires forcardiovascular diseases and clinical examination. The term health and familywelfare services cover a wide spectrum of personal and community services fortreatment of disease, prevention of illness and promotion of health.

1.1Machine Learning

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that which makes it more similar to humans: Theability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.

The term Machine Learning was coined by Arthur Samuel in 1959, an American pioneer in the field of computer gaming and artificial intelligence and in 1997, Tom Mitchell gave a "well-posed" mathematical and relational definition that "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

Consider trying to toss a paper to a dustbin.After first attempt, it is realizedthat too much force has been puton it. After second attempt, it is realized thatthe target is closer but still needed to increase the throw angle. What is happening here is basically after every throw, the process of learning something takes place and thus improving the end result. It is programmed to learn from the experience.

Within the field of data analytics, machine learning is used to devise complex models and algorithms that lend themselves to prediction. In commercial use, this is known as predictive analytics. These analytical models allow researchers,data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data set (input).

The highly complex nature of many real-world problems, though, often means that inventing specialized algorithms that will solve them perfectly every time is impractical, if not impossible. These problems are excellent targets for Machine Learning, and in fact machine learning has been applied such problems with great success.

1.2 Heart Disease for Diabetes Patients

According to the National heart, lung and blood institute (2011) heartdisease is a general name for a wide variety of diseases, disorders and conditionsthat affect the heart and sometimes the blood vessels as well. It is the greatest courage afflicting the world. The term heart disease is often used interchangeablywith cardio vascular disease that affects men and women. It accounts for theuniversal high mortality and morbidity.

Symptoms of heart disease vary depending on the specific type of heartdisease. A classic symptom of heart disease is the chest pain. It arises when theblood received by the heart muscles is inadequate. It is the main cause of death inpeople with diabetes (around 50%). People with type 2 diabetes are likely to die 5to 10 years earlier than people without diabetes. Most of these deaths are due tocardiovascular diseases. People with type 2 diabetes are more prone to have a heartattack or stroke, twice as likely as those without diabetes. It has been found that alarge part of the costs attributable to type-2 diabetes is due to the treatment ofcardiovascular diseases. The WHO points out that twelve million deaths occursworldwide due to Heart diseases. Many of the deaths occur in United States andother developed countries based on cardio vascular diseases. Heart disease was themajor causes of different countries include India. In every 34 seconds the heartdisease kills one person. There are different categories in Heart disease but itmainly focuses on three type's namely cardiovascular disease, cardiomyopathy andcoronary heart disease.CHD isa key reason of sickness and death in the modern society and it is caused thedecreased blood and oxygen supply to the heart due to the narrowing of thecoronary arteries. CHD includes myocardial infarctions, commonly called as a heartattacks, and angina pectoris, or chest pain.Two major types of heart and blood vessel disease, also calledcardiovascular disease, are common in people with diabetes are CAD and CVD.

Coronary artery disease, also called Ischemic Heart Disease, is caused by ahardening or thickening of the walls of the blood vessels that go to the heart. Ithasbeen defined as "impairment of heart function due to inadequate blood flow to theheart as compared to its needs, caused by obstructive changes in the coronarycirculation to the heart". Cerebral vascular disease affects blood flow to the brain,leading to strokes. It is caused by narrowing, blocking, or hardening of the bloodvessels that go to the brain or by high blood pressure. Another condition related toheart disease and common in diabetics is PAD. With this condition, the bloodvessels in the legs are narrowed or blocked by fatty deposits, decreasing blood flowto the legs and feet. PAD increases the chances of a heart attack or stroke. Many ofthe heart diseases can be prevented effectively with preventive measures includingregular exercise, abandonment of smoking or drinking and eating a healthy well-balanced diet. Heart disease is the leading cause of death all over the world in thepast ten years. The life expectancy of diabetics is reduced by nearly eight years dueto increased mortality.

In today's world, most deaths occur due to non-communicable diseasesand just over half of these are as a result of CVD. In developed countries, heartdiseases and stroke are the first and second leading causes of death for adult menand women. It is estimated that there were approximately 29.8 million patients withcardiovascular disease in India during the year 2003. An estimated 1.5 millionpeople die of CVD every year. Compared with all other countries, India suffersthe highest loss.

Diabetes mellitus is a long-term disease with variable clinical symptoms andsequences. It describes a metabolic disorder of multiple etiologies characterized bychronic hyperglycemia with disturbances of carbohydrate, fat and proteinmetabolism resulting from defects in insulin secretion, insulin action, or both.

Diabetes is the most common endocrine disease across all population and agegroups and this disease have become the fourth leading cause of death in developedcountries. It is also a chronic disease in which body cannot regulate the amount ofsugar, specifically glucose in the blood and causes serious health complicationsincluding renal (kidney) failure, heart disease, stroke, and blindness. At least 90%of patients with diabetes have type2 diabetes and it is typically recognized inadulthood when the body cannot effectively use the insulin produced.

The risk factors for type 2 diabetes are being 45 years of age or older, being
Over-weight, having a parent or sibling with diabetes (family heredity), having highblood pressure (140/90 or higher), having high cholesterol (High DensityLipoprotein 35 or lower; triglycerides 250 or higher) and acute stress. Over 80percent of people with type-2 diabetes are overweight and they are treated with dietand exercise and the blood sugar level is lowered with suitable drugs.

It can be classified into two main types. Type1 called as Insulin DependentDiabetes Mellitus is usually diagnosed in children and young adults, and was previously known as Juvenile diabetes due to deficient insulin production. In thiscase, patients require lifelong insulin injection for survival. Type 2 diabetes Non-Insulin Dependent Diabetes Mellitus is due to body's ineffective use of insulin andoften occurs in adulthood.

2. LITERATURE REVIEW

Aishwarya et al.[1]have illustrated in their paper, machine learning has been one of the standard and improving techniques with strong methods for classification and reorganization based on recursive learning. It allows training and test classification system, with Artificial Intelligence. Machine learning in recent years has been the evolving, reliable and supporting tools in medical domain and has provided greatest support for predicting disease with correct case of training and testing. Automatic learning has fetched a greater amount of interest in medical domain due to less amount of time for detection and less interaction with patient, saving time for patients care.

Machine learning provides technical basis for data mining which can be considered as a confluence of statistics and machine learning. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. This section provides an overview of data mining applications in healthcare, particularly the heart disease and diabetes.

Techniques for mining stream data are critically reviewed. At present, many data mining methods have been successfully applied to a variety of practical problems in clinical medicine. Data mining techniques are the result of a long process of research and product development and can yieldthe benefits of automation on existing software and hardware platforms to enhance the value of existing information resources. Due to the increase in database sizes, new algorithms have been proposed to deal with the scalability issue. The goal is to extract knowledge from different subsets of a dataset in order to build a global model of the whole dataset. There are various data mining techniques and classification is one of the most common tasks in machine learning and it enables us to categorize records in a large database into predefined set of classes.

According to Bhuvaneshwari et al. [2] machine learning is to build computer systems that can adapt and learn from their experience. It is the domain of research and recently it has developed in medical domain. The domain is automatically learn some task of healthcare information, medical management, patient health management etc., Application of machine learning methods to large databases is called data mining. But machine learning is not just a database problem. It is also a part of artificial intelligence. If the system can learn and adapt to such changes, the system designer need not foresee and provide solutions for all possible situations.

3. METHODS TO PREDICT CARDIO VASCULAR DISEASES

3.1 Logistic regression

LR is a widely used algorithm in epidemiological studies and was used as a reference for comparison with the other algorithms for analyzing data. The purpose of LR is to use the relationship between the dependent and independent variables, as detailed, for the purpose of general regression analysis[8] for future prediction models. The LR dependent variable can be understood as a classification technique because it divides the results into specific categories for the categorical data.

3.2 Linear discriminant analysis

LDA is the most commonly used algorithm in the field of machine learning. It is a method of classifying data by learning the distribution of the data and creating a decision boundary. When classifying the given data into K classes, it is directed to find a straight line where the center (average) of each class is distant and dispersion is small.

3.3 Quadratic discriminant analysis

QDA is a more flexible classification method than LDA, which can only identify linear boundaries, because QDA can also identify secondary boundaries. Both QDA and LDA assume that the observations of each class follow a normal distribution; however, QDA assumes that the covariance matrix of each class is different from LDA. This implies that the Bayesian theorem

assigns an initial estimate to the parameter. QDA assigns an observation to the class that maximizes the quantity of the parameter so that a quadratic function-type discriminant emerges.

3.4K-nearest neighbor

The KNN algorithm is a new method to predict new data with K neighbors from the existing data when new data is provided. This is a method of classification using only the instance, without a model generation process. The hyper-parameters (detailed tuning options for efficient learning of the model) of the KNN algorithm are the number of neighbors (K) to be searched and the distance measurement method. If K is small, it overestimates the regional characteristics of the data (Overfitting) [10], and if it is very large, the model tends to be over-normalized (Underfitting). Also, the KNN algorithm is one whose result is greatly affected by the distance measurement method chosen. In this study, we investigated the optimal K in the KNN analysis of the clinical medical data and verified the model performance according to each distance measurement method. The distance measurement method of KNN was evaluated for each city block, Euclidian, Cosine, Minkowski, Mahalanobis, Hamming, Jaccard, Correlation, Spearman, and Chebyshev models.

3.5 Statistical analysis

In this study, the comparison of 'continuous variables' between the two groups was evaluated by unpaired t-test or Mann-Whitney rank test and expressed as the mean±standard deviation (SD). Comparisons of categorical variables between the two groups were assessed by χ^2 or Fisher's exact test and expressed as a number and a percentage. Each parameter used to predict T2DM underwent a relative risk analysis. The performance evaluation of the learning model generated by machine learning was evaluated by the AUC of ROC analysis. The statistical significance in this study was $p<0.05$.

3.6 Support Vector Machine

SVM is a supervised learning method, a useful technique for data classification. In other terms, it is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. A classification task mainly involves separating the datasets into training and testing sets. It can be defined as a collection of systems which use hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. SVM, when used to build the regression models, is known as Support Vector Regression. 46 The SVM is very popular as a high-performance classifier in several domains in classification and is particularly suited to analyzing large amount of data, for example, thousands of predictor fields. It uses a supervised learning approach for classifying data. That is, SVM produces a model based on a given training data which is then used for predicting the target

values of the test data. Given a labelled training set (x_i,y_i), SVM requires the solution of the following optimization problem to perform classification.

$$min_{w.b.\in} \frac{1}{2} W^T \mathbf{W} + C\sum_{i=1}^{l} \in_i$$

Subject to,

$$y_i(W^T \phi(x_i) + b) \geq 1 - \varepsilon_i$$

Where, $\varepsilon_i \geq 0$, a slack variable to allow for errors in the classification.

x_i – Training vectors,

x_i R n - function mapping xi into a higher dimension space,

C – Penalty parameter of the error term (usually C>0) and

y_i – Class label,y_i {1,-1}

The goal of SVM is to produce a model which predicts target value of data instances in the testing set given only the attributes. SVM became famous when, using pixel maps as input; it demonstrated accuracy comparable to sophisticated artificial neural networks with elaborate features in a handwriting recognition task.

3.7 Naive Bayes method

Asha Rajkumar et al. (2010)[5] has illustrated that, naïve Bayes classifier as a term dealing with a simple probabilistic classifier based on application of Bayes theorem with strong independence assumptions. It assumes that the presence or absence of particular feature of a class is unrelated to the presence or absence of any other feature. It is based on conditional probabilities. It uses Bayes' theorem which finds the probability of an event occurring, given the probability of another event that has already occurred. If B represents the dependent event and A represents the prior event, Bayes' theorem can be stated as follows:

Prob (B given A) = Prob(A and B)/Prob(A)

3.8 Decision Classifier

Sudha et al. (2012)[11] have described the decision tree as a popular classifier and prediction method for handling high dimensional data and it looks like a tree structure. It is one of the successful data mining techniques used in the diagnosis of heart disease. It applies a straightforward idea to solve the classification problem and is a very simple and easy way for handling dataset. It divides the dataset into multiple groups by evaluating individual data record, which can be described by its 49 attributes. It is also simple and easy to visualize the process of classification where in the predicates return discrete values and can be explained by a series of nested if-then-else statements. The results obtained from decision trees are easy to read and interpret. The goal is to create a model that predicts the value of a target variable based on several input variables. Here, each internal node

denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. It is a popular classifier and prediction method for handling high dimensional data. There are many types of decision trees. The difference between them is the mathematical model that is used in selecting the splitting attribute in extracting the decision tree rules. The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets. It breaks down a dataset into smaller subsets while at the same time, an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision tree can easily be transformed into a set of rules by mapping from the root node to the leaf nodes one by one. By creating a decision tree, the data can be mined based on the past history to determine the likelihood a person may be having the risk of heart disease. The decision trees generated by C4.5 algorithm can be used for classification and it is often referred to as a statistical classifier. It builds decision trees from a set of training data using the concept of information entropy.

4. DISCUSSION

The emergency clinics give patients' anonymzed electronic wellbeing records (EHRs) that contain the majority of the data the clinic has about every patient, including socioeconomics, analyze, confirmations, strategies, imperative signs taken at specialist visits, prescriptions recommended, and lab results. We at that point release our calculations to anticipate who may must be hospitalized. This gives the medical clinic chances to intercede, treat the malady all the more forcefully in an outpatient setting, and maintain a strategic distance from an exorbitant hospitalization while improving the patient's condition.

The exactness rates of these expectations outperform what is conceivable with very much acknowledged hazard scoring frameworks, for example, the one that rose up out of the renowned Framingham Heart Study, the continuous long haul cardiovascular companion consider that is currently in its third era of members. Utilizing that framework, a specialist surveys the patient's age, cholesterol, weight, circulatory strain, and a few different components to land at the person's odds of creating cardiovascular malady throughout the following 10 years. Utilizing the Framingham Study 10-year cardiovascular hazard score, one can foresee hospitalizations with a precision of about 56%, which is considerably lower than the 82% rate we accomplished.

The potential advantages from applying AI examination in human services are tremendous. In view of an investigation of a year of clinic affirmations, the U.S. Organization for Healthcare Research and Quality (AHRQ) evaluated that 4.4 million of those affirmations in the United States, totaling $30.8 billion in expenses, could have been avoided. Of that $30.8 billion, $9 billion was for patients with heart sicknesses and $5.8 billion for patients with difficulties from diabetes. That is half of every single pointless hospitalization.

5. CONCLUSION

Analytics and data-driven personalized medicine and health monitoring present risks. These various algorithms can be used to predict the cardiovascular diseases in diabetic patients with at most accuracy.Be that as it may, if Patient is distinguished as diabetes right off the bat there is a need of discovering Control and Un-control state of diabetes. Provided that Patient has diabetes in Un-control condition, might be the patient has serious impact on Patient's Organ like Heart, Eye, Kidney and so forth. So there is need of finding early Severity which might be help tolerant for diminishing the Severity on Organ or Halting the Severe Effect on Organ.

REFERENCES

1. Akhil Jabba. M, Deekshatulu B.L. And Priti Chandra, “Classification of Heart Disease using Artificial Neural Network and Feature Subset selection”, Global Journal of Computer Science and Technology, Neural and Artificial Intelligence, version 1.0 year 2013; 13(3).
2. Ahmad Mohawish, Ragini Rathi and Vibhanshu Abhishek,”Predicting Coronary Heart Disease Risk Using Health Risk Assessment Data”, SSH 2015: The Third International Workshop on Service Science for E-Health, IEEE.
3. Sudhakar K.& Manimekalai Dr.M. : Study of Heart disease prodiction using data mining: International journal of Advanced Research in Computer Science and Software Engineering, Vol 4, Issues 1, ISSN: 2277 i28x, , January 2014; 1157-1160
4. Muhammad Arif Mohammad, Haswadi Hassan, Dewi Nasien & Habibollah Haron,:A Review on Feature Extraction and Feature selection for Handwritten Character Recognition: International Journal of Advanced Computer Science and Applications,2015; 6(2).
5. Chitra1R. And Seenivasagam V. : Review Of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques: Ictact Journal On Soft Computing, July 2013; 03(04)
6. Georgeena. S. Thomas, Siddhesh.S. Budhkar, Siddhesh.K. Cheulkar, Akshay.B.Choudhary, Rohan Singh: Heart Disease Diagnosis System Using Apriori Algorithm: International Journal of Advanced Research in Computer Science and Software Engineering, February 2015; 5(2)
7. Beant Kaur, Williamjeet Singh: Review on Heart Disease Prediction System using Data Mining Techniques: International Journal on Recent and Innovation Trends in Computing and Communication, October 2014; 2(10).
8. Institute for health metrics and evaluation- Article Deaths from cardiovascular disease increase globally while mortality rates decrease.

9. Ron Kohavi and Dan Sommerfield: Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology: KDD-95 Proceedings, 1995.
10. Azhagusundari.B, Antony Selvadoss Thanamani: Feature Selection based on Information Gain: International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, January 2013; 2(2)
11. Asha Gowda Karegowda A. S. Manjunath & M.A.Jayaram: Comparative Study Of Attribute Selection Using Gain Ratio And Correlation Based Feature Selection: International Journal Of Information Technology And Knowledge Management, July-December 2010; 2(2): 271-277.
12. Praveena Priyadarsini.R, Valarmathi M.L And S. Sivakumari: Gain Ratio Based Feature Selection Method For Privacy Preservation: Ictact Journal On Soft Computing, April 2011; 01(04)
13. URL:www.cs.cmu.edu/afs/cs/academic/class/15385-s12/...slides/lec18.ppt
14. URL:www.doc.ic.ac.uk/~dfg/Probabilistic Inference/IDAPILecture15. Pdf
15. Jayaraman M.A: comparative study of attribute selection using gain ratio and correlation based feature selection: Research Gate: Aug 2014.