CARDIAC DISEASES PREDICTION USING SVM WITH XG BOOST ALGORITHM

Mr.E. Loganathan[1], Mr.I.T.Saranraj[2],Mr.G.Vijayakumar[3],Ms.S.Sowndharya[4]

[1]Assistant Professor,Department of Computer Science and Engineering, Perundurai,Erode,Tamilnadu,India.

[2,3,4]Student,Department of Computer Science and Engineering,perundurai, Erode,Tamilnadu,India.

ABSTRACT

At present, a multifaceted clinical diseaseknown as heart failure disease can be affect a greater number of people in the world. In the early stages, to estimate and diagnose thedisease of heart failure, cardiac centers and hospitals are heavily grounded on ECG. The ECG can be considered in a regular tool. Heart disease early discovery is a critical concern in Health Care Services(HCS). Twoclassifiers similar as Support Vector Machine(SVM) with XG Boost with the best performance are selected for the classification in this system. The third one is the heart failure automatic identification system by using an bettered SVM grounded on the duality optimization scheme also anatomized. Eventually, for a ClinicalDecision Support System(CDSS),an effectiveHeart Disease Prediction Model (HDPN).

This is used, which includes viscosity- grounded spatial clustering of operations with noise (DBSCAN) for outlier detection and elimination, a Hybrid Synthetic Minority Over-Sampling Technique-Edited Nearest Neighbor (SMOTE-ENN) for balancing the training data distribution, and XG Boost for heart disease prediction. Machine learningcan be applied in the medical assiduity for disease opinion, discovery, and prediction. The major purpose of this paper is to give clinicians a tool to help them diagnose heart problems beforehand on. As a result, it'll be easier to treat cases effectively and avoid serious impacts. This study uses XG Boost to test indispensable decision tree classification algorithms in the expedients ofperfecting the delicacy of heart disease opinion. In terms of perfection, delicacy, f1- measure, and recall as performance is above system is to be

mentioned, four types of machine learning (ML) models are compared.

Keywords: Machine Learning, Heart Attack Prediction, SVM, Naive Bayes, Random Forest, XG Boost.

1.INTRODUCTION

Life is dependent on the competent functioning of heart, because heart is necessary part of our body. However, it'llaffect the other body corridor of mortal similar as brain, order etc, If function of heart isn't suitable. Heart disease is a disease that goods on the function of heart. There are number of factors which increases threat of heart disease. At the present day, in the world heart disease is the main cause of deaths.The World Health Organization (WHO) has anticipated that 12 milliondeaths do worldwide, every time due to the heart conditions. prediction by using data mining ways gives us accurate result of disease. IHDPS (Intelligent Heart Disease Prediction System) can find out and prize retired knowledge related with heart disease from a literal heart disease database. It can answer complex queries for diagnosing heartdisease and therefore help healthcare judges and interpreters to make intelligent clinical opinions which conventional decision support systems can not. A many kinds ofheart disease are cardiovascular conditions, heart attack, coronary heart disease and Stroke. Stroke is a type of heart disease; it's caused by narrowing, blocking, or hardeningof the blood vessels that go to the brain orby high blood pressure. System grounded onthe threat factors would not only helpmedical professionals but also it would give cases a warning about the probable presence of heart disease indeed before he visits asanitarium or goes for expensive medical checks. Hence this system presents a fashion for prediction of heart disease. These ways involve one successful data mining fashion named XG Boost algorithm.

1.1 Goal of the Project

In moment's ultramodern world cardiac disease is the most murderous one . This disease attacks a person so incontinently thatit hardly gets any time to get treated with. Sodiagnosing cases rightly on timely base isthe most grueling task for the medical fraternity. A wrong opinion by the sanitarium leads to earn a bad name and loosing character. At the same timetreatment of the said disease is relatively high and not affordable by utmost of thecases

particularly in India. The purpose of this paper is to develop a cost effective treatment using data mining technologies for easing data base decision is used to make support system. nearly all the hospitals use some sanitarium operation system to manage healthcare in cases. Unfortunately utmost of the systems infrequently use the huge clinical data where vital information is hidden. As these systems produce huge quantum of data in varied forms but this data is infrequently visited and remain untapped. So, in this direction lots of sweats are needed to make intelligent opinions. The opinion of this disease using different features or symptoms is a complex exertion. In this System using varied data mining technologies an attempt is made to help in the opinion of the disease in question.

2. LITERATURE SURVEY

2.1 Jyoti Soni Ujma Ansari Dipesh Sharma-The successful operation of data mining in largely visible fields likee- business, marketing and retail has led to its operation in other diligence and sectors. Among these sectors just discovering is healthcare. The healthcare terrain is still information rich but knowledge poor There'sa wealth of data available within the healthcare systems. still, there's a lack of effective analysis tools to discover retiredconnections and trends in data. This exploration paper intends to give a check of current ways of knowledge discovery in databases using data mining ways that are in use in moment's medical exploration particularly in heart disease prediction.

2.2 R. Chitra And Dr.V. Seenivasagam- Cardiovascular disease remain the biggest cause of death worldwide and the Heart Disease Prediction at the early stage is significance. In this paper Supervised Learning Algorithm is espoused for heart disease prediction at the early stage usingthe case's medical record is proposed and the results are compared with the known supervised classifier Support Vector Machine(SVM). The information in the case record is classified using a Protruded Neural Network(PNN) classifier. In the bracket stage 13 attributes are given as input to the CNN classifier to determine the threat of heart disease. The proposed system will give an aid for the physician to diagnosis the disease in a more effective way. The effectiveness of the classifier is tested using the records collected from 270 cases. The results show the CNN classifier canprognosticate the liability of cases with heartdisease in a more effective way. It's delicate to diagnose cardiac disease due to thepresence of numerous health problems similar as diabetes, high blood pressure, inordinate cholesterol, and an irregular palpitation rate. multitudinous data analysis and neural network styles have been used to

determine the inflexibility of cardiac diseasein people. The inflexibility of illness is distributed using a variety of ways, including the K- Nearest Neighbor(KNN) algorithm, DT, Genetic Algorithm(GA), andthe Naive Bayes(NB) algorithm. Due to the complexity of cardiac disease, it must betreated with caution. Failure to do so mayhas a mischievous effect on the heart or result in early death. Medical wisdom and statistical perspectives are employed to identify different types of metabolic diseases..

2.3 Suriya Begum- Due to heart disease in India nearly one person dies every day. A fashion should be developed to descry the heart disease to reduce the number of deaths which is handy and at the same time dependable also. In the health care sector, Machine literacy plays an important part in the health care Assiduity. This paper deals with exploring and probing different Machine Learning Algorithms. Also, it deal with applying multiple Algorithms on Heart Disease Dataset. In this study, Six models were trained and tested, which are Logistic Retrogression, Random Forest Classifier, XG Boost Classifier, Support Vector Machine Classifier, Artificial Neural Network Classifier, K Nearest Neighbors Classifier. The Machine Learning algorithm Random Forest Classifier has proven to be the most accurate and dependable algorithm and hence used in the proposed system. We present an algorithm that incorporates hunt constraints to find medically applicable association rules, and validates them with the well- known train and test approach toget rules with high prophetic delicacy. Search constraints include maximum association size, item filtering dependent on prophetic thing (absence or actuality of disease), trait grouping (discard inapplicable combinations), and antecedent/ consequentrule filtering (find prophetic rules). Support, confidence, and lift are the criteria used to estimate the medical significance andtrustability of association rules. trials study the significance of each constraint collectively.

3. EXISTING SYSTEM:

The existing system using naive bayes is that it requires a small quantum of training data to estimate the parameters. Naive bayes is used to cipher posterior chances given compliances. For illustration, a case may be observed to have certain symptoms. Bayes theorem can be used to cipher the probability that a proposed opinion is correct, given that observation. In simple terms, a naive Bayes classifier assumes that the presenc(or absence)

of a particular point of a class is unconnected to the presence (or absence) of any other point. Generally all machine learning algorithms need to be trained for supervised literacy tasks like vaticination. Then training means to train them on particular inputs in such a way that, if latterly on we may test them for unknown inputs (which they've noway seen ahead) for which they may prognosticate grounded on their literacy. According to Naive bayes algorithm first we've to convert the data set into a frequence table. produce a frequence table for all the features against the different classes. Liability table is created by chancing the chances. Naive Bayes Testing Phase will be used to cipher posterior chances. For illustration, a case may be observed to have certain symptoms. Bayes' theorem is used to cipher the probability that a proposed opinion is correct, given that observation. Naive Bayes fashion recognizes the characteristics of cases with heart complaint. It shows the possibility of each 15 input trait for the predictable state. The main thing of this system is to prognosticate heart complaint using data mining fashion similar as Naive Bayesian Algorithm. Raw hospital data set is used and then preprocessed and transformed the data set.Then apply the data mining technique such as Naive Bayes algorithm on the transformed data set. After applying the data mining algorithm, heart complaint is prognosticated and also delicacy is calculated.

4. PROPOSED SYSTEM:

In SVM classification is a supervised learning that can be used to design models describing important data classes, where class attribute is involved in the construction of the classifier. Support Vector Machine (SVM) is a machine learning tool that is based on the idea of large margin data classification Standard implementations, though provide good classification accuracy, are slow and do not scale well.Although Electronic Health Records (EHRs) have attracted increasing research attention in the data mining and machine learning communities. The approach is limited to a binary classification problem (using alive/deceased labels) and consequently it is not informative about the specific disease area in which person is at risk. Unlabeled data classification are commonly handled via Semi-Supervised Learning (SSL) that learns from both labeled and unlabeled data, and Positive and Unlabeled (PU)learning, a special case of SSL that learns from positive and unlabeled data alone.It is a class of ensemble machine learning algorithms which is mostly used to solve classification problems. The gradient is mainly useful for reducing loss function which is nothing but actual different between the

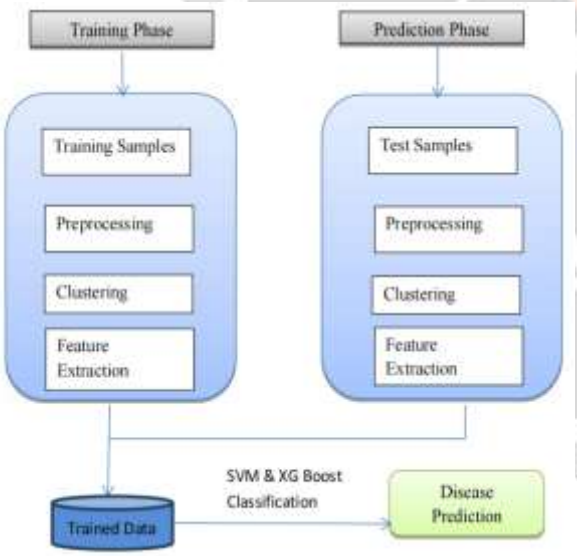

Figure 5.1 Data Flow Diagram

original values and predicted values. Gradient boost is a greedy algorithm thatcan easily over fit the training data set in a quick time which results in improving the performance of an algorithm. Boosting is nothing but a type of ensemble machine learning model in which new models of decision trees are added to correct the errors made by previous existing models. In our research we used the XG Boost machine learning algorithm. XG Boost stands for extreme Gradient Boost. It's an open-source library which provides specific implementation of a gradient boost technique. It's a more regularized form of the gradient boost. It is much better in performance than the Gradient boost algorithms because it uses more advanced regularization (L1 & L2). It's one of the most popular algorithms in the machine learning field because it delivers .

The highest performance in terms of accuracy most of the time then any other algorithm. It's execution speed is really fast.

5. SYSTEM ARCHITECTURE

The main goal of this system is to predict heart disease using data miningtechniques such as Naive Bayesian Algorithm. Raw hospital data set is used and then preprocessed to transform thedata set. Then apply the data mining technique such as Naive Bayes algorithm on the transform data set. After applying the data mining algorithm , heart diseaseis predicted and Then accuracy is calculated.

6. SYSTEM IMPLEMENTATIONMODULES:

- Data Set Acquisition
- Pre-processing
- Clustering
- Feature Selection
- Classification

6.2 MODULE DESCRIPTION:

6.2.1 Data set Acquisition

In this module, upload the datasets. Gatherthe data from hospitals, data centers, and cancer exploration centers. The collected data is pre-processed and stored in the knowledge base on making the model. The opinion traitis used to prognosticate the heart complaint with value 2 for case having heart complaint and 1 for case having no heart complaint. Thepatient ID trait is used as a key and other are input attributes.

6.2.2 Preprocessing

Data pre-processing is an important step of the data mining process. The expression scrap in, scrap out" is particularly applicable to data mining and machine systems. Data- gathering styles are frequently approximately controlled, performing in out- of- range values, insolvable data combinations, missing values, etc. assaying data that has not been
precisely screened for similar problems can produce deceiving results.

6.2.3 Clustering

Clustering is a fashion in data mining to findintriguing patterns in a given data set. The k-means algorithm an evolutionary algorithm that gains its name from its system of operation. The algorithm clusters information into k groups, where k is considered the input parameter. It also assigns each information to clusters grounded upon the observation's propinquity to the mean of the cluster. The cluster's mean is also more reckoned and the process begins again. The k- means algorithmis one of the simplest clustering ways and it's generally used in medical data and related fields. K- Means algorithm is a divisive, unordered system of defining clusters.

Algorithmic steps for k-means clustering

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers. Assign the data point to the cluster center whose distance from the cluster center is minimum of all thecluster centers..

3) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, *'ci'* represents the number of datapoints in i^{th} cluster.

4) Recalculate the distance between each data point and new obtained cluster centers.

5) If no data point was reassigned then stop, otherwise repeat from step 3.

6.2.4 Feature Selection

In this module is used to select the features of the given data set. Attribute selection was performed to determine the subset of features that were largely identified with the class while having low inter correlation.

6.2.5 Classification

Support Vector Machine (SVM) proposed by vapid and cortes have been successfully applied for gender bracket problems by numerous experimenters. An SVM classifier is a direct classifier where the separating hyperactive aero plane is chosen to minimize the anticipated bracket error of the unseen test patterns. SVM is a strong classifier that can identify two classes.SVM classifier the test image to the class which has the maximum distance to the closest point in the training. SVM training algorithm erected a model that prognosticate whether the test image fall into this class or another. SVM bear a huge quantum of training data to select an affective decision boundary and computational cost is veritablyhigh indeed if we circumscribe ourselves to single disguise(anterior) detection. The SVMis a learning algorithm for classification. It tries to find the optimal separating hyper plane such thatthe expected classification error for unseen patterns is minimized. For linearly non- separable data the input is mapped to high- dimensional feature space where they can be separated by a hyper plane. This projection into high dimensional feature space is efficiently performed by using kernels.More precisely, given a set of training samples and the corresponding decision values -1, 1 the SVM aims to find the best separating hyper plane given by the equation Wt x+b that maximizes the distance between the two classes.It is a class of ensemble machine learning algorithms which is mostly used tosolve classification problems. The gradient is mainly useful for reducing loss function which is nothing but actual difference between original values and predicted values. Gradient boost is a greedy algorithm that can easily overfit the training data set in a quick time which results in improving the performance of an algorithm.Exploring important features ofdiabetes through analytical methods of data mining is able to predict and prevent diabetes. This paper proposes a diabetes prediction algorithm based on XGBoost algorithm withthe numerical features being separated whilesome important features are extracted from the text features of experiment data. Experimentresults show that accuracy of diabetesprediction based the improved XGBoost algorithm with features combination is
93.70 %,which is feasible and effective method for diabetes prediction.

7.SCREEN SHOT

RESULT

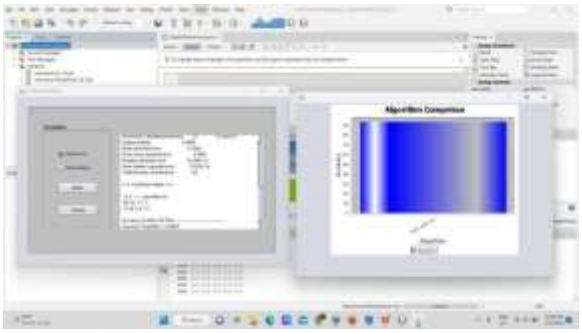

Figure7.1: Accuracy of SVM &XG Boost

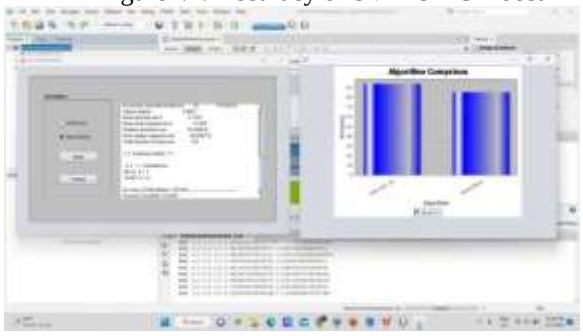

Figure7.2 : comparison of Naive bayes andSVM&XG Boost

Figure7.3 : Result is High Risk

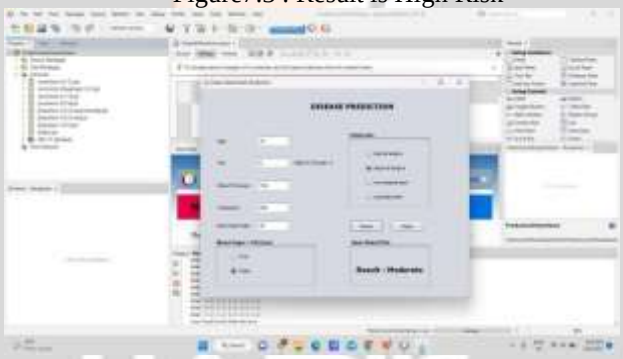

Figure7.4 : Result is Moderate

Figure7.5 : Result is Healthy

8.CONCLUSION

The aim was to design a predictive model for heart disease detection using data mining technique from tranthoracic echocardiograp report dataset that is capable of enhancing the reliability of heart disease diagnosis using echocardiography. The performance of the models was evaluated using the standard metrics of accuracy, precision, recall, and the F- measure. Most of the experiment conducted in this study was implemented with the default parameter of the algorithms, further investigations should be performed with different parameter settings to enhance

and expand the capabilities of the prediction models.

9.REFERENCE

[1] Reddy, K. S., Patel, V., Jha, P., Paul, V. K., Kumar, A. S., Dandona, L., & Lancet India Group for Universal Healthcare.(2011). Towards achievement of universal health care in India by 2020: a call to action. The Lancet, 377(9767), 760-768.

[2] Steen, V. D., & Medsger, T. A.(2007). Changes in causes of death in systemic sclerosis, 1972–2002. Annals ofthe rheumatic diseases, 66(7), 940-944.

[3] Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JG, Coats AJ, et al. 2016 ESC Guidelines for the diagnosis andtreatment of acute and chronic heart failure: the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur Heart J. 2016;37(27):2129–200.

[4] A.Hussain , Aljaaf AJ, Al-Jumeily D, , Dawson T, Fergus P, AlJumaily M. Predicting the likelihood of heart failure with a multi level risk assessment using decision tree. In: 2015 Third International Conference on Technological Advances in Electrical, Electronics and ComputerEngineering (TAEECE); 2015. p. 101– 106.IEEE.

[5] P. Croft, D. G. Altman, and J. J. Deeks, ‘‘The science of clinical practice: Disease diagnosis or patient prognosis? Evidence about ‘what is likely to happen’ should shape clinical practice,’’ BMC Med., vol. 13, no.1, p. 20, 2015.

[6] E. Choi, M. T. Bahadori, A. Schuetz, W.

F. Stewart, and J. Sun, ‘‘Doctor ai: Predicting clinical events via recurrent neural networks,’’ in Proc. Mach. Learn. Healthcare Conf., 2016, pp. 301–318.

[7] Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. InternationalJournal on Computer Science and Engineering (IJCSE), 2(02), 250-255.

[8] Kaur, S., Singla, J., Nkenyereye, L., Jha,S., Prashar, D., Joshi, G. P., & Islam, (2020). Medical diagnostic systems using artificial intelligence (ai) algorithms: Principles and perspectives. IEEE Access, 8, 228049-228069.

[9] Ntiloudi,Giannakoulas,Parcharidou, Panagiotidis,Gatzoulis,Karvounis, H.(2016). Adultcongenital heart disease: a paradigm ofepidemiologicalchange. Internationaljournal of cardiology, 218, 269-274.

[10] Yahaya, L., Oye, N. D., & Garba,

E. J. (2020). A comprehensive review on heart disease prediction using data miningand machine learning techniques. American Journal of Artificial Intelligence, 4(1), 20-29.


IJARIIE