International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 1313-1318 www.ijaem.net ISSN: 2395-5252

Identification of Heart Disease Using Machine Learning Classification

[1]KA Sudharsan, [2]N Giridharan
[1]Student, K S Rangasamy College Of Technology,TiruchengodeTamilnadu.
[2]Assistant Professor, CSE Dept,K S Rangasamy College Of Technology,TiruchengodeTamilnadu.

Submitted: 30-03-2021 Revised: 06-04-2021 Accepted: 09-04-2021

ABSTRACT:The application of disease prediction using machine learning in the medical diagnosis field is increasing successively. This can be contributed primarily to the improvement in the classification contributed primarily to the improvement in the classification and pinpointing systems used in disease identification which is able and recognition systems used in disease diagnosis which is able to provide data that aids medical experts in early identification of fatal diseases and therefore, raising the survival rate of patients importantly. We apply different types of algorithms, each with its own advantage on three separate databases of disease (Heart) available in UCI repository for disease prognosis. The feature selection for each dataset was accomplished by backward modelling using the value test. The results of the study strengthen the concept of the applying of machine learning in early detection of diseases. A support vector machine and artificial neural network, trained with dataset of spectra and algorithms, have been implemented forPODHeart disease prediction using data processing is one among the foremost interesting and challenging tasks. The shortage of specialists and high wrongly diagnosed cases has necessitated the necessity to develop a quick and efficient detection system. According to past system the mixing of clinical decision support with computerbased patient record can reduce medical errors, are often made more precise and hence enhance patient safety. The system helps in prediction of heart disease by considering risky factor associated with heart disease. Here system applies support vector machine algorithm on historical information/data of patient and it provides features like Age, Sex, Smoking, Overweight, Alcohol Intake, Bad Cholesterol, vital sign and Heart Rateto make prediction of CHD with higher accuracy.

KEYWORDS:POD-Prediction of Disease, CHD-Coronary Heart Disease, UCI, Heart Rate

I. INTRODUCTION

Heart is one in all the foremost in depth and organ of physique therefore the care of heart is crucial. Most of diseases square measure associated with heart therefore the prediction concerning heart diseases is important and for this purpose comparative study required during this field, these days most of patient square measure died as a result of their diseases square measure recognized finally stage thanks to lack of accuracy of instrument thus there's ought to realize the additional economical algorithms for diseases prediction. during this paper, we tend to calculate the accuracy of 4 totally different machine learning approaches and on the premise of calculation we tend to conclude that that one is best among them.

PROBLEM DEFINITION

Cardiovascular heart Diseases (CVD) are caused by disorders of the heart and blood vessels and result in coronary heart Disease**,** heart failure,cardiac arrest, ventricular arrhythmias and sudden cardiac death, ischemic stroke, transient ischemic attack, subarachnoid and intra cerebral hemorrhage, rheumatic heart Disease, abdominal aortic aneurysm, peripheral artery heart Disease and congenital Heart Disease. According to World Health Organization (WHO), 17.5 million people died from CVD in 2012 amounting to 31 % of all global deaths.CAD is a type of CVD in which presence of atherosclerotic plaques in coronary arteries, leads to myocardial infarction or sudden cardiac death. In order to diagnose positive sign of heart Disease and to assess the level of damage of heart muscles, certain tests may be prescribed by a medical practitioner including nuclear scan, angiography, echocardiogram, electrocardiogram (ECG), exercise stress testing, ECG is a non-invasive technique used to identify CAD case, though it could lead to undiagnosed symptoms of

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 1313-1318 www.ijaem.net ISSN: 2395-5252

CAD. This limitation leads to angiography which isan invasive diagnosis to confirm CAD cases and is considered as the gold standard for heart Disease detection and severity analysis. However, it is costly and requires high level of technical expertise. Researchers are, therefore, seeking less expensive and effective alternatives, say, using data mining for predicting CAD cases. During the past few decades, image processing, signal processing, statistical and machine learning techniques have been increasingly applied to assist medical diagnosis using ECG and echocardiogram.ECG and echocardiogram are specialized processes conducted by trained practitioners. Sometimes ECG is not able to confirm CAD cases. This process is complex, costly, involves lot of time and effort. To overcome these limitations many researchers used other risk factors excluding angiography to predict CAD cases. These methods are non-invasive, less complex, low cost, reproducible and objective diagnoses, can do automated detection of heart Disease and can be used for screening large number of patients based on clinical data easily obtained at hospitals.

II. PROPOSED SYSTEM

In the proposed work user will search for the heart Disease diagnosis (heart Disease and treatment related information) by giving symptoms as a query in the search engine. These symptoms are pre-processed to make the further process easier to find the symptoms keyword which helps to identify the Heart Disease quickly. The symptoms which keyword is matched with the stored medical input database to identify the multiple Heart Diseases related to that keyword. Multiple heart Diseases is identified, it'll make the pattern matching about the multiple heart Diseases and also find the probability of heart Diseases. Then the heart Disease will make a differential diagnosis to find the heart Disease accuracy.The keyword which is a pre-processed symptom is matched with the heart Diseases stored in the local database to identify the corresponding heart Disease related to those symptoms given by the user. This has to search a record database of more than 20000 heart Diseases and even more symptoms, which is very time consuming, so CFS+PSO classification was applied to classify Heart Diseases features into subgroups. If a gaggle of symptoms match higher preference is given thereto subgroup and searching therein new smaller subgroup thus reduces database access. In pattern recognition, CFS with PSO Feature Selection algorithm is a method for classifying objects based on closest training examples in the feature space. CFS+PSO are a kind of instance-based learning, or lazy learning where the function is merely approximated locally and every one computation is deferred until classification. This feature has been identified because the best suited for this system.

III. EXPERIMENTAL SETUP AND PROCEDURE

K-NEARESTNEIGHBOUR ALGORITHM

KNN for cardiopathy prediction. Compared the approach with different classifiers. Wittiness associate degreed J48 gain precision of 85.09%. Assessment of heart disease events risk factors was performed and planned by karallas. Authors investigated two varieties of risk factors particularly modifiable and no modifiable. We tend to collect 528 samples and data processing analysis was done victimization C4.5. The planned system turn out highest accuracy gain by their classifier was 75% can achieve for PCI and heart bypass surgery graft models. Authors used C4.5 classifier while not feature choice measures. The information turn out correctness gain by this model is a smaller amount compared with different approaches. Diagnosis of cardiopathy victimization regression trees was planned by ruler [12]. Authors have collecting 216 heart sound signals knowledge set and added KNN. This model have plan to analysis Phonocardiogram's (PCG) knowledge. Their methodology obtained associate degree accuracy of ninety nine. Within the year 2015 patient planned a structure to expect the heart disease victimization multilayer perceptron. The strategy will uses thirteen clinical parts as input associate degreed achieved an produce an 88% . The mentioned literature during this connected work has not useful effective feature to choice measures to boost the accuracy. We used low classifiers algorithm to predict the unwellness. During this paper, we've got integrative by PSO with KNN algorithm to get effective results. K Nearest neighbour (KNN) may be a straightforward, lazy and statistic classifier. KNN is most popular by all the options area unit continuous. KNN is additionally referred to as case primarily based reason has been employed in several applications like favourite things for user, applied mathematics estimation.Classification is achieved by distinctive the closest neighbour to work out the category of associate degree unknown sample. KNN is most popular formula different classification algorithms because of its high convergence speed and ease. Show nearest neighbour classification. KNN classification has 2 stages

1) Notice the K^{th} variety of instances within the dataset that can be nearest to instance S

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 1313-1318 www.ijaem.net ISSN: 2395-5252

2) K^{th} variety of instances will vote to work out the category of instance S

The Accuracy by the KNN depends on merit and K method to classify. Alternative ways of measurement the gap between 2 instances area unit trigonometric function, geometrician distance. Assess the obscene sample, KNN compute K nearest neighbours and assigning a category by node form the parameter.The k-nearest neighbor classifier can be viewed as assigning the k nearest neighbors a weight and all others 0 weight. This can be generalized to weight nearest neighbor classifiers. That is, where the ith nearest neighbor is assigned a weight, with. An analogous result on the strong consistency of weighted nearest neighbor classifiers also holds.Let denote the weighted nearest classifier with weights. Subject to regularity conditions on the class distributions the excess risk has the following asymptotic expansion

$$w_{ni}^{*} = \frac{1}{k^{*}}\left[1 + \frac{d}{2} - \frac{d}{2k^{*2/d}}\left\{i^{1+2/d} - (i-1)^{1+2/d}\right\}\right] \text{ for } i = 1, 2, \ldots, k^{*} \text{ and } w_{ni}^{*} = 0 \text{ for } i = k^{*}+1, \ldots, n.$$

k-NN has some strong consistency results. As the amount of data approaches infinity, the two-class k-NN algorithm is guaranteed to yield an error rate no worse than twice the Bayes error rate (the minimum achievable error rate given the distribution of the data). Various improvements to the k-NN speed are possible by using proximity graphs.For multi-class k-NNclassification, Cover and Hart (1967) prove an upper bound error rate of

$$R^{*} \leq R_{kNN} \leq R^{*}\left(2 - \frac{MR^{*}}{M-1}\right)$$

where $K^{th}$is the Bayes error rate (which is the minimal error rate possible), is the k-NN error rate, and M is the number of classes in the problem. For and as the Bayesian error rate approaches zero, this limit reduces to "not more than twice the Bayesian error rate".

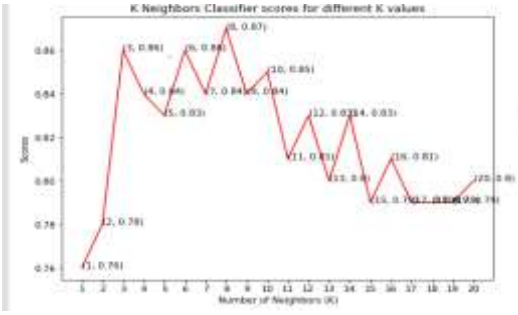

Fig: KNN Classifier scores for different K values

SUPPORT VECTOR MACHINE ALGORITHM

This survey is carried out by different techniques used in the detection of HD. The survey is available which is used to detect heart disease. Alka S. Barhatteproposed ECG signal analysis and classification method using wavelet energy histogram method and support vector machine (SVM). The classification of cardiac heart disease in the ECG signal consists of three stages including ECG signal pre-processing, feature extraction and heartbeats classification.The discrete wavelet transform is used as a pre-processing tool for signal demonizing and feature extraction such as R point location, QRS complex detection. We use Binary SVM as a classifier to classify the input ECG beat into four classes i.e. MITBIHarrhythmia database is used for performance analysis. A. OrozcoDuquegive premature ventricular contraction detection method based on Discrete Wavelet Transform for pre-processing, segmentation and feature extraction.

Discrete Wavelet Transform (DWT) is used to perform baseline wander and power line noise reduction algorithm. Testing by three different feature spaces based on wavelet coefficients. We applied principal Component Analysis (PCA) to reduce dimension into a lower feature space. KNN and SVM future recurrence of the same by alarming the doctor and caretaker on variation in risk factors of stroke disease. Decision making, by the real time health parameters of the patient, helps the doctor easy diagnosis followed by tailored restorative treatment of the disease. The classification algorithms uses for the proposed model diagnosis and prediction. The paper gives an approach about Intelligent Heart Disease Prediction System (IHDPS) used by data mining techniques, i.e. Decision Trees, Naïve Bayes, and Neural Network. Each method possesses to gain suitable results. The relationships and hidden patterns among them have been used to construct this system. The relationships and hidden pattern among them have been used to construct this system. The IHDPS is user friendly, web based, scalable, and reliable and expandable. Secondo Hadiyosoproposed a mini wearable graphical record device and real time cardiopathy detection supported humanoid mobile application. Graphical record signals will be captured by mistreatment the ECG's analogy forepart and sent to humanoid mobile through a Bluetooth module device.

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(\vec{w}\cdot\vec{x}_i - b)\right)\right] + \lambda\|\vec{w}\|^2,$$

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 1313-1318 www.ijaem.net ISSN: 2395-5252

The difference between the hinge loss and these other loss functions is best stated in terms of target functions - the function that minimizes expected risk for a given pair of random variables X and Y.

$$y_x = \begin{cases} 1 & \text{with probability } p_x \\ -1 & \text{with probability } 1 - p_x \end{cases}$$

The optimal classifier is therefore:

$$f^*(x) = \begin{cases} 1 & \text{if } p_x \geq 1/2 \\ -1 & \text{otherwise} \end{cases}$$

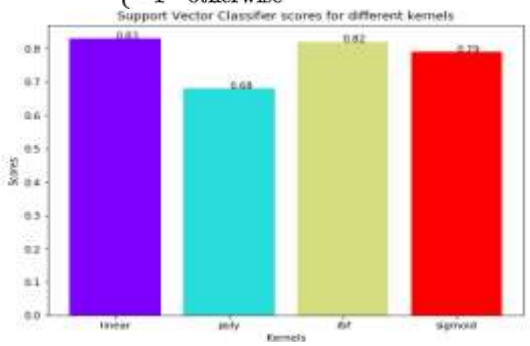

Fig: SVM scores for different kernels

DECISION TREE ALGORITHM

In 2005, SanhezOrdwenez, survey set of operational data that records of the global populace affected by heart disease and chest pain with quality by the risk area, estimation from heart provide the blood pumping through the glands and restricted passage.They are 3 interactive to conversant in decrease the number of problems, case study the following process:

1) The attributes have to be compile to show the fact and the pain by the body. 2) The set of rule and quality are enter into the uninteresting gatherings.

3) The quality rules that can control the ECG records the populace of heart pain are long problem facing globally.

The period consistence with the people they are suffering with chest pain. Data mining process might follow within the expectation of the surviving of suffering within the happiest of the session after. In 2006, Franck autoimmune disorder pudding call for planning an economical tree for capital punish procedure to the diagnosis. Data processing technique can work within the classifier of the living of the environment of people, the connection of ancient classifier and data processing system that add the variables.They provide the important information and volatile for building a root node. In 2007, Kiyondtg Noh et al, used classification of methodology for extract to form a patterns from information. In 2007, BoleeslawSzymeanski "Using economical star Kernel for heart condition Diagnosis "Economically the heart disease can be perform by the process using ECG test is used to test the heart performance. We tend to apply a mark state capital to agree a dataset to expose the patient was suffering from heart disease or chest pain. The activity that are supported by the area to` check the human heart performance. SVM produce the correct result for the attributes to the heart. In year 2009, SellappanPalaniappan, perform to figure, "Intelligent heart condition Prediction System exploitation data processing Techniques". He created Intelligent heart condition Prediction System (IHDPS) utilizing data processing strategies, i.e. call Trees, Naïve mathematician and Neural Network. Each strategy has its own explicit energy to extend acceptable outcomes. The examples and connections among them are utilized to make this framework. The IHDPS is simple to grasp, web based, scalable, versatile, and expandable. In year 2012, Chaitrali S. Dangare, perform to figure, "Improved Study of heart condition Prediction System exploitation data processing Classification Techniques". The planning system, has observed assumption planning frame for heart that can produce the data that can consist of precision and recall for each and every algorithm.

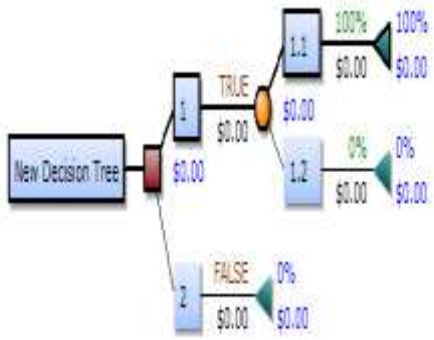

Fig: Decision Tree

Drawn from left to right, a decision tree has only burst nodes (splitting paths) but no sink nodes (converging paths). Therefore, used manually, they can grow very big and are then often hard to draw fully by hand. Traditionally, decision trees have been created manually as the aside example shows although increasingly, specialized software is employed.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 3 Mar. 2021, pp: 1313-1318 www.ijaem.net ISSN: 2395-5252

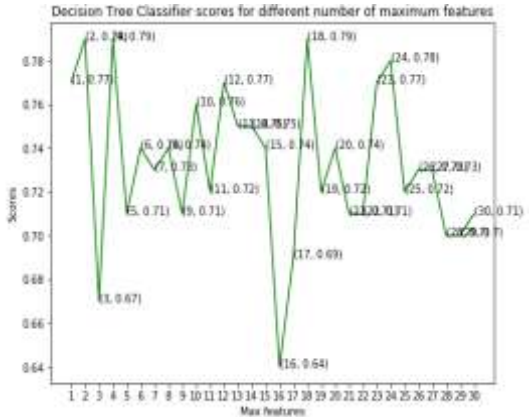

Fig: Scores for different number of maximum features.

RANDOM FOREST ALGORITHM

A singular task of any analysis system is that the method of making an attempt to see and/or determine a doable illness or mess and therefore the call have achieve by this method. For prediction, machine learning algorithm classifiers are wide used. For these machine learning techniques to be helpful in kind of medical diagnostic issues, they'll be characterised by high consummation, the power to traumatize missing information and therefore the filmy of analysis data, ability to clarify selections. As public are generating a lot of information everyday thus there's a requirement for such a classifier which may classify those freshly generated information accurately and expeditiously. This technique mainly focuses on the supervised learning technique said because the Random forest of data analysis by high-powered the values of assorted in Random Forests to induce correct classification results. Within the planned system, random forests classification rule, that meets the same characteristics, is self-addressed.This can be achieved by deciding mechanically by standardisation argument of the rule that is that the variety of base analysis ensemble and affects its performance. The planned methodology has some benefits over the same ways since it doesn't embrace any standardisation parameter, which may be associated with the amount of base classifiers, like the choice ways, Associate in Nursing it doesn't contain an over production section, like the post choice methods; so, it doesn't construct base classifiers beforehand which will not be required. The planned system determines the members dynamically taking under consideration the mix performance of the classifiers, in distinction to the ranking ways. The planned system provides optimum accuracy and correlation. We've got planned system criterion each the options that Associate in nursing ensemble classifier ought to fulfil: high precision and low recall. A lot of specifically. The development of the RF is initiating by adding a tree. Each new node is additional anytime, the new precision and therefore the new recall of the RF are computed, and a web proper procedure is added on the every curves are expressing the several of accuracy and precision, severally. The planned system process is terminated by the RF algorithm

1) Recall between the arc of the accuracy and precision is fitted for the curve.
2) Precision curve of the low process correlation and therefore the process of providing one meet a selected criterion.

Characteristics allow the planned methodology to be totally integrated into any analysis or therapy system since it improves RF rule, providing a classifier rule of the class will offer high consumption, time to process effective working severally by the medical drawback and therefore the nature of knowledge, it will handle screaming or missing information, a standard attributes of medical datasets, it doesn't need any human interruption the sole standardisation argument of the rule is set automatically.

$$\hat{y} = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} W_j(z_i, x')\, y_i = \sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m} W_j(z_i, x')\right) y_i.$$

This shows that the whole forest is again a weighted neighbourhood scheme, with weights that average those of the individual trees. The neighbours of x' in this interpretation are the points sharing the same leaf in any tree . In this way, the neighbourhood of x' depends in a complex way on the structure of the trees, and thus on the structure of the training set. Lin and Jeon show that the shape of the neighbourhood used by a random forest adapts to the local importance of each feature.

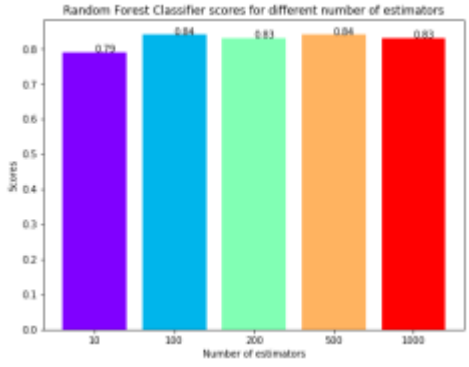

Fig: RF classifier scores for different number of estimator.

IV. RESULT AND DISCUSSIONS

The different statistical operations like removing attributes missing values, standard scalar (ss), min-max scalar, means, standard division are applied to the dataset. The results of these operations are reported in table 5. The processed dataset has 297 instances and 13 inputs attribute with one output label. Data visualization is that the presentation of knowledge in graphical format. It helps people understand the importance of knowledge by summarizing and presenting huge amount of knowledge during a simple and straightforward to-understand format and helps communicate information Clearly and Effectively.Data set represents the frequency of occurrence of specific phenomena which lie within a specific range of values and arranged in consecutive and glued intervals and figure 3 describes the co-relation among the features of the dataset using heat map. The heat map, which is a two-dimensional representation of data in which colors represent values. A single heat map provides a fast visual summary of data. More elaborate heat maps allow the viewer to know complex datasets. Furthermore, heat map can be super useful when we want to see which intersections of the categorical values have higher concentration of the data compared to the others.

V. CONCLUSION

In our human body, heart is an important organ. If the blood motion to the body is inadequate, the organs of the body that are brain and heart stop working and death occurs in few minutes. The most leading cause of death globally from past 15 years is known as heart disease. So, it is an important concept to predict Heart disease at an early stage to avoid human death. The importance of data mining in medical domain is realized and steps are taken to apply relevant techniques in the Disease Prediction.The parameter on which heart disease domain of some and steps are taken to apply relevant method in the Disease Prediction. Heart disease is dependent is extremely susceptible and variant. After achieving historical information about the patient, heart disease can be predicted. Here, the proposed method predicts the heart disease based on the historical clinical data of patient using RF algorithm.Furthermore, we know that irrelevant features also degrade the performance of the diagnosis system and increased computation time. Thus, another innovative touch of our study to used features selection algorithms to selects the appropriate features that improve the classification accuracy as well as reduce the processing time of the diagnosis system. In the future, we will use other features selection algorithms, optimization methods to further increase the performance of a predictive systemfor HD diagnosis.

REFERENCES

[1]. S. I. Ansarullah and P. Kumar, ‘‘A systematic literature review on cardiovascular disorder identification using knowledge mining and machine learning method,’’ Int. J. Recent Technol. Eng., vol. 7, no. 6S, pp. 1009–1015, 2019Ashhab, M-S; and Stefanopoulou, A., 2000, “Control of a Camless Intake Process – Part II,” ASME Journal of Dynamic Systems, Measurement, and Control – March 2000.

[2]. A. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad, G. A. Khan, and A. Ali, ‘‘Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data,’’ Sensors, vol. 20, no. 9, p. 2649, May 2020.Schechter, M.; and Levin, M., 1998, “Camless Engine,” SAE Paper No. 960581.

[3]. A. U. Haq, J. Li, M. H. Memon, M. H. Memon, J. Khan, and S. M. Marium, ‘‘Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection,’’ in Proc. IEEE 5th Int. Conf. Converg. Technol. (ICT), Mar. 2019, pp. 1–4.

[4]. U. Haq, J. Li, M. H. Memon, J. Khan, and S. U. Din, ‘‘A novel integrated diagnosis method for breast cancer detection,’’ J. Intell. Fuzzy Syst., vol. 38, no. 2, pp. 2383–2398, 2020.

[5]. S. Mohan, C. Thirumalai, and G. Srivastava, ‘‘Effective heart disease prediction using hybrid machine learning techniques,’’ IEEE Access, vol. 7, pp. 81542–81554, 2019.

International Journal of Advances in Engineering and Management
ISSN: 2395-5252

IJAEM

Volume: 03 Issue: 03 DOI: 10.35629/5252

www.ijaem.net
Email id: ijaem.paper@gmail.com