JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Low Risk and High Accuracy Heart Disease Prediction using Hybrid Naïve Bayes Machine Learning and PCA

Anesh Kumar, Dr. Vikas Gupta

M. Tech. Scholar, Head of Dept.

Department of Electronics and Communication Engineering

Technocrats Institute of Technology, Bhopal

Abstract: — This paper gives an endeavor to productively arrange and foresee heart illnesses at a beginning phase with high exactness and execution measures. The huge commitment of this exposition is isolated into two sections. Initial, a powerful way to deal with prior location and grouping of coronary illness is portrayed. Next, a fourier change based clinical proposal model is introduced for the previous conclusion of heart diesease. Supervised machine learning classifiers can be categorized into multiple types. These types include naïve Bayes, linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), generalized linear models, stochastic gradient descent, support vector machine (SVM), linear support vector classifier (Linear SVC) decision trees, neural network models, nearest neighbours and ensemble methods. The ensemble methods combine weak learners to create strong learners. In this paper the implemented result with the help of hybid Naïve Bayes Machine Learning and principal component analysis (PCA) algorithms.

Index Terms **–** Support Vector Machine, Neural Network, Classification, Heart Disease, Naïve Bayes, PCA

I. INTRODUCTION

Heart illness is a significant worldwide medical issue in current medication. The twenty-first-century is aphorism perfect expansion in future and a critical transaction in the reasons for coronary illness mourning all through the world. Today it is deciphered for roughly 30% diminishing over the globe remembering around 40% for the big league salary nation and 28 percent in low and center pay nations. Constrained by financial turn of events, suburbanization and related with circadian life changes this consistent progress is emerging far and wide among all races, ethnic gatherings, and countries at a significantly quicker rate than the only remaining century [1]. An ongoing improvement of current way of life dramatically expands the cardiovascular breakdown rates.

Ongoing examination indicated that the proof of cardiovascular breakdown is significantly increased in the last a quarter century. Ongoing examination expresses that Chronic noninfectious sickness like cardiovascular infection is one of the unmistakable reasons of downfall around the globe. Worldwide ascent in heart sickness impacts from an emotional assignment in the wellbeing status of people far and wide [2].

The heart infection turned into unquestionably the regular schedule of death around the world. The worldwide ascent in heart infection impacts from an emotional concession in the wellbeing status of people far and wide. Heart sicknesses are inauspiciously expanding step by step in the course of recent many years, and it has gotten one of the chief purposes behind mourning in the vast majority of the nations over the globe. Late cardiovascular wellbeing focused overview persuaded that practically 1.2 billion individuals die each year as a result of heart illnesses. There is no single answer for the rising weight of coronary illness, given the monstrous changes in cultural, ethnic, and financial environs. Generally cardiovascular breakdown anticipation is exceptionally an intriguing errand in the night before significant expense proportions [3].

The enormous and complex nature of the clinical consideration information put away across electronic-wellbeing information bases catches the eye of analysts towards clinical applications. This amend parts of clinical administrations with cutting edge electronic-wellbeing methods. Clinical applications, when all is said in done, contain six huge exercises, for example, screening, analysis, treatment, visualization, observing, and the executives [4]. This paper centers principally around the order cycle with wellbeing hazard expectation. The basic target of the section is to expand the forecast system characterized in past parts this proposition points to present scanty standard part investigation strategy for include decrease, and for order, the fluffy min-max neural organization (FMMN) with double cuckoo search is introduced for improvement. The hybridization strategies improve forecast exactness with information pre-preparing and highlight decrease procedures [5,6].

This part accepts the current information mining philosophies to expand upon, and the proposed calculation aids cardiovascular danger expectation and suggestion measures. The coupling of quick fourier changes with AI models is accepted to chip away at the premise of half breed order draws near. It is accepted that the utilization of quick fourier change aids time arrangement

investigation of the patient's information and the troupe model backings compelling forecast and clinical suggestion measure. Further, the information dataset related with the proposed framework is thought to be liberated from clamor and missing qualities [7].

II. TYPES OF CARDIAC DISEASE

There are a few classifications of heart sicknesses. Figure 2 shows the different kinds of coronary illness dependent on clinical conditions. These classifications are comprehensively delegated myocardial dead tissue, cardiovascular breakdown, heart arrhythmia, angina pectoris, cardiomyopathy, atrial fibrillation dependent on their clinical proof. Coronary illness has numerous highlights, which influence the capacity or structure of the heart [8].

Coronary Artery Disease

The coronary conduit infection is inconvenience prompt by drained course of blood .The consumption supply in corridors will harm the vein and produce the uneasiness to the standard systolic and diastolic capacity of the heart [9].

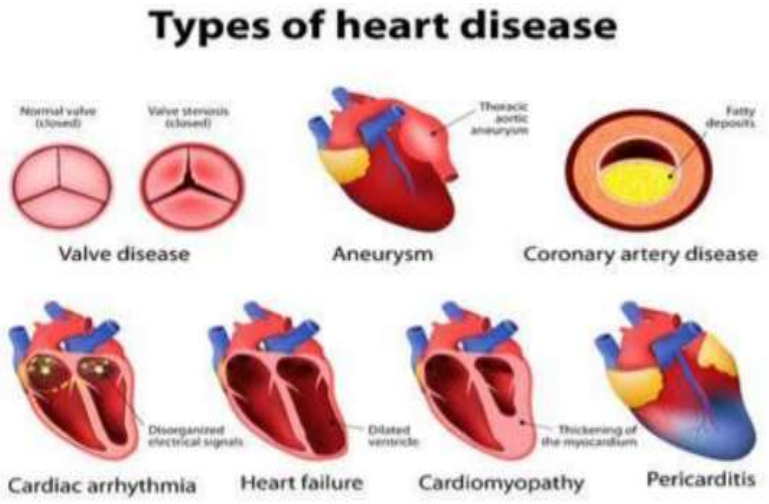

Fig. 1: Types of Cardiac Disease

Acute myocardial infarction

Clinical name for a heart failure is intense myocardial localized necrosis. A heart failure is a condition that greasy substances present in the blood esteem influence the pace of stream which results tissue harm on corridors. The blockage corridors will most likely be unable to supply the oxygenated blood supply to the body which will bring about the brokenness to different organs. Figure 2 clarifies a kind of heart capture brought about by extreme weight [10].

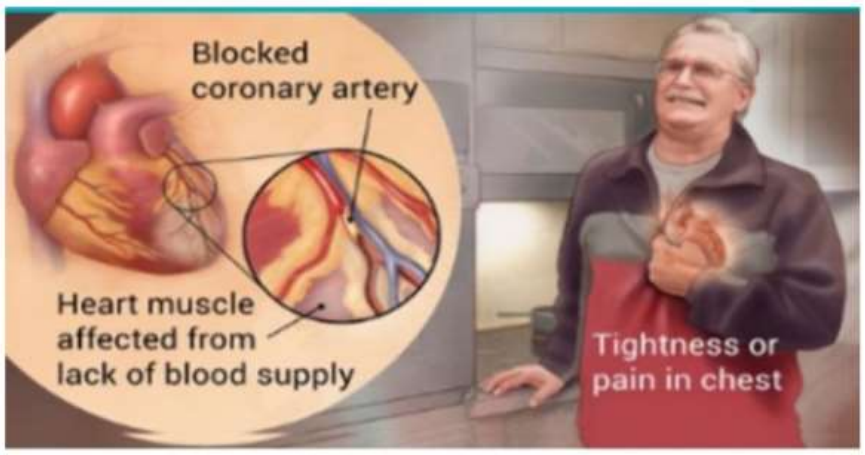

Fig. 2: Acute Myocardial Infarction

Chest Pain (Angina)

Clinical name of chest pressure is Angina. It is overwhelming clinical consideration need crisis treatment for the patients. Patients needs to treated with ventilators promptly on the off chance that we experience this sort of distress. Because of the helpless stock of blood stream will cause the tension on the blood dividers and influence the veins. Which will makes tension on the blood vessals results chest torment. Figure 4 shows common angina caused in the coronary vessel. Stable angina is the condition causes in peritoriam. Sporadic blood stream between the peritoris dividers. The fundamental reasons of flimsy angina are way of life adjustment, social propensities. Figure 3 shows run of the mill unstableangina caused in the coronary vessel [11].

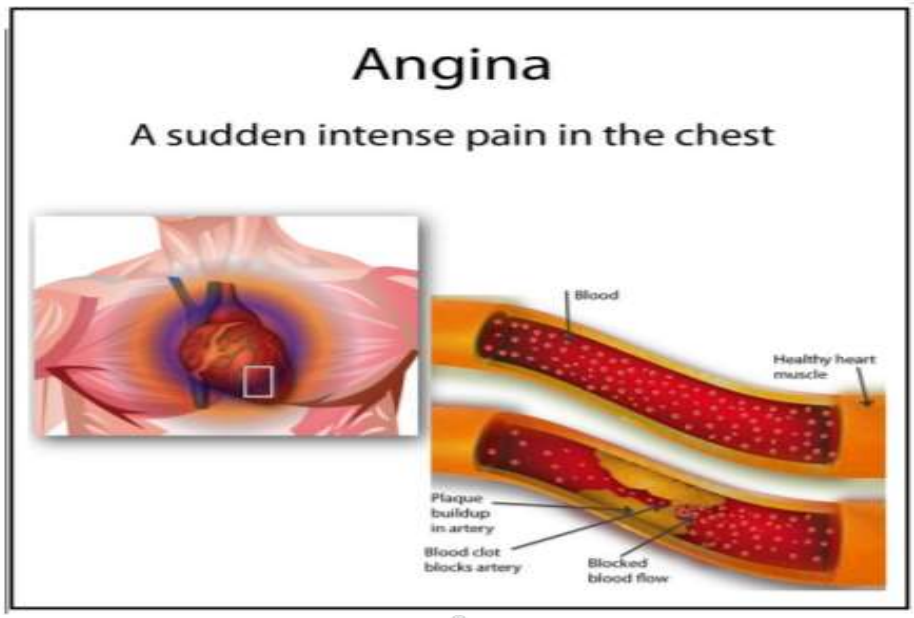

Fig. 3: Angina

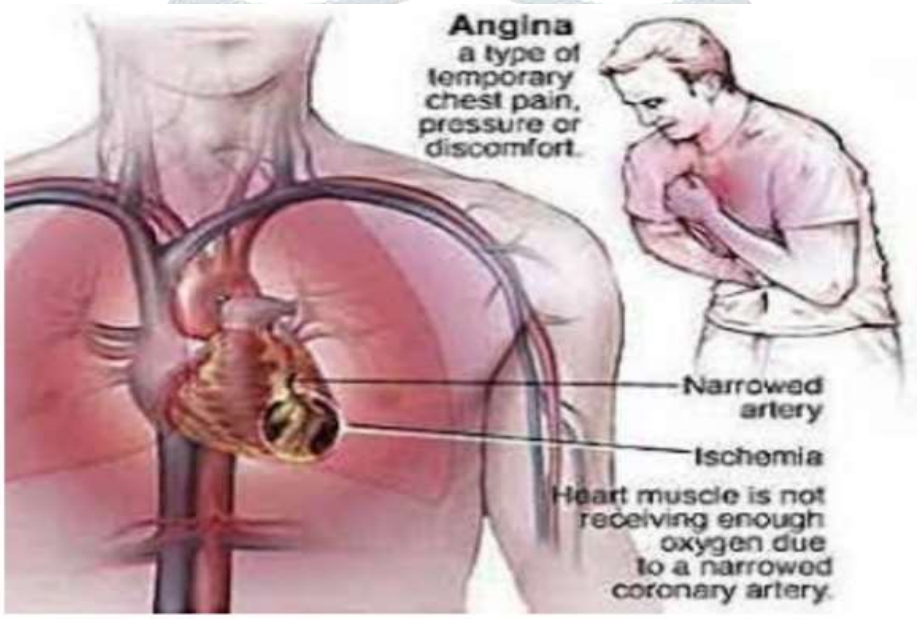

Fig. 4: Unstable Angina

III. **PROPOSED METHODOLOGY**

Naive Bayes Classifier technique is functioned based on Bayesian theorem. The designed technique is used when dimensionality of input is high. Bayesian Classifier is used for computing the possible output depending on the input. It is feasible to add new raw data at runtime. A Naive Bayes classifier represents presence (or absence) of a feature (attribute) of class that is unrelated to presence (or absence) of any other feature when class variable is known. Naïve Bayesian Classification Algorithm was introduced by Shinde S.B and Amrit Priyadarshi (2015) that denotes statistical method and supervised learning method for classification. Naive Bayesian Algorithm is used to predict the heart disease. Raw hospital dataset is employed. After that, the data gets preprocessed and transformed. Finally by using the designed data mining algorithm, heart disease was predicted and accuracy was computed.

Table 1: Heart Disease Attributes Datasets

Sr. No.	Attribute	Representative icon	Details
1	Age	AGE	Patient age (In years)
2	Sex	SEX	Gender of patient (male-0 female-1)
3	Chest Pain	CP	Chest pain type
4	Rest blood pressure	TRESTBPS	Resting blood pressure (in mm Hg on admission to hospital ,values from 94 to 200)
5	Serum cholesterol	CHOL	Serum cholesterol in mg/dl, values from 126 to 564)
6	Fasting blood sugar	FBS	Fasting blood sugar>120 mg/dl, true-1 false-0)
7	Rest electrocardiograph	RESTECG	Resting electrocardiographics result (0 to 1)
8	Max Heart rate	THALCH	Maximum heart rate achieved (71 to 202)
9	Exercise-induced angina	EXANG	Exercise included agina(1-yes 0-no)
10	ST depression	OLDPEAK	ST depression introduced by exercise relative to rest (0 to .2)
11	Slope	SLOPE	The slop of the peak exercise ST segment (0 to 1)
12	No. Of vessels	CA	Number of major vessels (0-3)
13	Thalassemia	THAL	Defect types; 3—normal; 2—fixed defect; 1—reversible defect
14	Target	TARGET	0 or 1

The dataset was divided into two datasets (70%/30%, training/testing) to avoid any bias in training and testing. Of the data, 70% was used to train the ML model, and the remaining 30% was used for testing the performance of the proposed activity classification system. The expressions to calculate precision and recall are provided in Equations (1) and (2).

$$\text{Precision} = TP\ TP + FP \times 100 \quad (1)$$

$$\text{Recall} = TP\ TP + FN \times 100 \quad (2)$$

PCA:-

PCA, also known as Hotelling Transform or Karhunen-Loeve expansion, is a well-known data representation and feature extraction technique widely used in the areas of pattern recognition, computer vision, etc. The purpose of PCA is to reduce the data dimensionality with reveal its essential characteristics i.e. to extract the relevant information from high dimension data set.

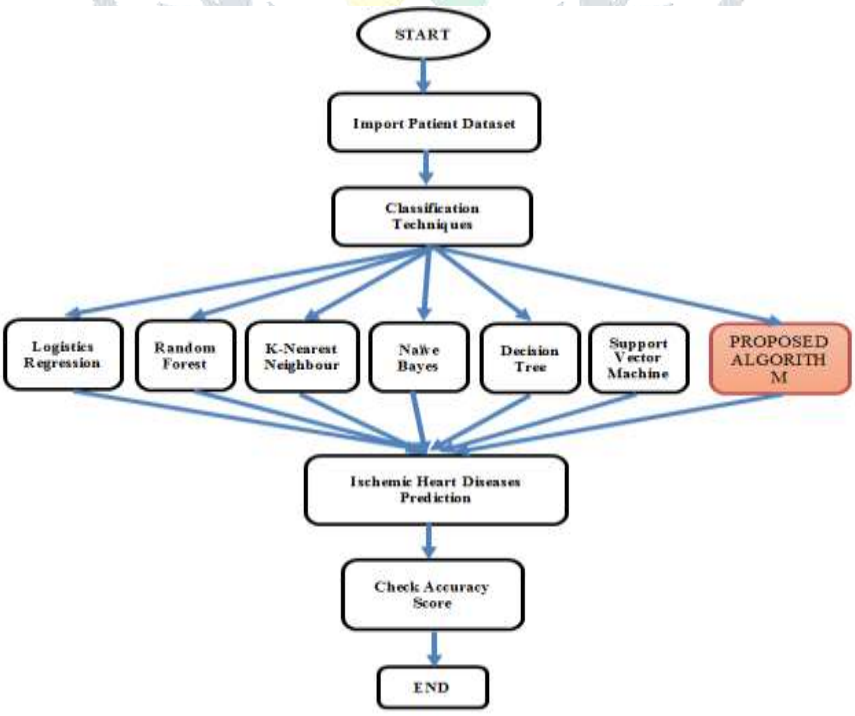

Fig. 5: Flow chart of Proposed Algorithm

The Eigenface (PCA) based Method of Turk and Pentland [4] is one of the foremost successful method applied in the literature which is based on the Karhunen-Loeve expansion and their study was motivated by the earlier work of Sirowich and Kirby [3] [5]. It is based on the application of Principal Component Analysis to the human faces. It treats the face images as 2-D data, and classifies the face images by projecting them to the eigenface space which is composed of eigenvectors obtained by the variance of

the face images. Eigenface recognition derives its name from the German prefix eigen, meaning own or individual. The Eigenface method of facial recognition is considered the first working facial recognition technology [6]. When the method was first proposed by Turk and Pentland [4], they worked on the image as a whole. Also, they used Nearest Mean classifier two classify the face images. By using the observation that the projection of a face image and non-face image are quite different, a method of detecting the face in an image is obtained. They applied the method on a database of 2500 face images of 16 subjects, digitized at all combinations of 3 head orientations, 3 head sizes and 3 lighting conditions. They conducted several experiments to test the robustness of their approach to illumination changes, variations in size, head orientation, and the differences between training and test conditions. They reported that the system was fairly robust to illumination changes, but degrades quickly as the scale changes [4]. This can be explained by the correlation between images obtained under different illumination conditions; the correlation between face images at different scales is rather low. The eigenface approach works well as long as the test image is similar to the training images used for obtaining the eigenfaces.

IV. **SIMULATION RESULTS**

There are two classes found in Scikit-learn machine learning library called LabelEncoder and OneHotEncoder. LabelEncoder basically transforms the categorical values into numbers which are ordinal in nature. In data set used for this study, there are categorical variables such as Cp, chest pain type which is represented as 1,2,3 and 4. 1,2,3 and 4 does not have ordinal relationship with each other therefore it gives wrong results when applied directly to machine learning algorithms. Thus, OneHotEncder is used to encode chest pain type values into binary values, this resolves the issue of ordinality. In this data set the dependent variable or the value to be predicted is multi class. It ranges from 0 to 4.

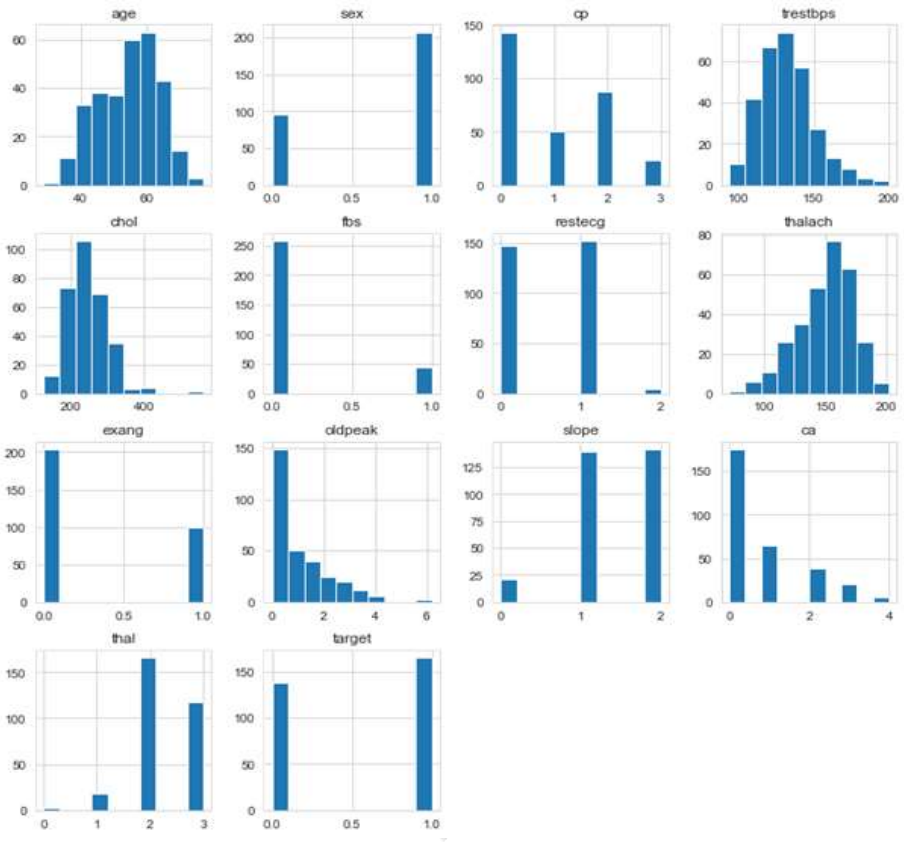

Fig. 6: Histogram of Dataset

Training set is the portion of data in which the model is trained. In this study, 70 percent of data was used for training. In general, in machine learning communities, it is a norm to used 60 to 70 percent of data for training but it varies diversely according to the need 31 and purpose of the experiment. In data training, often the accuracy of training is high, meaning the model shows high level of accuracy performance in the training set but when tested against the test set, the performance is poor. So to avoid performance error, k-fold cross validation was used. In k-fold cross validation, for example 10-fold cross validation, training set is split in 10 parts and and from each 10 part, training and test set is defined and model is employed and the result of all the 10 parts are averaged, this helps to minimize the over fitting and under fitting of the data.

Figure 6 shows the histogram of attributes shows the range of dataset attributes and code which is used to create it.

In proposed algorithm we used an ensemble of SVM, KNN and navies bayes to achieve an accuracy of 92.615%. The Majority vote-based model as demonstrated which comprises of random foreest, Decision Tree and Support Vector Machine classifiers, gave an accuracy of 83.51%, sensitivity of 72.52% and specificity of 82.41% for UCI heart disease dataset.

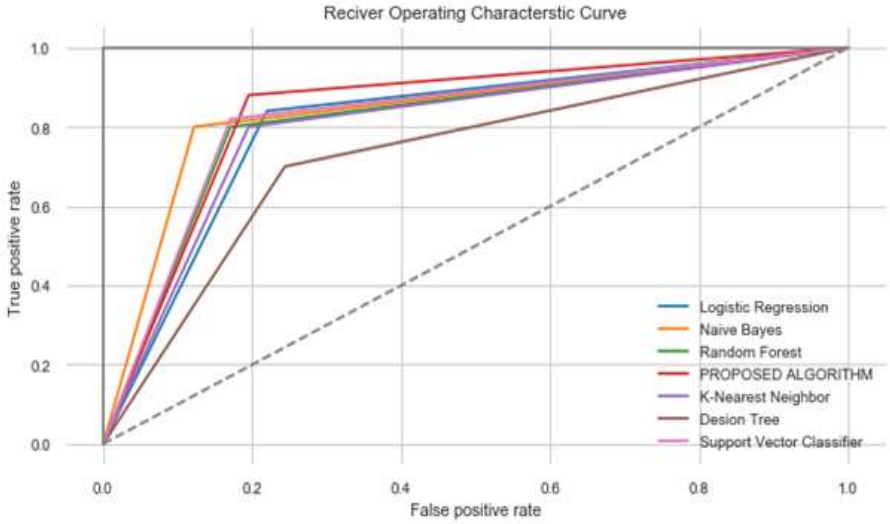

Fig. 7: Roc Curve for Accuracy

After performing the machine learning approach for testing and training we find that accuracy of the knn is much efficient as compare to other algorithms.

Table 2: Accuracy Comparison

Algorithm	Accuracy
Logistic Regression	81.31868131868131
Naive Bayes	83.51648351648352
Random Forest	79.31868131868131
K-Nearest Neighbor Classifier	80.21978021978022
Decision Tree Classifier	72.52747252747253
Support Vector Machine	82.41758241758241
PROPOSED ALGORITHM	**92.61538461538461**

Accuracy should be calculated with the support of confusion matrix of each algorithms as shown in Figures here number of count of TP, TN, FP, FN are given and using the equation (2) of accuracy, value has been calculated and it is conclude that proposed algorithm is best among them with 92.615% accuracy and the comparison is shown in Table 2.

V. CONCLUSION

This research work achieved the highest accuracy of 83.51% with navies bayes without PCA, 82.61% with navies bayes with PCA. Though ensemble classifiers using Boosted Tree, Bagged Tree, Subspace DA, and Subspace navies bayes without and with PCA are trained, it was observed that the ensemble classifiers did not perform better than the single classifiers in term of accuracy. In the future, intend to perform hyper parameter optimization and conduct more experiments by using feature selection algorithms on a dataset with more observations to improve the classifier's performance.

REFERENCES

[1] Karna Vishnu Vardhana Reddy, Irraivan Elamvazuthi, Azrina Abd Aziz, Sivajothi Paramasivam and Hui Na Chua, “Heart Disease Risk Prediction using Machine Learning with Principal Component Analysis”, International Conference on Intelligent and Advanced Systems (ICIAS), IEEE 2021.

[2] M.Ganesan and Dr. N. Sivakumar, “IoT based heart disease prediction and diagnosis model for healthcare using machine learning models”, International Conference on System, Computation, Automation and Networking (ICSCAN), IEEE 2019.

[3] Priyan Malarvizhi Kumar, Usha Devi Gandhi, "A novel threetier Internet of Thingsnarchitecture with machine learning algorithm for early detection of heart diseases", Computers and Electrical Engineering, Vol.65, pp. 222–235, 2018.

[4] Prabal Verma, Sandeep K. Sood, "Cloud-centric IoT based disease diagnosis healthcare framework", J, Parrallel Distrib. Comput., 2018.

[5] M.Ganesan, Dr.N.Sivakumar, “A Survey on IoTrelated Patterns”,International Journal of Pure and Applied Mathematics,Volume 117 No. 19, 365-369, 2017.

[6] R.Rajaduari, M.Ganesan,Ms.Nithya “A Survey on Structural Health Monitoring based on Internet of Things”International Journal of Pure and Applied Mathematics,Volume 117 No. 18, 389-393, 2017.

[7] Amin Khatami, AbbasKhosravi, C. L. (2017), 'Medical image analysis using wavelet transform and deep belief networks', Journal of Expert Systems With Applications 3(4), 190–198.

[8] Zhang, Shuai, Y.-L. S. A. (2017), 'Deep learning based recommender system: a survey and new perspectives', Journal of ACM Computing Surveys 1(1), 1–35.

[9] Zhiyong Wang, Xinfeng Liu, J. G. (2016), 'Identification of metabolic biomarkers in patients with type-2 diabetic coronary heart diseases based on metabolomic approach', 6(30), 435–439.

[10] Ashwini Shetty, Naik, C. (2016), 'Different data mining approaches for predicting heart disease', International journal of innovative research in science, engineering and technology 3(2), 277–281.

[11] Berikol, B. and Yildiz (2016), 'Diagnosis of acute coronary syndrome with a support vector machine', Journal of Medical System 40(4), 11–18.

[12] Chebbi, A. (2016), 'Heart disease prediction using data mining techniques', International journal of research in advent technology 25(3), 781–794.

[13] Cheng-Hsiung Wenga, Tony Cheng-Kui Huang, R.-P. H. (2016), 'Disease prediction with different types of neural network classifiers', Journal of Telematics and Informatics (4), 277–292.

[14] Ghadge, Prajakta, K. (2016), 'Intelligent heart attack prediction system using big data', International journal of recent research in mathematics computer science and information technology 2(2), 73–77.

[15] Lafta, R., Zhang, J. and Tao (2016), 'An intelligent recommender system based on short-term risk prediction for heart disease patients', Journal of web intelligence and intelligent agent technology (12), 102–105.