Analysis and Prediction of Cardiovascular Disease using Machine Learning Classifiers

[1]Prof. Vishal Shinde,[2]Mr. Hrithik Kedare, [3]Mr. Rahul Gupta, [4]Mr. Vinit Mahajan
[1]Asst.Professor,[2,3,4]UG Student,[1,2,3,4] Computer Engg. Dept. Shivajirao S.Jondhle College of Engineering & Technology, Asangaon, Maharshatra, India. [1]mailme.vishalshinde@gmail.com, [2]kedarehrithik@gmail.com, [3]rahul916g@gmail.com, [4]mahajanvinit410@gmail.com

Abstract- Cardiovascular disorder refers to a variety of issues that encompass narrowed or blocked veins, which can result in a heart assault, chest pain (angina), or stroke. The condition is anticipated via the gadget gaining knowledge of classifier based at the nation of the patient's facet impact. This research is used to look at the presentation of system mastering Tree Classifiers inside the prediction of Cardiovascular disorder (CVD). Random forest, choice Tree, Logistic Regression, assist vector system (SVM), and k-nearest neighbors (KNN) were used to break down system learning tree classifiers based on their precision and AUC ROC rankings. The Random forest device studying classifier performed an extra precision of eighty-five percent, ROC AUC rating of 0.8675, and execution time of one.09s in this examination of predicting Cardiovascular disease.

Keywords- Regression, K-nearest neighbors (KNN), disease prediction

I. INTRODUCTION

Cardiovascular Disease (CVD) is the most well-known deadly sickness in the world, claiming the lives of more people each year than any other disease. In, 17.9 million people died from Coronary Heart disease (CHD), responsible for 31 percent of the overall death worldwide. Heart stroke and heart failure account for 85 percent of these fatalities. More than three-quarters of CVD fatalities occur in low-yielding countries. In 2015, 82 percent of the 17 million less-than-ideal closures (younger than 70) due to non-infectious diseases occurred in low-yield countries, and 37 percent were caused by Cardiovascular Disease (CVD). Most cases of Cardiovascular Disease (CVD) can be avoided by avoiding known risk factors like cigarette use, poor eating habits, obesity, lack of exercise, and heavy alcohol consumption in social conditions. People with Cardio Vascular Disease (CVD) or who are at high heart disease risk (due to the presence of at least one risk factor, hyperlipidemia, such as diabetes, hypertension, or a well-established illness) require an early introduction and direction using short prescriptions, as indicated. The growth of blood clusters and the deposition of fatty deposits inside the conduits (atherosclerosis) characterize Cardiovascular Disease (CVD). Coronary episodes and strokes are most commonly caused by a blockage that prevents blood from flowing to the heart or brain during stressful situations. The accumulation of greasy stores in the veins' most inner dividers is the most well-known explanation for this. The presence of a mix of risk factors, such as obesity, poor eating habits, and cigarette use is typically the cause of cardiovascular failures and strokes.

II. AIMS AND OBJECTIVE

a) Aim

In human life, providing healthcare is an unavoidable task. A wide range of illnesses that damage the heart and veins are classified as cardiovascular illnesses. Early methods for predicting cardiovascular diseases aided in making decisions about the progressions that occurred in high-risk people, lowering their risks.

b) Objective

Considering informational collection from Kaggle in the proposed research, and it does not necessitate information pre-handling systems such as the noisy data removal, the evacuation of incomplete information, the filling of default esteems if acceptable, and the classification of attributes for prediction and decision making at various levels. Methods such as specificity analysis, sensitivity, accuracy, classification, and sensitivity are used to determine the diagnostic model's performance. This research provides a prediction algorithm for determining whether or not some person has a cardiovascular illness and provides details or a treatment. This is accomplished by comparing the accuracy of applying the rules to the individual performances of the SVM, Random forest, Naive Bayes classifier, and other algorithms. An accurate model for predicting cardiovascular disease was constructed using logistic regression on a dataset gathered in a specific place.

III. LITERATURE SURVEY

 DOI : 10.35291/2454-9150.2022.0171

Paper 1: ML Algorithms for Medical Incident Prediction Comparison Juan-Jose Beunza, Juan-Jose Beunza, Juan-Jose Beunza, Juan-Jose Be. The objective of this paper was to examine the internal appropriateness and accuracy of a few supervised ML algorithms for risk exists events. Two distinct mathematical software systems were used to employed the findings. The data for this investigation came from the Framingham Heart Experiment database, which began as a prospective study of heart disease risk factors in Framingham, Massachusetts in 1948. Three complete data models were created using data mining procedures, and also a comparison technique research of distinct ML methods - tree cutting, random forest, vector support equipment, sensory networks, and retransmission. The trendline was used as a global selection index for determining the appropriate number of parameters and kind of data processing (AUC).

Paper 2: Using Heart rate Speed Flexibility and Machine Learning to Improve Detection Accuracy for Clinical Heart Failure Lina Zhao, Lina Zhao, Lina Zhao, Lina Zhao, Lina Zhao, Lina Zhao, Lin The diversity of physiological signals can reveal a lot about cardiovascular function and clinical cardiovascular disorders. Two key time-series variabilities are the heart rate variable as well as the pulse transit time variable. But, combining HRV as well as PTV may improve heart failure classification accuracy, which is uncertain. A simultaneous study of HRV & PTTV was carried out on both normal participants and heart failure patients in this paper, with the goal of improving HRV-based heart failure identification with the use of PTTV analysis. A total of forty healthy people and forty heart failure patients were included in the study. The limb lead-II electrocardiogram as well as the radial vascular pressure patterns were recorded in synchrony.

The traditional time- (MEAN, SDNN, and RMSSD) and frequency- (LF, HF, and LF/HF) domain indices, as well as non-linear (SD1, SD2, sample entropy, and fuzzy measure entropy) domain indices, were used to analyse the obtained RR and PTT time series. Except for MEAN (P = 0.1) and LF/HF (P = 0.9), all HRV indices indicated significant differences (all P;0.01) between the two categories, whereas only MEAN in PTTV decreased considerably in heart failure patients (P; 0.01). Furthermore, when the HRV, PTTV data points, as well as expected scenarios produced from screened based 3D CNN models were combined, a SVM classifier produced the best classification results, with a responsiveness of 0.93, exactness of 0.88, as well as accuracy of 0.90. This paper demonstrated the potential of PTTV analysis for the detection of clinical heart failure.

Paper 3: Parallel Delta Modulations with Rotated Linear-Kernel SVM have been used in a real-time arrhythmia pulse classification algorithm. Xiaochen Tang, Xiaochen Tang, Xiaochen Tang, Xiaochen Tang, Among the most promising possibilities for helping cardiovascular disease detection is a real-time wearable ECG monitoring sensor. We introduce a revolutionary real-time machine learning technique for arrhythmia classification in this work. The parallel Delta modulation & QRS/PT wave detection algorithms are being used in the system. Presenting a patient-dependent rotational linear-kernel svm classifier that integrates local and global classifiers, as well as three kinds of feature vectors collected directly from Delta modulated bit-streams.

The MIT-BIH Arrhythmia database has been used to assess the performance of the proposed system. Two binary classifications, Supraventricular Ectopic Beat (SVEB) vs the remainder four classes, and Ventricular Ectopic Beat (VEB) versus the rest, are conducted and properly assessed as per the AAMI standard. The F1 score, sensitivity, specificity, and positive predictivity value of the preferred SkP-32 method for SVEB classification are 0.83, 79.3 percent, 99.6 percent, and 88.2 percent, respectively, while the numbers for VEB classification are 0.92 percent, 92.8 percent, 99.4 percent, and 91.6 percent, respectively. The findings of Project I: Analysis and Prediction of Cardiovascular Disease Using Machine Learning Classifiers reveal that our suggested technique performs similarly to published studies. The proposed low-complexity technique might be used as a machine learning solution on the sensor.

IV. EXISTING SYSTEM

Cardiovascular illness shown as a sophisticated, dynamic clinical framework using cosmology and machine learning. The approaches that have been used in present methodology include ontology and machine learning. As a result, it demonstrates a viable cardiovascular choice to assist instrument for preventing errors in the clinical hazard appraisal of chest torment patients and assisting clinicians in accurately distinguishing patients with severe angina/heart chest torment from those with other causes of chest torment. Another machine learning approach offered in this methodology is the Coronary Artery Disease method dubbed N2 Genetic optimizer agent (another hereditary prepared). These are aggressive results that are nearly similar to the best in the field.

V. COMPARTIVE STUDY

 DOI : 10.35291/2454-9150.2022.0171

SR NO.	PAPER TITLE	AUTHOR NAME	Publication	Technology	Purpose
1.	Comparison Of Machine Learning Algorithms For Clinical Event Prediction (Risk Of Coronary Heart Disease)	Beunza, JuanJose	ELSEVIER, 2019	ML,RStudio,Rapid	To compare the utility of several supervised machine learning (ML) algorithms.
2.	Enhancing Detection Accuracy for Clinical Heart Failure Utilizing Pulse Transit Time Variability and Machine Learning	Zhao, Lina	IEEE,2019	ML	To investigate the improvement of HRV-based heart failure detection with the assistant of PTTV analysis
3.	A Novel Ontology And Machine Learning Driven Hybrid Cardio vascular Clinical Prognosis As A Complex Adaptive Clinical System.	Farooq, Kamran, and Amir Hussain	SpringerOpen,2016	MLDPS	To develop a hybrid clinical decision support mechanism by combining evidence, extrapolated through legacy patient data to facilitate cardiovascular preventative care.

VI. PROBLEM STATEMENT

The identification of cardiac disease is a big difficulty. There are tools that can forecast heart disease, but they are either too high priced or ineffective for calculating the risk of heart disease in humans. The mortality rate and overall consequences of heart disorders can be reduced if they are detected early. However, it is not possible to correctly monitor patients every day in all circumstances, and a doctor's 24-hour consultation is not accessible since it needs more patience, time, and experience. Using various machine learning algorithms to evaluate the data for hidden patterns because it has so much data in today's environment. In pharmaceutical data, the hidden patterns might be employed for health diagnosis.

VII. PROPOSED SYSTEM

The data set includes features such as age, gender, cp, resting blood pressure, serum cholestrol, fbs, restecg, max heart rate achieved, ca, and pick out with 304 instances has been acquired from the UCI (University of California at Irvine) repository. The dataset is cleansed and processed at the first level utilising preprocessing technique like Data Integration, Data Transformation, Data Reduction, and Data Cleaning with the pandas tool. There were 304 patient records visualised using the suggested framework. The data scientist can use data visualisation tools to determine the dataset's viability. The sex and target attribute connection is depicted in a box plot. The histogram and correlation matrix were shown.

VIII. ALGORITHM

The general idea of working of proposed system algorithm is given as follow:

Step.1: Start

Step.2: Random Forest Algorithm Implementation

Importing Random Forest from sklearn library.

clf=Model(model=RandomForestClassifier(), X=X, y=y)

clf.crossValScore(cv=10)

clf.accuracy()

clf.confusionMatrix()

clf.classificationReport()

Step.3: Decision Tree Algorithm Implementation

Importing Decision Tree from sklearn library.

dt=Model(model=DecisionTreeClassifier(),X=X,y=y)

dt.crossValScore(cv=10)

dt.accuracy()

dt.confusionMatrix()

dt.classificationReport()

Step.4: SVM Algorithm

Importing SVM from sklearn library.

svm = Model(model=SVC(C=5, probability=True), X=X, y=y)

svm.crossValScore(cv=10)

svm.accuracy()

svm.confusionMatrix()

svm.classificationReport()

Step.5: KNN Algorithm

Importing KNN from sklearn library.

knn=Model(model=KNeighborsClassifier(n_neighbors=100), X=X, y=y)

knn.crossValScore()

knn.accuracy()

 DOI : 10.35291/2454-9150.2022.0171

knn.confusionMatrix()
knn.classificationReport()

Step.6: Logistic Regression

Importing Logistic Regression from sklearn library.
pipeline=make_pipeline(QuantileTransformer(output_distribution ='normal'), lr)
lg = Model(model=pipeline, X=X, y=y)
lg.crossValScore()
lg.accuracy()
lg.confusionMatrix()
lg.classificationReport()

Step.7: Exit

IX. MATHEMATICAL MODEL

1.SVM

Support Vector Machine (SVM) had gotten a lot of press recently after demonstrating their effectiveness in a range of pattern classification tasks. They've been used to solve a variety of challenges, including hand-written text detection, bioinformatics, and voice recognition (among others). Consider a binary classification task with a set of linearly separable training samples S = ©(a1, b1), (a2, b2)· · · (an, bn)[a], (1)where a ∈ Rd, i.e., a lies in a d-dimensional input space, and yi is the class label such that bi ∈ {−1, 1}. The label indicates the class to which the data belongs. A suitable discriminating function could then be defined as : f(a) = sgm(hw·ai + c) . (2) Where vector w determines the orientation of a discriminant plane (or hyperplane), hw·ai is the inner product of a vectors, w and a and c is the bias or offset. Clearly, there are an endless number of planes that could categorise the training data properly.

2.Logistic Regression

Logistic regression is a statistical and machine-learning approach for categorising records in a dataset depending only on their input field values. It predicts outcomes by using one or more independent factors to accurately predict the variable. The dependent variable in this algorithm is represented as a binary variable with values of 1 (yes) and 0 (no) (no). As function of X, the model forecasts. The following are the assumptions that are employed in log regression: Binary logistic regression necessitates binary dependent variables; for binary regression, the dependent variable's factor 1 level must indicate the intended outcome, and independent variables must be unrelated.

3. K nearest neighbor

K nearest neighbour (KNN) is a algorithm which saves all instances and categorizes newer cases based on similarity. 1) Case-based reasoning (KNN algorithm) 2) k nearest neighbour 3) reasoning from examples 4) Learning through examples 5) Reasoning from memory 6) Procrastination in learning. KNN is non classification algorithm with two kinds. 1) no-structure NN approaches; 2) structure-based NN techniques. The total information is split in testing and train sample data in structure less NN algorithms. The distance between the training and sample points is measured, and the point with smallest distance is referred to as the closest neighbour. Data structures such as the orthogonal structural tree (OST), k-d tree, ball tree, central line, closest future line, and axis tree are used in structure-based NN approaches.

4. Decision Tree

The decision tree is indeed a graph that employs a branching strategy to show each possible choice outcome. Decision trees could be formed or hand-drawn with specialized software or a graphics programme. When a group has to make a decision, decision trees can help concentrate the conversation. They could be used to give monetary, time, or other values to probable outcomes in order to automate choices.

X. SYSTEM ARCHITECTURE

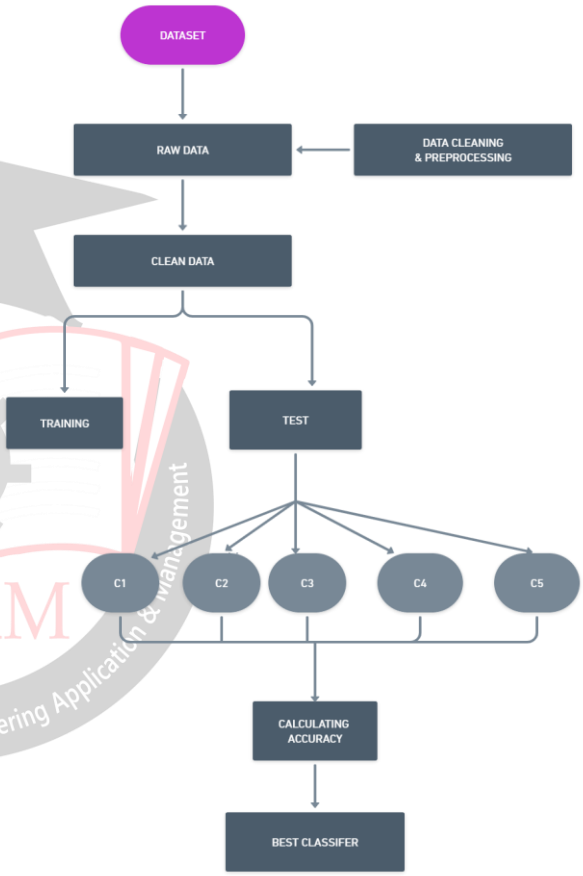

Fig.1: System Architecture

Description: There are 2 types of phases: 1. Training Phase 2. Testing phase

1)Training phase: The system collects five valid and five invalid signatures from the user, with three being used for training and two being used for testing. After that, some preprocessing operations are performed on the signature, followed by feature extraction operations using the ratio, centroidx, centroid y, and solidity extraction functions.

2) Testing phase: It verifies the new user's signature during the testing process. The system will examine five valid or

 DOI : 10.35291/2454-9150.2022.0171

invalid signatures from a new user and determine if they are authentic or counterfeit.

XI. ADVANTAGES

1. Permits solve complex actual-global issues with numerous constraints.
2. Address problems like having little or nearly no categorized records availability.
3. Ease of shifting knowledge from one model to some other based on domain names and tasks.
4. Ease of transferring expertise from one model to some other primarily based on domains and responsibilities.
5. It automatically detects the essential functions with none human supervision.
6. Much less computational energy or assets like RAM, CPU, GPU or TPU, and many others.
7. In less amount of information, we can gain extra accuracy.

XII. DESIGN DETAILS

Fig.2: Data Description

XIII. CONCLUSION

Thus, we have tried to implement the paper " A.Shaeen Sulthana, N. Komal Kumar, D.Krishna Prashanthi, G.Sarika Sindhu", "Analysis and Prediction of Cardiovascular Disease using ML Classifiers", ICACCS 2020, and according to this paper machine learning classifiers such as Decision Tree, Support vector machine (SVM), Random Forest, K-nearest neighbors (KNN), Logistic Regression, were used in the prediction of Cardio Vascular Disease (CVD).The proposed method using the random forest machine learning phase achieved 85.71% accuracy with the 0.8675 ROC AUC school performing the most well-evaluated design in classifying patients with Cardio Vascular Disease.

XIV. REFERENCE

[1] "N.Kumar", "G.Sarika Sindhu", "D.Krishna Prashanthi", "A.Shaeen Sulthana", et al. "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers." 2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS). IEEE, 2020.

[2] "Beunza", "Juan-Jose", et al. "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)." Journal of biomedical informatics 97 (2019): 103257.

[3] "Zhao", "Lina", et al. "Enhancing Detection Accuracy for Clinical Heart Failure Utilizing Pulse Transit Time Variability and Machine Learning." IEEE Access 7 (2019): 17716-17724.

[4] "Tang", "Xiaochen", et al. "A Real-time Arrhythmia Heartbeats Classification Algorithm using Parallel Delta Modulations and Rotated Linear-Kernel Support Vector Machines." IEEE Transactions on Biomedical Engineering (2019).

[5] "Kelly", "B. B.", & "Fuster", V. (Eds.). (2010). Promoting cardiovascular health in the developing world: a critical challenge to achieve global health. National Academies Press.

 DOI : 10.35291/2454-9150.2022.0171