International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.135

Cardio Vascular Disease Prediction Using Multiple Machine Learning Algorithms

Venkata Sai Ashrith Kona
Department of Data Science and Business Systems, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology,
Kattankulathur-603203,Tamil Nadu, India kv6209@srmist.edu.in

Maithili Saran Reddy Lingala *Department of Data Science and Business Systems, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology,*
Kattankulathur-603203,Tamil Nadu, India ls2204@srmist.edu.in

Rajasekar P
Department of Data Science and Business Systems, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology,
Kattankulathur-603203,Tamil Nadu, India rajasekp@srmist.edu.in

Hruday Vuppala
Department of Data Science and Business Systems, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology,
Kattankulathur-603203,Tamil Nadu, India hv6819@srmist.edu.in

Sravya Adapa
Department of Data Science and Business Systems, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology,
Kattankulathur-603203,Tamil Nadu, India na8385@srmist.edu.in

***Abstract*—This Cardio vascular diseaseis one of the serious issue that we are facing in current day it has become a massive challenge to try and analyse the cardiovascular disease survivors. Artificial intelligence is a component of machine learning, which is used to address several issues in data science. We can predict results based on past data which is a very frequently used application of machine learning for the machine to forecast predictions it has to identify patterns from the previous data and these patterns can be used on latest or new data to predict the outcome. The medical industry produces enormous amounts of unprocessed data, which data mining transforms into meaningful information that might aid in making decisions. Decision Tree (DT), Adaptive boosting classifier (AdaBoost), Logistic Regression (LR), RandomForest (RF), Gradient Boosting classifier (GBM), and K-NearestNeighbor (KNN)arethe classificationmethodsusedin this study**

Keywords—Cardiovascular disease, Machine learning (ML), Random Forest, Decision Tree, Adaptive boosting classifier, Gradient Boosting classifier, KNN

I. INTRODUCTION

Thebiggestcause ofdeathworldwide, asreportedbythe WHO, is heart disease. According to estimates, cardiac conditions account for 24% of deaths in India from non-communicable diseases. The cause of one-third of all fatalities worldwide is heart disease. Heart diseases are to blame for 50 percent of mortality in the United States and other industrialised nations. Every year, around 1crore 70 lakh people worldwide die from cardiovascular disease (CVD). It might be hard to identify (CVD) due to many contributing variables, including high BP, high cholesterol, diabetes, irregular pulse rate, and several other illnesses.The symptoms of CVD might occasionally vary based on a person's gender. For instance, a female patient may also suffer nausea, severe tiredness, and shortness of breath in addition to chest pain, but male patients are most likely to have chest pain. Researchers have investigated a variety of waystopredictcardiacdiseases,butpredictingsoatabeginningstageisnotparticularlysuccessfulforavarietyof reasons, including complexity, execution time, and method accuracy. Consequently, efficacious diagnosis and treatment can save a lot of lives. Between healthcare service guidelines, medications, and lost productiveness as a result of death, in 2014 and 2015 it cost roughly $219 billion annually. Heart failure, which can result in death, can also be avoided with early detection. Although angiography is thought to be the most exact and accurate procedure for predicting cardiac artery disease (CAD), it is quite expensive, making it lessaccessible tofamilies withlimited financial resources. Physical examination can cause few errors which might even lead to death of few patients as heart disease is a very complicated disease and we have to take at most care and here using machine learning based expert systems will help us to effectively diagnose Cardio Vascular Disease (CVD).DataMiningplays a major rolein many fields like engineering, business, and education to extract data and find interesting patterns out of those. Examining data to find hidden information that will be useful to take important decisions in the future is a process called as "data mining". By decreasing the error in factual resultsandforecast,Understandingthe complexityandnon-linear interplay between several components, a wide range of machine learning techniques have been used. Medical experts must employ ML and AI algorithms to analyse data and draw exact and detailed diagnostic judgments because the amount of medical data is always growing. Different categorizationalgorithmsareusedindataminingofmedical data to predict patients' CVD and deaths from heart attack.

II. LITERATURESURVEY

[1] Melillo et al. proposed a system that automatically distinguishes high-risk patients from low-risk individuals. Classification and regression tree (CART) (93.3% sensitivity, 63.5% specificity) performed better in their investigation.Only12little-riskand34huge-riskpatientswereexamined.Tofindoutiftheirproposedmethodis useful, a huge dataset must be carefully investigated Guidi et al examined the clinical support system (CDSS) for heart failure inspection. This model provided outputs such as HF(Heart Failure) sensitivity . They conducted study using various machine learning classifiers and

DOI: rp-9788770040723.135

compared the results. Random forest and CART performed best with 87.6% accuracy out of all classifiers.

[2] Parthiban and Srivatsa have done a extensive study and have conducted research to find out heart disease in those patients who have diabetes. They used many predictive features like blood pressure, blood sugar, and age there is a imbalance in data set and the writers ofhave not employed any strategy to address this issue. they were able to achieve an accuracy of 94.60% by using support vector machine (SVM) classifier.

[3] Al Rahhal*et al* have used a novel approach using deepneural network (DNN) they used raw ECG data to predictusinganunsupervisedlearningtechnique stackeddenoising autoencoders(SDAEs) to examine the highest level of features. They allowed expert engagement, which can induce biases, throughout each training cycle. It may bring about prejudice.

[4] Muthukaruppan and Er proposed a fuzzy expert system for the identification of CVD that is based on Particle Swarm Optimization (PSO). Fuzzy rules were created when rules from the decision tree were retrieved. Their accuracy using the fuzzy expert system was 93.27%. On the short dataset used in their investigation, a few rules were extracted. Alizadehsani and others

[6,7]Alizadehsani*et al.* utilised a group-based learning strategy. They utilised a dataset with 303 cases that they acquired from the "Rajaie Cardiovascular Medical and Research Centre" for their study. For CVD prediction, authors employed the introductory C45 ensemble learning approach. Left circumflex stenosis, left anterior descending stenosis, and right coronary artery (RCA) stenosis were accurately identified with 68.96%, 61.46%, and 79.54%, respectively (LAD). By using the SVM model,the results were improved and "80.50% accuracy for RCA, 86.14% accuracy for LAD, and 83.17% accuracy for LCX" were reached by a new team of researchers.

[8] Tama*etal.* presentedtheideaofa two-tierensemble paradigm, where certain classifiers serve as the basis for anotherensemble.UsingclasslabelsfromExtremeGradient Boosting (EGB), Random Forest (RF), and Gradient Boosting Machine (GBM), the proposed stacking architecture isconstructed (XGBoost). Four unique datasets areusedtoassesstheirproposeddetectionmodel.Moreover, theyusedfeatureselectionmethodsbasedonparticleswarm optimization. With a k value of 10, their suggested model faredbetter inthe k-fold cross-validation. Onlythe stacking of tree-based models was considered by the authors. Additionalstatisticalandregression-basedtechniquesmight be used to improvise model outcome.

[9] Abdar et al. established the N2Genetic optimizer, a novel optimization approach. The patients were then identifiedashavingCHDornotusingthenuSVM.OntheZ-Alizadeh Sani dataset, the proposed detection approach had an accuracy of 93.08% when compared to earlier works. Raza proposedan ensemble architecture with majorityvote. To forecast heart illness in a patient, it incorporated logistic regression, multilayer perceptron, and naive Bayes. A classification accuracy of 88.88% was attained, surpassing all base classifiers combined.

[10] Mohan et al developed a hybrid approach based on combining a linear model with a random forest to predict cardiac disease (HRFLM). On the Cleveland dataset, the suggested technique raised performance levels and had an accuracy rate of 88.7%.

[11] Soni and Vyas they used WARM, and their degree of confidence was 79.5%. dependent on age, smoking behaviours, BMI range and Hypertension their research assigned weights. Soni et al. on the other hand, gave each qualityaweightdependingontheadvicetheyobtainedfrom the medical experts. By attaining a maximum score of 80% confidence, Using a weighted associative classifier, they demonstrated a bright and effective cardiovascular attack prediction system.

[12] GannaA,MagnussonPKandteam.Effortonusing machinelearningalgorithmstoidentifycardiovascularheart disease has had a substantial effect on this work. In this paper, a summary of the literature is presented. Using a varietyofmethods,aneffectivepredictionofcardiovascular disease has been achieved. Logistic Regression, KNN, Random Forest Classifier, etc. are a few of them. The outcomes demonstrate the capability of each algorithm to register the given objectives. The findings indicate that every algorithm is capable of registering the given objectives, with KNN displaying the greatest performance (88.52%).

III. PROPOSEDSYSTEM

In this literature we have proposed multiple ML algorithmstofindifapersonhasCardiovasculardiseaseor not. Building, training, testing and validating the architecture for a specified challenge is a complex process. "Decision Tree, Adaptive boosting classifier, Logistic Regression, Gradient Boosting classifier and K-Nearest Neighbor" are the classificationmethodsusedinthisstudy. Google colab was used to run the experiment. In this study the data is collected from 1025 patients which consists of both healthy and patients suffering from cardio vascular disease and we use attributes like age, sex, chol, cp(chest pain) etc to predict if a person is healthy or suffering from cardio vascular disease and this data set contains a total of 14attributestheabovementionedalgorithmsareconsidered to be best for predicting the cardio vascular disease as they are all supervised learning algorithms.

A. Overviewofarchitecture

Fig 1 consists of the overall architecture of the cardio vasculardiseasepredictionusingmultiplemlmodelsandthe main parts of this architecture is data collection, data preprocessing and predicting the data using the given algorithms.Our technique uses the data of patient to predict thepatient'sheartconditionweatherthepersonhastheheartdisea se or not. And these predictions are made by the best algorithm of all the ml algorithms used and the model is trainedbeforehandwithagenuinedatasettomakeaccurate predictions.

DOI: rp-9788770040723.135

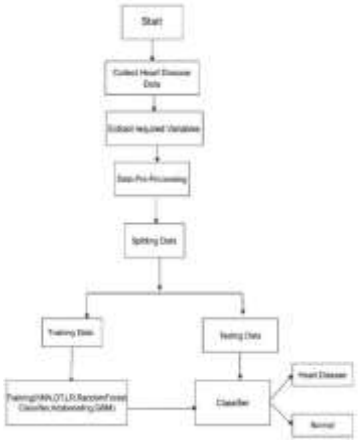

Fig.1.FlowchartforHeartdiseaseprediction

B. *RelatedWork*

Apurv Garg et al have proposed a model that forecasts the chances of getting cardio vascular disease utilizing the two ML algorithms that are KNN and Random forest they have compared these two models in order to get the best accuracy possible out of which KNN yielded a prediction accuracyof 86.88% where as the RF yieldedanaccuracyof 81.96% [15]

Archana singh and Ramesh kumar have proposed a prediction based model with multiple machine learning algorithms like SVM, DT, LR, KNN out of which KNN yielded highest accuracy of 87% they have first collected data then selected the required attributes then the data is preprocessedandthenbalancedthedatatheyhaveusedUCI repository dataset [16]

In this study we have used similar approaches and we were able to get better accuracies for the models by using different dataset with more number of data points. We were able to achieve better accuracy for our proposed modelKNN

C. *DataCollection*

We took our data from kaggle website for free and our dataiscalledheart.csvthisdatasetcontainsof1025patients records.Outofthis1025people499people arenormaland526 people have heart disease and this data set has 14 attributes and out of these people there are 713 male and312 female. And out of people that have heart disease 300 are male and 226 female

```
.: Dataset Info :.
********************************
Total Rows: 1025
Total Columns: 14
********************************


.: Dataset Details :.
********************************
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1025 non-null   int64
 1   sex       1025 non-null   int64
 2   cp        1025 non-null   int64
 3   trestbps  1025 non-null   int64
 4   chol      1025 non-null   int64
 5   fbs       1025 non-null   int64
 6   restecg   1025 non-null   int64
 7   thalach   1025 non-null   int64
 8   exang     1025 non-null   int64
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64
 11  ca        1025 non-null   int64
 12  thal      1025 non-null   int64
 13  target    1025 non-null   int64
dtypes: float64(1), int64(13)
```

Fig.2.Datavaluesandtheattributes

D. *DataPreprocessing*

In this step we use one-hot encoding technique to transform categorical values to numerical values then we drop of unnecessary variables then we separate features, now we normalize the data using min-max method now we splitthedataintotwopartstrainingandtestingoutof which the ratio of training is 80 and testing is 20 and now the data is ready and can be used in any model.

IV. MACHINELEANINGALGORITHMS

A. *LogisticRegression(LR)*

Logistic Regression is one of the best available tough classifier among the supervised ml algorithms. It is an elongation of general regression model it reflects the likelihood of the occurrence or nonoccurrence of a certain instance.Logisticregressionisusedtodescribedataandthe connection between a dependent variable and one or more independentvariables.Nominal,ordinal,orperiodtypesare all acceptable for the independent variables. LR calculates the chance that a new observation will belong to a specific class, with the result lying between 0 and 1 because it is a probability.

B. *DecisionTree(DT)*

DecisiontreeamongsttheoldestMLalgorithms.Used for issuesrelatedtoclassification and regressionwe have a best supervised algorithm that can deal with them and that algorithm is Decision tree and most of the times it is used for classification problems. It is basically a Tree shaped classifier root node is the top node while others are child nodes. Internal nodes represent the features of datasets while leaf node consist result Decision node and the leaf node are the nodes that make up decision tree. Decision node generallymakesupdecisionsasithasmanybranches whereas leaf node can't make any decisions

C. *K-NearestNeighbor(KNN)*

KNN is among the very few oldest algorithms or statistical learning technique. In KNN K is basically to represent the total number of nearest neighbors used which is directly mentioned in the object builder. As a result,relatedsituationsareclassifiedsimilarly,andanew instance is classified by comparing it to each of the existing examples. KNN method will search the pattern space for k training samples adjacent to the supplied unique sample when one is provided. Two distinct methodsareofferedtotranslatethedistanceintoaweight so that predictions from many neighbors of the test instance may be calculated based on their distance.

D. *Adaboost*

An ensemble method in machine learning is called AdaBoost, also known as Adaptive Boosting. The most popular AdaBoost algorithm is a decision tree with one level, or a decision tree with only one split. A model is createdviaAdaBoost,andallthedatapointsaregiventhe same weight. After that, it gives points with incorrect classifications more weight. The following model now accords greater relevance to each and every point with greater weights. As long as no low errors are received, it will continue training models.

DOI: rp-9788770040723.135

E. RandomForest

A ML method that utilizes many numerous decision treestomakeadecisionisknownasRandomforest.Itisa ensemble learningbased technique. While it is in the training stage, it

Producesmany trees and a forest of decision trees. Eachandeverytree, acomponentoftheforest, predicts a class label for each and every occurrence during the testing period. The model will take the class with the highest votes and makes it as prediction. The individual tree makes a class prediction from a very large independent tree models working together will give out the best result.

F. GradientBoosting

Using boosting, weak learners may become strong learners.Everylatesttreecreatedbyboostingisfittedtoa updated version of the original data set. Then it is anticipatedthatwhenmergedwitholdermodels,thebrand newmodelwillproduceforecastswithreducedfaultrates. The major aim is to set objectives for this next model to reducemistakes.GradientBoostinginagradual,additive, andlinearfashiontrainsmanymodels.Inviewoffactthat eachandeveryindividualcase goalresultsare decidedby the gradient's deviation respective to the forecasts, the phrase "gradient boosting" came into popularity. Every model picks up speed in a correct procedure by reducing theforecast errors.

E. ProposedAlgorithm

In this study the best out of all the algorithms is KNN whichhasachievedanaccuracyof97%whichisconsidered as one of the best algorithm in supervised classification algorithm and other than that it is simple KNN is non-parametricandlazy,whichmeansitdoesnotassumeanything about the distribution of the underlying data and does not create a model from the training set. As an alternative,itmemorisesthefulltrainingdatasetandutilises it to make predictions when presented with fresh test cases. For many applications, KNN is a straightforward and efficientmethod,althoughitcanhavelargecomputingcosts and be sensitive to the choice of K and the distance metric usedtocompare instances. KNN is a flexible technique that may be used to solve a variety of issues since it can be applied to both classification and regression jobs.

V. EVALUATION

For the machine learning models, there are some approachesforperformanceevaluation.Itisanticipatedthat the blending of several assessment tools will support the advancement of analytical research. Four fundamental measures(‘accuracy,precision,recall,andF-Score’)willbe looked at in this study to see how machine learning-based algorithms differ from one another.

Using the confusion matrix, we may assess the four measures. The Confusion Matrix's constituents are True Positive(TP),TrueNegative(TN),FalsePositive(FP),and False Negative(FN).Inthemedicaldatathemostimportant thing is to find out (FN). The performance metrics are provided below

$$\text{Accuracy}=\text{correctlyclassifiedpredictions} \quad (1)$$

$$/\text{Totalnumberofpredictions}$$

$$\text{Precision}=TP/TP+FP \quad (2)$$

$$\text{Recall}=TP/TP+FN \quad (3)$$

$$F\text{–Score}=2*\text{precision}*\text{recall}/\text{precision}+\text{recall} \quad (4)$$

The total collection of features in the heart disease dataset have been exposed to comparison analysis of supervised machine learning classifiers. Some classifiers performedwell onevaluation measures, whereasothersdid not. In order to predict heart failure survival, this work employedtree-based,statistical-basedandregression-based models. The DT, RF ensemble models are tree-based. AdaBoost and GBM are two tree-based boosting methods. Statistically-basedmodelswhereasregression-basedmodels include LR and KNN

Model	Accuracy
K-Nearest Neighbour	95.121951
Random Forest	89.268293
Gradient Boosting	88.780488
Logistic Regression	82.439024
AdaBoost	81.463415
Decision Tree	79.024390

Fig. 3.Differentaccuracycomparison

Asper thetablewe have KNN with thebest accuracyof 97.02%, Random forest with an accuracy of 90.16%, Gradintboostingwithaaccuracyof88.7%,LRwithaaccuracyof 82.43%,Adaboost with a accuracyof 81.46% and with the least accuracy is the decision tree algorithm with an accuracy of 79%

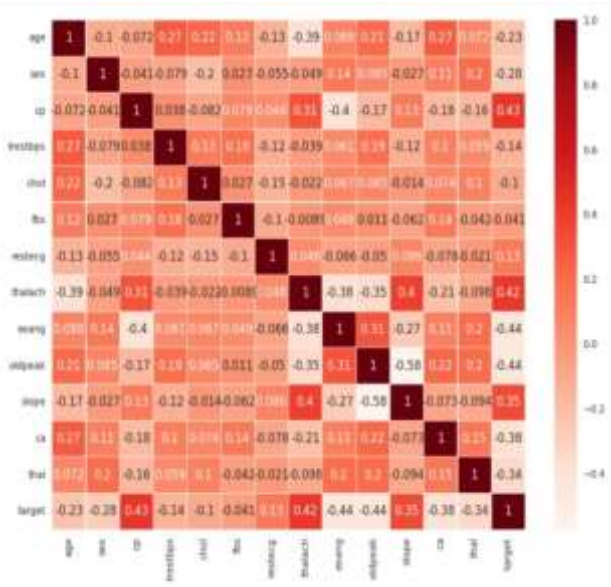

Fig..4.Correlationmatrixofvariables

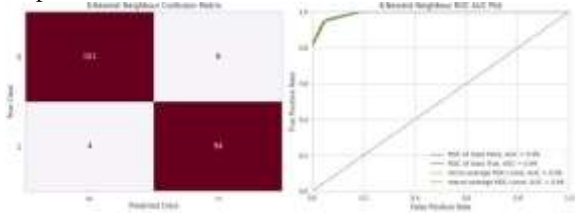

Fig.5.RocandconfusionmatrixofKNNthebestalgorithm

TABLE.1. PRECISION,RECALL ANDF-MEASURES

Algorithm	Precision	Recall	F-1
KNearest Neighbor	0.97	0.97	0.97
Random forest	0.91	0.90	0.90
Gradient Boosting	0.89	0.89	0.89
Logistic Regression	0.84	0.82	0.82
Adaboost	0.82	0.81	0.81
Decision Tree	0.80	0.79	0.79

TABLE.2.VALUEOF AREA UNDERROC

Algorithm	AUROC
KNearest Neighbor	0.99
Randomforest	0.96
Gradient Boosting	0.95
LogisticRegression	0.91
Adaboost	0.87
DecisionTree	0.86

In conclusion, a dataset on heart illness was gathered, preprocessed as needed, and then analysis was done to better understand the dataset. Following the application of six machinelearningalgorithmsAdaboost,LR,Gradientboost, KNN,DT,andRFweassessedthepredictionsusingtheF-1 Measure, ROC curve, recall, accuracy, and precision. We discovered that all of the used algorithms performed well, with KNN demonstrating the greatest performance with97% accuracy, showing that these algorithms are the most effective at predicting cardiac disease.

VI. CONCLUSION

Heartpatients'liveswill be savedthroughtheprocessing of raw health data of heart information using machine learning algorithms. By identifying risk factors for heart failure, preventive steps can be taken to lower mortality rates. In this study, a machine learning-based technique for predicting the survival of heart patients is suggested. The following ML methods are used: LR, AdaBoost, RF, GBM, DT and KNN. KNN with a accuracy of 97% the highest of all algorithms with precision score 0.97 recall 0.97 F-1 0.97 and AUROC 0.99 the work done here has the potential to advance the medical field and help doctors foresee how much time a patient with heart condition will live. Additionally, it will aid doctors in realizing that if a heart failure patient survives, they can concentrate on key risk factors. To gain from their combined advantages, the research can employ a range of machine learning model combinations in the future. To better the efficiency of ML models, better feature selection methods may be created. Due to the fact that these feature selection issues are NP-hard, meta-heuristics can be used.

REFERENCES

[1] P. Melillo, N. De Luca, M. Bracale and L. Pecchia, "Classification tree for risk assessment in patients suffering from congestive heart failure via long- term heart rate ariability", IEEEJ. Biomed. HealthInformat.,vol.17, no. 3, pp. 727-733, May 2013.

[2] Gomathy, V., Janarthanan, K., Al-Turjman, F., Sitharthan, R., Rajesh, M., Vengatesan, K., &Reshma, T. P. (2021). Investigating the spread of coronavirus disease via edge-AI and air pollution correlation. ACM Transactions on Internet Technology, 21(4), 1-10.

[3] G. Parthiban and S. K. Srivatsa, "Applyingmachine learningmethodsin diagnosing heart disease for diabetic patients", Int. J. Appl. Inf.Syst., vol. 3, no. 7, pp. 25-30, Aug. 2012**.**

[4] M.M.A.Rahhal,Y.Bazi,H.AlHichri,N.Alajlan,F.Melgani andR.R. Yager, "Deep learning approach for active classification ofelectrocardiogram signals", Inf. Sci., vol. 345, pp. 340-354,Jun. 2016

[5] Rajesh, M., &Sitharthan, R. (2022). Introduction to the special section on cyber-physical system for autonomous process control in industry 5.0.Computers and Electrical Engineering, 104, 108481.

[6] Z. Sani, R. Alizadehsani, J. Habibi, H. Mashayekhi, R. Boghrati, A.Ghandeharioun, et al., "Diagnosing coronary artery disease via datamining algorithms by considering laboratory and echocardiographyfeatures", Res. Cardiovascular Med., vol. 2, no. 3, pp. 133, 2013.

[7] R. Alizadehsani, M. H. Zangooei, M. J. Hosseini, J. Habibi, A.Khosravi, M. Roshanzamir, et al., "Coronary artery disease detectionusing computational intelligence methods", Knowl.-Based Syst., vol.109, pp. 187-197, Oct. 2016.4

[8] B. A. Tama, S. Im and S. Lee, "Improving an intelligent detectionsystem for coronary heart disease using a two-tier classifierensemble", BioMed Res. Int., vol. 2020, Apr. 2020.

[9] M. Abdar, W. Książek, U. R. Acharya, R. S. Tan, V. Makarenkov,and P. Pławiak, "A new machine learning technique for an accuratediagnosis of coronary artery disease," Computer Methods andPrograms in Biomedicine, vol. 179, article 104992, 2019.

[10] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart diseaseprediction using hybrid machine learning techniques," IEEE Access,vol. 7, pp. 81542–81554, 2019.

[11] Soni S, Vyas OP. Using associative classifiers forpredictiveanalysisinhealthcaredatamining. IntJComput Appl.2010;4(5):33–7.

[12] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M,Ärnlöv J & Ingelsson E (2013). Multilocus genetic risk scores forcoronary heart disease prediction. Arteriosclerosis, thrombosis, andvascular biology, 33(9), 2267-7

[13] https://www.kaggle.com/datasets/johnsmith88/heart-disease-datas

[14] Syed Nawaz Pasha, Dadi Ramesh, Sallauddin Mohmmad, A.Harshavardhan and Shabana "Cardiovasculardisease prediction usingdeep learning techniques "OP Conference Series: Materials Scienceand Engineering, Volume 981, International Conference on RecentAdvancements in Engineering and Management (ICRAEM-2020) 9-10 October 2020, Warangal, India Citation Syed Nawaz Pasha et al2020 IOP Conf. Ser.: Mater. Sci. Eng. 981 022006

[15] Garg, Apurv& Sharma, Bhartendu& Khan, Rizwan. (2021). Heartdiseaseprediction using machine learningtechniques. IOPConferenceSeries:Materials Scienceand Engineering.1022.012046.10.1088/1757-899X/1022/1/012046

[16] Singh, A., & Kumar, R. (2020).Heart Disease Prediction Using Machine Learning Algorithms. 2020 International Conference on Electrical and Electronics Engineering (ICE3).doi:10.1109/ice348803.2020.9122958