MACHINE LEARNING-BASED PREDICTION OF CARDIOVASCULAR DISEASE RISK

[1]Upputuri Prathibha,[2]P China Yalamanda Rao,[3]Y. Lakshmi Annapurna,[4]Raju Pothana
[1,2,3]Assistant Professor,[4]Student
Department of CSE
G V R & S College of Engineering & Technology,Guntur,AP

ABSTRACT

The healthcare sector manages billions of people globally and generates enormous amounts of data. Better insights are being produced by the machine learning-based algorithms as they analyse the multidimensional medical information. In this work, many cutting-edge Supervised Machine Learning algorithms that are specifically employed for illness prediction are applied to classify a cardiovascular dataset. According to the findings, Decision Tree classification model outperformed Naive Bayes, Logistic Regression, Random Forest, SVM, and KNN based methods in its ability to predict cardiovascular illnesses. The Decision Tree delivered the best outcome with a 73% accuracy rate. This method could assist medical professionals anticipate the onset of cardiac problems and provide the proper therapy.

1. INTRODUCTION

1.1 BRIEF INFORMATION

According to WHO, the leading cause of mortality globally is cardiovascular disease (CVD). Nearly 17.9 million deaths annually are attributed to CVD, which accounts for 31% of all fatalities worldwide. Any abnormality of the heart's blood vessels, including narrowing of the arteries and veins that convey blood to and from the organ, is referred to as cardiovascular disease (CVD). This makes blood flow more difficult and may even result in blockages, which may lead to a heart attack or stroke. High blood pressure, poor diet, physical inactivity, elevated blood cholesterol, alcohol and cigarette use, obesity, and genetic mutations are risk factors for CVD. Early prognostication may help to prevent the mortality brought on by these variables. On the other hand, the installation of the Internet of Things has led to a constant improvement in data gathering methods. Terabytes of data are being produced daily by healthcare organisations thanks to these innovations. A human being cannot compile many data points and determine a specific patient's ailment. However, machine learning might be used to detect patterns in the data as a prediction method.

The variables are examined and used to forecast who is most likely to suffer from heart disease using machine learning. Methodologies for machine learning may analyse vast amounts of data and spot patterns that might not be obvious to humans. The efficiency and precision of processing the ever-growing volumes of data are often improved. Additionally, it enables immediate modification without the need for human involvement. Supervised machine learning is the process of giving the system labelled data and output patterns in order to complete a job. The programme looks for data patterns during training that correspond to the intended output. The

supervised learning model can predict the proper label for freshly given input data after training.

1.2 PURPOSE

By choosing the appropriate subset of characteristics that have a substantial influence on the prediction outcomes, it is possible to improve model performance and save a large amount of runtime. It is feasible to achieve both of these objectives using a technique known as feature selection. The terms "filters," "wrappers," and "embedding" refer to the three most popular techniques for selecting features. The embedded strategy known as GBDT was used in the study to choose the feature variables. This is because embedded approaches are substantially faster than wrapper methods and provide better prediction performance when compared to filter methods. To accomplish learning, GBDT uses a forward stepwise algorithm and an additive model. Together, these two elements help to achieve this. The relevance of the characteristics grows correspondingly with the size of the weighted impurity reduction that takes place during splitting for non-leaf nodes. Due to this, a thorough description of the part that each characteristic plays in determining the overall accuracy of the predictions given by the integrated GBDT is not feasible. We employ a method called feature imputation to address this problem, in which the explanatory model is a linear function of the data generated by feature imputation.

1.3 SCOPE

The dataset, which includes 76 features, comprises the predicted characteristics that contribute to heart disease in individuals, and 14 significant characteristics are chosen from them to help assess the system. If all the characteristics are taken into account, the creator receives a less efficient system. Attribute selection is carried out to improve efficiency. In this case, n characteristics must be chosen in order to evaluate the model that provides more accuracy. Some dataset characteristics have virtually equal correlations, thus they are eliminated. The efficiency significantly declines if all the characteristics in the dataset are taken into consideration. A prediction model is created after comparing the accuracy of each of the seven machine learning techniques. Thus, the objective is to use a variety of assessment measures, such as the confusion matrix, accuracy, precision, recall, and f1-score, which accurately predicts the illness. The extreme gradient boosting classifier has the greatest accuracy of 81% when comparing all seven. The dataset, which has 76 characteristics, contains the anticipated factors that influence heart disease in patients, and 14 significant variables that are essential for assessing the system are chosen from them. If all the characteristics are taken into account, the creator receives a less efficient system. Attribute selection is carried out to improve efficiency. In this case, n characteristics must be chosen in order to evaluate the model that provides more accuracy. Some dataset characteristics have virtually equal correlations, thus they are eliminated. The efficiency significantly declines if all the characteristics in the dataset are taken into consideration. A prediction model is created after comparing the accuracy of each of the seven machine learning techniques. Thus, the objective is to use a variety of assessment measures, such as the confusion matrix, accuracy, precision, recall, and f1-score, which accurately predicts the illness. The extreme gradient boosting classifier has the greatest accuracy of 81% when comparing all seven.

1.4 MOTIVATION

Weng et al. (31) used clinical data from more than 300,000 households in the UK to evaluate four alternative models. The results showed that with the bigger quantity of data analysed, NN was the approach that generated the most precise CVD prediction results. K-Nearest Neighbour (KNN), Random Forest (RF), and Decision Tree were the three conventional machine learning models Dimopoulos et al. tested and assessed using ATTICA data with 2020 samples for the small CVD dataset. The HellenicSCORE tool, which is a calibration of the ESC Score, was used in comparison to indicate that RF had provided the best results. Mohan et al. have suggested a hybrid HRFLM approach as a way to further improve the accuracy of the model predictions in light of the aforementioned popularity of machine learning methods. This is due to the expanding use of machine learning techniques in IoT applications. In order to forecast the state of the human body's cardiovascular system, Akash et al. looked at an IoT-ML approach. After gathering crucial information from the human body, the algorithm model employs machine learning (ML) methods to calculate and forecast the patient's cardiovascular health. The patient's heart rate, ECG signal, and cholesterol are all included in this data. Using more than 200,000 high-risk participants in eastern China, LR was used to assess 30 cardiovascular disease-related variables within the scope of Yang et al.'s assessment of local areas using independent prediction models. The trials' findings led to the creation of an RF model that is better appropriate for eastern China. Yang et al. introduced the concept of a stacking model for the very first time in the study of CVDs. To further understand how the stacking model affects the daily hospitalisation rate for CVDs, statistics on air pollution and weather were taken into account. A grassroots level of five basic learners was originally built to help with the stacking model creation.

2. LITERATURE SURVEY

The most crucial stage of the software development process is the literature review. Determine the time factor, economics, and corporate strength prior to building the tool. The following stages are to decide which operating system and language were utilised to construct the tool if these requirements have been met. Once the programmers begin creating the tool, they need a lot of outside assistance. This assistance was gathered from senior programmers, books, or websites. The aforementioned factors were taken into account before constructing the suggested system.

1) P-reserved Ejection Fraction-Based Machine Learning Prediction of Mortality and Hospitalisation in Heart Failure.

Angraal, S., Mortazavi, B., Gupta, A., Khera, R., Ahmad, T., Desai, N., Jacoby, D., Masoudi, F., Spertus, J., and Krumholz, H.

In the TOPCAT (Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist) trial, models for predicting mortality and heart failure (HF) hospitalisation for outpatients with HF with preserved ejection fraction (HFpEF) were developed. Although there are models for HF patients with a decreased ejection fraction, few have examined the risks of hospitalisation and mortality in HFpEF patients. The models were trained using the following 5 techniques: logistic regression with a forward selection of variables; logistic regression with a lasso regularisation for variable selection; random forest (RF); gradient descent boosting; and support vector machine. The models were validated using 5-fold cross-validation. Using Brier scores and receiver operating characteristic curves, respectively, model discrimination and calibration were computed.

2) A method based on machine learning for forecasting the onset of cardiovascular illnesses in dialysis patients.

Authors: Giacomo Fiumara, Pasquale De Meo, Ant onio Vilasi, Claudia Torino, Sabrina Mezzatesta,

End-stage kidney disease (ESKD) patients have a special cardiovascular risk. This research tries to accurately predict cardiovascular illnesses and mortality in dialysis patients.

Machine learning approaches have been used to accomplish our goal. Two datasets were taken into account: one was an American dataset provided by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) repository, and the other was an Italian dataset obtained from the Istituto di Fisiologia Clinica of Consiglio Nazionale delle Ricerche of Reggio Calabria. Depending on the result of interest, we were able to extract 5 datasets from each one. Both linear and non-linear algorithms were explored, but Support Vector Machine was ultimately chosen. The non-linear SVC with RBF kernel technique, which we improved using Grid Search, gave us the best results in particular. In order to increase the accuracy of the process, the final algorithm may be used to identify the ideal pair of hyper-parameters (in our instance, to find the best pair (C,)).

3) Machine learning technique-based predictive model of cardiac arrest in smokers based on Heart Rate Variability parameter.

Shashikant R. and Chetankumar P.

Cardiac arrest is a serious cardiac condition that claims billions of lives every year. Although smoking is a known risk factor for cardiovascular pathology, including coronary heart disease, there has been little research on the link between smoking and cardiac mortality. The Heart Rate Variability (HRV) characteristics employed in this paper's machine learning approach to predict cardiac arrest in smokers. Machine learning is a technique for computing that improves performance to improve prediction. It is based on automated learning. In order to predict cardiac arrest in smokers, this research compares how well logistical regression, decision trees, and random forest models work. In this study, a machine learning approach was used to a dataset that was provided by the Indian data science research team MITU Skillogies in Pune. Three prediction models with 19 input features of HRV indices and two output classes were constructed to determine if the patient has a possibility of going into cardiac arrest or not. The accuracy, precision, sensitivity, specificity, F1 score, and Area under the curve (AUC) of these models were examined. The logistic regression model has an 88.50% accuracy, an 83.11% precision, a 91.79% sensitivity, an 86.03% specificity, an F1 score of 0.87, and an AUC of 0.88. With an accuracy of 92.59%, precision of 97.29%, sensitivity of 90.11%, specificity of 97.38%, F1 score of 0.93, and AUC of 0.94, the decision tree model has been developed. The random forest model has a 93.61% accuracy rate, a 94.59% precision rate, a 92.11% sensitivity rate, a 95.03% specificity rate, an F1 score of 0.93, and an AUC of 0.95. Following the decision tree, the random forest model had the highest classification accuracy, while logistic regression had the lowest classification accuracy.

3. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

The variables are examined and used to forecast who is most likely to suffer from heart disease using machine learning. Methodologies for machine learning may analyse vast amounts of data and spot patterns that might not be obvious to humans. The efficiency and precision of processing the ever-growing volumes of data are often improved. Additionally, it enables immediate modification without the need for human involvement. Supervised machine learning is the process of giving the system labelled data and output patterns in order to complete a job. The programme looks for data patterns during training that correspond to the intended output. The supervised learning model can forecast the proper label for freshly provided input data after training. By comparing the classification accuracy of several supervised machine learning algorithms, this research seeks to uncover a successful strategy.

Disadvantages of Existing System

- There will be instances when they must wait for fresh data to be created since it takes additional data sets to train.
- It requires a lot of time and money to accomplish its goals accurately and pertinently.
- It must to be capable of correctly interpreting the outcomes.
- High sensitivity to errors.

3.2 PROPOSED SYSTEM

This study concentrated on using a few categorization algorithms and contrasting the results. The training and testing sections of the dataset were split 70/30. To predict CVD, classification models such as Naive-Bayes, Decision Tree, Logistic Regression, Random Forest, SVM, and KNN were utilised. The confusion matrix is used to pinpoint labelling or prediction errors. Four elements—True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN)—are used to match the actual and projected values. False Positive and False Negative values serve as the seeds for Type-I and Type-II mistakes. The calculation of Precision, Recall, F1-score, and Accuracy may be done extremely quickly using the confusion matrix.

Advantages of Proposed System

- The performance of Naive Bayes is good for higher dimensional data.

- Decision Tree can tolerate missing values and does not need data normalisation.
- Logistic regression is efficient and does not need input characteristics to be scaled.
- On reduced data and unbalanced datasets, Random Forest performs well.
- SVM uses very little memory.
- KNN is a model that is always changing and does not make any assumptions about the data.

3.5 HARDWARE REQUIREMENTS

The physical computer resources, sometimes known as hardware, are the most typical set of specifications given out by any operating system or software programme. The following sections go into detail about the different hardware requirements.

- System : CORE i3 Processor.
- Hard Disk : 40 GB.
- RAM : 4 GB.

3.6 SOFTWARE REQUIREMENTS

Software requirements are concerned with specifying the software resources and prerequisites that must be installed on a computer to provide the best possible performance of a programme. These prerequisites must be installed individually before the programme can be installed since they are often not included in the software installation package.

- Operating system : Windows 7 Ultimate(min)
- Coding Language : Python
- Front-End : Python, Django
- Designing : HTML, CSS, JavaScript.
- Data Base : MySQL
- Dataset : Kaggle

4. SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE

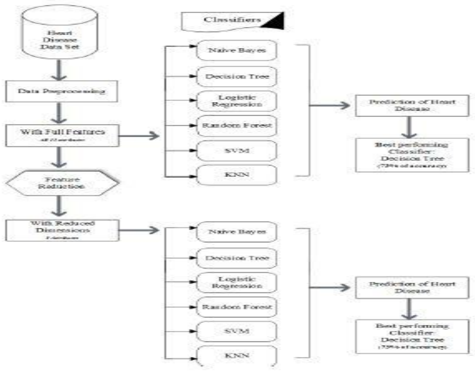


Fig: 4.1 System Architecture

4.2 MODULES

The step of implementation is when the theoretical design is translated into a programmatically-based approach. The application will be divided into a number of components at this point and then programmed for deployment. The following modules make up the bulk of the application. They are listed below.

- Run the decision tree algorithm,
- Generate the train and test data,
- Upload the cardiac dataset, and
- Run the svm method.
- Run the knn algorithm,
- Compare the graph,
- Then run the logistic regression method.

MODULES DESCRIPTION

UPLOAD CARDIAC DATASET MODULE

This module provides information about a patient's id, age, gender, height, weight, blood pressure, cholesterol levels, genetic mutations, smoking, alcohol, etc.,

GENERATE TRAIN AND TEST

This module provides information on the total number of records that were located in the dataset. 80% of the data are utilised to train the machine learning algorithms, whereas 20% of the records are used in the training process.

RUN SVM ALGORITHM

SVM is only 50% accurate, thus proceeding with the other modules as-is.

RUN DECISION TREE ALGORITHM

In the cardiac dataset, our decision tree technique outperforms naive bayes with 90% accuracy.

RUN LOGISTIC REGRESSION

For the cardiac dataset, this Logistic Regression provides 50% accuracy.

RUN KNN ALGORITHM

TFor a cardiac data collection, his KNN technique provides 62% accuracy.

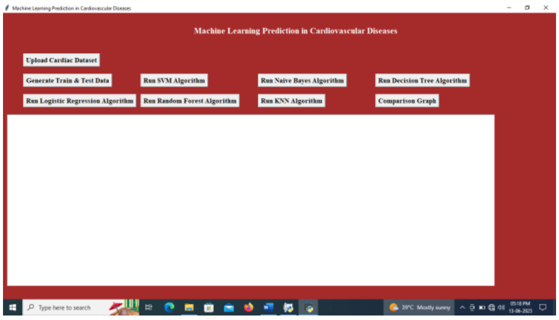
COMPARISION GRAPH

We used a cardiac dataset to train all the algorithms, and the Decision Tree approach had the greatest accuracy. In the comparison graph, the Y-axis represents accuracy, precision, recall, and F-score, and the X-axis the name of the method.

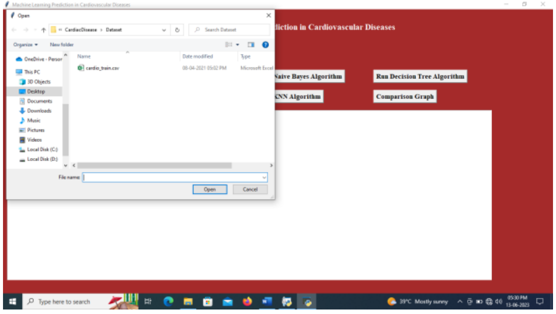
5.4 OUTPUT SCREENS

To reach the screen below, double-click the 'run.bat' file to launch the project.

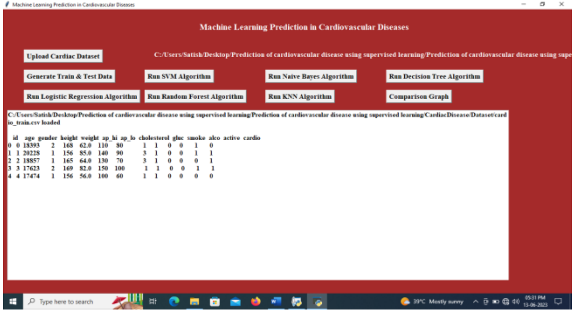
Using the 'Upload Cardiac Dataset' button on the


Machine Learning Prediction in Cardiovascular Diseases
Upload Cardiac Dataset
Generate Train & Test Data
Run SVM Algorithm
Run Naive Bayes Algorithm
Run Decision Tree Algorithm
Run Logistic Regression Algorithm
Run Random Forest Algorithm
Run KNN Algorithm
Comparison Graph

previous screen, upload the dataset.

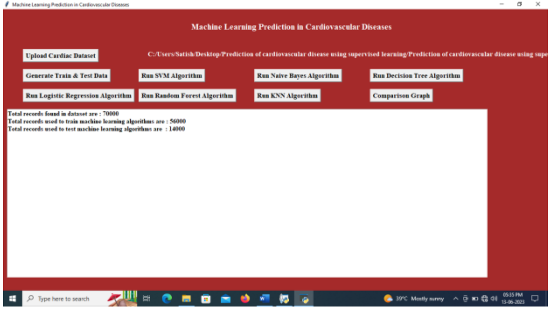


choosing and adding the "cardiac_train.csv" file to the above-mentioned screen, and then clicking the "Open" button to load the dataset and bring up the page below.



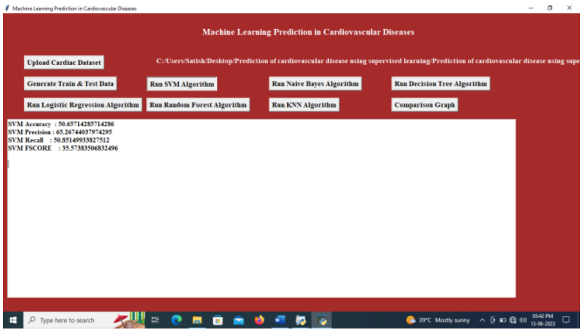
```
C:/Users/Satish/Desktop/Prediction of cardiovascular disease using supervised learning/Prediction of cardiovascular disease using supervised learning/CardiacDisease/Dataset/card
io_train.csv loaded

   id   age  gender  height  weight  ap_hi  ap_lo  cholesterol  gluc  smoke  alco  active  cardio
0   0  18393       2     168    62.0    110     80            1     1      0     0       1       0
1   1  20228       1     156    85.0    140     90            3     1      0     0       1       1
2   2  18857       1     165    64.0    130     70            3     1      0     0       0       1
3   3  17623       2     169    82.0    150    100            1     1      0     0       1       1
4   4  17474       1     156    56.0    100     60            1     1      0     0       0       0
```

The dataset is imported in the page above, and we can see some of its entries. Next, click the "Generate Train & Test Data" button to split the dataset into two parts: the train portion, which the programme used to train machine learning algorithms, and the test part, which it used to determine how accurate those algorithms were.



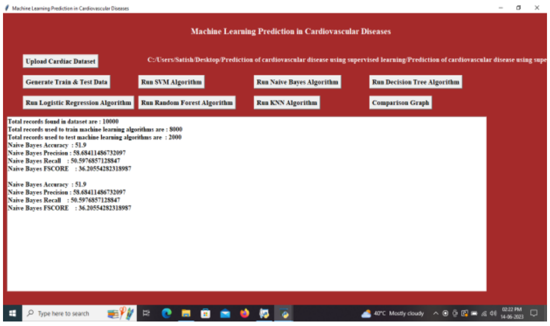
```
Total records found in dataset are : 70000
Total records used to train machine learning algorithms are : 56000
Total records used to test machine learning algorithms are  : 14000
```

The programme uses 8000 records for training and 2000 records for testing from the dataset of 10,000 records shown in the above screen. The features graph is shown below. Any characteristic with a correlation value near to 1 will be considered relevant on the significance graph.



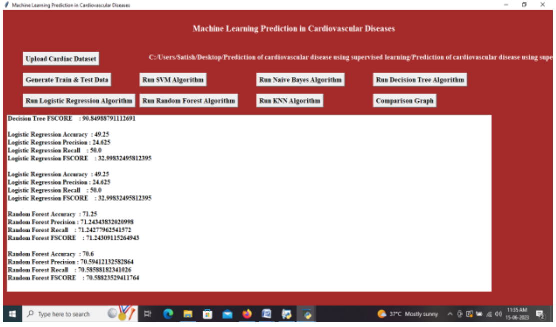
```
SVM Accuracy  : 50.65714285714286
SVM Precision : 65.26744037974295
SVM Recall    : 50.85149933827512
SVM FSCORE    : 35.57383506832496
```

SVM had a 50% accuracy rate on the screen above; to see its accuracy, click the buttons for Nave Bayes and Decision Tree.



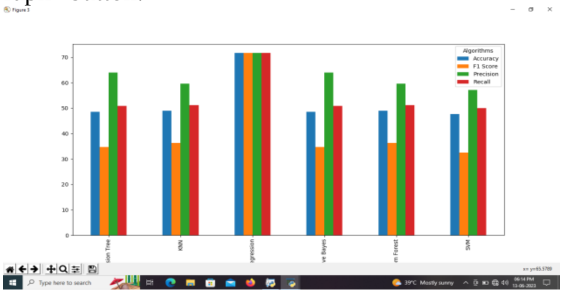
```
Total records found in dataset are : 10000
Total records used to train machine learning algorithms are : 8000
Total records used to test machine learning algorithms are  : 2000
Naive Bayes Accuracy  : 51.9
Naive Bayes Precision : 58.68411486732097
Naive Bayes Recall    : 50.5976857128847
Naive Bayes FSCORE    : 36.20554282318987

Naive Bayes Accuracy  : 51.9
Naive Bayes Precision : 58.68411486732097
Naive Bayes Recall    : 50.5976857128847
Naive Bayes FSCORE    : 36.20554282318987
```

Click the buttons for Logistic Regression, Random Forest, and KNN Algorithms to see their prediction accuracy on the screen above. With Nave Bayes, we achieved 51% accuracy, while with Decision Tree, we achieved 90% accuracy.



```
Decision Tree FSCORE    : 90.84988791112691

Logistic Regression Accuracy  : 49.25
Logistic Regression Precision : 24.625
Logistic Regression Recall    : 50.0
Logistic Regression FSCORE    : 32.99832495812395

Logistic Regression Accuracy  : 49.25
Logistic Regression Precision : 24.625
Logistic Regression Recall    : 50.0
Logistic Regression FSCORE    : 32.99832495812395

Random Forest Accuracy  : 71.25
Random Forest Precision : 71.24343832020998
Random Forest Recall    : 71.24277962541572
Random Forest FSCORE    : 71.24309115264943

Random Forest Accuracy  : 70.6
Random Forest Precision : 70.59412132582864
Random Forest Recall    : 70.58588182341026
Random Forest FSCORE    : 70.58823529411764
```

The accuracy of the logistic regression was 50%, random forest was 71%, and KNN was 62% in the images above. The decision tree method, which we trained on the cardiac dataset on the previous screen to attain the maximum accuracy possible, is shown below in the comparison graph after clicking the "Comparison Graph" button.



The algorithm name is shown on the x-axis of the above graph, while the accuracy, precision, recall, and

FSCORE are shown on the y-axis. We may infer from the graph above that decision trees provided improved prediction accuracy.

7. CONCLUSION

This illness is now one of the most prevalent and harmful to human health. Based on clinical information about a patient's prior heart disease diagnosis, this heart disease detection system helps the patient. The following methods were used to create the provided model: decision tree, Naive Bayes, SVM, Random Forest Classifier, and Logistic Regression. This is a project with a bright future since it can assist people of any age who are experiencing the symptoms forecast a cardiac arrest. Decision Tree has a high level of accuracy.

REFERENCES

[1] WHO (World Health Organizat ion): Cardiovascular Diseases - ht tps://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1

[2] Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, Jacoby DL, Masoudi FA, Spertus JA, KrumholzHM , “Machine Learning Prediction of Mortality and Hospitalizat ion in Heart Failure Wit h P reserved Ejection Fraction”, JACC : Heart Failure, vol. 8, Issue 1, January 2020.

[3] Sabrina Mezzatesta , Claudia Torino, Pasquale De Meo, Giacomo Fiumara , Ant onioVilasi, “ A machine learning-based approach for predict ing the outbreak of cardiovascular diseases in pat ients on dialysis” Comput er Met hods and P rograms in Biomedicine, Elsevier, vol. 177, pp. 9-15, August 2019

[4] Shashikant R,Chetankumar P, “Predict ive model of cardiac arrest in smokers using machine learning technique based on Heart Rate Variability parameter”, Applied Computing and Informatics, June 2019

[5] Ahmed M. AlaaI, Thomas Bolton, Emanuele Di Angelantonio, James H. F. Rudd, Mihaela van der Schaar, “Cardiovascular disease risk predict ion using automated machine learning: A prospective study of 423,604 UK Biobank participants”, P loS One 14 (5): e0213653, May 2019.

[6] Runchuan Li, Shengya Shen, Xingjin Zhang, Runzhi Li, Shuhong Wang, Bing Zhou and Zongmin Wang, “ Cardiovascular Disease Risk P redict ion Based on Random Forest ”, P roceedings of t he 2nd International Conference on Healthcare Science and Engineering, vol. 536, pp. 31-43, May 2019.

[7] Amin Ul Haq , Jian Ping Li , Muhammad Hammad Memon , Shah Nazir and Ruinan Sun, “ A Hybrid Intelligent System Framework for the Predict ion of Heart Disease Using Machine Learning Algorithms”, Hindawi , Mobile Information Systems, vol. 2018, pp. 1-15, December 2018

[8] Alexandros C. Dimopoulos, Mara Nikolaidou, Francisco Félix Caballero, WorrawatEngchuan, Albert Sanchez-Niubo, Holger Arndt , José Luis Ayuso-Mateos, Josep Maria Haro, Somnath Chat terji, Ekavi N. Georgousopoulou, Christos Pitsavos and Demosthenes B. Panagiotakos, “Machine learning m

[9] Guixia Kang, Bo Yang, Dongli Wei, and Ling Li , “The Applicat ion of Machine Learning Algorithm Applied to 3Hs Risk Assessment ”, Big Data – BigData 2018, pp.169-181, June 2018

[10] Stephen F. Weng, Jenna Reps, Joe Kai, Jonathan M. Garibaldi, Nadeem Qureshi, “Can machine-learning improve cardiovascular risk predict ion using rout ine clinical dat a?”, PLoS One 12(4): e0174944, April, 2017.

[11] Ashok Kumar Dwivedi, “Performance evaluation of different machine learning t echniques for prediction of heart disease”, Neural Computing and Applicat ions, vol. 29, pp. 685–693, September 2016

[12] A. J. Aljaaf, D. Al-Jumeily, A. J. Hussain,T. Dawson, P Fergus and M. Al-Jumaily, “ Predict ingt he Likelihood of Heart Failure wit h a Multi Level Risk Assessment Using Decision T ree”, 2015 Third InternationalConference on Technological Advances in Elect rical, Elect ronics and Computer Engineering, IEEE xplore, pp. 101 - 106, June 2015.

[13] ht tps://www.kaggle.com/sulianova/cardiovascular-disease-dataset