International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.030

Cardio Vascular Disease Prediction Using Multiple Machine Learning Algorithms

VenkataSaiAshrith Kona
Department of Data Science and Business Systems
SRM Institute of Science and Technology
Chennai kv6209@srmist.edu.in

Maithili Saran Reddy Lingala
Department of Data Science and Business Systems
SRM Institute of Science and Technology
Chennai ls2204@srmist.edu.in

Rajasekar P
Department of Data Science and Business Systems
SRM Institute of Science and Technology
Chennai rajasekp@srmist.edu.in

HrudayVuppala
Department of Data Science and Business Systems
SRM Institute of Science and Technology
Chennai hv6819@srmist.edu.in

SravyaAdapa
Department of Data Science and Business Systems
SRM Institute of Science and Technology
Chennai na8385@srmist.edu.in

***Abstract*— This Cardiovascular disease is 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 health business generates enormous amounts of raw data, which data mining transforms into meaningful information that might aid in making decisions. Decision Tree (DT), Adaptive boosting classifier (AdaBoost), Logistic Regression (LR), Random Forest (RF), Gradient Boosting classifier (GBM), and K-Nearest Neighbor (KNN) are the classification methods used in this study.**

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

I. Introduction

The biggest cause of death worldwide, as reported by the 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 difficult to identify (CVD) due to several 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 ways to predict cardiac diseases, but predicting so at a beginning stage is not particularly successful for a variety of reasons, including complexity, execution time, and method accuracy. Consequently, efficacious diagnosis and treatment can save a lot of lives. Between healthcareservice 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 less accessible to families with limited 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). Data Mining plays a major role in 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 results and forecast, Understanding the complexity and non- 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 categorization algorithms are used in data mining of medical data to predict patients' CVD and deaths from heart attack.

II. Literature Survey

[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. Only 12 little-risk and 34 huge-risk patients were examined. To determine whether their suggested method is beneficial, a larger dataset must be examined.

[2] Guidi et al examined the clinical support system (CDSS) for heart failure analysis. This model provided outputs such as HF(Heart Failure) sensitivity . They conducted study using various machine learning classifiers and compared the results. Random forest and CART performed best with 87.6% accuracy out of all classifiers.

[3] 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 of have 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. Al Rahhal*et al* have used a novel approach using deep neural network (DNN) they used raw ECG data to predict using an unsupervised learning technique stacked denoisingautoencoders (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 aquired 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 *et al.* presented the idea of a two-tier ensemble paradigm, where certain classifiers serve as the basis for another ensemble. Using class labels from Extreme Gradient Boosting (EGB), Random Forest (RF), and Gradient Boosting Machine (GBM), the suggested stacking architecture is constructed (XGBoost). Four different types of datasets are used to evaluate their suggested detection model. Moreover, they employed feature selection methods based on particle swarm optimization. With a k value of 10, their suggested model fared better in the k-fold cross- validation. Only the stacking of tree-based models was considered by the authors. Additional statistical and regression-based techniques might be used to improve model results.

[9] Abdar et al. established the N2Genetic optimizer, a novel optimization approach. The patients were then identified as having CHD or not using the nuSVM. On the Z- AlizadehSani dataset, the proposed detection approach had an accuracy of 93.08% when compared to earlier works. Raza proposed an ensemble architecture with majority vote. 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 quality a weight depending on the advice they obtained from 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] Ganna A, Magnusson P K and team. Effort on using machine learning algorithms to identify cardiovascular heart disease has had a substantial effect on this work. In this paper, a summary of the literature is presented. Using a variety of methods, an effective prediction of cardiovascular 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. Proposed System

In this literature we have proposed multiple machine learning algorithms to detect if a person has Cardio vascular disease or 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 classification methods used in this study. 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 14 attributes the above mentioned algorithms are considered to be best for predicting the cardio vascular disease as they are all supervised learning algorithms.

A. *Overview of architecture*

Fig 1 consists of the overall architecture of the cardio vascular disease prediction using multiple ml models and the 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 the patient’s heart condition weather the person has the heart disease or not. And these predictions are made by the best algorithm of all the ml algorithms used and the model is trained before hand with a genuine data set to make accurate predictions.

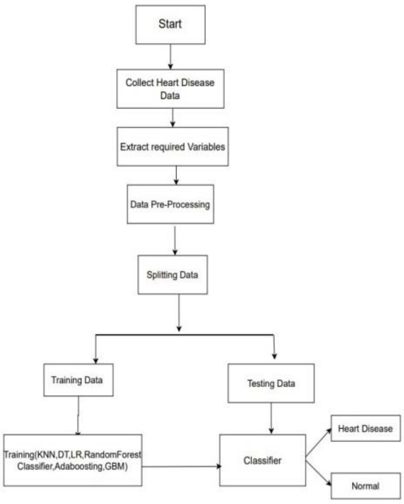

Start
Collect Heart Disease Data
Extract required Variables
Data Pre-Processing
Splitting Data
Training Data
Testing Data
Training(KNN,DT,LR,RandomForest Classifier,Adaboosting,GBM)
Classifier
Heart Disease
Normal

Fig..1. Flowchart for Heart disease prediction

B. Related Work

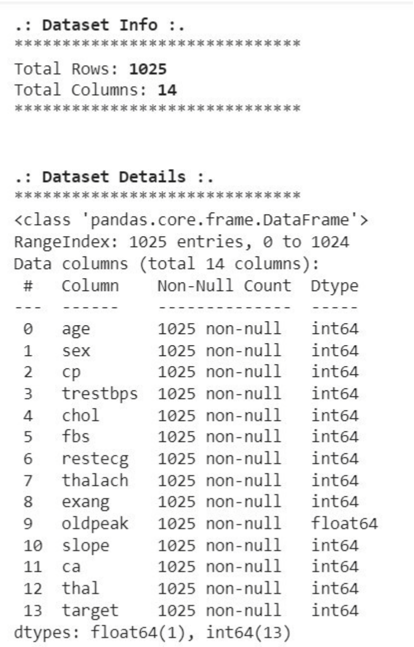
ApurvGarg et al have proposed a model that predicts the chances of getting heart disease using two machine learning algorithms that are KNN and Random forest they have compared these two models in order to get the best accuracy possible out of which KNN yielded a prediction accuracy of 86.88% where as the RF yielded an accuracy of 81.96% [15]

Archanasingh 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 preprocessed and then balanced the data they have used UCI 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 model KNN

C. Data Collection

We took our data from kaggle website for free and our data is called heart.csv this data set contains of 1025 patients records. Out of this 1025 people 499 people are normal and526 people have heart disease and this data set has 14 attributes and out of these people there are 713 male and 312 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. Data values and the attributes

D. Data Pre processing

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 split the data into two parts training and testing out of 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. Machine Leaning Algorithms

A. Logistic Regression (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. Logistic regression is used to describe data and the connection between a dependent variable and one or more independent variables. Nominal, ordinal, or period types are all acceptable for the independent variables. The likelihood that a new observation will fall into a particular class is determined by LR, with the result falling between 0 and 1 because it is a probability

B. Decision Tree (DT)

Decision tree is one of the oldest ml algorithms. For issues related to classification and regression we 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 generally makes up decisions as it has many branches whereas leaf node can’t make any decisions.

C. K-Nearest Neighbor (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, related situations are classified similarly, and a new 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 methods are offered to translate the distance into a weight 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 created via AdaBoost, and all the data points are given the same weight. After that, it gives points with incorrect classifications more weight. The following model now accords greater relevance to all of the points with higher weights. As long as no low errors are received, it will continue training models.

E. Random Forest

A ML technique that uses many numerous decision trees to make a decision is known as Random forest. It is a ensemble learning based technique. While it is in the training stage, itProduces many trees and a forest of decision trees. Each and every tree, a component of the forest, 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. Gradient Boosting

Using boosting, weak learners may become strong learners. Each new tree created by boosting is fitted to a modified version of the initial data set. It is anticipated that when merged with older models, the new model will produce forecasts with lower error rates. The major goal is to set objectives for this next model to reduce mistakes. Gradient Boosting in a gradual, additive, and linear fashion trains many models. Because each case's goal results are decided by the gradient's deviation relative to the predictions, the phrase "gradient boosting" came into popularity. Every model picks up speed in the correct way by reducing the prediction errors.

E. Proposed Algorithm

In this study the best out of all the algorithms is KNN which has achieved an accuracy of 97% which is considered as one of the best algorithm in supervised classification algorithm and other than that it is simple KNN is non-parametric and lazy, which means it does not assume anything about the distribution of the underlying data anddoes not create a model from the training set. As an alternative, it memorises the full training dataset and utilises it to make predictions when presented with fresh test cases. For many applications, KNN is a straightforward and efficient method, although it can have large computing costs and be sensitive to the choice of K and the distance metric used to compare 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 approaches for performance evaluation. It is anticipated that the blending of several assessment tools will support the advancement of analytical research. Four fundamental measures (accuracy, precision, recall, and F-Score) will be 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), True Negative (TN), False Positive (FP), and False Negative (FN).In the medical data the most important thing is to find out (FN). The performance metrics are provided below

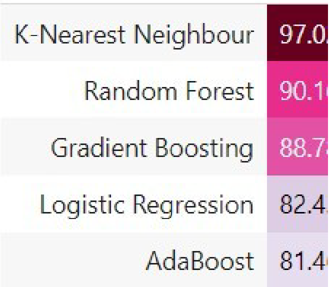
$$\text{Accuracy} = \text{Number of correctly classified predictions} / \text{Total predictions} \tag{1}$$

$$\text{Precision} = TP / TP + FP \tag{2}$$

$$\text{Recall} = TP / TP + FN \tag{3}$$

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

The total collection of features in the heart disease dataset have been exposed to comparison analysis of supervised machine learning classifiers. Some classifiers performed well on evaluation measures, whereas others did not. In order to predict heart failure survival, this work employed tree-based, statistical-based and regression-based models. The DT, RF ensemble models are tree-based. AdaBoost and GBM are two tree-based boosting methods. Statistically-based models whereas regression-based models include LR and KNN



K-Nearest Neighbour	97.0
Random Forest	90.1
Gradient Boosting	88.7
Logistic Regression	82.4
AdaBoost	81.4

Fig. 3. Different accuracy comparison

As per the table we have KNN with the best accuracy of 97.02%, Random forest with an accuracy of 90.16%, Gradint boosting with a accuracy of 88.7%, LR with a accuracy of 82.43%,Adaboost with a accuracy of 81.46%and with the least accuracy is the decision tree algorithm with an accuracy of 79%

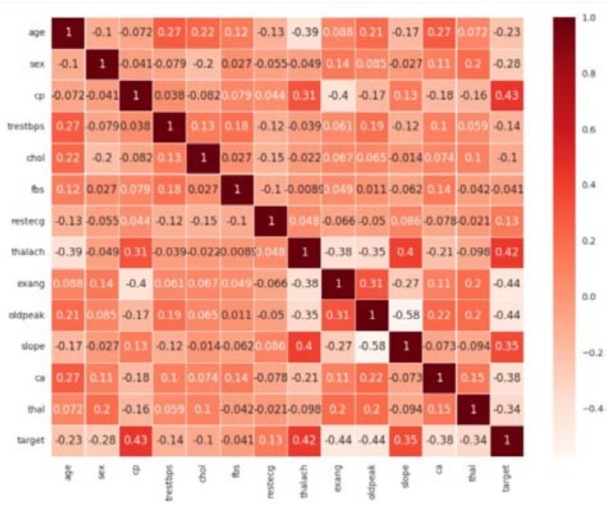


Fig. 4. Correlation matrix of variables

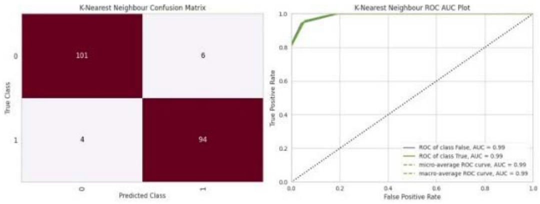


Fig. 5. Roc and confusion matrix of KNN the best algorithm

TABLE.1. PRECISION, RECALL AND F-MEASURES

Algorithm	Precision	Recall	F-1
K Nearest 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
Ada boost	0.82	0.81	0.81
Decision Tree	0.80	0.79	0.79

TABLE.2. VALUE OF AREA UNDER ROC

Algorithm	AUROC
K Nearest Neighbor	0.99
Random forest	0.96
Gradient Boosting	0.95
Logistic Regression	0.91
Ada boost	0.87
Decision Tree	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 machine learning algorithms Ada boost, LR, Gradient boost, KNN, DT, and RF we assessed the predictions using the F-1 Measure, ROC curve, recall, accuracy, and precision. We discovered that all of the used algorithms performed well, with KNN demonstrating the greatest performance with 97% accuracy, showing that these algorithms are the most effective at predicting cardiac disease.

VI. CONCLUSION

Heart patients' lives will be saved through the processing 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 machine learning methods are used: LR, AdaBoost, RF, GBM, DT and KNN. KNN with a accuracy of 97% the highest of all algorithms with precision score0.97 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 forecast how long a patient with heart failure 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 variability", IEEE J. Biomed. Health Informat., vol. 17, no. 3, pp. 727-733, May 2013.

[2] G. Guidi, M. C. Pettenati, P. Melillo and E. Iadanza, "A machine learning system to improve heart failure patient assistance", IEEE J. Biomed. Health Informat., vol. 18, no. 6, pp. 1750-1756, Nov. 2014.

[3] G. Parthiban and S. K. Srivatsa, "Applying machine learning methods in 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 and R.R. Yager, "Deep learning approach for active classification of electrocardiogram signals", Inf. Sci., vol. 345, pp. 340-354, Jun. 2016

[5] S. Muthukaruppan and M. J. Er, "A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease", Expert Syst. Appl., vol. 39, no. 14, pp. 11657-11665, Oct. 2012.

[6] Z. Sani, R. Alizadehsani, J. Habibi, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, et al., "Diagnosing coronary artery disease via data mining algorithms by considering laboratory and echocardiography features", 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 detection using computational intelligence methods", Knowl.-Based Syst., vol. 109, pp. 187-197, Oct. 2016.4

[8] 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.

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

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

[11] Sitharthan, R., Vimal, S., Verma, A., Karthikeyan, M., Dhanabalan, S. S., Prabaharan, N., ...&Eswaran, T. (2023). Smart microgrid with the internet of things for adequate energy management and analysis. Computers and Electrical Engineering, 106, 108556.

[12] A. Ganna, P. K. Magnusson, N. L. Pedersen, U. de Faire, M. Reilly, J. Ärnlövand E. Ingelsson,"Multilocus genetic risk scores for coronary heart disease prediction," Arteriosclerosis, thrombosis, and vascular biology, vol. 33, no. 9, pp. 2267-7, 2013.

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

[14] Syed Nawaz Pasha, Dadi Ramesh, SallauddinMohmmad, A. Harshavardhan and Shabana "Cardiovascular disease prediction using deep learning techniques "OP Conference Series: Materials Science and Engineering, Volume 981, International Conference on Recent Advancements in Engineering and Management (ICRAEM-2020) 9-10 October 2020, Warangal, India Citation Syed Nawaz Pasha et al 2020 IOP Conf. Ser.: Mater. Sci. Eng. 981 022006

[15] Garg, Apurvand Sharma, Bhartenduand Khan, Rizwan. (2021). Heart disease prediction using machine learning techniques. IOP Conference Series: Materials Science and Engineering. 1022. 012046. 10.1088/1757-899X/1022/1/012046

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