Identification of Cardiovascular Disease Patients

Tavleen K. Nagi [1], Abhishek Tomar [2], Deepanshi Jain [3], Surinder Kaur [4,*]

[1,2,3,4] Bharati Vidyapeeth's College of Engineering, GGSIPU, Delhi, INDIA
Emails: tavleennagi15@gmail.com; tomar4349@gmail.com; jaindeepanshi04@gmail.com; kaur.surinder@bharatividyapeeth.edu

* Correspondence: kaur.surinder@bharatividyapeeth.edu

Abstract

For the prevention and treatment of illness, accurate and timely investigation of any health-related problem is critical. The prevalence of cardiovascular illnesses is rising among Indians. Aging has long been recognized as one of the most significant risk factors for heart attacks, affecting men and women aged 50 and up. Cardiovascular attacks are increasingly becoming more common in people in their 20s, 30s, and 40s.. To detect and predict cardiovascular disease patients, starting with a pre-processing step in which we used feature selection to pick the most important features, we tested the accuracy of different models on a dataset with features like gender, age, blood pressure, and glucose levels. The model predicts whether a patient is likely to suffer from cardiovascular disease based on their medical records. Finally, we performed hyperparameter tuning to find the best parameter for the models. In comparison to the other algorithms, the XGBoost model produced the best results with an accuracy of 75.72%

Keywords: Cardiovascular disease; Machine Learning; Disease Prediction

1.Introduction

With 1.3 billion people living in cities and villages across the country, India presents a unique healthcare problem. In order for India to achieve its aim of universal health coverage (UHC), technology and healthcare must work together effortlessly. Cardiovascular diseases (CVDs), commonly known as heart diseases, are a set of heart and blood vessel abnormalities. Atherosclerosis (blockages in coronary arteries), the most frequent cause of CVD, is mostly brought on by lifestyle factors. According to WHO, bad nutrition, physical inactivity, smoking, and problematic alcohol consumption are the most significant behavioral risk factors for heart disease and stroke. Individuals may experience elevated blood pressure, elevated blood glucose, elevated blood lipids, as well as overweight and obesity as a result of behavioral risk factors. Adults today suffer from a wide range of chronic illnesses that limit their independence and damage their health. The prevalence of cardiovascular illnesses is rising among Indians. According to WHO country-level statistics on non-communicable diseases, NCDs account for 53% of all deaths in India, with CVDs accounting for 24% of all fatalities [1]. This is frequently the result of a lack of timely health checkups and poor lifestyle choices. Early detection of abnormal variations in health indices could lead to earlier detection of chronic diseases and, as a result, better medical decision-making and planning. Aging has long been recognized as one of the most significant risk factors for heart attacks, affecting men and women aged 50 and up. Cardiovascular attacks are increasingly becoming more common in people in their 20s, 30s, and 40s. This research revolves around the detection of whether the person is likely to suffer from cardiovascular diseases according to their medical records.

Doi : https://doi.org/10.54216/FPA.100101
Received: May 10, 2022 Accepted: October 14, 2022

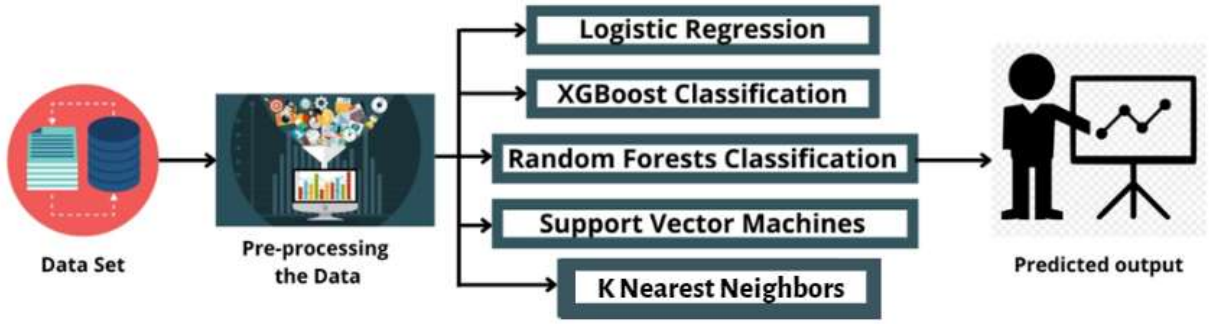

Figure 1: Research Project Roadmap

This is the diagrammatic representation of the working of our research. After the raw dataset was pre-processed data was passed through six different models and accuracy was calculated. Further hyper-parameter tuning was performed to give an accuracy of about 75.72% in XGBoost and 75.56% in Random Forest.

2. 2. Literature review

• Currently, there are a variety of chronic conditions that limit senior citizens' freedom and damage their health [2]. This frequently occurs as a result of their failure to undergo prompt health assessments. Most of the time, this happens because elders try to hide their issues or even their fear of being institutionalized; any change is considered a normal part of becoming older. [3]. Early detection of abnormal variations in health indices may lead to earlier detection of chronic diseases and, as a result, better medical decision-making and planning.

• Heart disease has been one of the two top causes of mortality since 1975, accounting for 633,842 deaths, or 1 in every 4 deaths. Cancer fatalities (595,930) were the second-highest cause of death in 2015. [4]. The burden of CVD further extends as it is considered the costliest disease even ahead of Alzheimer's disease and diabetes with calculated indirect costs of $237 billion dollars per year and a projected increase to $368 billion by 2035.[5]

• Researchers in 2021, used the correlation matrix to choose the most important features, and then we used three data analytics approaches (neural networks, SVM, and KNN) on data sets of various sizes. Using numerous machine learning techniques, they created a disease prediction system. There were about 230 diseases in the dataset that needed to be processed.[6] In comparison to the other algorithms, the weighted KNN method produced the best results. The prediction accuracy of the weighted KNN method was 93.5 percent.

• Logistic Regression, Random Forest Classifier, and KNN are the algorithms utilized to create the provided model, the accuracy of the model was 87.5%. The accuracy of KNN, which is 88.52%, was the highest of the three algorithms that they used [7].

• Researchers also identified the many data mining strategies that can be used to effectively forecast cardiac disease. The information was pre-processed before being utilized in the model. The most efficient algorithms are Random Forest (86.89 percent) and XGBoost (78.69 percent) [8].

• Research conducted in 2016, there were 346,201 records in all, which corresponded to 346,201 patients. In this study, five AI algorithms were built and analyzed, including four standard machine learning methods (logistic regression [LR], random forest [RF], extra trees [ET], and gradient boosting trees [GBT]) and a deep learning algorithm (a densely connected neural network [DNN]). In terms of model performance, GBT and RF had the highest area under the receiver operating characteristic curve (97.8% and 97.7%, respectively) in terms of discrimination, followed by Extra Trees (96.8%) and Logistic Regression (96.4%), and DNN had the least discriminative model (95.3%) [9].

Doi : https://doi.org/10.54216/FPA.100101
Received: May 10, 2022 Accepted: October 14, 2022

• The proposed study project considered 14 important factors. In this work, machine learning techniques including Random Forest (RF), Logistic Regression, Support Vector Machine (SVM), and Nave Bayes are compared for the classification of cardiovascular disease. Random Forest, a machine learning algorithm, is used in the suggested fix since it has been shown to be the most reliable and accurate algorithm in comparison (84.41 percent). [10].

3. Methodology

3.1 Dataset

The dataset comprises values collected at the moment of the medical examination of a patient. There were three types of input variables given in the dataset:

- The objective feature that gives factual information;
- The examination feature gives the results of the medical examination;
- The subjective feature gives the information given by the patient.

Table 1: Dataset and Features

S. No.	FEATURES	TYPE OF VARIABLE
1	Age	Objective Feature
2	Height	Objective Feature
3	Weight	Objective Feature
4	Gender	Objective Feature
5	Systolic blood pressure	Examination Feature
6	Diastolic blood pressure	Examination Feature
7	Cholesterol	Examination Feature
8	Glucose	Examination Feature
9	Smoking	Subjective Feature
10	Alcohol intake	Subjective Feature
11	Physical activity	Subjective Feature
12	Presence or absence of cardiovascular disease	Target Variable

3.2 Data Pre-Processing

Data pre-processing is a method for converting unclean data into a clean data set. The first step was to check whether the dataset had any null values, fortunately, the dataset didn't have any. Then we had to delete these outliers and filter the dataset. The data rows with blood pressure higher than 400 and negative blood pressure values were deleted. The values of low blood pressure greater than 300 were also deleted. The features such as high blood pressure had seemingly large and unrealistic values or even negative numbers. Further, one new column was created called BMI. Since our dataset had variables that are label encoding, i.e., each label is issued a unique integer. Labels are given according to alphabetical order, quite often, the machine interprets the relationship as 1<2<3, where 1, 2,3 are labels for categories. This might be misleading therefore; we perform one hot encoding. One hot encoding is done as follows, based on the number of unique values in the categorized characteristic, which provides extra

Doi : https://doi.org/10.54216/FPA.100101
Received: May 10, 2022 Accepted: October 14, 2022

features. Every category's unique value will be added as a feature. Feature selection is the process of minimizing the number of input variables while relating a predictive model. Model performance can be harmed by features that are irrelevant or only partially relevant. Therefore, we rejected all features which had p-values less than 0.05. Finally, we standardized the dataset. Standardization refers to when the features are being rescaled to ensure the mean and the standard deviation are 0 and 1, respectively.

3.3 Models

3.3.1 Logistic Regression

This type of statistical analysis usually referred to as a logit model, is frequently used for machine learning applications as well as predictive analytics and modeling. In this analytics technique, the dependent variable is either binary (A or B) or categorical (A, B, C, or D), or it can be a range of finite possibilities (multinomial regression). Instead of using probabilities to predict group membership, LR uses the log odds ratio, and to fit the final model, iterative maximum likelihood is used instead of least squares. A logistic regression equation is used in statistical software to understand the relationship between the dependent variable and one or more independent variables by calculating probabilities.

3.3.2 K Nearest Neighbors

The k-nearest neighbor's algorithm, also referred to as KNN or k-NN, is a supervised learning classifier that generates predictions or classifications about the grouping of individual data points based on proximity. It can be applied to classification and regression problems, although it is most frequently used as a classification technique because it is based on the assumption that similar points can be found close together. KNN determines how far away every point is from the unknown data and then eliminates those with the closest distances.

3.3.3 Random Forests Classification

A popular supervised machine learning technique for solving classification and regression issues is Random Forest. Using the average for regression and the majority vote for classification, it builds decision trees from a variety of samples. The Random Forest Algorithm's ability to handle data sets with both continuous and categorical variables, as in regression and classification, is one of its key features. Multiple decision trees serve as the fundamental learning models in Random Forest. The fundamental idea is to integrate many decision trees rather than relying just on one decision tree to determine the outcome. We create sample datasets for each model by randomly selecting rows and features from the dataset. This component is known as Bootstrap.

3.3.4 Xgboost Classification

Extreme Gradient Boosting or XGBOOST contains parallel tree boosting and is the best machine learning tool for tasks including regression, classification, and ranking. Gradient boosting, decision trees, ensemble learning, and supervised machine learning are the machine learning concepts and techniques that must first be understood in order to fully comprehend XGBoost. Weights are highly important in XGBoost. Weights are assigned to each independent variable, which is then used to inform the decision tree's prediction of results. The second decision tree is supplied with the variables that the first decision tree mistakenly anticipated and its weight is increased.

3.3.5 Support Vector Machine

A supervised machine learning method called SVM can be applied to both classification and regression problems. Even though we could also point out problems with regression, categorization fits the bill the best. Finding a hyperplane in an N-dimensional space that categorizes data points clearly is the aim of the SVM method. The quantity of features determines the size of the hyperplane. The hyperplane is essentially a line if the input qualities are limited to only two. The hyperplane transforms into a two-dimensional plane when there are three input features. When there are more than three features, it is impossible to imagine.

Doi : https://doi.org/10.54216/FPA.100101
Received: May 10, 2022 Accepted: October 14, 2022

4. Result

A graph demonstrating the effectiveness of a classification model at every level of categorization is called a receiver operating characteristic curve (ROC curve). Parameters that are plotted on this curve:

- True Positive Rate

$$TPR = \frac{TP}{TP + FN}$$

Where TPR stands for a true positive rate

- False Positive Rate

$$FPR = \frac{FP}{FP + TN}$$

Where FPR stands for a false positive rate

"Area under the ROC Curve" is the abbreviation for "Area under the ROC Curve." AUC, in other words, assesses the complete two-dimensional area beneath the entire ROC curve from (0,0) to (1,1), using integral calculus. The closer a ROC curve is to the top left corner of the figure, the better the model is in categorizing the data. To quantify this, we may use the AUC (area under the curve) formula, which tells us how much of the plot lies beneath the curve. For the supplied input dataset, various machine-learning models were applied to investigate disease prediction. For the prediction, we used 5 different machine-learning models**.**

Table 2: Accuracy of all models after hyper-parameter tuning on test data

Model	**Accuracy**
Logistic Regression	74.61%
Random Forest	75.56%
XGBOOST	75.72%
Support Vector Machine (SVM)	75.22%
K Nearest Neighbors	73.06%

Doi : https://doi.org/10.54216/FPA.100101
Received: May 10, 2022 Accepted: October 14, 2022

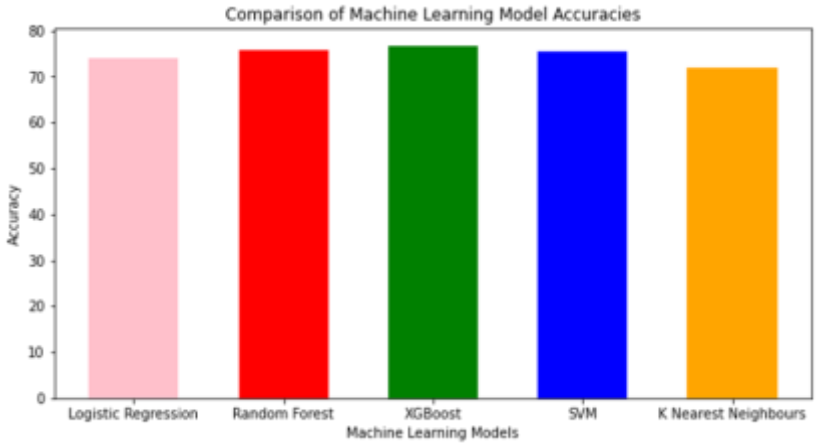

Figure 2: Comparison of ML Model Accuracies

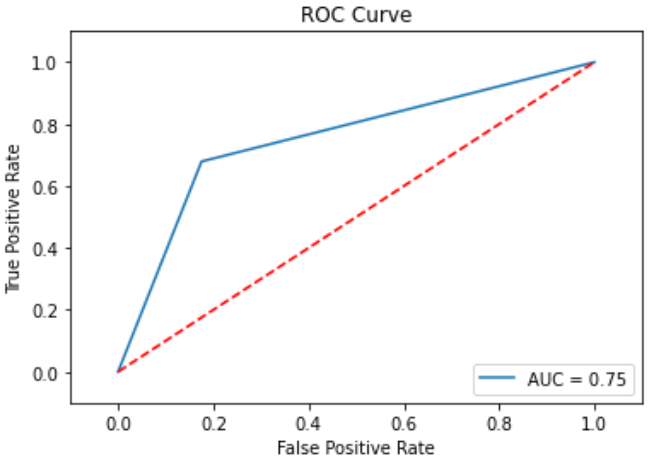

Figure 3: Logistic Regression ROC-AUC Curve

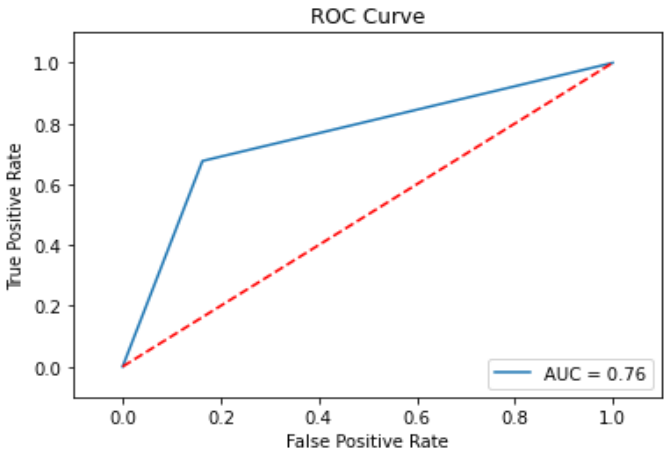

Figure 4: XGBoost ROC-AUC Curve

Doi : https://doi.org/10.54216/FPA.100101
Received: May 10, 2022 Accepted: October 14, 2022

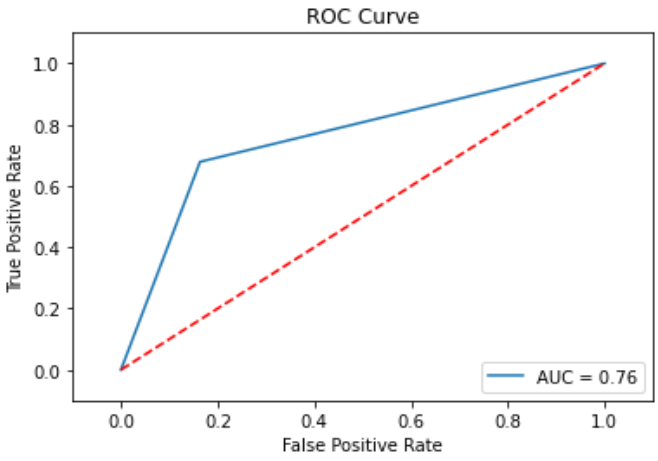

Figure 5: Random Forest ROC-AUC Curve

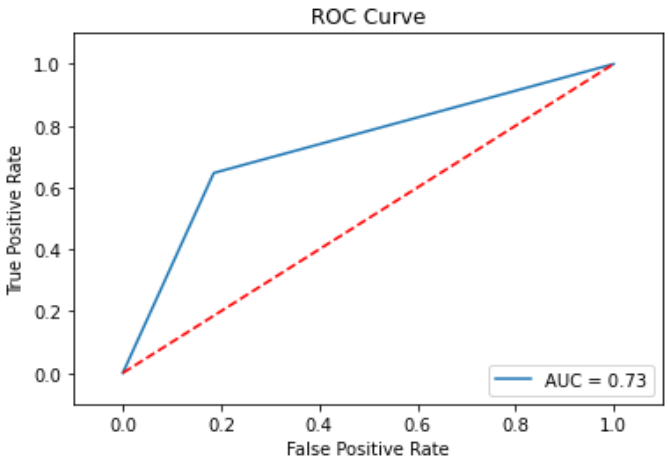

Figure 6: KNN ROC-AUC Curve

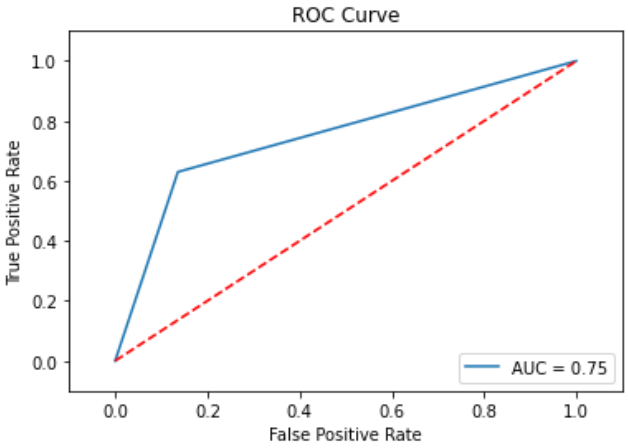

Figure 7: SVC ROC-AUC Curve

We were able to get 70% or higher accuracy for all our models. According to the research (Fig 4.3), the best-performing model is XGBoost, which uses an implementation of gradient-boosted decision trees designed for speed and performance with 75.72%. The other model that performed well was Random Forest Classification (Fig 4.4) model with accuracies 75.56%. K Nearest Neighbors (Fig 4.5) gave the worst result, therefore clustering wouldn't work in this form of classification. Also, the most important feature (Fig 4.7) or the feature that has the highest significance in the model is ap_hi or high blood pressure value. Patients with high blood and high cholesterol have higher chances of cardiovascular diseases. It is also interesting to notice that gender has no or minimal relationship with cardiovascular diseases in middle-aged adults.

Doi : https://doi.org/10.54216/FPA.100101
Received: May 10, 2022 Accepted: October 14, 2022

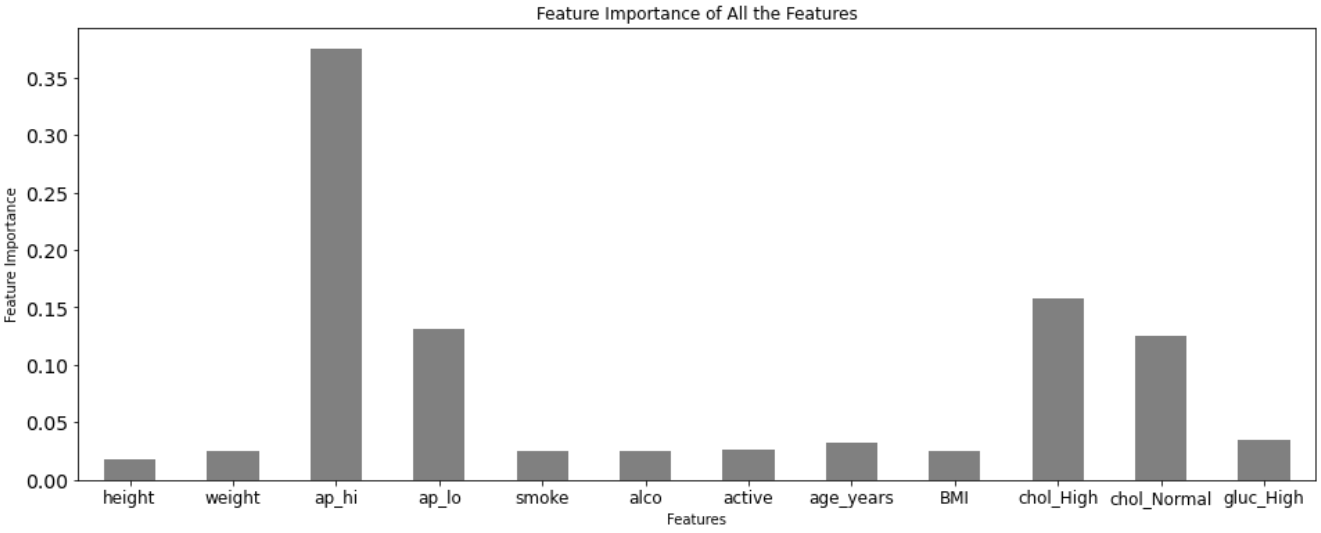

Figure 8: Feature Importance of XGBoost ML Model

5. Future Scope

The current model takes into consideration 12 attributes to predict whether a patient suffers from cardiovascular disease, we can perform further feature selection based on different metrics like domain knowledge or graphical visualization. We can use ensemble methods to increase model accuracy. Ensemble methods are distinct from general modeling techniques that model the data using a variety of weak models and then integrate the results. The reason for being more accurate is the results are combined. Cross-validation is a technique that can be used to improve model performance. It works best when there is an issue with overfitting the model during the modeling process. There are various techniques of cross-validation such as K-fold, leaving one group out, leaving P groups out, etc.

6. Conclusion

People are becoming increasingly afflicted with heart disease, especially middle-aged adults. As a result, predicting the disease before infection lowers the risk of death. This is a well-researched prediction. In this study, we predicted cardiovascular disease based on a patient's medical records like gender, age, high blood pressure, low blood pressure, glucose levels, alcohol intake, and smoking history. For disease prediction utilizing the above-mentioned parameters, the XGBoost model had a maximum accuracy of 76.67% percent. Almost all of the machine learning models produced accurate results. We could easily manage the medical resources required for treatment once the condition is predicted. This methodology would help to reduce the costs associated with treating the disease while also improving the recovery process.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] Cardiovascular Disease Article https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Antonis S. Billis and Panagiotis D. Bamidis, "Employing time-series forecasting to historical medical data: an application towards early prognosis within elderly health monitoring environments", AIAM'14: Proceedings of the 3rd International Conference on Artificial Intelligence and Assistive Medicine - Volume 1213, August 2014; ISBN:16130073

Doi : https://doi.org/10.54216/FPA.100101
Received: May 10, 2022 Accepted: October 14, 2022

[3] Hayes TL, Pavel M, Kaye JA, "An unobtrusive in-home monitoring system for detection of key motor changes preceding cognitive decline". Proc. of the 26th Annual Intl. Conf. of the IEEE EMBS, pp. 2480-2483, San Francisco, CA. (2004)

[4] Dunbar SB, Khavjou OA, Bakas T, Hunt G, Kirch RA, Leib AR, Morrison RS, Poehler DC, Roger VL, Whitsel LP., American Heart Association. "Projected Costs of Informal Caregiving for Cardiovascular Disease: 2015 to 2035: A Policy Statement From the American Heart Association. Circulation". 2018 May 08;137(19):e558-e577.

[5] T. Hayes, M. Pavel, and J. Kaye, "An approach for deriving continuous health assessment indicators from in-home sensor data. In Technology and aging: Selected papers from the 2007 international conference on technology and aging", vol. 21, pp. 130–137, 2008

[6] Keniya, Rinkal and Khakharia, Aman and Shah, Vruddhi and Gada, Vrushabh and Manjalkar, Ruchi and Thaker, Tirth and Warang, Mahesh and Mehendale, Ninad and Mehendale, Ninad, "Disease Prediction From Various Symptoms Using Machine Learning" (July 27, 2020). Available at SSRN: https://ssrn.com/abstract=3661426 or http://dx.doi.org/10.2139/ssrn.3661426

[7] Harshit Jindal et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1022 012072

[8] Pooja Anbuselvan, "Heart Disease Prediction using Machine Learning Techniques", Student Bangalore Institute of Technology Bengaluru, Karnataka, India. International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181, Published by: Vol. 9 Issue 11, November 2020

[9] Tran L, Chi L, Bonti A, Abdelrazek M, Chen Y, "Mortality Prediction of Patients With Cardiovascular Disease Using Medical Claims Data Under Artificial Intelligence Architectures: Validation Study", JMIR Med Inform 2021;9(4):e25000

[10] Rubini P. E., Dr. C. A. Subasini, Dr. A. Vanitha Katharine, V. Kumaresan, S. Gowdham Kumar, T. M. Nithya. (2021). "A Cardiovascular Disease Prediction using Machine Learning Algorithms", Annals of the Romanian Society for Cell Biology, 25(2), 904–912.

Doi : https://doi.org/10.54216/FPA.100101
Received: May 10, 2022 Accepted: October 14, 2022