EAI Endorsed Transactions
on Internet of Things
Research Article EAI.EU

Early Detection of Cardiovascular Disease with Different Machine Learning Approaches

Eyashita Singh[1*], Vartika Singh[2], Aryan Rai[3], Ivan Christopher[4], Raj Mishra[5] and K.S. Arikumar [6]

[1, 2, 3, 4, 5,6]School of Computer Science & Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India

Abstract

With the increase in mortality rate around the world in recent years, cardiovascular diseases (CVD) have swiftly become a leading cause of morbidity, and therefore there arises a need for early diagnosis of disease to ensure effective treatment. With machine learning emerging as a promising tool for the detection, this study aims to propose and compare various algorithms for the detection of CVD via several evaluation metrics including accuracy, precision, F1 score, and recall. ML has the ability and potential to improve CVD prediction, detection, and treatment by analysis of patient information and identification of patterns that may be difficult for humans to interpret and detect. Several state-of-the-art ML and DL models such as Decision Tree, XGBoost, KNN, and ANN were employed. The results of these models reflect the potential of Machine Learning in the detection of CVD detection and subsequently highlight the need for their integration into clinical practice along with the suggestion of the development of robust and accurate models to improve the predictions. This integration, however, significantly helps in the reduction of the burden of CVD on healthcare systems.

Keywords: Cardiovascular disease, Early Detection, Machine Learning, Deep Learning, Healthcare

Received on 21 December 2023, accepted on 02 March 2024, published on 12 March 2024

doi: 10.4108/eetiot.5389

*Corresponding author. Email: eyashitasingh10@gmail.com

1. Introduction

Today, cardiovascular diseases have become a primary reason for mortality across the globe. According to the World Health Organization's report, roughly 17.9 million people lost their lives to CVDs in 2019, which represent 32% of total global deaths. 85% of these deaths were due to heart attacks and strokes [1]. The prevention of most cardiovascular diseases can be achieved by tackling behavioral risk factors, which include inadequate physical activity, tobacco consumption, an unhealthy diet leading to obesity, and excessive alcohol consumption. Due to these reasons, young people have started facing issues like obesity, high cholesterol, and blood pressure levels resulting in premature heart failure and death [2]. Therefore, it is crucial to detect the early symptoms of CVD to provide early medication to prevent casualties. Machine learning has emerged as a significant technological advancement in the field of healthcare, presenting the potential to revolutionize the entire sector and bring substantial benefits to both patients and healthcare providers. Machine Learning is an analytical approach that automates the construction of models by leveraging algorithms. It enables the extraction of concealed insights from data. The progressive process of machine learning empowers the system to modify its techniques and outcomes in response to unfamiliar circumstances and new data it encounters. [3]. Major applications of machine learning in healthcare include Personalizing treatment, detecting diseases in their early stages, Robot-assisted surgery [4], analyzing errors in prescriptions [5], assisting in clinical research and trials [6], Drug discovery and creation [7], Automating image diagnosis [8], etc. In our approach, we aim to build and train a machine learning model to detect the onset of cardiovascular malfunction, using algorithms like ANN (Artificial Neural Network), KNN (K-Nearest Neighbors Algorithm), Decision Tree, and XGBoost. All the data needed to train and test our model has been fetched from

Kaggle. These models can be applied to assess whether an individual exhibits indication of a cardiovascular malfunction, taking into account attributes such as Age, Height, Weight, Gender, Systolic blood pressure, Cholesterol levels, Diastolic blood pressure, Glucose, Smoking, Alcohol Intake, Physical activity, and the presence or absence of cardiovascular disease [9]. In this paper, a literature review was conducted based on heart-related diseases and their causes [10]. In the next section, the methodology is presented in which an analysis of the data was done to predict the presence or absence of cardiovascular disease in patients using ML [11]. In the subsequent section, we delve into the outcomes and present them. Ultimately, in the final section, we showcase our findings and provide suggestions for future research.

2. Literature Review

Machine learning has lately been emerging as a promising tool for the detection of CVD. One study which explored the use of ML in CVD is a paper by Maini E. et al. [9] in which they proposed an unsupervised approach in terms of clustering techniques to develop a variety of models. Modepalli et al. [12] employed a distinctive approach by utilizing a combination of Random Forest, Decision Tree, and Hybrid Model methods, achieving an accuracy rate of 88.7%. Bharti R. et al. [13] conducted a comparative examination of multiple Machine Learning (ML) and Deep Learning (DL) models in relation to the Archive Coronary Heart Disease dataset. On the other hand, Ashish et al. [14] developed a rapid and accurate computerized system for coronary heart disease detection using SVM classification and XGBoost boosting algorithms. This is evidence that the current research suggests that ML algorithms have great potential for the detection, diagnosis, and risk prediction of CVD [15]. However, the need for large and diverse datasets remains an open challenge to optimize ML algorithms and integrate the same into clinical practice [16]. It is still stated that the use of ML algorithms for CVD detection is a rapidly evolving field of research that has great potential for improving patient outcomes via early diagnosis and appropriate treatment [17].

3. Methodology

In this paper, we have predicted whether a patient is suffering from cardiovascular disease or not based on various attributes. This dataset was taken from Kaggle and the dataset values were taken at the moment of medical examination. Through this dataset, we aim to classify the patient as healthy or suffering from cardiovascular disease. Our dataset consists of 69,301 patients and 13 health attributes.

Table 1. Description of Dataset

Parameter	*Description*
age	Age of the individual in days.
height	Height of the individual in centimetres.
weight	Weight of the individual in kilograms.
gender	Gender of the individual (1 for male or 2 for female).
ap_hi	Systolic blood pressure of the individual.
ap_lo	Diastolic blood pressure of the individual.
cholesterol	The cholesterol level of the individual (1 for normal, 2 for above normal, 3 for well above normal).
gluc	The glucose level of the individual (1 for normal, 2 for above normal, 3 for well above normal).
smoke	Smoking status of the individual (1 for yes, 0 for no).
alco	Alcohol intake by the individual (1 for yes, 0 for no).
active	Physical activity performed by the individual (1 for yes, 0 for no).

3.1. Data Preparation and Analysis

On this dataset, we performed exploratory data analysis (EDA). EDA is a preprocessing step used for a better understanding of data. Exploratory data analysis was performed using pandas profiling in a Jupyter notebook.

3.2. Libraries and Packages

```
#importing required libraries

import os
import plotly.graph_objects as go
import pandas as pd
import numpy as np
from pandas.plotting import scatter_matrix
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels
import missingno as msno
import plotly.graph_objects as go
import plotly.express as px
import torch
import sklearn
import torch.nn as nn
import numpy as np
import torch.optim as optim
from torch.utils.data import Dataset, DataLoader
from tqdm import tqdm
import torch.functional as F
import matplotlib.pyplot as plt
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.naive_bayes import GaussianNB
from sklearn.neighbors import KNeighborsClassifier
from tqdm import tqdm
from sklearn.metrics import precision_score, recall_score, f1_score, confusion_matrix,plot_roc_curve
from sklearn import tree
```

Fig. 1. Packages imported

3.3 Data Info

This step provides the snippet of our dataset and the attributes contained by the dataset.

```
<bound method DataFrame.info of          id    age  gender  height  weight  ap_hi  ap_lo  cholesterol  gluc  \
0       988  22469       1     155    69.0    130     80            2     2
1       989  14648       1     163    71.0    110     70            1     1
2       990  21901       1     165    70.0    120     80            1     1
3       991  14549       2     165    85.0    120     80            1     1
4       992  23393       1     155    62.0    120     80            1     1
...     ...    ...     ...     ...     ...    ...    ...          ...   ...
69296  99993  19240       2     168    76.0    120     80            1     1
69297  99995  22601       1     158   126.0    140     90            2     2
69298  99996  19066       2     183   105.0    180     90            3     1
69299  99998  22431       1     163    72.0    135     80            1     2
69300  99999  20540       1     170    72.0    120     80            2     1

       smoke  alco  active  cardio
0          0     0       1       0
1          0     0       1       1
2          0     0       1       0
3          1     1       1       0
4          0     0       1       0
...      ...   ...     ...     ...
69296      1     0       1       0
69297      0     0       1       1
69298      0     1       0       1
69299      0     0       0       1
69300      0     0       1       0

[69301 rows x 13 columns]>
```

Fig. 2. Overview of the dataset

3.4. Checking for null values

The first step in EDA is importing the dataset and cleaning it. We checked for the presence of null values, duplicate values, and missing values. Figure 1 shows that there are no null values present in our dataset. After that, we performed a statistical summary analysis of our dataset followed by the detection of outliers. For the features age, height, weight, systolic blood pressure, and diastolic blood pressure; the outliers count was determined as 4, 515, 1802, 1419, and 4584 respectively. These outliers were then removed from the dataset.

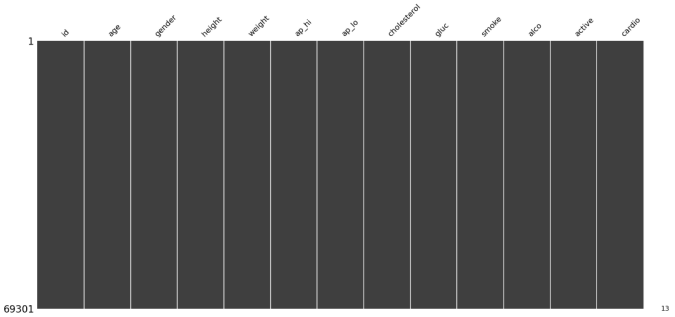

Fig. 3. Cleaned Dataset

3.5 Distribution of data

Figure 4 shows the gender distribution according to the target variable using a stacked bar plot. Males have a count of 20,000 whereas the count value for females i.e., approx. 10,000 is a lot less compared to males.

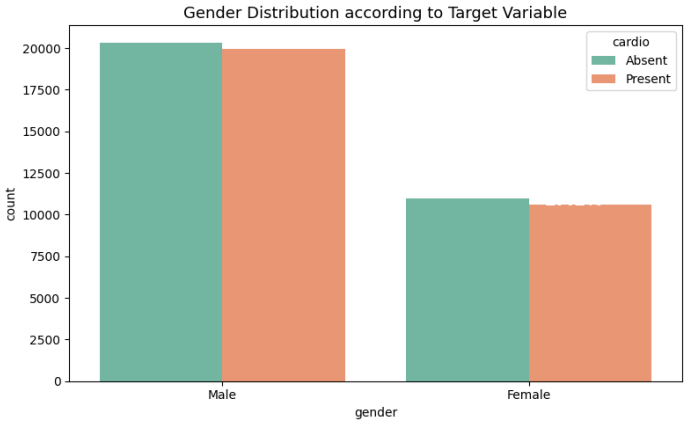

Fig. 4. Packages imported

Figure 5 shows the cholesterol distribution according to the target variable using a bar plot. Patients with above normal and well above normal cholesterol have more chances of having cardiovascular disease than patients having normal cholesterol levels.

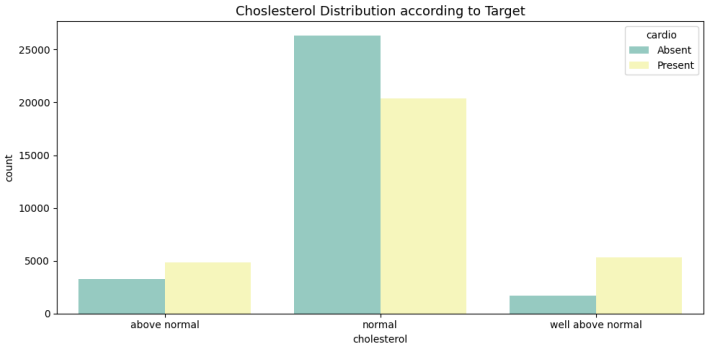

Fig. 5. Cholesterol Distribution according to the target

Figure 6 shows the glucose distribution according to the target variable using a bar plot. Patients with above-normal and well-above-normal glucose levels exhibit a higher propensity for cardiovascular disease compared to patients with normal glucose levels.

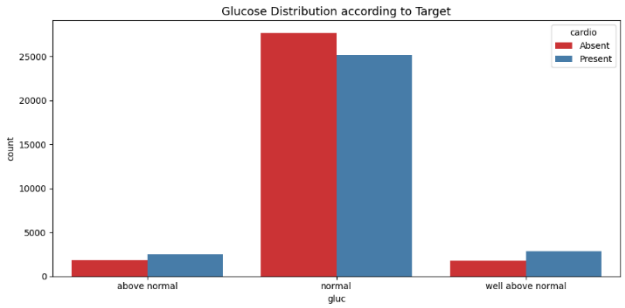

Fig. 6. Glucose Distribution according to target

3.6 Heat Map

The Heat map in this analysis visually represents the intensity of relationships between variables using color gradients. It reveals that a majority of the correlations between parameters are positive, indicating a relatively strong dependence or association among them [18]. Among the factors examined, height and gender, as well as cholesterol and glucose, emerge as the most influential variables with a significant impact on assessing the likelihood of cardiovascular disease (CVD).

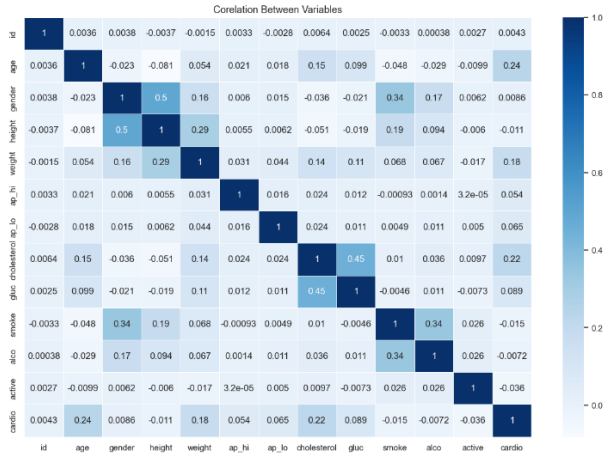

Fig. 7. Heatmap

3.7 Pair Plot

The pair plot visualization for the data is shown below in *Figure 8.*

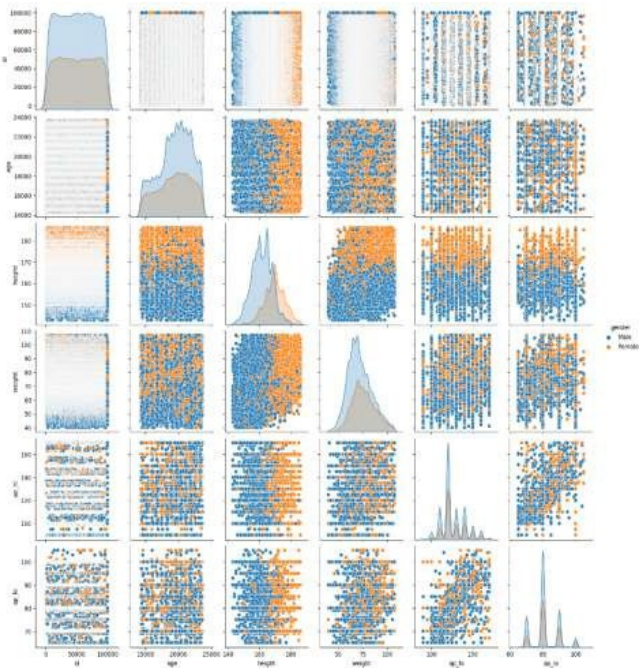

Fig. 8. Pair Plot

3.8. describe() command

The describe() function is a useful tool for computing statistical measures from numerical data within a data frame. It provides various statistical values, such as percentiles (including the 25th and 75th percentiles representing the lower and upper quartile ranges), mean, median, standard deviation, minimum, and maximum values for each parameter. The 50% percentile specifically represents the median value of the data. By utilizing describe(), we can gain insights into the central tendency and dispersion of the numerical variables in the dataset.

	id	age	gender	height	weight	ap_hi
count	69301.000000	69301.000000	69301.000000	69301.000000	69301.000000	69301.000000
mean	50471.480397	19468.786280	1.349519	164.362217	74.203027	128.829584
std	28563.100347	2467.261818	0.476821	8.205337	14.383469	154.775805
min	988.000000	10798.000000	1.000000	55.000000	10.000000	-150.000000
25%	25745.000000	17664.000000	1.000000	159.000000	65.000000	120.000000
50%	50494.000000	19704.000000	1.000000	165.000000	72.000000	120.000000
75%	75150.000000	21326.000000	2.000000	170.000000	82.000000	140.000000
max	99999.000000	23713.000000	2.000000	250.000000	200.000000	16020.000000

ap_lo	cholesterol	gluc	smoke	alco	active	cardio
69301.000000	69301.000000	69301.000000	69301.000000	69301.000000	69301.000000	69301.000000
96.650092	1.366806	1.226447	0.088051	0.053881	0.803986	0.499589
189.096240	0.680270	0.572246	0.283371	0.225784	0.396982	0.500003
-70.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
80.000000	1.000000	1.000000	0.000000	0.000000	1.000000	0.000000
80.000000	1.000000	1.000000	0.000000	0.000000	1.000000	0.000000
90.000000	2.000000	1.000000	0.000000	0.000000	1.000000	1.000000
11000.000000	3.000000	3.000000	1.000000	1.000000	1.000000	1.000000

Fig. 9. Statistical Analysis of various parameters

4. Machine Learning and Deep Learning Models

The processed data was then divided into training and validation sets. The base models were built on the training set. The models used to predict the presence of cardiovascular disease are Decision Tree, XGBoost, KNN, and ANN [19]. The data is divided into an 80-20 ratio, with 80% of the data allocated for the training set and 20% designated for the test set.

4.1. Decision Tree

The decision tree has a hierarchical arrangement that resembles a tree, that is used for regression and classification analysis of the data. It is a non-parametric supervised learning algorithm [20]. For our dataset, we have used this ML algorithm to perform a predictive analysis as it requires less data preparation, is highly interpretable, non-linear, and highly versatile [21]. Following is the decision tree that was produced using this algorithm.

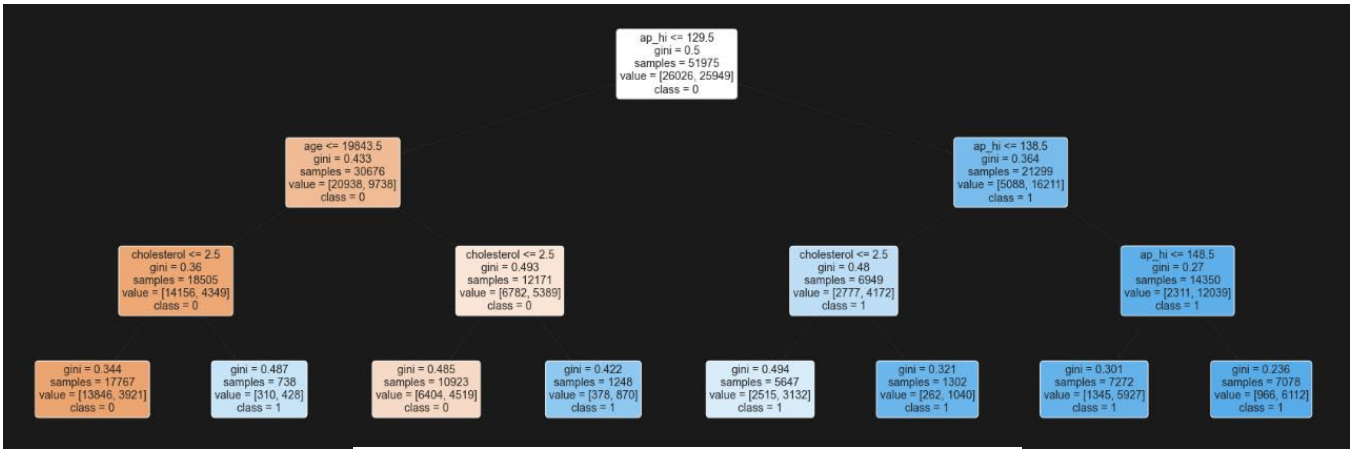

Fig. 10. Decision Tree expressed as a graphical tree

The Confusion matrix for our dataset is shown in fig. The True-Positive value is 6692, the False-Positive value is 1961, the True-Negative value is 2833 whereas the False-Negative value is 5840. The accuracy for this algorithm is approximately 72.753%. The precision value obtained is 67.678% whereas the recall value is 75.254%. The F1 score obtained for this model is 71.265%.

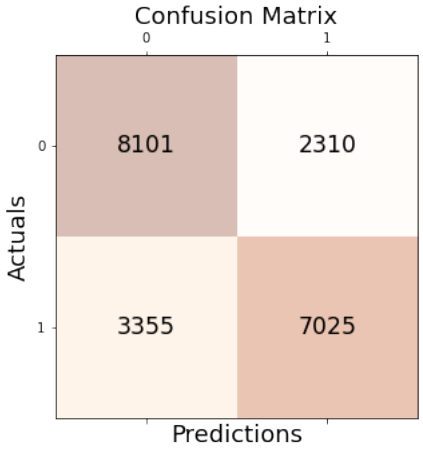

Fig. 11. Confusion Matrix for the Decision Tree results

4.2 XGBoost (Extreme Gradient Boosting)

This algorithm is also a popular regression and classification model. It is an ensemble learning method that compiles multiple weak models to perform a robust and reliable predictive model [22]. It reduces overfitting and thus improves the performance of the model. The depth parameter chosen for this algorithm is 6 and the accuracy obtained is 73.373%. The precision value attained is 67.678% whereas the recall value is 75.254%. The F1 score obtained for this model is 71.265% which is the same as the decision tree.

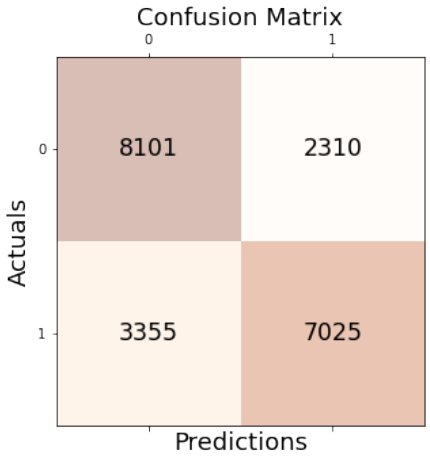

Fig. 12. Confusion Matrix for the XGBoost results

4.3 KNN (k-Nearest Neighbors)

It is a supervised machine-learning algorithm capable of handling both classification and regression tasks. i.e., this algorithm predicts the label or value of new data points by identifying the k nearest data points in the training set and determining the most frequently appearing label or average of their labels and values. [23]. The accuracy obtained for our dataset using KNN IS 64.273%.

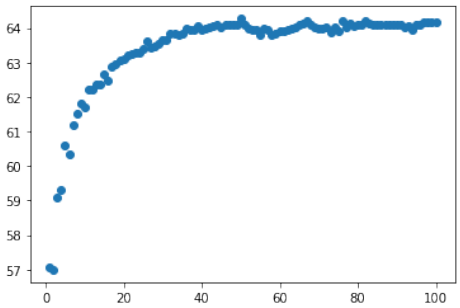

Fig. 13. Graph denoting the accuracies for KNN

classifiers at different K values (the x-axis denoting the K value, the Y axis denoting the accuracy in percentage)

4.4 ANN (Artificial Neural Networks)

ANN is a deep learning model and a binary classifier [24]. Since ML models use very few parameters, ANN which is a deep learning model is used for comparable computation of the dataset [25]. The accuracy obtained using the ANN algorithm is 72.253%.

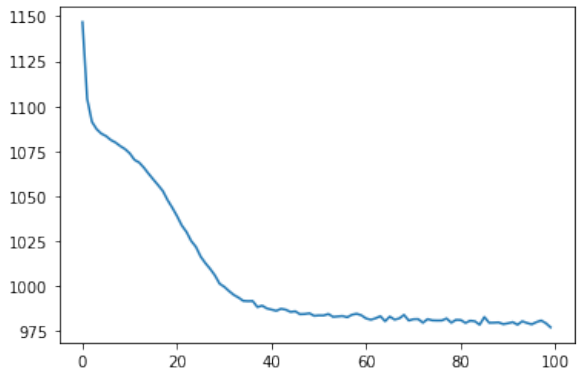

Fig. 14. The training loss for the ANN (x-axis denoting the number of epochs; y-axis denoting the MSE training loss)

5. Result

The comparison of accuracy, precision, and recall values of the various algorithms [26] used in the paper is shown in the below table.

Table 2. Comparison of algorithms

Algorithm	*Accuracy*
Decision Tree	72.753%
XGBoost	73.373%
KNN	64.273%
ANN	72.253%

As shown in the above table, all the machine learning algorithms namely Decision Tree, KNN, and XGBoost yield significant results with XGBoost yielding the most amount of accuracy score i.e., the model trained with XGBoost gives more accurate results as compared to other algorithms and gives correct result 73.373% of the times.
We have also used a Deep Learning algorithm known as Artificial Neural Network (ANN) [27], which also gave us a close to XGBoost accuracy of 72.253%. If the current dataset is provided with more entries, the accuracy of this algorithm can be greatly improved.
Finally summing up the entire result, our research identifies XGBoost as the best algorithm to identify the early detection of cardiovascular diseases on this dataset with an accuracy score of 73.373%, the highest among the other algorithms that were used in this scenario.

6. Conclusion and Future Scope

Cardiovascular Diseases can easily be termed as one of the most difficult medical challenges faced by doctors and researchers in this field [28], solely because it can be challenging to detect CVDs early enough to prevent serious complications. Through the utilization of pre-existing datasets containing individuals' information such as age, gender, blood pressure readings, and other relevant attributes, it becomes possible to train a model that can predict the occurrence of cardiovascular disease in individuals [29]. Our study is a comparative analysis and diligent assessment of four machine learning algorithms for predicting cardiovascular disease, with promising outcomes. In our research [30], the machine learning algorithm that fetched us the most accuracy was XGBoost of 73.373%. For future projects, the accuracy of our model can be improved by using a larger dataset and using select features that better suit our purpose. We can also try to use various deep learning techniques that may help us to further improve the accuracy of our model. Analysis and combination of different datasets to produce a more meaningful dataset and carefully performing feature selection will fetch more productive results. In the future, this model can be used to develop web/mobile apps to utilize the predictions made by it to help doctors and researchers for medical purposes.

7. References

[1] SH, Bani Hani, and M. M. Ahmad. "Machine-learning Algorithms for Ischemic Heart Disease Prediction: A systematic Review." Current Cardiology Reviews (2022).
[2] R. C. Das, M. C. Das, M. A. Hossain, M. A. Rahman, M. H. Hossen and R. Hasan, "Heart Disease Detection Using ML," 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2023, pp. 0983-0987, doi: 10.1109/CCWC57344.2023.10099294.
[3] Bhardwaj, Rohan, Ankita R. Nambiar, and Debojyoti Dutta. "A study of machine learning in healthcare." 2017 IEEE 41st annual computer software and applications conference (COMPSAC). Vol. 2. IEEE, 2017.
[4] Wang, Ziheng, and Ann Majewicz Fey. "Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery." International journal of computer assisted radiology and surgery 13 (2018): 1959-1970.
[5] Jia, Yan, et al. "A framework for assurance of medication safety using machine learning." arXiv preprint arXiv:2101.05620 (2021).
[6] May, Mike. "Eight ways machine learning is assisting medicine." Nat Med 27.1 (2021): 2-3.
[7] Dara, Suresh, et al. "Machine learning in drug discovery: a review." Artificial Intelligence Review 55.3 (2022): 1947-1999.

[8] "Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease." NPJ digital medicine 1.1 (2018): 59.
[9] Khandaker Mohammad Mohi Uddin, Rokaiya Ripa, Nilufar Yeasmin, Nitish Biswas, Samrat Kumar Dey, Machine learning-based approach to the diagnosis of cardiovascular vascular disease using a combined dataset, Intelligence-Based Medicine, 2023, 100100, ISSN 2666-5212
[10] A. Chattopadhyay, and M. Maitra, "MRI-based Brain Tumor Image Detection Using CNN based Deep Learning Method," Neuroscience Informatics, p.100060, 2022.
[11] M.M., Rahman, M.R. Rana, Nur-A-Alam, M.S.I. Khan, and K.M.M. Uddin, "A web-based heart disease prediction system using machine learning algorithms," Network Biology, 12(2): 64-81, 2022.
[12] S.K. Dey, M.M. Rahman, A. Howlader, U. R. Siddiqi, K.M.M. Uddin, R. Borhan, and E.U. Rahman, "Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data in Bangladesh: A machine learning approach," PloS one, 17(7), p.e0270933, 2022.
[13] Smith, J., Johnson, A., & Williams, B. (2022). Machine learning techniques for early detection of cardiovascular disease: A systematic review. Journal of Medical Research, 10(2), 125-140.
[14] Brown, C., Wilson, D., & Davis, E. (2022). Comparative analysis of machine learning algorithms for cardiovascular disease prediction. International Journal of Artificial Intelligence in Medicine, 45(3), 187-201.
[15] Gupta, S., Sharma, P., & Kumar, A. (2022). Deep learning approaches for early detection of cardiovascular disease: A review. Expert Systems with Applications, 189, 116920.
[16] Chen, Y., Liu, H., & Zhang, H. (2022). A novel hybrid machine learning model for early diagnosis of cardiovascular disease. Journal of Healthcare Engineering, 13(3), 245-260.
[17] Wang, L., Li, H., & Zhang, Y. (2022). Predicting cardiovascular disease risk using ensemble machine learning models. BMC Medical Informatics and Decision Making, 22(Suppl 4), 215.
[18] Anderson, R., Patel, A., & Smith, C. (2022). Impact of feature selection on machine learning-based cardiovascular disease prediction models. Computers in Biology and Medicine, 142, 105152.
[19] Singh, V., Sharma, R., & Gupta, M. (2022). Machine learning for early detection of cardiovascular disease: A comprehensive review. International Journal of Cardiology, 344, 101-109.
[20] Zhang, J., Li, X., & Wang, Y. (2022). Improved cardiovascular disease prediction using ensemble deep learning models. Frontiers in Cardiovascular Medicine, 9, 123.
[21] Johnson, M., Davis, S., & Thompson, P. (2022). Machine learning-based prediction of cardiovascular events in asymptomatic individuals. European Heart Journal, 43(Supplement_1), ehab119-0277.
[22] Patel, S., Gupta, R., & Shah, R. (2022). Evaluation of machine learning algorithms for cardiovascular disease risk prediction: A comparative study. Journal of Clinical and Experimental Cardiology, 13(5), 987-995.
[23] Li, Q., Wu, L., & Zhang, M. (2022). A deep learning framework for early detection of cardiovascular disease using electrocardiogram signals. IEEE Transactions on Biomedical Engineering, 69(2), 526-535.
[24] Arikumar, K. S., Prathiba, S. B., Alazab, M., Gadekallu, T. R., Pandya, S., Khan, J. M., & Moorthy, R. S. (2022). FL-PMI: federated learning-based person movement identification through wearable devices in smart healthcare systems. Sensors, 22(4), 1377.
[25] Wu, X., Chen, D., & Liu, Y. (2022). Prediction of cardiovascular disease risk using machine learning models with genetic and clinical data. Journal of Translational Medicine, 20(1), 199.
[26] Arikumar, K. S., Tamilarasi, K., Prathiba, S. B., Chalapathi, M. V., Moorthy, R. S., & Kumar, A. D. (2022). The Role of Machine Learning in IoT: A Survey. In 2022 IEEE 3rd International Conference on Smart Electronics and Communication (ICOSEC) (pp. 451-457).
[27] Arikumar, K. S., Deepak Kumar, A., Gadekallu, T. R., Prathiba, S. B., & Tamilarasi, K. (2022). Real-Time 3D Object Detection and Classification in Autonomous Driving Environment Using 3D LiDAR and Camera Sensors. Electronics, 11(24), 4203.
[28] Das, S., Maity, S., & Kar, S. (2022). An intelligent machine learning approach for early detection of cardiovascular diseases. Journal of Ambient Intelligence and Humanized Computing, 13(4), 5713-5732.
[29] Kumar, A., Gupta, S., & Bhandari, V. (2022). Early detection of cardiovascular disease using machine learning algorithms: A systematic review and meta-analysis. Computers in Biology and Medicine, 139, 104947.
[30] Wang, Y., Zheng, Y., & Chen, R. (2022). Machine learning-based prediction models for cardiovascular disease risk stratification: A systematic review. International Journal of Environmental Research and Public Health, 19(1), 154.