# Multimodal Prediction of Cardiovascular Diseases

Jaskaran Singh
2020306

Aditya Ahuja
2020275

Pratyush Kumar
2020454

Vedant Gupta
2020261

## Abstract

*Cardiovascular diseases (or heart diseases) are the leading cause of death worldwide, with around 18 million deaths yearly. Several diagnosis methods are available worldwide, but they are expensive in terms of time, money and personnel. The motivation for our project is to conduct data analysis and build classification models using machine learning techniques to help practitioners make more accurate decisions regarding classifying heart diseases. We aim to combine multiple models for data analysis and perform relevant model evaluation techniques to present a proper, functioning model. Github link for the project: Link.*

## 1. Introduction

Cardiovascular Diseases (CVDs) are a group of diseases related to blood vessels and muscles around the heart and can cause heart attacks and strokes. Both are cases of interrupted blood flow from the heart to the brain. CVDs are the leading cause of death globally, representing 32% of all global deaths. Most CVDs can be prevented via early detection and management, and studies show that 80% of CVDs can be prevented through early diagnosis and intervention [1].

This project provides multiple preliminary prediction models which label an individual as either 1 (Likely to be suffering from a CVD) or 0 (Not likely to be suffering from a CVD). This classification can further be used to decide whether the subject is in need of further medical attention or not. We hope that this research assists medical practitioners in deciding the best course of action for their patients and further reduces the mortality rate of CVDs.

## 2. Literature Review

### 2.1. Implementation of a Heart Disease Risk Prediction Model Using Machine Learning [3]

This paper does a multi-modal classification on a subset of our dataset to determine heart diseases with a chi-square statistical test performed on certain attributes. Six algorithms were used for the classification, namely SVM, Gaussian Naive Bayes, Logistic regression, LightGBM, XGBoost and Random Forest, with varying but largely similar accuracy save for random forest, which came out to be the highest with 88.5%.

The paper has sections with various causes, current prediction methods (largely practitioner's evaluation), and the benefits of using ML techniques. It also goes through multiple related works with various ML techniques with similar levels of accuracy. For its own study, the paper made feature selection with K-best attributes per the sklearn library and many data visualization techniques we will implement to infer important attributes and patterns. We will keep SVM and boosting approaches like LightGBM and XGBoost for later and continue preliminary model testing on our data with other techniques illustrated in this paper and the next one.

The data given is not wholly separable with a linear boundary as per the visualizations, so aggregation techniques like Random Forest, XGBoost and LightGBM have almost no training error compared to the other two; however, only Random Forest performs better than LR, GNB and SVM on testing data.

### 2.2. Analysis and Prediction of Cardiovascular Disease using Machine Learning Classifiers [4]

The previous paper didn't include a few major classification algorithms, which this paper talks about. It also has a similar introduction and background on cardiovascular diseases. In addition, it also talks about artificial neural networks (ANN) with respect to various diseases, in general, relating to kidneys, liver, lungs and so on and how a certain level of caution must be maintained when it comes to heart diseases.

The paper also goes into detail about how tree-based algorithms perform well on this type of data, as well as the great performance of ANNs in various medical methodologies. The data was cleaned and scaled, and differentiated into multifold validation sets before training the algorithms. It used five classification algorithms: Logistic Regression, Decision Trees, Linear SVM, K-Nearest Neighbours and Random forest, and performance metrics on all of those were mentioned along with ROC AUC analysis. Random forests once again came out to be the best performing algorithm on the given dataset, followed by Decision Tree, Logistic Regression, SVM and K-Nearest Neighbors, which was significantly worse than other algorithms.

# 3. Dataset details & Pre-processing Techniques

## 3.1. Dataset information

The original dataset is from 1988 and contains 76 attributes, but all experiments use a subset of 14 of them. We combined 4 datasets from Cleveland, Hungary, Switzerland, and the VA Long Beach and removed the duplicates. The combined dataset has 12 common features and 918 records.

The 12 attributes used to detect heart diseases are as follows:

- **Age** - Age of the patient
- **Sex** - Gender, classified as Male or Female
- **ChestPainType** - Classify chest pain as Typical Angina (TA), Atypical Angina (ATA), Non-Anginal Pain (NAP) and Asymptomatic (ASY)
- **RestingBP** - Measure the resting blood pressure in [mm Hg]
- **Cholesterol** - Measure serum cholesterol in [mm/dl]
- **FastingBS** - Measure fasting blood sugar, classified as '1' if FastingBS $\geq$ 120 mg/dl and '0' otherwise
- **RestingECG** - Measure resting blood pressure, classified as 'Normal', 'ST' and 'LVH'
- **MaxHR** - Measure the maximum heart rate achieved
- **ExerciseAngina** - Check exercise-induced angina and, classified as Yes (Y) and No (N)
- **Oldpeak** - Measure ST depression induced by exercise relative to rest
- **ST_Slope** - Measure the slope of the peak exercise ST segment, classified as 'Up' for upsloping, 'Flat', and 'Down' for downsloping
- **HeartDisease** - Target variable, classified as 0 and 1

## 3.2. Data Preprocessing Techniques

The non-numerical columns, viz., Sex, ChestPainType, RestingECG, ExerciseAngina, ST_Slope and Cholesterol, were encoded using LabelEncoder to convert them into numerical data to be used for data visualization and model training. The encoding has been tabulated in the code file.

## 3.3. Data Analysis

The number of males (725) in this dataset is higher than that of females (193), with 63.17% of males testing positive and 25.90% of females testing positive.

The analysis of a histogram plot of individuals suffering from heart disease shows that individuals between 50 and 60 years are more likely to suffer from some form of CVD.

Upon analyzing the maximum heart rate of individuals diagnosed with CVDs, we found that most had a maximum heart rate between 110 and 140. The average maximum heart rate of individuals below 70 yrs is above 150 bpm. [2]

Using SelectKBest with k value set to 3 resulted in 'ChestPainType, 'MaxHR' and 'ExerciseAngina'. This was somewhat expected as all three are commonly used by medical practitioners to diagnose an individual's cardiovascular condition accurately.

Below is a plot of the correlation matrix obtained from the data.

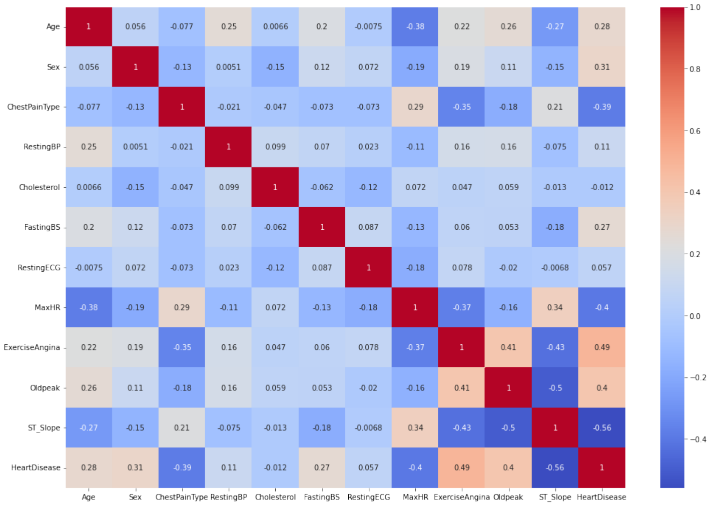


Figure 1. Heatmap of pairwise correlation between features.

The t-SNE graph tells us that the data is not separable. The majority of features have low correlation, as seen from the heatmap. Most of the points are very close to the points of other classes.

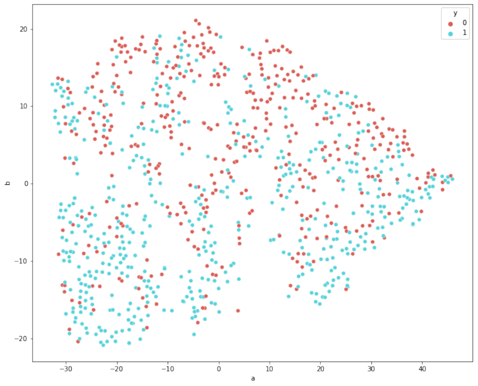


Figure 2. t-SNE plot for the data.

PCA analysis was done on the dataset with first 3 components being plotted below. As we can see the data points are quite intermingled even in the highest variance axes.

# 4. Methodology & Model details

We have used sklearn library to create and train the following supervised learning models: Logistic Regression, Naive Bayes, Random Forest Classifier, Support

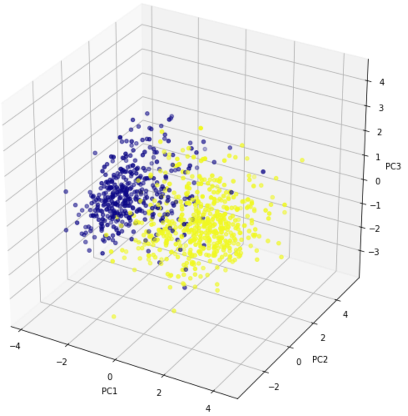


Figure 3. PCA plot for the data.

Vector Machine (SVM) and K-Nearest Neighbour Clustering. In addition to these models, two more boosting frameworks, LightGBM and XGboost, were also used to implement gradient boosting decision trees. Keras was used to implement an Artificial Neural Network classification model.

### 4.1. Logistic Regression

Since our dataset is small, we decided to train a Logistic Regression model with bootstrapping by using resample. We have created 100 bags (bootstrapped datasets) sampled from 80% of the original dataset. Each of these samples was then used to train a Logistic Regression Classifier, with the final accuracy scores being the mean of these classifiers. Hyperparameter tuning has been done using GridSearchCV and the model with optimal parameters were used to train the 100 classifiers.

### 4.2. Naive Bayes

We trained the Naive Bayes model with bootstrapping similar to the other models. Gaussian and Bernoulli priors distributions were used to train two different models using GaussianNB and BernoulliNB with optimal parameters which we obtained through hyperparameter tuning.

### 4.3. k-Nearest Neighbours

k-Nearest Neighbours classification was used to gauge raw performance even though t-SNE plot suggested that the data is quite mixed. Hyperparameter tuning was done to get the best performance which resulted in metric being Manhattan distance. It can be seen in the same t-SNE plot that even though data is mixed, there are patches where similar kind of data is clustered and the distance is rather uniform between them.

### 4.4. Support Vector Machine

SVM was used to classify the patients with heart diseases with a linear boundary along with optimum margin. Different kernels were utilised to transform the data since it was not linearly separable. Hyperparameter tuning gave us ’rbf’ as the best kernel.

### 4.5. Artificial Neural Network

Artificial Neural Network via TensorFlow sequential framework was used to check performance of neural networks on the given dataset. The literature review suggested neural network to be one of the better performing models and so various combinations of hidden layers and neurons were tested along with various optimizers. Early stopping condition was also used to stop the training once the validation loss stagnated. From testing these parameters, best sizes of layers came out to be [11, 24, 16, 8, 4, 1] with leaky relu being used as activation and sigmoid being used for the final layer. 250 epochs with batch size of 25 were run with early stopping condition and reusing best weights combination.

### 4.6. Random Forest Classifier

Random Forest Classifier with decision tree classifier is used along with bootstrapping. We used AdaBoost to improve the accuracy obtained from the ’vanilla’ random forest model.

### 4.7. Boosting Models

LightGBM and XGBoost frameworks were used to implement gradient boosting decision trees. Both models were trained with bootstrapping and gave optimal results with an average accuracy score of 91.70 %. LightGBM has some advantages over XGBoost, in that it is faster than the other and has lower memory overhead.

## 5. Results & Analysis

The dataset was bootstrapped 100 times, with each iteration split 80:20 into training and testing sets. Four performance metrics from sklearn metrics, viz., Accuracy, Recall, Precision and F1 scores, were calculated for each model. ROC-AUC analysis was done based on the confusion matrix obtained on the test set. The table at the end shows each tested model’s corresponding scores and AUC values.
Linear classifiers that were tested i.e. Logistic Regression, Naive Bayes and SVM didn’t perform as good as the following models. The data analysis had suggested the given data points were not linearly separable and intertwined. Both Bayes models we tested performed worse than Logistic regression model with corresponding accuracies being 82.74, 83.74 and 84.46%. SVM in particular performed the worst out of all models. We believe this is because a lack of hyperplane in any of the available dimensions. Even with best parameters chosen, the accuracy of SVM could only go up to 69.1%. Any kernel applied upon the dataset failed to provide a hyperplane and so the linear models were out of question for this dataset. The following is the ROC-AUC curve for Logistic Regression model, the best performing linear model:

Next on, the k-Nearest Neighbour classifier used had best parameters taken via GridSearchCV which came

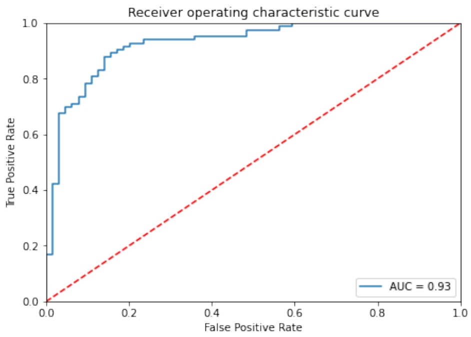


Figure 4. ROC-AUC Curve for Logistic Regression.

out to be the Manhattan Distance metric and 49 neighbours to be used for queries. The model notably performed well on the dataset with more than 85% accuracy score which was more than any linear classifier tried above.

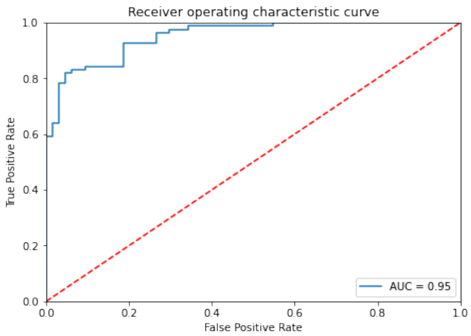


Figure 5. ROC-AUC Curve for k-Nearest Neighbour classifier.

A big part of model training and tuning were done for Artificial Neural Network Model with different combinations of layers and activation functions. The best parameters were chosen and testing accuracy on it didn't prove to be much better than linear classifiers and on a similar level to the kNN classifier with accuracy being 85%.

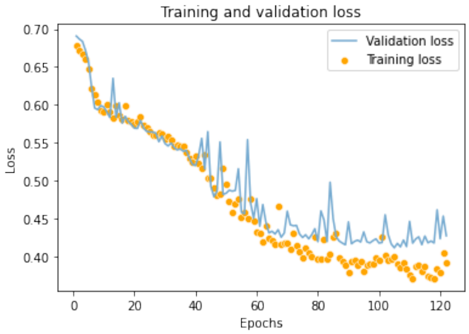


Figure 6. Loss Curve for ANN.

We can see that the tree-based algorithms, namely Random Forest and AdaBoost, perform particularly well with the given data, with them having 85.87% and 88.04% accuracies. The AUC is also the highest in both algorithms. The Confusion Matrix and ROC-AUC curve for AdaBoost can be seen below:

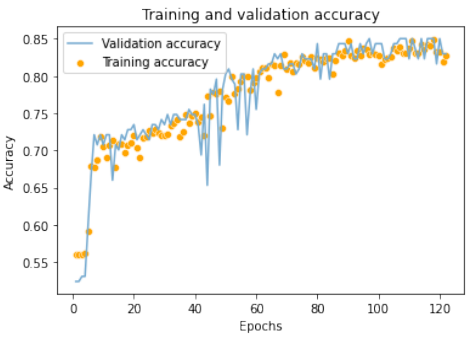


Figure 7. Accuracy curve for ANN.

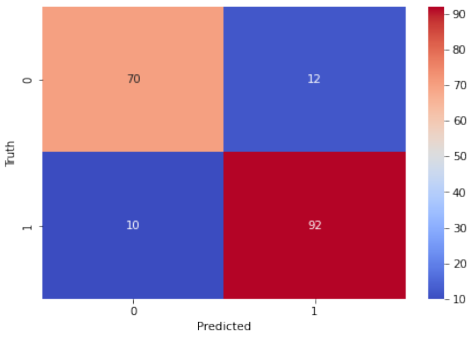


Figure 8. Confusion Matrix for AdaBoost model on the testing dataset.

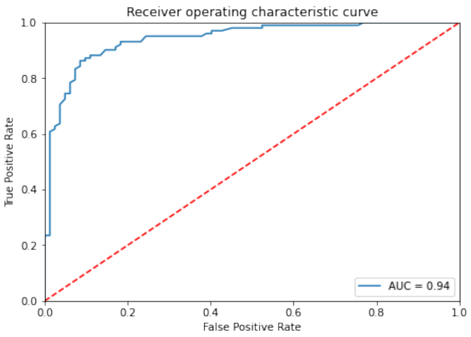


Figure 9. ROC-AUC Curve for the AdaBoost model on the testing dataset.

Feature importance was also calculated for the random forest model. We can see how ST_Slope has the most information gain in Random Forest, which follows with the highest correlation value in the heatmap. The next page contains a table for accuracy values corresponding to all the previous models.

We can see how ST_Slope has the most information gain in Random Forest, which follows with the highest correlation value in the heatmap. The next page contains a table for accuracy values corresponding to all the previous models.

Following the success of tree based algorithms, other boosting techniques were used to further improve the

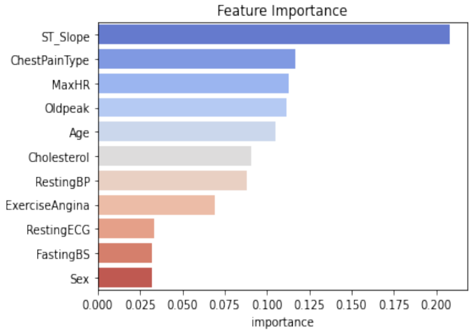


Figure 10. Feature importance in Random Forest model.

model. XGBoost with 500 estimators and each estimator being 3 nodes deep performed really well on the given data with 91.6% accuracy score. Another boosting technique used was LightGBM which gave the similar accuracy score of 91.6% with 100 estimators being 10 nodes deep and with 20 leaves after parameter tuning. Both of these methods performed better than AdaBoost and significantly better than Random Forest classifier. The ROC-AUC curves for the same can be seen below:

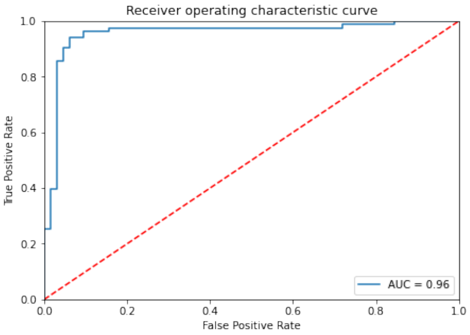


Figure 11. ROC-AUC Curve for XGBoost.

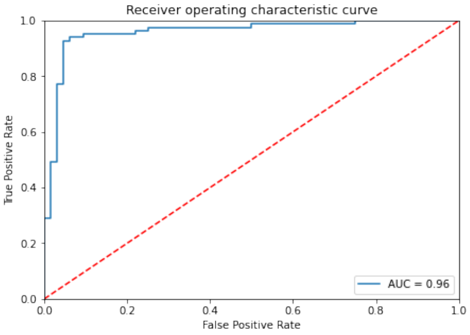


Figure 12. ROC-AUC Curve for LightGBM.

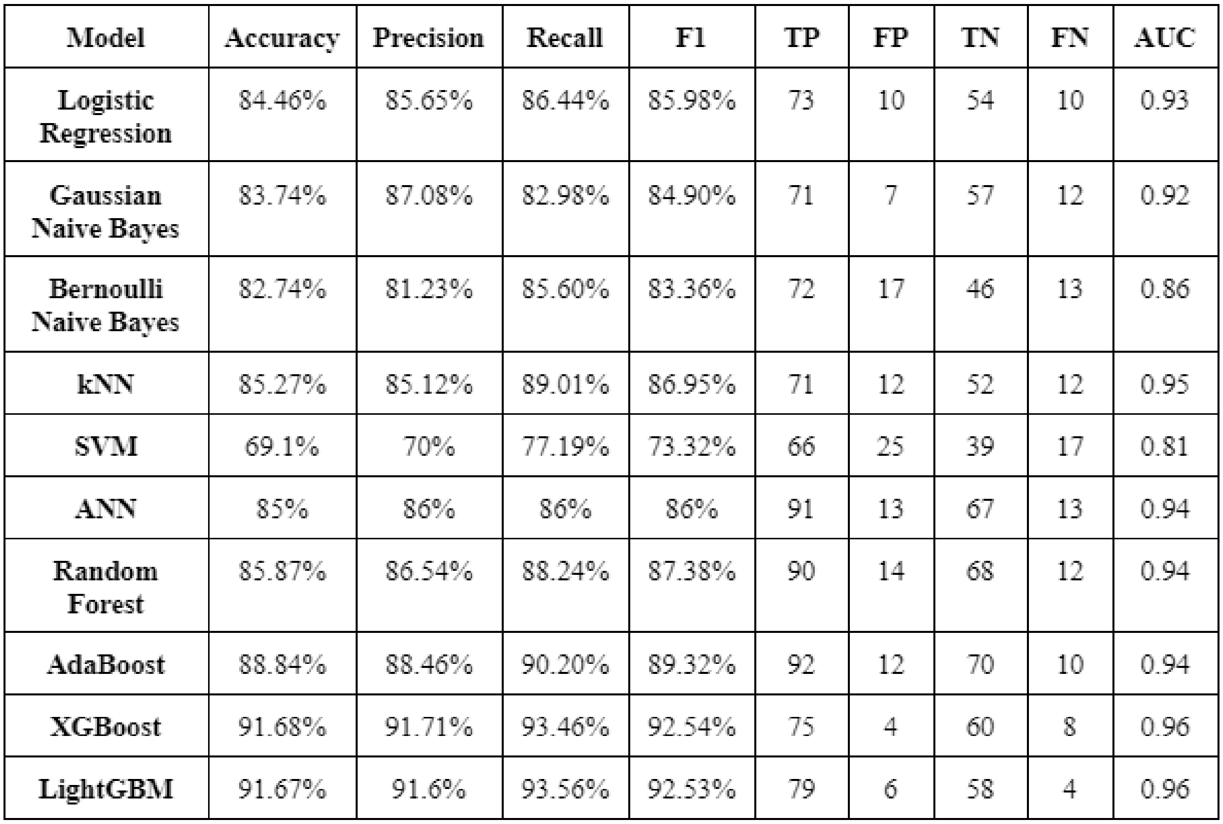
## 6. Conclusion

In this project, we tested various machine learning models on the dataset containing various features of patients and classifying them as patients with heart conditions or not. The models used were Logistic Regression, Naive Bayes, Support Vector Machine, k-nearest Neighbours, Artificial Neural Network, Random Forest, AdaBoost, XGBoost and LightGBM. With poor performance of SVM and relatively lower performance of linear classifiers suggested a lack of a linear hyperplane in the data. All classifiers had accuracy scores ranging from 82% to 84% with the exception of SVM with 69%. The kernelisation of data didn't prove to be fruitful with similar accuracy on SVM. This led to us switching on model philosophy.

In the t-SNE plot, even though the points weren't separable, there were a significant number of patches where similar points were gathered. This was confirmed by relatively high accuracy of k-Nearest Neighbours Model. However, it was only marginally better than Logistic Regression Model.

Seeing the lack of hyperplane in the dataset, we decided to use Artificial neural network to create a complex decision boundary. Hyperparameter tuning led us to the aforementioned prameters but the results were not much better than kNN Model either with 85% accuracy being reported.

Finally we tried the tree-based algorithms which outperformed the linear models as well as ANN by some margin. Random Forest with 100 decision trees didn't perform as expected but boosting with AdaBoost came out to be the third best model on the dataset. From these findings, we can safely assume that tree-based models will perform better on the data. Seeing this result, we tried two popular boosting techniques XGBoost and LightGBM which uses gradient boosting with a multitude of trees. Both of them came out to be the best models with difference in accuracy being in the margin of error with 91.6%. The results conclude that a systematic analysis of features and tree-based predictions were the most helpful instead of fitting a function through the data points. Further study on improving the model can be done but the lack of latest compiled data hindered us. Data collection on this subject was very scarce and so that remains to be looked at.



| Model | Accuracy | Precision | Recall | F1 | TP | FP | TN | FN | AUC |
|---|---|---|---|---|---|---|---|---|---|
| Logistic Regression | 84.46% | 85.65% | 86.44% | 85.98% | 73 | 10 | 54 | 10 | 0.93 |
| Gaussian Naive Bayes | 83.74% | 87.08% | 82.98% | 84.90% | 71 | 7 | 57 | 12 | 0.92 |
| Bernoulli Naive Bayes | 82.74% | 81.23% | 85.60% | 83.36% | 72 | 17 | 46 | 13 | 0.86 |
| kNN | 85.27% | 85.12% | 89.01% | 86.95% | 71 | 12 | 52 | 12 | 0.95 |
| SVM | 69.1% | 70% | 77.19% | 73.32% | 66 | 25 | 39 | 17 | 0.81 |
| ANN | 85% | 86% | 86% | 86% | 91 | 13 | 67 | 13 | 0.94 |
| Random Forest | 85.87% | 86.54% | 88.24% | 87.38% | 90 | 14 | 68 | 12 | 0.94 |
| AdaBoost | 88.84% | 88.46% | 90.20% | 89.32% | 92 | 12 | 70 | 10 | 0.94 |
| XGBoost | 91.68% | 91.71% | 93.46% | 92.54% | 75 | 4 | 60 | 8 | 0.96 |
| LightGBM | 91.67% | 91.6% | 93.56% | 92.53% | 79 | 6 | 58 | 4 | 0.96 |

Figure 13. Performance of various tested models

## References

[1] Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts) was Developed with the special contribution of the European Association for Cardiovascular Prevention and Rehabilitation (EACPR). Eur Heart J. (2016) 37:2315–2381. doi: 10.1093/eurheartj/ehw106

[2] Hirofumi Tanaka, Kevin D Monahan, Douglas R Seals, Age-predicted maximal heart rate revisited, Journal of the American College of Cardiology, Volume 37, Issue 1, 2001, Pages 153-156, ISSN 0735-1097, https://doi.org/10.1016/S0735-1097(00)01054-8.

[3] Karthick K, Aruna SK, Samikannu R, Kuppusamy R, Teekaraman Y, Thelkar AR. Implementation of a Heart Disease Risk Prediction Model Using Machine Learning. Comput Math Methods Med. 2022 May 2;2022:6517716. doi: 10.1155/2022/6517716. PMID: 35547562; PMCID: PMC9085310.

[4] N. K. Kumar, G. S. Sindhu, D. K. Prashanthi and A. S. Sulthana, ”Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers,” 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 15-21, doi: 10.1109/ICACCS48705.2020.9074183.

[5] Repository

[6] Dataset Link

The Individual contribution of each team member is as follows:

**Jaskaran Singh:** Data preprocessing, Data Analysis, Model Training and Tuning, Report, Slides.

**Aditya Ahuja:** Data preprocessing, Data Analysis, Model Training, Report.

**Pratyush Kumar:** Literature Review, Model Analysis, Report, Slides.

**Vedant Gupta:** Domain study, Data Analysis, Feature Engineering, Slides.