Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method

Rahma Atallah
Communications Engineering Department
Princess Sumaya University for Technology
Amman, Jordan
r_rahma@hotmail.com

Amjed Al-Mousa
Computer Engineering Department
Princess Sumaya University for Technology
Amman, Jordan
a.almousa@psut.edu.jo

***Abstract*—This paper presents a majority voting ensemble method that is able to predict the possible presence of heart disease in humans. The prediction is based on simple affordable medical tests conducted in any local clinic. Moreover, the aim of this project is to provide more confidence and accuracy to the Doctor's diagnosis since the model is trained using real-life data of healthy and ill patients. The model classifies the patient based on the majority vote of several machine learning models in order to provide more accurate solutions than having only one model. Finally, this approach produced an accuracy of 90% based on the hard voting ensemble model.**

Keywords—Machine learning; Majority Voting ensemble method; heart disease; UCI dataset; classification.

I. INTRODUCTION

In the present era, heart disease rates have dramatically increased to become the leading cause of death in the United States upon adults due to the widespread of unhealthy habits [1]. These include a declination in physical activity since the technology trend is moving towards replacing human physical activity and unhealthy eating habits which are directly linked to increasing the risk of having heart diseases.

Starting off with the definition of a Heart Disease, according to [2] the National Heart, Lung, and Blood Institute states that heart disease is a disruption to the heart's normal electrical system and pumping functions. Where the disease makes it harder for the heart muscle to pump blood efficiently.

Furthermore, according to the World Health Organization (WHO), 17.9 million people die each year from cardiovascular diseases which correspond to 31% of all deaths around the world [3]. This incurs the need of having an affordable system that is able to give a preliminary assessment of a patient based on relatively simple medical tests that are affordable to everyone.

To conduct the training and testing of the machine learning model, the Cleveland dataset from the well-known UCI repository was used since it is an authenticated dataset that is widely used for training and testing in machine learning models [4]. The dataset contains 303 instances and 14 attributes that are based on well-known factors that are thought to correlate with risks of heart diseases.

The approach presented in this paper uses the hard voting ensemble method which is a technique where multiple machine learning models are combined and the prediction result is based on the majority vote from all models. This technique is used in order to improve the overall prediction results since the combination of models produces a powerful collaborative overall model.

Section II of this paper presents a review of related work, then Section III introduces the intricate details of the dataset, data preprocessing and the machine learning techniques used. Moreover, the results of each model along with the overall accuracy of the hard voting model are presented in Section IV. Finally, a conclusion is outlined in section V.

II. RELATED WORK

In the field of heart disease detection, a variety of techniques regarding data preprocessing and model variation has been used. The work presented in [5] used the same dataset as this paper but different machine learning models were implemented. Three discrete classifier models were built which included Support Vector Machine (SVM) classifier, naïve Bayes algorithm, and C4.5. The prediction of the heart disease was conducted based on each of these models discretely and produced a maximum accuracy of 84.12% in the SVM machine learning model.

The work in [6] also used the Cleveland heart disease dataset but the classification models that were implemented involved only Tree algorithms. Those included J48, Logistic Model, and Random Forest algorithm. A comparison of the three methodologies was conducted and the highest accuracy achieved was 84% using the J48 algorithm.

Furthermore, the work in [7] presents a prediction system of coronary artery heart diseases using four different datasets including the Cleveland dataset. The algorithm used for prediction involved only decision tree techniques that included C4.5 and Fast Decision Tree. At first, the model is trained based on each dataset using all features. Then the best features from each dataset are selected and used for training the model. This technique improved the accuracy of prediction of the model from 76.3% to 77.5% using C4.5 (this accuracy represents the average accuracy from all datasets) and for the Fast Decision Tree, the average accuracy improved from 75.48% to 78.06%.

The work in [8] uses data mining techniques where the large Cleveland dataset with all 76 attributes is investigated in order to extract hidden and previously unknown patterns. This allows the prediction to utilize the most dominant and effective attributes provided in the dataset. The machine learning algorithm consists of different Decision Tree methods (J48, Logistic Model Tree Algorithm, Random Forest Algorithm.) The highest accuracy is obtained from the

978-1-7281-2882-5/19/$31.00 ©2019 IEEE

J48 model that is 56.76% with a total build model time of 0.04 seconds.

Finally, the work in [9] deploys various machine learning models in order to investigate the highest performance metric (Accuracy, Sensitivity, Specificity, and Kappa). The machine learning algorithms involve Random Forest, Logistic Regression and Artificial Neural Network. Cross-Industry Standard Process for Data mining technique (CRISP-DM) is used to find insights and meaningful information from the data. The CRISP-DM involves six stages that were followed in this research. Moreover, the accuracies obtained from the models used were as follows; 80.9 % for the Random forest, 79.78% for the Artificial Neural Network, and 85.39% for the Logistic Regression.

III. Experimental Setup

The objective of this paper is to produce a heart disease prediction system using the aforementioned dataset. This dataset represents real-life data which serves the purpose of this paper and allows the prediction system to generalize to any new data.

TABLE I. Attribute Information

Attributes	Type	Description
Age	Continuous	Age in years
Sex	Discrete	0=female 1=male
Cp	Discrete	Chest pain type: 1=typical angina 2=atypical angina 3=non-anginal pain
Trestbps	Continuous	Resting blood pressure (mm/Hg)
Chol	Continuous	Cholesterol (mg/dl)
Fbs	Discrete	Fasting blood sugar > 120 (mg/dl): 1=true 0=false
Restecg	Discrete	Resting electrocardiographic result: 0=normal 1=having ST-T abnormality 2= probable left ventricular hypertrophy
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise-induced angina: 1=yes 0=no
Old peak ST	Continuous	ST depression induced by exercise relative to rest
Slope	Discrete	Peak exercise slope segment: 1=up sloping 2=flat 3=down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy that ranges from 0-3
Thal	Discrete	Heart rate: 3=normal 6=fixed defect 7=reversible defect
Target	Discrete	Diagnosis classes: 0=healthy 1=possible heart disease

A. Dataset Attribute Information

The UCI repository was used to retrieve the heart disease database. The original database contains 76 attributes, but based on extensive experiments it was found that the most effective attributes were 14. The Cleveland database contains the most dominant 14 attributes, which why they were chosen for training the model. Table I presents the attribute's name, type, and description.

In order to analyze the data, a correlation value was calculated between each of the values and the Target diagnosis. It can be noted that the highest correlated features with the target attribute were Cp, Thalach, Exang, and Oldpeak. This helps in forming an overview of the data that is being dealt with.

TABLE II. Correlation with Target Diagnosis

Attribute name	Correlation value
Cp	0.433798
Thalach	0.421741
Slope	0.345877
Restecg	0.137230
FBS	-0.028046
Chol	-0.085239
Trestbps	-0.144931
Age	-0.225439
Sex	-0.280937
Thal	-0.344029
Ca	-0.391724
Oldpeak	-0.430696
Exang	-0.436757

Moreover, to further form a clear overview of the feature correlation between each of the attributes, a heat map showing the correlations between all features is shown in Figure 1.

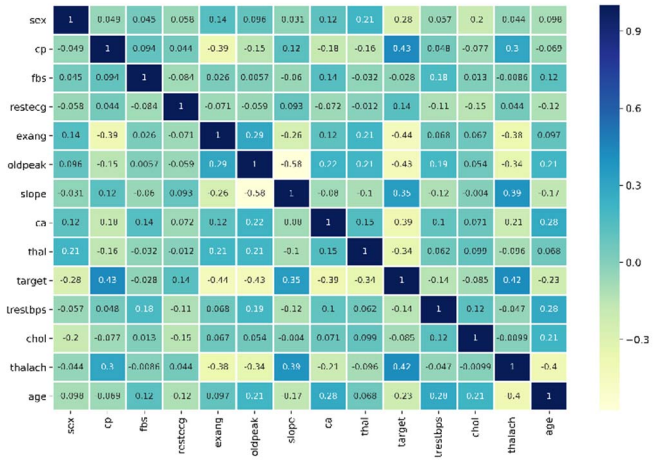

Figure 1: Heat map of cross-correlation values

Also, a pie chart as shown in Figure 2 displays the gender distribution of the instances in the Cleaveland data set. It is clear that the dataset had more males 68% than females 32%.

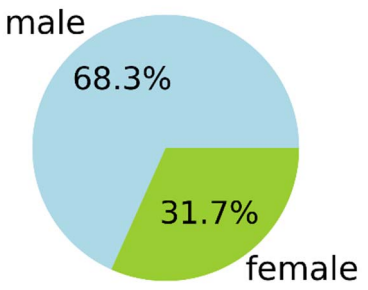

Figure 2: Gender distribution within the dataset

Furthermore, for continuous attributes data visualization histograms are plotted to preview the data distribution as shown in Figures 3-6. It can be noted that all of the continuous attributes have a normal distribution.

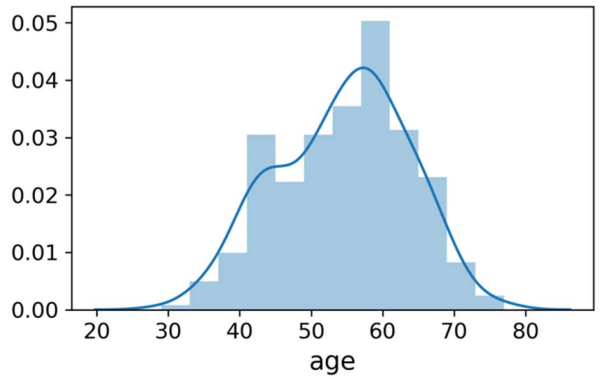

Figure 3: Age distribution

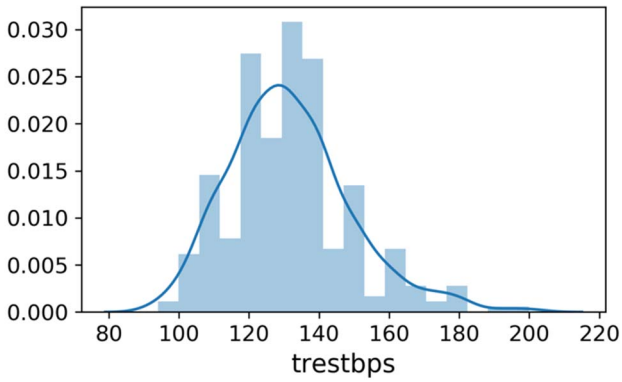

Figure 4: Resting blood pressure distribution

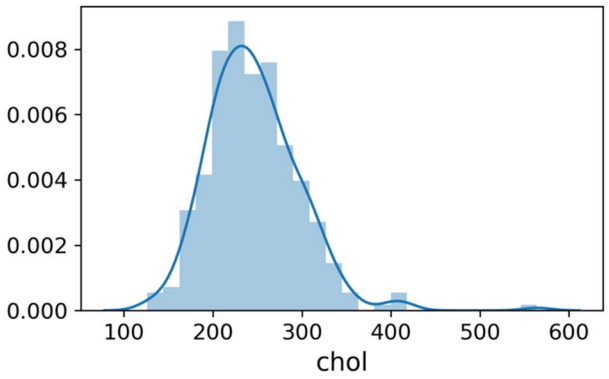

Figure 5: Cholesterol distribution

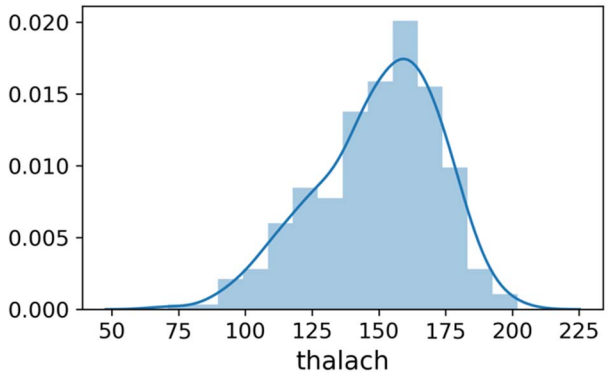

Figure 6: Maximum heart rate achieved

For the age attribute in Figure 3, it can be seen that most observations lie between 47-61 years old. To further investigate if age has a relation to having heart diseases, Figures 7 & 8 show the age distribution for people with no heart disease and people with heart disease respectively. It is observed that people with heart diseases had a major concentration in the age range from 51-53 and 41.

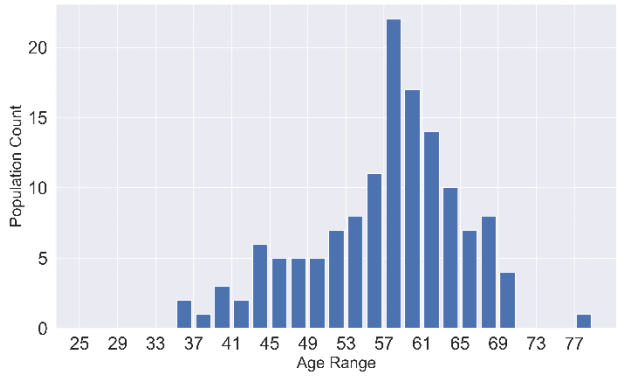

Figure 7: Age Distribution for people with No heart disease

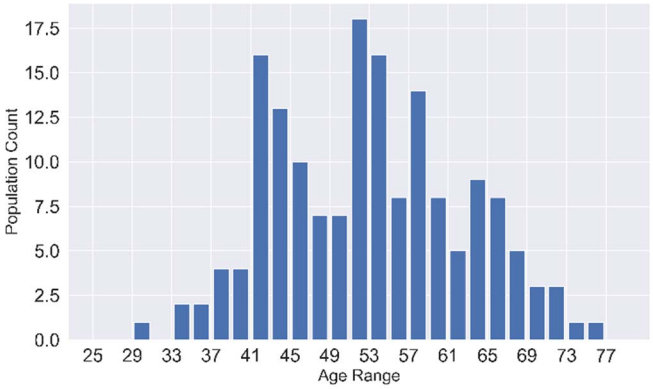

Figure 8: Age Distribution for people with heart disease

In addition, the highest correlated continuous attribute (thalach) is plotted against age, as shown in Figure 9 to examine if there is any relation. It is noticed that for people with heart disease at all age ranges, the heart rate was generally higher than that for people with no heart disease. In addition, in both groups as age increased the maximum heart rate decreased leading to a negative correlation of -0.4 with age shown earlier in Figure 1.

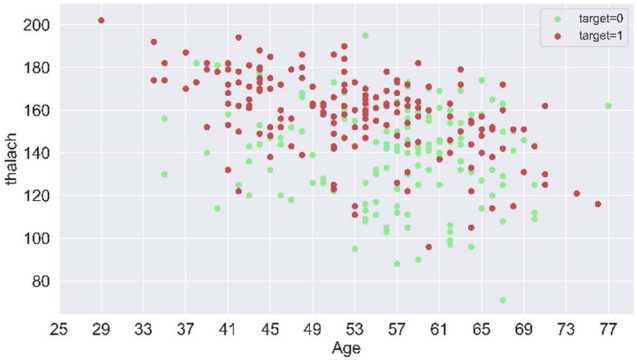

Figure 9: Max heart rate distribution Vs Age

B. Data preprocessing

The data in the Cleveland dataset had different scales which led to the need to scale the large continuous data using the Min-Max normalization strategy. This strategy involves linearly transforming the data by subtracting the minimum and dividing over the data range as shown in equation (1). Thus, the data is mapped to a range between 0 and 1 which helps the machine learning model to form a clearer trend between data and normalize the impact of different parameters.

$$Y = \frac{X - min}{\max - min} \quad (1)$$

IV. Machine Learning Algorithm

After analyzing the data, the data was split into training and testing sets into a ratio of 80% training data and 20% testing data. This is needed to validate the ability of the model to generalize to new data. Several Classifier models have been tested as follows:

A. Stochastic Gradient Descent (SGD) Classifier

Starting off with the first model, a binary classifier that uses the SGD approach was built. The SGD approach picks random instances in the training set and computes the gradient-based on that single instance in order to reach the minimum value of the cost function. Then based on the parameters chosen to minimize the cost function, classification occurs based on the simple binary classifier built that is able to identify whether heart disease is present or not.

B. K-Nearest Neighbor Classifier

The second model that was built is the K-Nearest Neighbor classifier. The algorithm in this classifier involves finding the distances between the new instance and all of the training instances, then from a predefined K number it selects the nearest K data points to the new instance. Finally, classification occurs based on the majority class of the K data points selected. The K number in this project was chosen to be 7 since it produced the best results based on the GridsearchCV.

C. Random Forest Classifier

The third model that was built is the Random Forest Classifier. This model involves building multiple decision trees and combines them together in order to obtain a more accurate and stable prediction. In this project, a number of 1000 trees worked best according to the GridsearchCV.

D. Logistic Regression Classifier

The fourth model built was the Logistic Regression Classifier. According to [10] the Logistic Regression Classifier computes a weighted sum of the input features and outputs the logistic of this result. The logistic is a sigmoid function that outputs a number between 0 and 1. Then based on the estimated probability, the classification occurs.

E. Ensemble Classifier

Finally, the four models mentioned in this section are combined in an ensemble method where the classification is done based on the majority vote of the models (hard voting.) The voting occurs when each model makes a prediction for each instance and the output prediction is the one that receives more than half of the votes.

V. Results and Analysis

Starting off with the SGD classifier, the prediction was run on the test set which is considered unseen data that the model has never prevailed. The first test was run on the default parameters of the classifier and produced an accuracy of 80%. Then after running a GridsearchCV, the optimized parameters based on cross-validation were found and the accuracy increased to 88%. Figure 10 shows the confusion matrix obtained from this model.

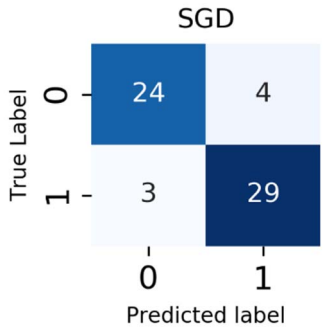

Figure 10: SGD classifier confusion matrix

Moving on to the second model, that is the K-Nearest Neighbor classifier. The model was built with the default parameters and was run on the unseen test set. The accuracy came out to be 82% and after running GridsearchCV to find the optimized parameters the accuracy went to 87%. Figure 11 shows the confusion matrix obtained from the results.

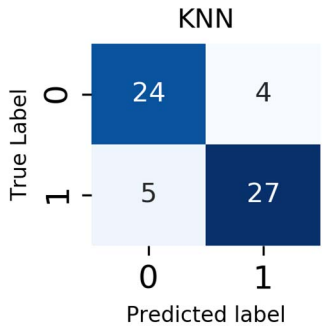

Figure 11: KNN classifier confusion matrix

Moreover, the third model that was built was the Random Forest Classifier. The model was built using the default parameters and conducted predictions on the unseen test set. The accuracy came out to be 85%, then a GridsearchCV was deployed and built the model using the optimized parameters to produce an accuracy of 87%. Also, feature importance was computed in this classifier and the top three features were (oldpeak, ca, thalach). Figure 12 shows the confusion matrix obtained from this model.

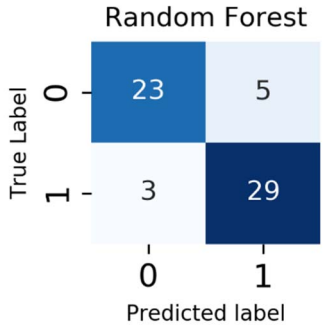

Figure 12: Random Forest classifier confusion matrix

Furthermore, the last model that was built was the Logistic Regression classifier. The model was built using the default parameters and the classification occurred based on the unseen test set. The accuracy came out to be 87% and after conducting GridsearchCV the accuracy remained the same since the default parameters came out to be the same as the optimized parameters. Figure 13 shows the confusion matrix of this model.

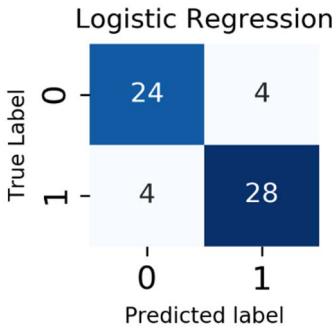

Figure 13:Logistic Regression classifier confusion matrix

Table III shows the overall final accuracies of the four models.

TABLE III. ACCURACY OF THE MODELS

Model Name	Accuracy
SGD Classifier	88%
KNN Classifier	87%
Random Forest Classifier	87%
Logistic Regression Classifier	87%
Hard Voting Ensemble Method	90%

Also, Figure 14 shows how running the GridsearchCV which is based on the cross-validation technique improves the accuracy of every model. This shows the need to fine-tune the parameters of any machine learning algorithm.

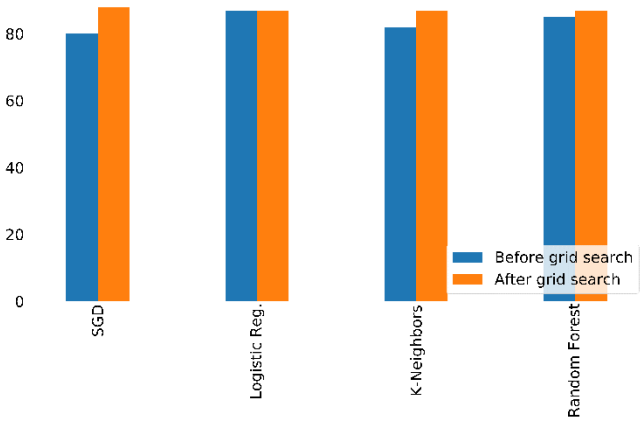

Figure 14: GridsearchCV accuracy improvement

To further investigate the models built, a receiver operating characteristic curve (ROC) was plotted as shown in Figure 16 for all of the models involved in this project. The ROC

represents the diagnostic ability of the classifier and the area under each curve is calculated and displayed in Figure 15. The closer the area value of the ROC curve to one, the better the diagnostic ability of the model.

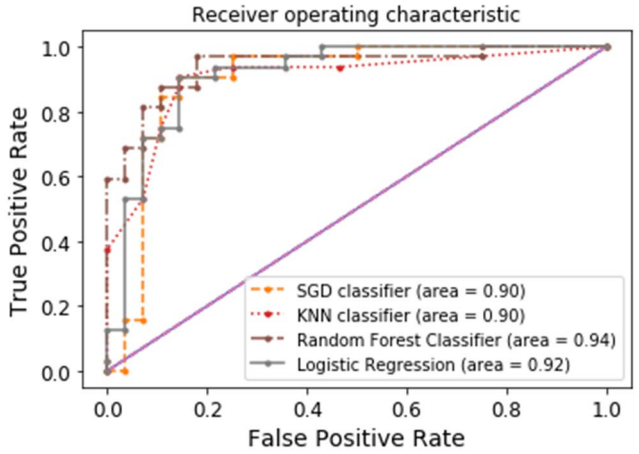

Figure 15: ROC curve for all models

Finally, the overall accuracy of this project after conducting the hard voting ensemble method came out to be 90% which is considered a fairly adequate accuracy that can be further built upon in the future.

VI. Conclusion

In conclusion, this paper presented a machine learning ensemble technique that combined multiple machine learning techniques in order to provide a more accurate and robust model for predicting the possibility of having a heart disease. The Ensemble model achieved 90% accuracy, which exceeds the accuracy of each individual classifier. The model can be used to assist doctors in analyzing patient cases in order to validate their diagnosis and help decrease human error.

References

[1] "Heart Disease Facts & Statistics," *Centers for Disease Control and Prevention*.[Online].Available:https://www.cdc.gov/heartdisease/facts.htm. [Accessed: 27-Apr-2019].

[2] Nhlbi, Nih. Anatomy of the Heart. 2011 [updated 2011 November 17; cited 2015 January 10]. Available from: http://www.nhlbi.nih.gov/health/health-topics/topics/hhw/anatomy

[3] "Cardiovascular diseases (CVDs)," *World Health Organization*, 26-Sep2018.[Online].Available:https://www.who.int/cardiovascular_diseases/en/. [Accessed: 27-Apr-2019].

[4] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[5] D. Chaki, A. Das, and M. Zaber, "A comparison of three discrete methods for classification of heart disease data," *Bangladesh Journal of Scientific and Industrial Research*, vol. 50, no. 4, pp. 293–296, 2015.

[6] R. G. Saboji, "A scalable solution for heart disease prediction using classification mining technique," *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 2017.

[7] El-Bialy, R., Salamay, M., Karam, O. and Khalifa, M. (2015). Feature Analysis of Coronary Artery Heart Disease Data Sets. *Procedia Computer Science*, 65, pp.459-468.

[8] Patel, J., TejalUpadhyay, D. and Patel, S., 2015. Heart disease prediction using machine learning and data mining technique. International Journal of Computer Science & Communication, 7(1), pp.129-137. DOI: 10.090592/IJCSC.2016.018.

[9] Ghosh, S. (2017). *Application Of Various Data Mining Techniques To Classify Heart Diseases*. [online] Pdfs.semanticscholar.org. Available at: https://pdfs.semanticscholar.org/dbe6/7e47cb35edc283cebd5cf06dd67faf1ad100.pdf [Accessed 13 Jul. 2019].

[10] A. Géron, Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. Beijing: OReilly, 2018.