# Identifying the Risk of Cardiovascular Diseases From the Analysis of Physiological Attributes

Nafis Mostafa*, Muhammad Anwarul Azim†, Md Rayhan Kabir‡ and Rasif Ajwad§
Department of Computer Science and Engineering, Brac University, Dhaka, Bangladesh
Email: *nafis.mostafa@g.bracu.ac.bd, †muhammad.anwarul.azim@g.bracu.ac.bd,
‡rayhan.kabir@bracu.ac.bd, §rasif.ajwad@bracu.ac.bd

***Abstract*—In the last few years, cardiovascular diseases have been increasing at an alarming rate and in most cases, this disease has not been detected at an early stage. In our study, we have analyzed some common physiological attributes to identify a pattern among the people having a cardiovascular disease which, in further, has been used to distinguish whether a person has a risk of developing cardiovascular disease or not. To enhance the performance of the algorithm models, we have generated a secondary dataset based on the output of the classification model, pushing the accuracy of the model to 97.03%. We have also evaluated the correlation of the attributes to the chance of having cardiovascular disease and found some general observation. Producing a secondary dataset, the analysis leading to the observable patterns among the attributes and, defining general observation for cardiovascular disease using machine learning models make this study unique.**

***Index Terms*—Cardiovascular Disease, machine learning, classification model.**

## I. Introduction

Cardiovascular disease(CVD) is the prime cause for death around the world, which amounts to an astounding number of about 18 million annually, according to the World Health Organization(2017). [1] Furthermore, this number is likely to increase even more as there is a significant rise in the level of obesity around the globe and the escalation is more common among middle and lower-income groups [2], [3]. So, now, one of the major challenges faced by the healthcare facilities is to provide proper and satisfactory treatment at an affordable cost. In America alone, the healthcare cost behind CVDs amounted to more than 32 billion USD and is predicted to hit 818 billion USD by 2030 [4]. Therefore, the hospitals must ensure affordable treatments and prevent poor clinical services as it can lead to a disastrous consequence.

In this study, we have aimed to analyze some of the common physiological attributes like age, sex, blood pressure, blood sugar etc, to identify whether a person has a risk of CVD or not. We have applied different classification, ensemble and deep learning models on the dataset to find a pattern among the people having CVD. To increase the performance of the model, we have created a secondary dataset with the help of the generated output of the classification model on a primary dataset. Finally, we have drawn two general conclusions from the statistical analysis on the mass population regarding cardiovascular disease.

## II. Literature Review

Many researchers have been drawn to physiology related dataset to find generalized rules or design high accuracy model for heart disease prediction. Proceeding on this track, different types of machine learning algorithm have been designed, tested and explained.

Detrano et al. [5] were one of the first researchers who tested classifying algorithms over the Cleveland dataset. Their regression-based algorithm was able to predict coronary artery diseases successfully with an accuracy of 77.0%. Kahramanli et al. [6] designed a classifying algorithm by combining Artificial Neural Network and Fuzzy Neural Network. This Hybrid Neural Network was able to classify heart disease data at 86.8% accuracy. Palaniappan et al. [7] developed the Intelligent Heart Disease Prediction System (IHDPS) with a view to finding hidden patterns from large health data. IHDPS was based on three data mining modeling techniques - Naïve Bayes Neural Network and Decision Trees whose overall accuracy were 86.12%, 86.12% and 80.40% respectively. More importantly, this system could draw the relationship between different medical attributes and heart disease along with a clear interpretation. Olaniyi et al. [8] tested backpropagation Neural network and Support Vector Machine (SVM), achieving the accuracy of 85% and 87.5% respectively.

Pouriyeh et al. [9] tested seven classifiers both in a separate and a collective manner. In the case of testing classifiers individually, the accuracy values were ranging from 69.96% to 84.1%. Furthermore, this highest accuracy value, which was obtained using SVM, remained the same even after using ensemble techniques like bagging or stacking with different combinations of classifiers.

D'Agostino et al. [10] have made another study regarding sex-specific heart diseases concluded that there is a positive correlation between coronary heart disease and different sex and cast. Yazid et al. [11] proposed a framework to tune the parameters of ANN and by using those parameters they analyzed two different datasets [11]. Their method observed a certain increase in performance to predict heart disease.

This review points the gradual improvement of classifying algorithms and motivates for proceeding to a similar study.

## III. Methodology

Our experiment has been done in two phases. At first, we have deployed different existing classifiers in order to identify

978-1-7281-7366-5/20/$31.00 ©2020 IEEE

whether the person has cardiovascular diseases or not. In the second phase, we have generated a secondary dataset using the output of all the algorithms and experimented them with the intermediate dataset to predict the chances of having CVD with a higher accuracy. The whole process has been shown in figure 1

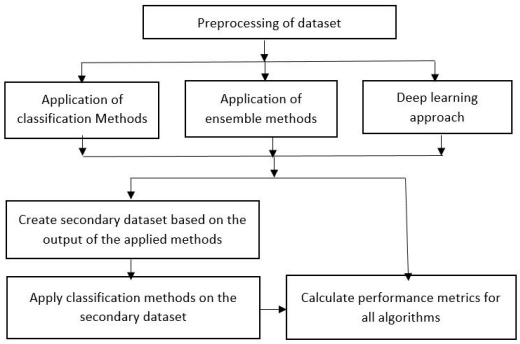


Fig. 1. Methodology

### A. Classification, ensemble and deep learning methods

As a first step, we have applied different classification methods. We have deployed common classification methods, one of which is the k nearest neighbors algorithm. [13]. Then we have applied decision tree classifier [14], followed by The random forest classification methods [15], support vector methods [16] and naive bayes classifier [17]. Furthermore, we have used different ensemble methods of classification like bagging, adaboost, gradient boost, bagging with extra tree, followed by implementing two different deep neural model using LSTM and Dense layers [18], [19] to analyze the dataset to identify the risk of developing cardiovascular diseases. In addition to this, we have calculated performance metrics for all the models and compared them.

### B. Creating secondary dataset

In the next step, we have created a secondary dataset based on the output of all the classification models which we have found in the first step. After generating the dataset, we have deployed the classification models again and measured the performance metrics with the secondary dataset and for each of the models, we have made our observation regarding the change in the performance by the techniques. Detailed discussion of the usage of the secondary dataset and how the models are trained and tested will be discussed in the result with secondary dataset section.

Finally, we have identified some general conclusion on the risk of having cardiovascular disease on the mass population.

## IV. Dataset and Experimentation

### A. Dataset Description

In our experiment, the dataset was collected from the UCI data repository platform [12]. This dataset contains 13 physiological attributes of 303 persons and the target value. The target value defines whether the person has heart disease or not.

The attributes are age, sex, resting blood pressure, fasting blood sugar, chest pain, serum cholesterol, electrocardiograph result, exercise-induced angina, maximum heart rate achieved, ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, the number of major vessels colored by fluoroscope, and Thal. Among these attributes, sex, Chest Pain, Number of major vessels colored by fluoroscope Exercise induced angina and Thai are categorical values. The rest of the attributes are numerical values.The standard deviation of the numerical values are shown in Figure 2.

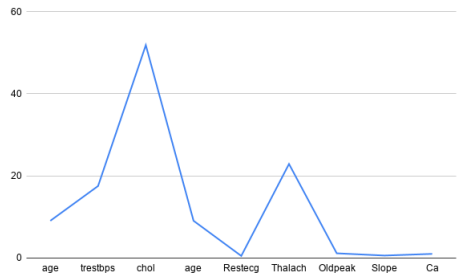


Fig. 2. St Deviation of the numerical values

### B. Experimentation

At first, all the categorical values were converted to numerical values before the experiment. 75 percent of the whole dataset was used to train the models and the rest are used to test the performance of the model. The dataset has been split randomly using the scikit learn library. Then we followed our methodology for the rest of the experiment.

## V. Result Analysis

### A. Output

The most common performance metrics of an algorithm are used to evaluate all the models. We have measured the accuracy as a primary metric for performance evaluation. Then, to have more insight, we have further measured the precision, recall and f1 score [20] for all the classification models.

### B. Result with the primary dataset

At first, we have applied KNN classifier to our dataset. To obtain the optimized value of K , we have used different values for k. The result for different values of K has been shown in figure 3. From this, we can see the maximum accuracy is acquired when the value of k is 7 or 8 and the accuracy is 86.84% Then we have applied the decision tree classifier and the random forest classifier. The accuracy of the decision tree classifier is 78.95% whereas in the random forest, the accuracy has been increased by 6.58%. Then we have applied the naive Bayes algorithm. It provides an accuracy of 82.89%. In addition to this, the support vector machine gives an accuracy of 85.53%.

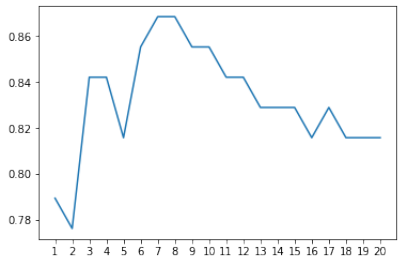


Fig. 3. Accuracy for different values of K

After this, the ensemble methods were applied and their performance has been measured. The bagging classifier provides an accuracy of 82.89%. The extra tree classifier provides an accuracy of 84.21%. The adaboost and gradient boost classifier provides an accuracy of 84.21% and 81.58% respectively.

TABLE I
ACCURACY OF DIFFERENT CLASSIFICATION MODELS TRAINED USING PRIMARY DATASET

| Algorithm | Accuracy |
|---|---|
| KNN | 86.84% |
| Random Forest | 85.53% |
| DecisionTree | 78.95% |
| Naive Bayes | 82.89% |
| SVMK | 85.53% |
| Bagging | 82.89% |
| ExtraTrees | 84.21% |
| Ada Boosting | 84.21% |
| Gradient Boosting | 81.58% |
| LSTM | 81.58% |
| ANN | 85.26% |

Finally, we have designed two different deep learning models using LSTM and dense layers. The first model consists of two LSTM layers and one dense layer with a dropout layer of 0.3. Each of the intermediate layers has 16 nodes. We have divided the data set in a batch size of 10 and run the algorithm for 100 epoch. An accuracy of 81.58% has been found for this model. The other model was designed using only the dense layers. Four dense layers were used for this model. This model has provided 84.21%. The accuracy of all the model has been summarized to Table I

### C. Result with Secondary dataset

After this, we have created a secondary dataset based on the output of the algorithms applied in the test case of the original dataset. Now, the newly generated dataset is split and 75% of it is used to train all the previously used machine learning classifiers except LSTM. Then another temporary dataset has been formed using the entire primary UCI dataset by first standardizing it and then passing it through the previously trained models. The temporary dataset is then used as the test case to generate the final output using the subsequent trained models. From the result of the secondary dataset, we have found an increase in performance for all the techniques. Among them, the accuracy of the support vector machine (SVM) and k-nearest neighbor (KNN) algorithm performances have been increased from 85.53% to 97.03% and 86.84% to 97.03% respectively. The accuracy of the decision tree algorithm, which was lowest in the primary dataset, has been increased from 78.95% to 92.41%.

TABLE II
ACCURACY OF DIFFERENT CLASSIFICATION METHODS TRAINED USING SECONDARY DATASET

| Algorithm | Accuracy |
|---|---|
| KNN | 97.03% |
| Random Forest | 96.70% |
| DecisionTree | 92.41% |
| Naive Bayes | 96.04% |
| SVMK | 97.03% |
| Bagging | 92.08% |
| ExtraTrees | 95.01% |
| Ada Boosting | 91.42% |
| Gradient Boosting | 94.06% |
| ANN | 86.4% |

### D. Overall Observation

From the overall observation of our experiment we found 2.63% instances of the dataset are wrongly classified by all the classification models in case of both the primary dataset and secondary dataset because from the dataset, we have found the attribute of those instances are almost similar to their opposite class and the rest 97.37% attributes of the dataset have been correctly identified by at least one of the classification model. In order to identify those 2.63% instances correctly, further investigation on the connectivity of the attributes should be made. In addition to this, we have also analyzed the correlation

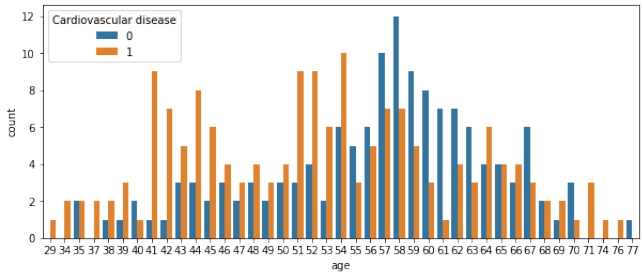


Fig. 4. Risk of cardiovascular disease with age

of the risk of cardiovascular disease with other factors. From the analysis, we have seen the risk of having CVD is highly correlated with chest pain and the slope of the peak exercise ST segment. We have also made an analysis of the relation of cardiovascular disease with age. From figure 4, we can see the ratio of having a cardiovascular disease is higher between the age of 29 to 54. After this range, the ratio of having cardiovascular diseases decreases. From this, we can say if someone doesn't have any cardiovascular disease within the age of 60, his risk of getting the disease decreases. This is because this type of disease greatly depends on the lifestyle and food habits of a person [21]. If someone maintains healthy lifestyle and food habit up to the age of 60 it is more unlikely to change after that age. The number of people diagnosed with

cardiovascular diseases increases up to age 58 and after that, the number of diagnosed people decreases.

We have also analyzed the relation of cardiovascular diseases between males and female which can be seen from figure5. In that figure, males and females have been expressed using 0 and 1 respectively. We have observed that, 75% of the males and 49.20% of the females have cardiovascular diseases. From this, we can say males have a higher risk of cardiovascular diseases than females. Finally to sum up,

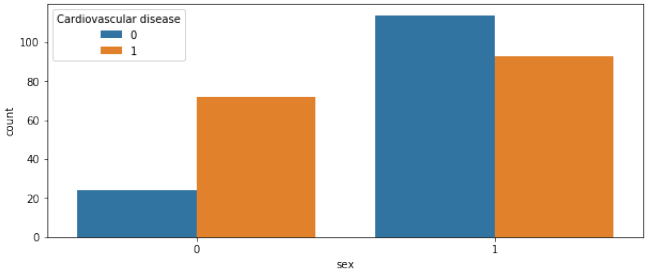


Fig. 5. Gender wise analysis of cardiovascular disease

we can say the risk of cardiovascular disease increases with the increase of age to a certain age. But if no cardiovascular disease is identified to a person within that age, then the chance of having cardiovascular disease in the future gets diminished even further. Along with this, the probability of developing cardiovascular disease is found to be higher in males than in females.

## VI. Conclusion

In this paper, we have made an analysis of the physiological attributes of people to identify the risk of cardiovascular disease. To analyze the dataset, we have applied different classification models, ensemble model and deep learning models and evaluated their performance. Then we have created a secondary dataset based on the output of the classification model on the primary dataset. Furthermore, we have applied the classification models on the secondary dataset and found that this approach greatly enhances the performance of the algorithms and also their precision of identifications. With the help of the models, we have discovered a general pattern among the attributes which have a greater risk of developing cardiovascular disease.

Finally, from the observation of the result and the dataset, we have identified a very small portion of the dataset which were failed to be identified by all of the classification models both in the primary and secondary datasets. To identify those instances accurately, more attributes are needed to be evaluated. Further study should be made to identify the attributes by which those instances can be classified with even greater accuracy and precision.

## References


[1] K. Samuelson, "Total Heart Disease Deaths on The Rise," News Center, 09-Sep-2019. [Online]. Available: https://news.feinberg.northwestern.edu/2019/09/total-heart-disease-deaths-on-the-rise/. [Accessed: 20-Feb-2020].

[2] S. Radcliffe, "Obesity Increasing: Solutions?," Healthline, 14-Nov-2017. [Online]. Available: https://www.healthline.com/health-news/obesity-rising-can-we-reverse-this-deadly-trend2. [Accessed: 21-Feb-2020].

[3] M. Lemstra, M. Rogers, and J. Moraros, "Income and heart disease: Neglected risk factor," Canadian family physician Medecin de famille canadien, Aug-2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4541436/. [Accessed: 25-Feb-2020].

[4] Giedrimiene, Dalia, et al. "Abstract 207: Burden of Cardiovascular Disease (CVD) on Economic Cost. Comparison of Outcomes in US and Europe."

[5] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," The American Journal of Cardiology, vol. 64, no. 5, pp. 304–310, Aug. 1989.

[6] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," Expert Systems with Applications, vol. 35, no. 1–2, pp. 82–89, Jul. 2008, doi: 10.1016/j.eswa.2007.06.004.

[7] S. Palaniappan and R. Awang, quot;Intelligent heart disease prediction system using data mining techniques,quot; 2008 IEEE/ACS International Conference on Computer Systems and Applications, Doha, 2008, pp. 108-115.

[8] E. O. Olaniyi, O. K. Oyedotun, and K. Adnan, "Heart Diseases Diagnosis Using Neural Networks Arbitration," International Journal of Intelligent Systems and Applications, vol. 7, no. 12, pp. 75–82, Nov. 2015, doi: 10.5815/ijisa.2015.12.08.

[9] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, quot;A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease,quot; 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 204-207.
.

[10] D'Agostino, Ralph B., et al. "Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation." Jama 286.2 (2001): 180-187.

[11] Yazid, Mohamad Haider Abu, et al. "Artificial Neural Network Parameter Tuning Framework For Heart Disease Classification." Proceeding of the Electrical Engineering Computer Science and Informatics 5.5 (2018): 674-679.

[12] Janosi, A., Steinbrunn, W., Pfisterer, M. and Detrano, R. (1988). UCI Machine Learning Repository: Heart Disease Data Set. [online] Archive.ics.uci.edu. Available at: https://archive.ics.uci.edu/ml/datasets/Heart+Disease [Accessed 26 Feb. 2020].

[13] Cunningham, Padraig, and Sarah Jane Delany. "k-Nearest neighbour classifiers." Multiple Classifier Systems 34.8 (2007): 1-17.

[14] Safavian, S. Rasoul, and David Landgrebe. "A survey of decision tree classifier methodology." IEEE transactions on systems, man, and cybernetics 21.3 (1991): 660-674.

[15] Oshiro, Thais Mayumi, Pedro Santoro Perez, and José Augusto Baranauskas. "How many trees in a random forest?." International workshop on machine learning and data mining in pattern recognition. Springer, Berlin, Heidelberg, 2012.

[16] Diehl, Christopher P., and Gert Cauwenberghs. "SVM incremental learning, adaptation and optimization." Proceedings of the International Joint Conference on Neural Networks, 2003.. Vol. 4. IEEE, 2003.

[17] Saritas, Mucahid Mustafa, and Ali Yasar. "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification." International Journal of Intelligent Systems and Applications in Engineering 7.2 (2019): 88-91.

[18] Rokach, Lior. "Ensemble-based classifiers." Artificial Intelligence Review 33.1-2 (2010): 1-39.

[19] Liu, Weibo, et al. "A survey of deep neural network architectures and their applications." Neurocomputing 234 (2017): 11-26.

[20] Goutte, Cyril, and Eric Gaussier. "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation." European Conference on Information Retrieval. Springer, Berlin, Heidelberg, 2005.

[21] Gupta, Rajeev, et al. "Correlation of regional cardiovascular disease mortality in India with lifestyle and nutritional factors." International journal of cardiology 108.3 (2006): 291-300.