# Further Study on the Identification of Predictive Capability of Classifiers for Early Heart Disease Detection Using Machine Learning

Aquila Peeran[1], Brinda U. Kumar[2], N. Neha[3*], Nikita Ravi[4]

[1,2,3,4]*Student, Department of Computer Science and Engineering, Dayananda Sagar College of Engineering, Bengaluru, India*

***Abstract*: Amongst all fatal diseases, cardiovascular diseases are considered the most prevalent. Due to the increase in workload, unhealthy diets and fast paced lifestyles, the younger generations have also fallen victim to heart complications. However, the early diagnosis of cardiovascular diseases can help in making decisions on lifestyle changes in high-risk patients and reduce the complications associated with it. The proposed system addresses this issue by using data collected by the Healthcare industries around the world and are used to effectively make predictions. The results from prediction of the system are used to prevent the disease and thereby reduce the cost for surgical treatments and other expensive tests associated with it. This prediction system aims to aid such individuals in addition to issues such as lack of physicians in rural areas and places with low quality of healthcare. By providing a prediction model for heart diseases at an early stage, this system helps reduce the cost of medical tests and the errors associated with it are also considerably reduced compared to manual testing. The added feature of instant diagnosis can be very useful in case of an emergency. The accuracy of machine learning algorithms is checked and the capability of deep learning classifiers are checked for cardiovascular disease identification and prediction along with a rigorous process of data mining to remove noisy data for a better decision making system with an extremely effective accuracy**

***Keywords*: Cardiovascular, Deep Learning, Prediction model, Data Mining.**

## 1. Introduction

We propose to implement a prediction system for the early detection of heart diseases. The Machine Learning classifiers and its capabilities are checked for the identification and prediction of cardiovascular diseases at an early stage. This helps reduce the cost of medical tests and prevents the inaccuracy of manual tests. Families with lower income cannot afford the cost of these expensive tests. The proposed work aims to aid such individuals in addition to issues such as lack of physicians in rural areas and places with low healthcare quality. The physical examination of the patient gives certain signs and symptoms that lead to the diagnosis of the presence of a heart Disease. Smoking, cholesterol level, family history of heart disease, obesity, blood pressure, and lack of physical exercise are some of the major factors that contribute to a cardiovascular disease. However, the early diagnosis of heart disease will help the patients with high risk to make suitable lifestyle changes in order to reduce the complications related to it. Large amounts of medical data with hidden information about the patient is collected by the health care industries around the world. This is useful in providing appropriate results and making effective decisions. The data collected can be used effectively to make predictions. The results of the predictions are used in order to prevent the disease thereby reducing the cost of surgical treatment and other expensive tests associated with it.

## 2. Methodology

The dataset used is the well verified UCI repository dataset. The proposed system is used for heart disease identification and prediction. The proposed system methodology has modules of Data preparation, Data mining, Splitting Data, Classifier Performance, Clinical Decision support system. The dataset undergoes a rigorous process of data mining in order to remove noisy data for a better decision making system with an extremely effective accuracy.

*1) Data preparation and exploration*

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data.

*2) Data cleaning*

Categorizing into sub groups: attributes like blood pressure, chest pain. Checking for null values, duplicate values.

*3) Data collection*

The features used in this dataset for the prediction of heart disease are:

*4) Data visualization*

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. On visualization the different attributes we found out that:

*Corresponding author: neha.nghode@gmail.com

*5) Target exploration*

- 56% with heart disease, 44% with no heart disease.
- It can be seen that, if the Target = 1 then the person has no heart disease and Target = 2 indicates a heart disease.

*6) Sex exploration*

- Sex=1 means a male patient, Sex=0 means a female patient.
- We can see that heart disease in male is more than in females.

*7) Age exploration*

- Min age of people who do not have heart disease: 29
- Max age of people who do not have heart disease: 76
- Min age of people who have heart disease: 35
- Max age of people who have heart disease: 77
- We can see that both young and old people are affected by heart diseases, but the probability of old people getting a heart disease is higher.

*8) Chest pain type (CP) exploration*

- We can see that most of the people with heart disease have asymptomatic chest pain.

*9) Electrocardiographic exploration*

Usually the people who do not have heart disease have normal electrocardiographic, whereas the people who have heart disease have probable or left ventricular hypertrophy.

*10) Attributes corelation*

- The number of major vessels (0-3) colored by flourosopy (ca) and the age.
- ST depression induced by exercise relative to rest (oldpeak) and the slope of the peak exercise ST segment (slope).
- The chest pain type (cp), exercise induced angina (exang).maximum heart rate (thalch) and the age.

## 3. Accuracy Results

The accuracy of each of the five algorithms namely Random Forest, Logistic Regression, Gradient Boosting, Artificial Neural Networks and K-Nearest Neighbor is calculated for the Heart Disease Prediction System. The accuracy of Random Forest was found to be 78% followed by Logistic Regression with an accuracy of 87%. Gradient Boosting algorithm got an accuracy of 74%. ANN resulted in 56% and KNN with 69%. Even though Logistic regression has the highest accuracy, it is not considered the most effective algorithm due to its misclassification. Hence, the second most accurate algorithm i.e. the Random Forest algorithm is considered the most effective algorithm amongst all.

## 4. Conclusion

In today's fast-paced society, both the young and the elderly have been profoundly impacted by current lifestyles and are at risk of developing cardiovascular disease; nevertheless, the risk of heart disease in the elderly is higher. The proposed working model helps in reducing treatment costs by providing Initial diagnostics in time. The model can also serve the purpose of training tools for medical students and will be a soft diagnostic tool available for physicians and cardiologists. By using the data collected effectively we can make predictions and the results can be used to prevent and thus reduce cost for surgical treatment and other expensive tests. Blood pressure, smoking, and cholesterol levels, have all been found to play a role in the prognosis of heart disease. The electrocardiographic readings of persons who do not have heart illness are usually normal, but those who do have heart disease exhibit likely or left ventricular hypertrophy. Further studies show that Heart Disease is more prevalent in males than in females. Through the project, Random Forest algorithm was found to be the most effective and accurate algorithm in the prediction of the presence of heart diseases in this particular system, opposed to Gradient Boosting, Logistic Regression, KNN and ANN algorithms, which can also be used in the prediction.

## Acknowledgement

We would like to thank our Professors, Bhavana B. R and Vindhya. M of Dayananda Sagar College of Engineering for offering their invaluable guidance and support throughout the duration of this project.

## References

[1] Amin Ul Haq et al, "Heart Disease Prediction System Using Model of Machine Learning and Sequential Backward Selection Algorithm for Features Selection", IEEE 5th I2CT 2019 Pun, 705 international conference, India, 2019.

[2] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chest X-ray 8: hospital- scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in Proceedings of IEEE CVPR 2017, Honolulu, HI, USA, 2017.

[3] Lamlili, E. N., Boutayeb, A., Derouich, M., Boutayeb, W., Moussi, A. Fish Consumption 710 Impact on Coronary Heart Disease Mortality in Morocco: A Mathematical Model with Optimal Control. Engineering Letters, vol. 24, no. 3, 2016.

[4] Shigei, Noritaka, et al."Land-Use Classification Using Convolutional Neural Network with Bagging and Reduced Categories." Proceedings of the International Multi Conference of Engineers and Computer Scientists 2019.

[5] Tu, M. C., D. Shin, et al. "Effective Diagnosis of Heart Disease through Bagging Approach." Biomedical Engineering and Informatics, IEEE, 2009.

[6] Enriko, I., Wibisono, G., Gunawan, D. Designing machine-to-machine (M2M) system in health-cure modeling for cardiovascular disease patients: Initial study. In Information and Com- munication Technology (ICoICT), 3rd International Conference pp. 528-532, IEEE (2015).

[7] K. Srinivas, K. Raghavendra Kao, and A. Govardham, "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," -Pronk, John M. Clymer and the Community Preventive Services Task Force (May 2017)

[8] Rivero, Jesus E., et al. "Thermal neutron classification in the hohlraum using artificial neural networks." Engineering Letters, vol. 23, no. 2, pp.87-91, 2015.

[9] Rivero, Jesus E., et al. "Thermal neutron classification in the hohlraum using artificial neu- ral networks." Engineering Letters 23.2 (2015): 87-91. Prediction of Parkinson Disease, 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing,14-16Dec.2018.

[10] Rukchart Prasertpong, Manoj Siripitukdet, "Rough Set Models induced by Serial Fuzzy Re- 730 lations Approach in Semigroups." Engineering Letters 27.1, 2019.