Ahmed Muntasir Hossain
Class of 2023
Computer Science
Research and Development of a Machine Learning Algorithm to detect Coronary Heart Disease
Mentor: Dr. Stephanie Gillespie
Department of Engineering and Applied Science

Heart disease, also referred to as Coronary Heart Disease (CHD), is defined as the condition when the heart does not receive an adequate supply of oxygen and nutrients due to the blood flow in the coronary arteries being restricted by plaque in them [1]. CHD has been the leading cause of death for the past 15 years in nearly all socioeconomic groups [2]. Currently, Machine Learning (ML) can be used to detect the disease in people by training computers to recognize patterns in medical data from patients affected and not affected by CHD.

The majority of ML algorithms have been tested for their ability to predict CHD using the Heart Disease Data Set, hosted by the UCI Machine Learning Repository [3]. Prior research using the dataset has investigated 14 attributes out of 75. However, in our research the entire dataset was used. There is a current lack of exploration of ML algorithms when using a large number of attributes. By using the entire dataset, along with state-of-the-art ML algorithms, we aimed to identify patterns to predict CHD [4].

The purpose of the study is to develop three machine learning classifiers that would predict heart disease with reliable accuracy, and to identify the optimal number of attributes required to produce a comprehensive diagnosis, using the Heart Disease Data Set. The research goals are 1) to determine the impact and characteristics of these medical attributes on CHD, and 2) to evaluate the accuracy of the different models being implemented.

The dataset was cleaned and irrelevant features were extracted converting the dataset from 75 attributes to 25. It was then uploaded on Weka and 12 classifiers, including Support Vector Machine, K-Nearest Neighbor, and Decision Trees, were applied to establish baseline accuracy. We tested four algorithms for feature selection and selected ReliefF as it was the most reliable algorithm for ranking the attributes. The attributes were divided into three feature sets consisting of the top 5, 10, and 15 attributes ranked by the algorithm.

All the initial classifiers were applied to the three feature sets to produce the initial results before optimization. The top six classifiers were then selected to be optimized by tuning their hyperparameters. The best tuned classifiers were chosen for the final results after optimization from which the top three classifiers were selected. These classifiers were applied onto the validation set of 30 instances to get the final absolute results. The order of the classifiers with their corresponding number of attributes in the validation set were as follows: Naïve Bayes (10), SVM (10) and Logistic regression (5).

During testing, we calculated six different performance measures; the most important one, however, was the F-measure as it considered both precision and recall. The F-measure values for the top three classifiers were 0.901, 0.864, and 0.834. The most important attributes

that predicted heart disease, as indicated by ReliefF, were thalassemia, sex, number of major blood vessels colored by fluoroscopy, chest pain type, and exercise induced angina.

These promising results have implications for lower medical expenses for patients as well as a lower rate of medical misdiagnosis of CHD [5]. To accomplish these, the most important next steps require testing the models with multiple tuned hyperparameters on more recent and larger datasets to compare their performance. The completion of these steps would allow diagnosticians to detect CHD at an earlier stage and treat it.

References

[1] I. Criteria, "Ischemic Heart Disease", *Ncbi.nlm.nih.gov*, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK209964/.

[2] "The top 10 causes of death," *World Health Organization*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

[3] *UCI Machine Learning Repository: Heart Disease Data Set*. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Heart Disease.

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

[5] Y. Yan, J.-W. Zhang, G.-Y. Zang, and J. Pu, "The primary use of artificial intelligence in cardiovascular diseases: what kind of potential role does artificial intelligence play in future medicine?," *Journal of geriatric cardiology : JGC*, Aug-2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6748906/.