Research and Development of Machine Learning Algorithms to detect Coronary Heart Disease

Ahmed Muntasir Hossain
University of New Haven
West Haven, United States of America
ahoss1@unh.newhaven.edu

Stephanie Gillespie
Tagliatela College of Engineering
University of New Haven
West Haven, United States of America
sgillespie@newhaven.edu

***Abstract*—This paper investigates and analyzes the results from six different machine learning algorithms, before and after hyperparameter tuning, on predicting heart disease using the Cleveland Heart Disease Data Set from the UCI repository. The algorithms were tested and trained using Weka, an open source machine learning software. The dataset underwent feature selection using the Relief-F algorithm to obtain the most optimal subset of attributes. The results were used to determine the top three classifiers that would provide a comprehensive diagnosis using the minimum number of attributes. Naïve Bayes, Support Vector Machine and Logistic Regression outperformed all the other classifiers using 10, 5, and 10 attributes, achieving F-measure accuracies of 0.901, 0.864, and 0.834, respectively.**

Keywords—Machine Learning, SVM, Heart Disease Data Set

I. Introduction

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 or blocked. This is most often due to the buildup of plaque in them. CHD results in angina (chest pain) and with further narrowing of the arteries may cause a heart attack [1].

CHD has been the leading cause of death for the past 15 years. It was responsible for 15.2 million deaths out of 56.9 million worldwide in 2016, and has been the leading cause of death in nearly all economy-income countries [2] because of the high rate of misdiagnosis and medical expenses for patients [3]. In 2017, the percentage of total deaths due to heart disease was 23.5% in the US alone [4]. These severe numbers support the urgency in developing cost effective means of detection and prevention of heart disease.

Currently with advances in technology, new methods for detection of CHD are being implemented. Machine learning (ML) can be used to detect the disease in people by training computers to recognize patterns in patient data and medical profiles from patients affected and not affected by heart disease. The attributes currently studied, such as age, sex, number of vessels colored, resting ECG and many more, do not require extensive and expensive tests [5], and could lead to the patient receiving a dependable diagnosis quickly. ML could also provide the possibility of early detection of heart disease with reasonable precision in patients depending solely on the physical symptoms.

This paper analyses and compares the results from six different ML algorithms including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Decision Trees before and after hyperparameter tuning. The algorithms were trained and tested on the Cleveland Heart Disease Data Set from the UCI repository using 10-Fold Cross Validation as the model validation technique. The results were used to determine the top three classifiers that would provide a comprehensive diagnosis (presence or absence of heart disease) using the minimal number of attributes.

II. Related Works

In recent research, many ML algorithms have been tested in their ability to predict CHD upon the Heart Disease Data Set, hosted by the UCI Center for Machine Learning and Intelligent Systems [6]. The dataset consists of 75 medical attributes in each of the four subset databases it contains; however only 14 attributes in one of the databases (Cleveland) are usually used by researchers. This is due to there being fewer missing attributes and more records in the Cleveland database compared to other databases in the dataset [7]. Conclusions by Palaniappan and Awang suggest that incorporating those additional attributes and databases in the prediction systems could increase accuracy in predicting CHD [8].

The ML algorithms primarily used by researchers were Naïve Bayes (NB), Artificial Neural Network (ANN), and Decision Trees [8-12]. However, multiple other types of ML algorithms have been developed - Support Vector Machine, K-Nearest Neighbor - that could be promising based on their success detecting other diseases such as diabetes and chronic kidney failure [12-14]. There is a current lack of exploration of these ML algorithms when using a larger number of attributes, and the effect it may have on overall accuracy. By using the larger datasets with state of the art ML algorithms, we may be able to better learn patterns to predict CHD, such as characteristics of patients with CHD and the impact and relationship of medical attributes and CHD [8].

III. Research Methodology

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 Cleveland Heart Disease Data Set. The research questions being investigated here are to determine the impact and characteristics of these medical attributes on CHD, and the accuracy of the different models being implemented.

A. Data Preprocessing

The Cleveland Heart Disease Data Set contains 303 instances and 75 attributes excluding the class label. The attributes include age (ranging from 29 to 77 years), sex (91 females and 191 males), chest pain type, thalassemia, exercise induced angina, resting blood pressure, resting electrocardiographic results, and more. The class label "num" is the diagnosis for heart disease and is indicated by value 0 (157 instances) - absence of CHD if there is less than 50% narrowing in all major blood vessels – and value 1 (125 instances) – presence of CHD if there is more than 50% narrowing in any major blood vessel.

Initially, the dataset was uploaded and cleaned in Excel. Irrelevant attributes unrelated to CHD that were present in the dataset were removed to prevent overfitting. Such attributes would not help the algorithm to predict CHD and instead may reduce its performance when tested on newer datasets. Furthermore, attributes missing approximately 20% or more data were extracted, as the data would otherwise be unreliable. This reduced the dataset from 75 attributes to 25. The dataset was then divided into two sections: "testing and training set", and "validation set". The "testing and training set" consisted of 243 instances and the "validation set" consisted of 30 instances. The "validation set" was kept separate to be used during the final testing of the top three classifiers.

B. Baseline Values for Classifiers

The "testing and training set" was uploaded onto Weka, an open source machine learning software. 10-Fold Cross Validation was implemented during testing and training to generate average and unbiased results compared to a test/train split. 12 classifiers were selected due to their common usage such as Support Vector Machine, K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Random Forest (RF), PART, J48, Linear Regression, Logistic Regression (LR), and more. The results were used to establish a baseline accuracy for the classifiers before feature selection and hyperparameter optimization as shown in Fig. 1.

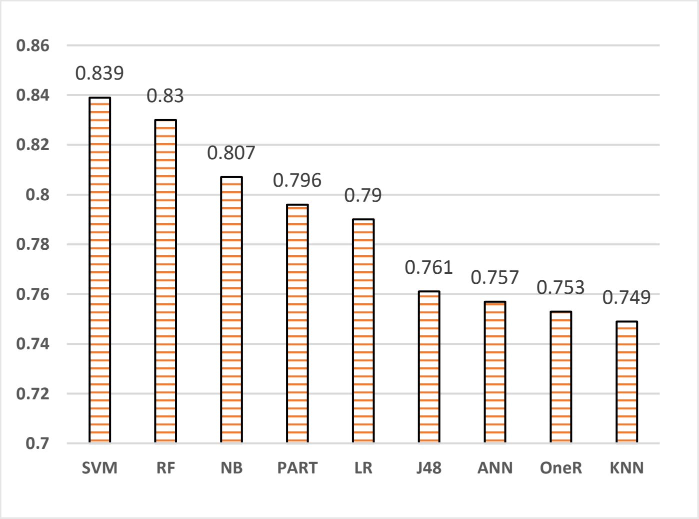

Fig 1. Baseline F-Measure Values of the Initial Classifiers

C. Feature Selection

Four algorithms for feature selection were tested including Information Gain, Gain Ratio, Relief-F, and Principal Components. Relief-F was selected due to common use of the algorithm in prior work by other researchers, and due to its ability to rank attributes based on their significance in distinguishing between two classes. It does so by sampling instances and measuring the distance from the selected instance to the nearest instance of the same (near-hit) and different (near-miss) class. If the difference in the attribute value between the sampled instance and the near-miss instance is larger than the difference in the attribute value between the sampled instance and near-hit instance, the weight of the attribute increases and vice versa.

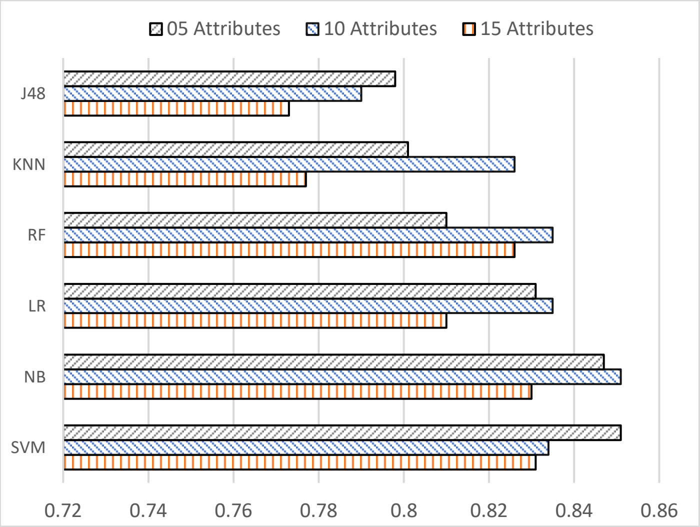

Fig 2. Comparison of the F-Measure Values of the six chosen Classifiers across the three Feature Sets after Feature Selection.

For the dataset, the algorithm assigned weights to each of the attributes. A vector of weights was formed which was then divided by the number of iterations, m. This is known as the relevance vector. After computing the relevance vector, the attributes were chosen based on a threshold, r. That is, if the weight was greater than the threshold r then the attributes were selected. The attributes were then ranked by the algorithm and the top 5, 10, and 15 attributes were selected to form three feature sets.

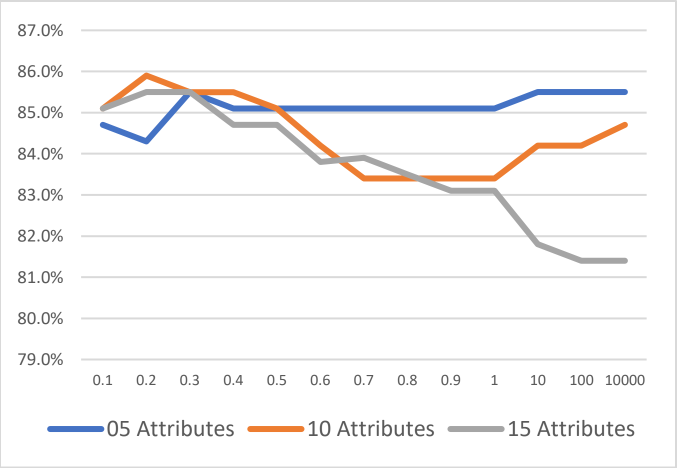

Fig. 3 Trend in F-Measure value of SVM for different values of the Complexity Parameter (C) across the three Feature Sets.

D. *Preliminary Results before Hyperparameter Optimization*

All the initial classifiers from the baseline values were applied to the three feature sets to produce the initial results before optimization as shown in Fig. 2. The top six classifiers were then selected to be optimized by tuning their hyperparameters. The non-optimized results of the selected classifiers were compared with their optimized values after hyperparameter tuning to determine the most reliable classifiers.

E. *Hyperparameter Optimization of the Selected Classifiers*

The top six classifiers were tuned by changing a single parameter value per classifier. The parameter value to be hypertuned was selected based on previous research and common use in the field. For instance, the complexity parameter when training SVM, as shown in Fig. 3, the number of randomly selected features when training random forest, the number of iterations when training logistic regression, and the minimum number of instances per leaf when training J48 were varied as their respective parameter values.

The tuning value and performance measure of each classifier was obtained by training and testing them on the three differently sized feature sets. The best hypertuned value of each of the classifiers for each feature set were selected as displayed in Fig. 4. This was done to compare and contrast the optimal parameter value for those classifiers in those feature sets.

After hyperparameter optimization, the top three classifiers were selected across all the feature sets and applied onto the validation set to obtain the final results.

IV. Experimental Results and Analysis

A. *Results before Optimization*

As can be seen from Fig. 1, the F-measure values of the single classifiers ranged from 0.749 to 0.839 before any optimization, using 25 attributes. From the results it was observed that the F-measure values of SVM, Random Forest, and Naïve Bayes were above 0.8 and performed better than the remaining classifiers.

After establishing the baseline, feature selection was performed from which the top six classifiers were selected as seen in Fig. 2. Four classifiers showed the most improvements in their F-measure values using 10 attributes. Naïve Bayes and SVM performed the best followed by Random Forest and Logistic Regression.

The classifiers were optimized by tuning one hyperparameter per classifier, as seen in Fig. 3 with SVM. The complexity parameter (C) was changed to the following values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 10, 100, 10000. This provided us with a wide range of numbers and gave insight into the performance of SVM for large values of c. It was observed that usually for smaller values of C (0.1 – 0.4) for any feature set, the classifier performed optimally.

B. *Final Results after Optimization*

After comparing the accuracies of each of the hypertuned classifiers across all the three feature sets, it was seen that three classifiers (Naïve Bayes, Logistic Regression, and Support Vector Machine) outperformed the other classifiers. The parameter and attribute numbers (feature sets) were then compared for each of the chosen classifiers, and the classifiers with the best performance results and fewest number of attributes were chosen. The results were as follows: SVM with a complexity parameter (C) of 0.2 using 10 attributes, Logistic Regression with a maximum number of iterations of 1 using 5 attributes and Naïve Bayes without any optimization using 10 attributes.

C. *Results from the Validation Set*

The three optimized classifiers were tested on the validation set of 30 instances and their results are displayed in Table I. Naïve Bayes performed the best with an accuracy of 90.00% and F-measure of 0.901, followed by Support Vector Machine with an accuracy of 86.67% and F-measure of 0.864, and lastly Logistic Regression with an accuracy of 83.33% and F-measure of 0.834.

During testing, we calculated six different performance measures: simple accuracy, kappa statistic, precision, recall, F-measure, and ROC area. Each performance measure gave insight as to how well the algorithms performed, and these measures were necessary so that we could compare algorithms that produced approximately similar results. However, the most important metric was F-measure because it is a mathematical expression that inputs both the values of precision and recall. Precision identifies the fraction of predicted positive results which are actually positive, and recall identifies the fraction of actual positive results which have been predicted correctly. This gave us a complete understanding of how well the algorithms were able to classify both the classes, in contrast to only one of the classes with the other performance measures.

Additionally, we were able to determine the most important attributes which predicted heart disease were thalassemia (an inherited blood disorder where the body does not produce natural levels of hemoglobin), sex, number of major blood vessels colored by fluoroscopy, chest pain type, and exercise induced angina (a type of chest pain that is caused due to lack of blood flow to the heart).

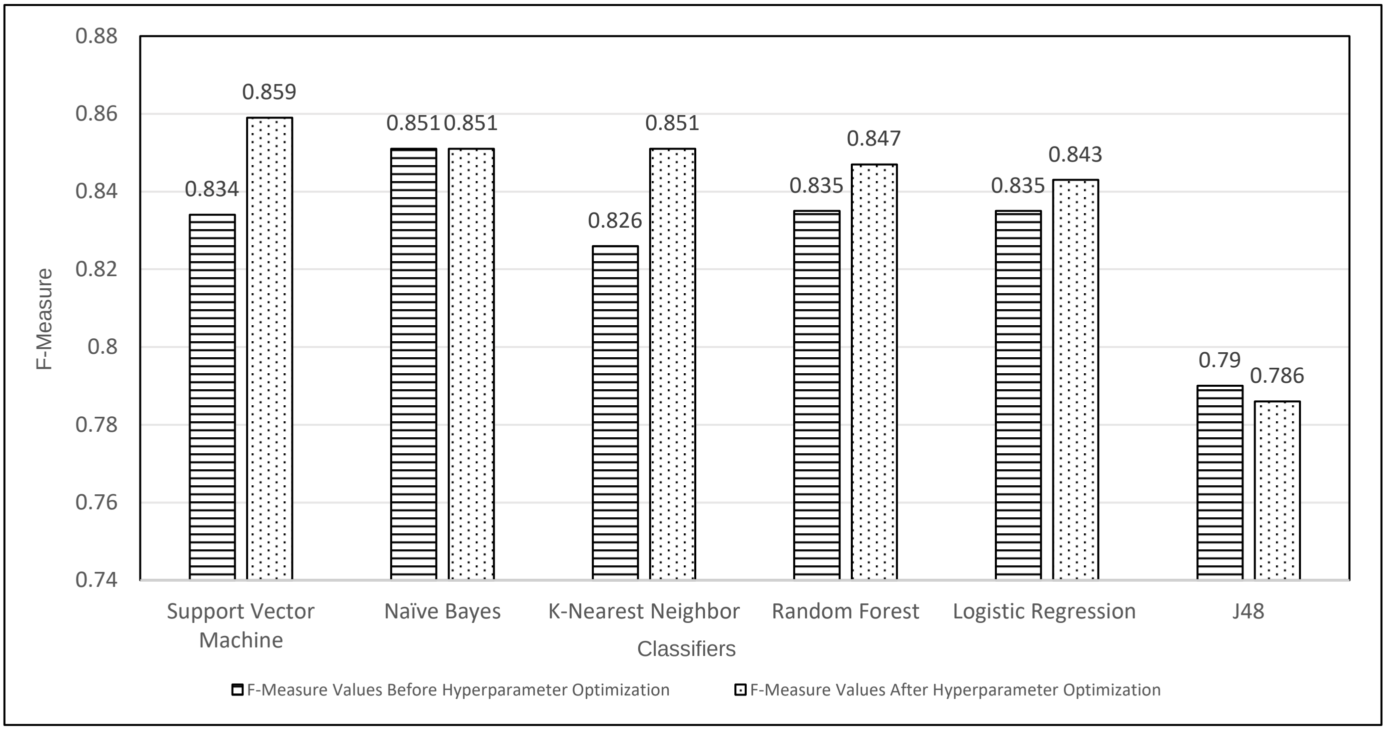

Fig. 4. F-Measure Values of Classifiers before and after Hyperparameter Optimization on the 10 Attribute Feature Set during Testing Phase

TABLE I
PERFORMANCE MEASURES OF THE TOP THREE CLASSIFIERS ON VALIDATION SET

Number	Hypertuned Classifiers	Simple Accuracy	Kappa Statistic	Precision	Recall	F-Measure	ROC Area	Attributes
1	Naïve Bayes	90.00%	0.7945	0.903	0.900	0.901	0.963	10
2	Support Vector Machine	86.67%	0.7143	0.870	0.867	0.864	0.847	10
3	Logistic Regression	83.33%	0.6575	0.837	0.833	0.834	0.905	5

This was indicated by the Relief-F algorithm, as it ranked the attributes based on their predictive ability of CHD.

Work by Pouriyeh et al. on the Heart Disease Data Set shows that they obtained accuracies of 84.15% and F-measure values of 0.860 for SVM, and 83.49% and 0.851 for Naïve Bayes using a single classifier. They also implemented boosting techniques which improved their accuracy of SVM to 84.81% [12]. Furthermore, research by C. B. C. Latha et al. shows that they were able to reach accuracies of 83.17% using Naïve Bayes (single classifier) and approximately 84% using bagging techniques. Similarly, there was a 0.99% increase with boosting techniques [9]. Our results are in a similar range and from previous research it can be concluded that the performance measures can be improved further by using ensemble techniques.

V. CONCLUSIONS

These promising results suggest the future ability to lower medical expenses for patients and rate of medical misdiagnosis of CHD [3]. To accomplish these, the most important next steps require testing the models on more recent and larger datasets since the Heart Disease Data Set is well-established but not recent, created in July, 1988, and the validation set consisted of only 30 instances. This would allow us to obtain more reliable results based on medical attributes affecting the current population and different demographic. Furthermore, the classifiers should be optimized further by tuning more hyperparameters per classifier, compared to only one as done in this research. Lastly, ensemble techniques such as bagging, and boosting have been used in prior research, as seen in Section IV, and can be implemented here to improve the classifiers. The completion of these steps would allow diagnosticians to detect CHD at an earlier stage and prevent 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] 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/.

[4] M. Heron, "Deaths: Leading Causes for 2017", National Vital Statistics Reports, 2019.

[5] B. Kolukisa et al., "Evaluation of Classification Algorithms, Linear Discriminant Analysis and a New Hybrid Feature Selection Methodology for the Diagnosis of Coronary Artery Disease," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 2232-2238.

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

[7] M. I. Al-Janabi, M. H. Qutqut, and M. Hijjawi, “Machine Learning Classification Techniques for Heart Disease Prediction: A Review,” *International Journal of Engineering & Technology*, vol. 7, pp. 5373–5379.

[8] 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.

[9] C. B. C. Latha and S. C. Jeeva, “Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques,” *Informatics in Medicine Unlocked*, vol. 16, p. 100203, 2019.

[10] A. M. Mahmood and M. R. Kuppa, "Early Detection of Clinical Parameters in Heart Disease by Improved Decision Tree Algorithm," 2010 Second Vaagdevi International Conference on Information Technology for Real World Problems, Warangal, 2010, pp. 24-29.

[11] P. Singh, S. Singh, and G. S. Pandi-Jain, “Effective heart disease prediction system using data mining techniques,” *International Journal of Nanomedicine*, vol. Volume 13, pp. 121–124, 2018.

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

[13] S. Tayeb et al., "Toward predicting medical conditions using k-nearest neighbors," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, 2017, pp. 3897-3903.

[14] S. Karatsiolis and C. N. Schizas, "Region based Support Vector Machine algorithm for medical diagnosis on Pima Indian Diabetes dataset," 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, 2012, pp. 139-144.