A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease

Seyedamin Pouriyeh*, Sara Vahid*, Giovanna Sannino†, Giuseppe De Pietro†, Hamid Arabnia*, Juan Gutierrez*

*Department of Computer Science, University of Georgia, Athens, USA
{pouriyeh, svahid, hra, jgutieri}@uga.edu
†Institute of High Performance Computing and Networking (ICAR - CNR), Naples, Italy
{giovanna.sannino, giuseppe.depietro}@icar.cnr.it

Abstract—This paper aims to investigate and compare the accuracy of different data mining classification schemes and their combinations through Ensemble Machine Learning Techniques for predicting heart disease. The Cleveland dataset for heart diseases, containing 303 instances, has been used in this study. Due to the limited number of samples, 10-Fold Cross-Validation is applied in order to portion the data into training and testing datasets. In this research, different machine learning classifiers such as Decision Tree (DT), Naïve Bayes (NB), Multilayer Perceptron (MLP), K-Nearest Neighbor (K-NN), Single Conjunctive Rule Learner (SCRL), Radial Basis Function (RBF) and Support Vector Machine (SVM) are utilized. Moreover, the ensemble prediction of classifiers, including bagging, boosting, and stacking are applied to the dataset. The result of the experiment indicates that SVM method using boosting technique outperformed among the aforementioned methods.

Index Terms—Machine Learning Techniques, Heart Disease Classification, Decision Support Systems.

I. INTRODUCTION

Heart disease has been the most significant cause of death in the world during the past 10 years [1]. The use of heart monitoring systems, such as mobile monitoring [2], and heart disease classification methodologies for decision support systems has been increasing accordingly. Unfortunately, many different factors can influence and complicate the detection of possible heart anomalies and can result in an inaccurate diagnosis or in a delay in a correct diagnosis. According to [3], due to the many and uncertain risk factors, sometimes heart disease diagnosis is difficult even for experts, who frequently require accurate tools that consider all these risk factors and give a clear result in a specific time period.

Motivated by the need to acquire such an indispensable instrument for diagnosis and by the importance of avoiding any unwanted biases, errors and excessive medical costs that might affect the quality of treatment provided to patients [4], many researchers have tried to find the most accurate machine learning techniques to discover the relationships between different heart disease and patient attributes in order to assist physicians [5], [6].

This paper provides a comparison of different machine learning classification techniques, such as Decision Tree (DT), Naïve Bayes (NB), MultiLayer Perceptron (MLP), K-Nearest Neighbor (K-NN), Single Conjunctive Rule Learner (SCRL), Radial Basis Function (RBF) and Support Vector Machine (SVM), and their use in combination, through bagging, boosting and stacking on a heart disease dataset. The dataset used is the Cleveland Heart Disease dataset taken from the University of California, Irvine (UCI) learning dataset repository, donated by Detrano.

The rest of the paper is organized as follows. Related works are presented in section II. In section III, the research methodology is described. In section IV, the experimental results and analysis are presented and discussed in detail. In section V, a comparison of data mining techniques is provided. The study is finally concluded in section VI.

II. RELATED WORKS

In the last few years, several studies have been dedicated to an evaluation of the classification accuracies of different classification algorithms applied to the Cleveland heart disease dataset [7] freely available at an online data mining repository of the UCI. Since its creation, this dataset has been used by many researchers investigating different classification problems with various classification algorithms.

Detrano in [8] used a logistic regression algorithm and obtained a 77.0% classification accuracy.

In [9], the author worked on the Cleveland dataset with a focus on comparing global evolutionary computation approaches, and observed some prediction performance improvements when using a new approach. However, the performance of his proposed technique is dependent on the attributes selected by the algorithm.

Gudadhe et al. [5] realized an architecture base with both the MLP network and the SVM approach. This architecture achieved an accuracy of 80.41%. On the other hand, Humar Kahramanli and Novruz Allahverdi [10] obtained an accuracy of 87.4% by using a hybrid neural network that combines a fuzzy neural network (FNN) with an artificial neural network (ANN).

In [11] an Intelligent Heart Disease Prediction System (IHDPS) was presented. The IHDPS system uses data mining
techniques including DT, NB and Neural Network (NN). The tests performed by the authors indicated that the NB model achieved the best performance in terms of correct predictions (86.12%). The second best model was NN with an accuracy of 86.12% and the third, DT, with a score of 80.4% of correct predictions.

In our current research, our goal has been to highlight a comparison of almost all the aforementioned techniques that have been employed in previous studies and their combinations, in order to be able to select the most appropriate prediction method.

III. Research Methodology

A. The Cleveland Dataset

The dataset used in the current research contains 303 instances with a total number of 76 attributes. However, the majority of the studies use a maximum of 14 attributes [11] as these are closely linked to heart disease [12]. The features included are age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise induced angina, oldpeak, slope, number of vessels colored and thalassemia, respectively. The class has two values, “False” and “True”, corresponding to the absence or presence, respectively, of any heart disease.

B. Data Portioning Technique

Due to the limitation in the amount of data, the K-Fold Cross Validation technique has been used for data portioning. The main reason for choosing this technique is that it has a lower variance in comparison with other estimators like the single hold-out approach. In addition, extensive tests on numerous different datasets with different techniques have also shown that K=10 is about the right number of folds to obtain the best estimate of accuracy [13].

C. Machine Learning Techniques

In the current study, seven classifiers, namely DT, NB, MLP, RFB, SCRL, K-NN and SVM, and combinations of these classifiers, using ensemble learning methods such as bagging, boosting and stacking, are discussed. In each scenario, the performance is calculated using the standard metrics, namely accuracy, precision, recall and F-measure. In addition, the Receiver Operation Characteristic (ROC) curve area has been employed to compare the performance of each classifier.

1) Decision Tree (DT): A Decision Tree is a flow chart-like structure that includes a root node, branches, and leaf nodes. The dataset attributes are defined through the internal nodes. The branches are the outcome of each test against each node. It is a popular classifier because it is simple, fast, and easy to interpret, explain and implement. It requires no domain knowledge or parameter setting.

2) Naive Bayes (NB): The Naive Bayes classifier is based on Bayes Theorem. In this classifier the independency between the attributes of the dataset is the main assumption and the key point in order to make a prediction. It is easy to implement and particularly useful for very large datasets. In addition to its simplicity, this model is shown to outperform even highly sophisticated classification methods

3) Multilayer Perceptron (MLP): The MLP classifier contains multiples of nodes arranged in layers. This makes a directed graph which covers input, hidden, and output layers and each layer is fully connected to the next. This classifier has a clear architecture and simple algorithm; therefore, it is one of the most famous neural network models.

4) K-Nearest Neighbors (K-NN): K-Nearest Neighbors classifies an object by the majority vote of its closest neighbors. In other words, based on some distance metrics, the class of a new instance will be predicted. The distance metric used in nearest neighbor methods for numerical attributes can be a simple Euclidean distance.

5) Single Conjunctive Rule Learner (SCRL): This method aims to infer a set of rules from a dataset that captures all the generalizable knowledge within those data. The rules in this model can be various and are usually in the order from the most general to the most specific.

6) Radial Basis Function (RBF): Similar to MLP, RBF is also another popular classifier based on ANN. Both MLP and RBF are non-linear layered networks which use universal approximation properties. In comparison with MLP, RBF networks take a slightly different approach in terms of the number of hidden layers, linearity of the output layer, local and global approximations, and number of parameters.

7) Support Vector Machine (SVM): Support Vector Machine models are defined as finite-dimensional vector spaces in which each dimension represents a ‘feature” of a particular object. It has been shown to be an effective approach in high-dimensional space problems. Due to its computational efficiency on large datasets this technique is usually used in document classification and sentiment analysis.

8) Ensemble Learning Methods: This is an aggregation of the prediction of several classifiers. It has three different models: bagging, boosting and stacking [13]. Bagging is the combining of predictions of exactly the same type by voting. Boosting is the same as bagging except that the performance of previous models has an effect on the new models. Stacking is the combining of models of different types.

IV. Experimental Results and Analysis

The results of applying the seven aforementioned methods (DT, NB, K-NN, SVM, MLP, RFB and SCRL, individually and in combination) using 10-Fold Cross Validation are presented in this section. The main performance measurements used in this study are Precision, Recall, F-Measure and ROC. Precision (also called positive predictive value) is a measure of result relevancy, while recall (also known as sensitivity) is a measure of how many truly relevant results are returned. As both measures, precision and recall, are important, we usually evaluate our classifier with the F-measure, which is the combination of precision and recall. ROC is another technique to examine and visualize the performance of a classifier. The ROC graph shows the trade-off between the ability of a classifier to correctly identify positive cases.
TABLE I
STANDARD METRICS FOR 10-FOLD CROSS VALIDATION TECHNIQUE

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>ROC Area</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree (DT)</td>
<td>0.774</td>
<td>0.830</td>
<td>0.801</td>
<td>0.800</td>
<td>77.55</td>
</tr>
<tr>
<td>Naive Bayes (NB)</td>
<td>0.836</td>
<td>0.867</td>
<td>0.851</td>
<td>0.904</td>
<td>83.49</td>
</tr>
<tr>
<td>K Nearest Neighbor (K-NN, K=1)</td>
<td>0.782</td>
<td>0.782</td>
<td>0.782</td>
<td>0.752</td>
<td>76.23</td>
</tr>
<tr>
<td>K Nearest Neighbor (K-NN, K=3)</td>
<td>0.821</td>
<td>0.836</td>
<td>0.829</td>
<td>0.838</td>
<td>81.18</td>
</tr>
<tr>
<td>K Nearest Neighbor (K-NN, K=9)</td>
<td>0.848</td>
<td>0.842</td>
<td>0.845</td>
<td>0.898</td>
<td>83.16</td>
</tr>
<tr>
<td>K Nearest Neighbor (K-NN, K=15)</td>
<td>0.847</td>
<td>0.836</td>
<td>0.841</td>
<td>0.904</td>
<td>82.83</td>
</tr>
<tr>
<td>MultiLayer Perceptron (MLP)</td>
<td>0.824</td>
<td>0.824</td>
<td>0.824</td>
<td>0.894</td>
<td>82.83</td>
</tr>
<tr>
<td>Radial Basis Function (RBF)</td>
<td>0.845</td>
<td>0.861</td>
<td>0.853</td>
<td>0.892</td>
<td>83.82</td>
</tr>
<tr>
<td>Single Conjunctive Rule Learner (SCRL)</td>
<td>0.734</td>
<td>0.703</td>
<td>0.718</td>
<td>0.707</td>
<td>69.96</td>
</tr>
<tr>
<td>Support Vector Machine (SVM)</td>
<td>0.827</td>
<td>0.897</td>
<td>0.860</td>
<td>0.836</td>
<td>84.15</td>
</tr>
</tbody>
</table>

and the number of negative cases that are incorrectly classified.

1) First Experiment: In our first experiment we used the whole dataset and applied the DT, NB, MLP, SVM, K-NN, RBF and SCRL classifiers while using the 10 fold cross validation technique. For the K-NN classifier, we experimented with K=1,3,9 and 15 and concluded that the best performance for K was 9. The accuracy of K-NN with different numbers of K is depicted in figure 1. Additionally, for the MLP method, we tried a combination of different numbers for the input, hidden and output and finally arrived at the best combination, namely 13 inputs, 7 hidden layers and 2 outputs. This combination allowed us to achieve the best results.

![Fig. 1. The accuracy of K-NN based on different values of K.](image1)

As figure 2 shows, DT has a 77.55% accuracy, NB 83.49%, K-NN 83.16%, MLP 82.83%, RBF 83.82%, SCRL 69.96% and SVM 84.16%. Among them, SVM reaches the highest percentage of accuracy while RBF, NB and K-NN are in the next places. SCRL among all single classifiers has the lowest accuracy of 69.96%. Table I gives more details about the standard metrics of accuracy for this experiment.

2) Second Experiment:

- **Bagging:** Based on the experimental results, we explored the ability of bagging techniques to improve the estimation performance. As figure 3 shows, bagging improved the accuracy of some of the classifiers. For instance, DT improved from 77.55% to 78.54% and, more significantly, SCRL increased from 69.96% to 80.52%. SVM did not improve but, nonetheless, maintained its accuracy level.

- **Boosting:** Boosting means applying a weak classifier, running it multiple times on the training data and then allowing the learned classifier to vote. In this experiment, we applied boosting on DT, NB, K-NN, MLP, RBF, SCRL and SVM, and, as we expected, it worked well on the weakest classifiers. As figure 4 shows, it increased the accuracy of DT from 77.55% to 82.17% and of SCRL from 69.96% to 81.18%. The others remained approximately the same.

- **Stacking:** Stacking is defined as a way of combining multiple machine learning models. It usually combines models of different types of classifiers. As the table II shows, a combination of SVM and MLP has the best accuracy of 84.15%.
The accuracy of the different methods with the Bagging technique

The accuracy of the different methods with the Boosting technique

TABLE II
THE PERFORMANCE AFTER APPLYING THE STACKING TECHNIQUE

<table>
<thead>
<tr>
<th>Stacking</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB, K-NN, MLP, RBF, SVM</td>
<td>78.54</td>
</tr>
<tr>
<td>NB, MLP, RBF, SVM, SCRL</td>
<td>82.17</td>
</tr>
<tr>
<td>NB, MLP, RBF, SVM</td>
<td>82.17</td>
</tr>
<tr>
<td>SVM, MLP, DT</td>
<td>83.16</td>
</tr>
<tr>
<td>SVM, MLP</td>
<td>84.15</td>
</tr>
</tbody>
</table>

V. A COMPARISON OF MACHINE LEARNING TECHNIQUES

The results of the experiments reveal that SVM with 84.15% and SCRL with 69.96% have the highest and lowest accuracy of all the classifiers tested. After bagging is applied, SVM, with the same accuracy percentage, is still the best approach while DT, with 78.54%, is the worst. When we use the boosting technique SVM wins again with 84.81%. The stacking technique also confirms that the combination of SVM and MLP has the highest accuracy with 84.15%.

VI. CONCLUSIONS

In this paper, different machine learning techniques including Decision Tree (DT), Naïve Bayes (NB), Multilayer Perceptron (MLP), K-Nearest Neighbor (K-NN), Single Conjunctive Rule Learner (SCRL), Radial Basis Function (RBF) and Support Vector Machine (SVM) have been applied, individually and in combination, using ensemble machine learning approaches, on the Cleveland Heart Disease dataset in order to compare the performance of each method. As the number of instances in this experiment was limited to 303, the 10-Fold Cross Validation for partitioning the data to the training and testing datasets has been employed. Each technique has been run with different parameters to receive the highest accuracy before the comparison has been made. The efficiency of each individual classifier, and also such classifiers in combination, by employing the bagging, boosting, and stacking techniques, has been evaluated. In general, we have observed some improvements (ranging from almost zero to quite significant improvements), in some cases after applying the bagging, boosting, and stacking approaches. Among the seven different techniques, SVM outperformed all the others when the boosting method was applied. Although our principal goal in this paper has been simply to compare the different machine learning techniques on a small dataset, we have tried to also improve the accuracy of the aforementioned techniques in order to achieve a better comparison.

REFERENCES