Prediction System for Heart Disease Based on Ensemble Classifiers

Joshua Emakhu and Sujeet Shrestha
Department of Industrial and Systems Engineering
Wayne State University
Detroit, MI 48201, USA
fw7471@wayne.edu, sujeet.shrestha@wayne.edu

Suzan Arslanturk
Department of Computer Science
Wayne State University
Detroit, MI 48201, USA
gl9815@wayne.edu

Abstract

The heart is an essential organ in the human body. On the off chance that this organ gets influenced, at that point, it equally influences the other fundamental pieces of the body. Heart diseases are the front runner in terms of death worldwide, making the need for an effective prediction system a source of high demand in treating affected patients. This study aims to analyze prediction systems, thereby designing an automated medical diagnosis system that takes advantage of the collected database. For this study, ensemble classifiers were implemented for classification of data of a medical database with discretization used during the preprocessing phase. The data employed in this research was obtained from the University of California (UCI) machine learning repository. The dataset utilized was the Statlog heart disease. Performance measures, such as accuracy, sensitivity, and specificity, were used to evaluate the proposed methods' performance. The proposed method achieved an accuracy of 87.04%. Based on the results obtained, we observed that it is amongst one of the best in comparison with other studies reported on the UCI website.

Keywords
Heart disease, Data mining, Prediction, Healthcare and Quality

1 Introduction

Heart disease is a condition that influences the heart negatively. As the "coronary arteries narrows, bloodstream to the heart can back off or quit, causing chest torment, cardiovascular failure"(Jabbar *et al.* 2013b). Diagnosing heart diseases requires profoundly gifted and experienced doctors (Jabbar *et al.* 2016).The heart is a primary organ in the human body. On the off chance that this organ gets influenced, at that point, it equally influences the other fundamental pieces of the body. The significant risk factors for heart diseases are age, sexual orientation, hypertension, diabetes mellitus, tobacco smoking, processed meat utilization, unnecessary liquor consumption, sugar consumption, family ancestry, weight, absence of exercise, psychosocial factors, and air contamination. Heart diseases are the front runner in terms of death around the world, in any case, since the 1970s, the death rate due to heart-related infections has declined in some high-income nations. Simultaneously, heart-related mortalities and infections have expanded at a quick rate in low and average-income nations. Although heart disease typically influences more older adults, the indications may start in early life, putting forth essential prevention efforts is essential from childhood. Consequently, risk factors might be adjusted by having good dieting propensities, exercising routinely, and abstaining from smoking tobacco.

The significant reason that the data mining has pulled in a lot of consideration in the information sector in the ongoing years, is due to the full accessibility of large disposal of data and the requirement to transform such data into valuable information and knowledge. The information garnered can be utilized for applications running from business management, production control, and market analysis to developing designs and scientific investigation and data analysis of health (Jabbar *et al.* 2012). In this day and age, the more significant part of the hospitals keeps up their

patient information in electronic formats through some hospital database management framework. These frameworks create colossal amounts of data consistently. This data might be a free text, organized as in databases or as pictures. This data might be utilized to obtain helpful information that might be helpful for decision making. This requirement has prompted the utilization of Knowledge Discovery in Databases (KDD), which is liable for changing low-level information into high-level knowledge for decision-making.

Data mining, which is an incorporation of numerous disciplines, is to obtain informed information from large measures of data. Data mining helps acquire healthcare knowledge and produce a theory from extensive clinical information (Jabbar *et al.* 2013a). Data mining helps in healthcare to aid in active treatment, healthcare management, client relation management, fraud and abuse detection, and decision making. A significant test confronting healthcare organization (emergency clinics, hospitals) is the establishment of quality services at a reasonable cost. Quality service suggests diagnosing patients accurately and controlling effective treatments. Poor clinical choices can prompt poor outcomes, which are, in this manner, unsuitable. Hospitals should likewise limit the expense of clinical tests. They can accomplish these outcomes by utilizing proper PC based information and decision support systems.

The healthcare industry gathers tremendous measures healthcare data which, lamentably, are not "mined" to find shrouded information for effective decision-making. Clinical choices are frequently dependent on specialists' intuition and experience as opposed to the information-rich information hidden in the database. This training prompts undesirable predispositions, mistakes, and high clinical costs, influencing the nature of services provided to patients. For example, it may be possible for the physicians to contrast diagnostic information of different patients with the same conditions. Moreover, physicians can likewise affirm their discoveries with the similarity of different physicians managing a similar case from everywhere throughout the world. Medical diagnosis is considered as a noteworthy yet complicated task that should be done precisely and proficiently. The objective of this study is to analyze prediction systems, thereby designing an automated medical diagnosis system that takes advantage of the collected database.

2 Literature Review

Data mining is valuable for varied types of problems. Forecasting a dependent variable from the significances of independent variables is one of the applications of this method. Healthcare is one of the fields with tons of data available publicly from numerous sources, and data mining could be the best approach to handle it automatically. According to the Centers for Disease Control and Prevention (CDC), heart disease is the number one cause of death in the US, with 647,457 deaths (2017) [Source: CDC, National Center for Health Statistics]. One of the reasons for fatality due to heart disease is that risks are either not identified or identified only at a later stage. In this situation, machine learning could be the best approach to overcome this problem and foretell at an early stage.

Various machine learning techniques have already been applied for prediction. Some of the techniques used for such prediction problems are the Support Vector Machines (SVM), Neural Networks, Decision Trees, Regression, and Naïve Bayes classifiers. SVM was recognized as the best predictor and accuracy, accompanied by neural networks with improved accuracy, and decision trees presented a reduced accuracy. Chest pain, age, smoking, hypertension, and diabetes were measured to be the risk aspects of heart disease (Latha and Jeeva 2019). Systematic studies on data mining techniques for heart disease prediction expose that neural networks, decision trees, Naïve Bayes, and associative classification are influential in predicting heart disease. Associative classification produces a high accuracy and substantial flexibility as associated with traditional classifiers, even in handling unstructured data.

A comparative analysis of classification techniques has shown that decision tree classifiers are modest and precise. Naïve Bayes was found to be the most exceptional algorithm, followed by neural networks and decision trees. Artificial neural networks are also engaged in the prediction of diseases. Supervised networks have been used for diagnosis, and they can be trained using the Back-Propagation Algorithm. The experimental results have shown satisfactory accuracy.

According to a research carried out by (Miao *et al.* 2016), they developed an advanced ensemble machine learning technology, utilizing an adaptive Boosting algorithm, developed for precise heart disease diagnosis and aftermath predictions. The developed ensemble learning classification and prediction models were applied to 4 different data sets for heart disease diagnosis, including patients diagnosed with heart disease from Cleveland Clinic Foundation (CCF), Hungarian Institute of Cardiology (HIC), Long Beach Medical Center (LBMC), and Switzerland University Hospital (SUH). The testing results showed that the advanced ensemble learning classification and prediction models accomplished model accuracies of 80.14% for CCF, 89.12% for HIC, 77.78% for LBMC, and 96.72% for SUH. These results exceeded the accuracy of previously published research. According to the ensemble method applied for identifying valvular heart disease (Das and Sengur 2010), a higher sensitivity rate (97.3%) is obtained using the FCM-CHMM. However, FCM–CHMM created the 92% specificity value; the ensemble method produced the highest specificity rate (100%). SVM technique improvements the worst specificity rate (90%). ANN, WPNN, and LDA–ANFIS approaches achieve the same 94% specificity rate, and ANN, LDA–ANFIS, and PCA–AIS

and fuzzy k-NN produce the same sensitivity rate (95.9%). Therefore, in the end, SVM and WPNN produce a lesser sensitivity rate (94.5%).

Nikookar and Naderi 2018 in their paper, examined the procedure of a hybrid ensemble model in which a new dependable ensemble than basic ensemble models are anticipated and leads to improved performance than further heart disease prediction models . A dataset containing 278 samples from the SPECT heart disease database is used to appraise the proposed models' performance. After applying the model on the data, they achieved 96% of classification accuracy, 80% sensitivity, and 93% of specificity. These results designate the acceptable performance of the proposed hybrid ensemble model in appraisal with a basic ensemble model and other states of the art models.

The proposed system by (Dua and Graff 2019) uses the heart Statlog dataset from the UCI repository. The set is pre-processed to remove irrelevant and missing data. Further, the predominant features get tested on ensemble classifiers for various performance measures. Steps proposed by the author are to build using algorithm performance that represents flow and builds an optimum classification model to solve the problems. The output improves the accuracy by combining ensemble models and Particle Swarm Optimization, which indirectly helps for early detection of heart disease. Accuracy of the proposed method for feature selection using PSO + Bagged Tree gave 100%, PSO + Random Forrest gave 90.37%, and PSO + AdaBoost gave 88.89%.

There are almost no studies depicting capabilities of hybrid ensemble models in which a more reliable ensemble, than basic ensemble models, are proposed. This approach leads to better performance than other heart disease prediction models based on findings from (Nikookar and Naderi 2018).

Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques by (Latha and Jeeva 2019) analyzed the accuracy of prediction of heart disease using an ensemble of classifiers. The Cleveland heart dataset from the UCI machine learning repository was utilized for training and testing purposes. The ensemble algorithms bagging, boosting, stacking, and majority voting got employed for experiments. With the use of bagging, there was an improvement in accuracy by a maximum of 6.92%. With the use of boosting, there was an improvement in accuracy by a maximum of 5.94%. When the weak classifiers are ensemble with majority voting, there was an improvement in accuracy by a maximum of 7.26%, and stacking improved the accuracy by a maximum of 6.93%. A comparison of results showed that the majority of voting produces the most significant improvement in the percentage of accuracy. The performance was further enhanced using feature selection techniques. The feature selection technique helped to improve the accuracy of the ensemble algorithms. The highest accuracy was obtained with majority voting with a feature set FS2.

Prediction of Heart Disease Using Machine Learning Algorithms by (Nichenametla *et al.* 2018) proposed the Naive Bayes classification technic and decision tree construction in this decision tree construction and this decision tree construction ID3 algorithm. ID3 algorithm is one of the old algorithms which is used for building decision trees in the process of building a decision tree. It handles missing values and removes outliers. The authors used their proposed approach that Naïve Bayes is more accurate if the input data is cleaned and well maintained. Even though ID3 can clean itself, it cannot give accurate results every time. In this same way, Naïve Bayes will not give accurate results every time; they need to consider different algorithms' results. By all its results, if a prediction is made, it will be accurate. Therefore, they can also use Naïve Bayes and K-means to get accuracy.

The significance of medical and demographic data in envisaging the presence of heart disease in an individual is investigated. According to a paper by (Dun *et al.* 2016), a multiplicity of ensemble and deep learning techniques with hyperparameter tuning and feature selection ultimately results in an extreme test accuracy of 78%. Model averaging does not expressively improve prediction accuracy. The same points are inclined to be misclassified by all the models that do not overfit. This recommends that the mistakes in this author's data can mainly be indorsed to a complicated error in the problem.

3 Methodology

In this section, the technical aspect of our approach is discussed. The main objective of this study is to analyze prediction systems, thereby designing an automated medical diagnosis system that takes advantage of the collected database. Thirteen attributes from medical terms, such as sex, blood pressure, cholesterol, are utilized to build up this system. The data mining classification technique used is the ensemble classifier comprising of AdaBoost, Bagging, Random Forest, and Voting Ensemble (Decision Trees, Logistic Regression and Support Vector Machines).

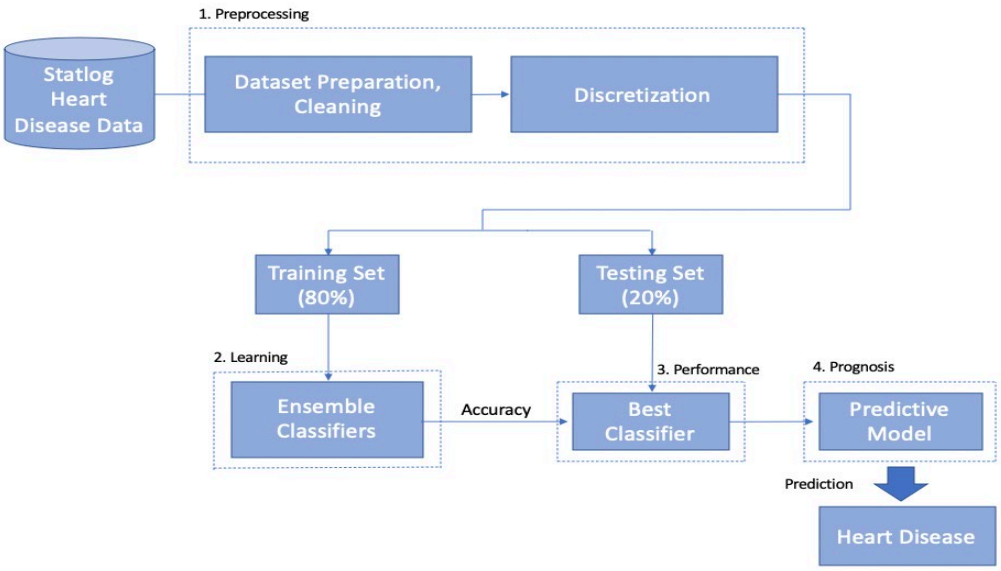

Figure 3-1: Architectural Framework for Heart Disease Prediction

3.1 Dataset

The publicly available Heart Disease dataset from the UCI Machine Learning Repository is used. The Statlog heart disease database (Dua and Graff 2019) consisting of 270 records and 13 attributes is utilized. The target output label has two classes to represent a patient with or without heart disease.

3.1.1 Preprocessing

This phase involves two steps – Dataset preparation and cleaning, and discretization. The *Python software* and the *Waikato Environment for Knowledge Analysis, WEKA* software for analysis get put in use. This software has a sophisticated graphical user interface which compiles and runs on a wide variety of Windows and macOS.

Step 1: Dataset preparation and cleaning – This involves procuring the data and carrying out the necessary tuning parameter for preprocessing. Upon extraction of the data, the data had no missing values, and it contained real, ordered binary and nominal data. Other visualization techniques were carried out to determine the presence of outliers. Also, the descriptive statistics of the overall data was done to determine the summary of the data.

Step 2: Discretization – This is one of the most influential data preprocessing tasks in knowledge discovery and data mining. Discretization is a data reduction mechanism. It diminishes data from a large domain of numeric values to a subset of categorical values. When discretization is applied to a dataset, it slows the machine learning method applied to present remarkable improvements in learning speed and accuracy.

After this process gets done, that wraps up the preprocessing phase. The data is further split into training and test data with an 80:20 split.

.

3.1.2 Learning

This process entails applying the ensemble classifier – AdaBoost, Bagging, Random Forest, and Voting Ensembles. These classifiers are then implemented on 80% of the entire dataset in other to train the model.

Step 3: Ensemble Classifiers – In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone (Opitz and Maclin 1999; Rokach 2010). Unlike a statistical ensemble in statistical mechanics, which is usually infinite, a machine learning ensemble consists of only a particular finite set of alternative models. However, it typically allows for a much more flexible structure to exist among those alternatives. The classifiers used are stated in the paragraph mentioned above.

3.1.2.1 AdaBoost

AdaBoost, short for Adaptive Boosting, is a machine learning meta-algorithm. It is sometimes used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum representing the final output of the boosted classifier. AdaBoost is adaptive in that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers (Gómez-Ríos *et al.* 2017). In some problems, it can be less susceptible to the overfitting problem than other learning algorithms. The individual learners can be weak, but as long as each one's performance is slightly better than random guessing, the final model can be proven to converge to a keen learner.

3.1.2.2 Bagging

Bootstrap aggregating also called bagging (from bootstrap aggregating), is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting. Although it is usually applied to decision tree methods, it can be used with any method. Bagging is a particular case of the model averaging approach.

3.1.2.3 Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression. It involves other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of classification or regression of the individual trees (Ho 1995, 1998). Random decision forests correct for decision trees' habit of overfitting to their training set (Friedman *et al.* 2001).

3.1.2.4 Voting Ensembles

Similarly, to the Random Forest, the Voting Ensemble estimates multiple base models. It uses voting to combine the individual predictions to arrive at the final ones. However, the critical difference lies in the base estimators. Models such as Voting Ensemble (and Stacking Ensemble) do not require the base models to be homogenous. In other words, we can train different base learners, for example, a Decision Tree, Logistic Regression, & Support Vector Machines, and then use the Voting Ensemble to combine the results.

3.1.3 Performance

In this phase, the best classifier is chosen and then applied to the remaining 20% of the data.

Step 4: Best Classifier – After this, performance evaluation metrics such as accuracy, sensitivity, specificity, and AUC curve we determined. However, for this project, the main emphasis is placed on the accuracy to determine which classifier is best suited for predicting the state of a patients' health as having heart disease or not.

3.1.4 Prognosis

In this phase, we use the results of making predictions for new patients with heart disease-related traits.

Step 5: Predictive Model – The predictive model is based on the prognosis derived from the machine learning best classifier. A prognosis is based on the ordinary course of the diagnosed disease, the individual's physical and mental condition, the available treatments, and additional factors ('Prognosis' 2018). A complete prognosis includes the expected duration, function, and description of the course of the disease, such as progressive decline, intermittent crisis, or sudden, unpredictable crisis.

.

4 Results & Evaluation

Classification of patients that has heart disease and no heart disease:

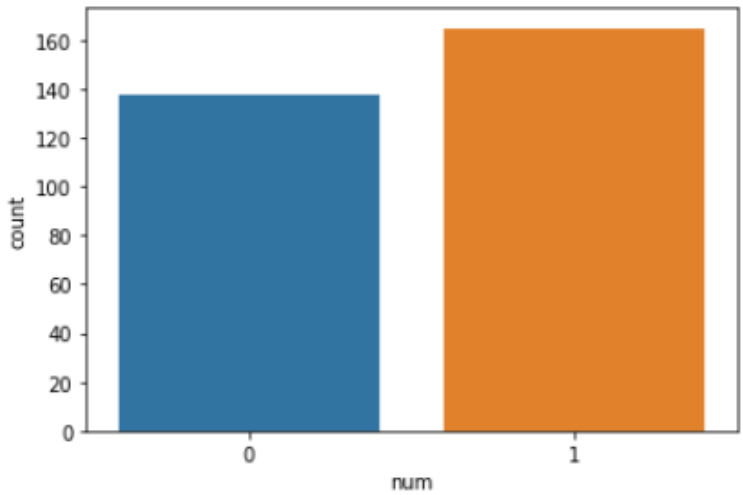

Figure 4-1: Classification of patients with Heart disease and No Heart Disease

Figure 4-1 shows the classification of the patients having No Heart Disease (0) with the number of 165 and the patients having Heart Disease (1) with 138.

Classification of patient in different age categories:

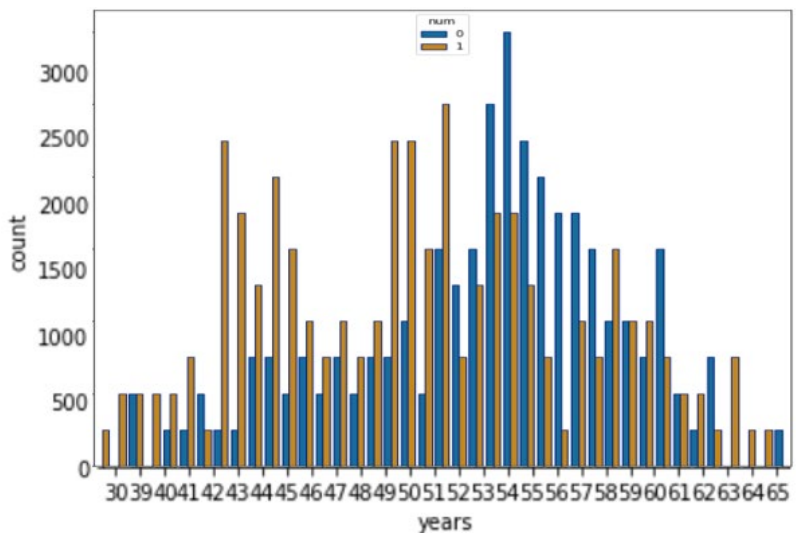

Figure 4-2: Classification of patients with Heart disease and No Heart Disease

Figure 4-2 shows the results after the preprocess step, classifying our dataset in terms of patients with heart disease versus patients without heart disease. 56% of patients do not have heart disease, and 44% of patients have heart disease. The bar diagram shows the break-even age of patients with heart disease versus Not having heart disease. The ratio of No-heart disease patients is higher than heart disease in every age category from age 30 thru 51. Once the age exceeds 52, the ratio of heart disease patients gets higher than No-heart disease in all age categories till age 64.
Note: *Dataset has the patients between 30 thru 65 age only*

Correlation Heat Map of all the attributes:

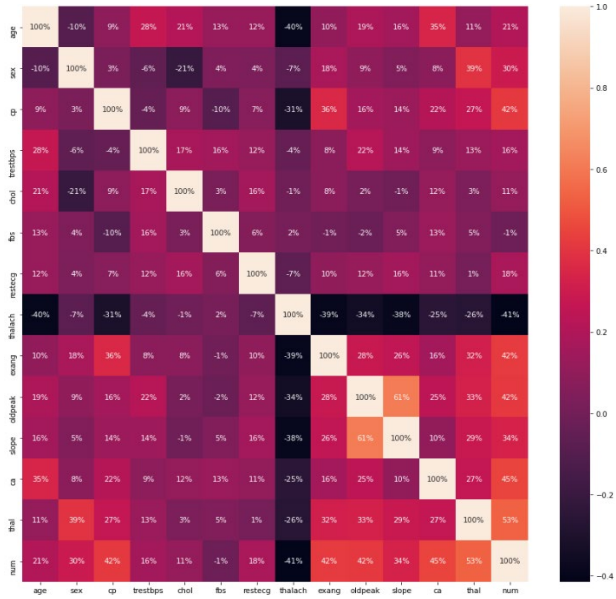

Figure 4-3: Correlation heat map to measure accuracy, sensitivity, and specificity

Figure 4-3 is the correlation heat map showing the relationship between 14 attributes or variables directly related to the cause of heart disease. Out of 14 variables, three variables stood out in terms of severity – Thalassemia, Calcium, and Chest pain.

- Thalassemia: It has a 53% impact on severity. It tracks the patients that have a maximum heart rate. It means that patients having a higher heart rate have a 53% chance of getting heart disease.
- Calcium: It has a 45% impact of severity for the cause of heart disease that measures the patient's significant vessels 0 through 3 levels.

Chest Pain: It has the 42% impact of severity, which means patients with chest pain type I through IV have the 42% chance of getting the heart disease

Confusion Matrix:

Table 4-1: Confusion Matrix of Heart Disease vs No Heart Disease patients

	Actual Heart Disease	Actual No Heart Disease	Total
Predicted Heart Disease	TP	FP	TP+FP
Predicted No Heart Disease	FN	TN	FN+TN
Total	TP+FN	FP+TN	TP+FP+FN+TN

The confusion matrix is used to measure the accuracy, sensitivity, and specificity. The matrix gives information about correctly classified as Normal (true) and misclassified as Abnormal (false). Evaluation of the Matrix is done on the Ensemble classifier (Bagged Tree, AdaBoost, Voting Ensemble, and Random Forest) (Yekkala *et al.* 2017).

Where TP – True Positive correctly labels by the classifier. TN – True Negative correctly labels by the classifier. FN – False Negative incorrectly labels by the classifier. FP – False Negative incorrectly labels by the classifier. To evaluate the performance of ensemble methods, we have used the following classification measures:

1) Sensitivity = TP/(TP+FN)
2) Specificity = TN/(FP+TN)
3) Precision = TP/(TP+FP)
4) Probability Misclassification Error (PME) = FN+FP/(TP+FN+FP+TN)
5) Positive predictive value (PPV) = TP/(TP+FP)

6) Negative predictive value (NPV) = TN/(FN+TN)
7) Model Accuracy = (TP+TN)/(TP+TN+FP+FN)

Detailed Accuracy by Class of each Ensemble Classifier:
1) AdaBoost: TP = 0.769, FP = 0.214, FN = 0.231, TN = 0.786
2) Bagging: TP = 0.885, FP = 0.214, FN = 0.115, TN = 0.786
3) Voting Ensemble: TP = 0.808, FP = 0.143, FN = 0.192, TN = 0.857
4) Random Forest: TP = 0.846, FP = 0.107, FN = 0.154, TN = 0.893

Table 4-2: Confusion matrix accuracy of each classifiers

	AdaBoost	**Bagging**	**Voting Ensemble**	**Random Forest**
Sensitivity	0.769	**0.885**	0.808	0.846
Specificity	0.786	0.786	0.857	**0.893**
Precision	0.782	0.805	0.849	**0.887**
PME	0.223	0.165	0.167	**0.131**
PPV	0. 782	0.805	0.849	**0.887**
NPV	0.773	0.872	0.817	**0.853**
Model Accuracy	0.778	0.835	0.832	**0.869**

Sensitivity is the percentage of the patients that have the disease and identified as having the disease. Bagging gave the highest of all with 89%. Specificity is the percentage of healthy patients who recognized as not having the disease, and Random Forest gave the best with 88%. Similarly, regarding precision, Random Forest gave 88% (highest). Likewise, positive predictive value (PPV) is the probability that disease identified when the diagnostic test is positive. The negative predictive value (NPV) is the probability that the disease does not get seen when the diagnostic test is Negative. On both, Random Forest gave 88% and 85%, respectively. Therefore, as Model Accuracy, the best model became Random Forest with ~87%.

Comparison Various Approaches with our Proposed Approach

Table 4-3: Comparison of various approaches with our proposed approach

S/N	**Author**	**Methodology**	**Accuracy**
1	K. Srinivas, B.K. Rani, and A. Gavrdhan	**Naïve Bayes**	83.70%
2	K. Polat and S. Gunes	**RBF Kernel F-score + LS-SVM**	83.70%
3	D. Tomar and S. Agarwal	**Feature Selection - based LSTSVM**	85.59%
4	W. Duch, R. Adamczak, and K. Grabczewski	**KNN Classifier**	85.60%
5	A.G. Karegowda, A.S. Manjunath, and M.A. Jayaram	**GA + Naïve Bayes**	85.87%
6	H. Kahramanli and N. Allahverdi	**Hybrid Neural Network Method**	86.80%
7	Our proposed approach	**Random Forest**, AdaBoost, Bagging, Voting Ensemble, SVM, Logistic Regression, Decision	87.04%

Classification accuracies obtained by other approaches are shown in Table 4-3 and Figure 4-4. The proposed approach obtained an accuracy of 87.04% when compared to other approaches. In papers by (Srinivas *et al.* 2010) and

(Polat and Güneş 2009) in their approach obtained 83.70%, which is 3.34% less than our approach. An accuracy of 85.59% was obtained by (Tomar and Agarwal 2014) in their work. The approach illustrated by (Duch *et al.* 2001) obtained an accuracy of 85.60%, which is 1.44% less than our proposed approach. Similarly, (Karegowda *et al.* 2010) obtained 85.87% and (Kahramanli and Allahverdi 2008) obtained 86.80%. Based on the comparison, we can deduce that our proposed approach achieved the highest accuracy compared to other approaches.

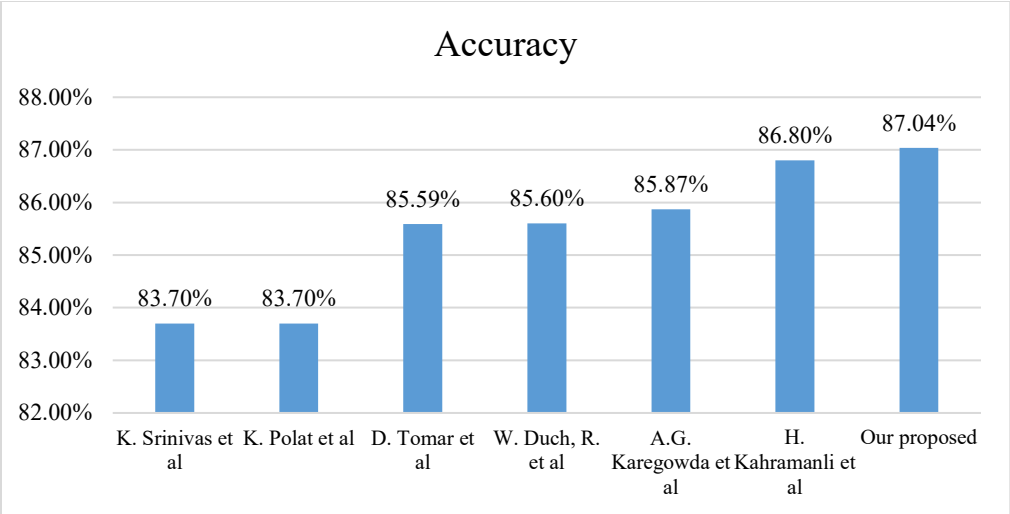

Figure 4-4: Comparison of various approaches with our proposed approach

5 Conclusion

In this paper, we studied the prediction of heart disease based on ensemble classifiers. Our proposed approach uses discretization to improve the signal to noise ratio. Fitting a model to bins reduces the impact that small fluctuations in the model's data and these small fluctuations are just noise. So, each bin smooths out the fluctuates or noise. Then applied ensemble methods as a classifier to reduce the misclassification rate and to improve the classification performance.

From the experimental results, it has been proven that the learning accuracy can be significantly improved by using Random Forest Ensemble Classifier. This model will help to accurately predict and early diagnosis of heart diseases using a subset of features. We want to integrate other machine learning algorithms to develop a model for early diagnosis of heart disease as our future work.

References

Das, Resul, and Abdulkadir Sengur. 2010. 'Evaluation of ensemble methods for diagnosing of valvular heart disease', *Expert Systems with Applications*, 37: 5110-15.

Dua, D., and C. Graff. 2019. "UCI Machine Learning Repository." In. Irvine, CA: University of California, School of Information and Computer Science.

Duch, Wlodzislaw, Rafal Adamczak, and Krzysztof Grabczewski. 2001. 'A new methodology of extraction, optimization and application of crisp and fuzzy logical rules', *IEEE Transactions on Neural Networks*, 12: 277-306.

Dun, Boyang, Eric Wang, and Sagnik Majumder. 2016. 'Heart Disease Diagnosis on Medical Data Using Ensemble Learning'.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning* (Springer series in statistics New York).

Gómez-Ríos, Anabel, Julián Luengo, and Francisco Herrera. 2017. "A study on the noise label influence in boosting algorithms: AdaBoost, GBM and XGBoost." In *International Conference on Hybrid Artificial Intelligence Systems*, 268-80. Springer.

Ho, Tin Kam. 1995. "Random decision forests." In *Proceedings of 3rd international conference on document analysis and recognition*, 278-82. IEEE.

———. 1998. 'The random subspace method for constructing decision forests', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20: 832-44.

Jabbar, M Akhil, BL Deekshatulu, and Priti Chandra. 2013a. 'Classification of heart disease using artificial neural network and feature subset selection', *Global Journal of Computer Science and Technology Neural & Artificial Intelligence*, 13: 4-8.

Jabbar, M Akhil, Bulusu Lakshmana Deekshatulu, and Priti Chandra. 2013b. "Heart disease prediction using lazy associative classification." In *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, 40-46. IEEE.

———. 2016. 'Prediction of heart disease using random forest and feature subset selection.' in, *Innovations in bio-inspired computing and applications* (Springer).

Jabbar, MA, BL Deekshatulu, and Priti Chandra. 2012. "An evolutionary algorithm for heart disease prediction." In *International Conference on Information Processing*, 378-89. Springer.

Kahramanli, Humar, and Novruz Allahverdi. 2008. 'Design of a hybrid system for the diabetes and heart diseases', *Expert Systems with Applications*, 35: 82-89.

Karegowda, Asha Gowda, MA Jayaram, and AS Manjunath. 2010. 'Feature subset selection problem using wrapper approach in supervised learning', *International journal of Computer applications*, 1: 13-17.

Latha, C Beulah Christalin, and S Carolin Jeeva. 2019. 'Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques', *Informatics in Medicine Unlocked*, 16: 100203.

Miao, Kathleen H, Julia H Miao, and George J Miao. 2016. 'Diagnosing coronary heart disease using ensemble machine learning', *International Journal of Advanced Computer Science and Applications*, 7: 30-39.

Nichenametla, Rajesh, T Maneesha, Hafeez Shaik, and Krishna Hari. 2018. 'Prediction of Heart Disease Using Machine Learning Algorithms', *International Journal of Engineering & Technology*, 7: 363-66.

Nikookar, Elham, and Ebrahim Naderi. 2018. 'Hybrid Ensemble Framework for Heart Disease Detection and Prediction', *IJACSA) International Journal of Advanced Computer Science and Applications*, 9.

Opitz, David, and Richard Maclin. 1999. 'Popular ensemble methods: An empirical study', *Journal of Artificial Intelligence Research*, 11: 169-98.

Polat, Kemal, and Salih Güneş. 2009. 'A new feature selection method on classification of medical datasets: Kernel F-score feature selection', *Expert Systems with Applications*, 36: 10367-73.

'Prognosis'. 2018. Nature Publishing. https://www.nature.com/subjects/prognosis.

Rokach, Lior. 2010. 'Ensemble-based classifiers', *Artificial Intelligence Review*, 33: 1-39.

Srinivas, K, B Kavihta Rani, and A Govrdhan. 2010. 'Applications of data mining techniques in healthcare and prediction of heart attacks', *International Journal on Computer Science and Engineering (IJCSE)*, 2: 250-55.

Tomar, Divya, and Sonali Agarwal. 2014. 'Feature selection based least square twin support vector machine for diagnosis of heart disease', *International Journal of Bio-Science and Bio-Technology*, 6: 69-82.

Yekkala, Indu, Sunanda Dixit, and MA Jabbar. 2017. "Prediction of heart disease using ensemble learning and Particle Swarm Optimization." In *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, 691-98. IEEE.

Bibliographies

Joshua Emakhu is a Ph.D. student, and Graduate Teaching Assistant in the Department of Industrial and Systems Engineering at Wayne State University, Michigan, USA. He earned B.Sc. in Civil Engineering from Covenant University, Ogun, Nigeria, and a Master's in Industrial Engineering also from Wayne State University, Michigan, USA. He has published journal and conference papers. Joshua has writing funded proposal with the National Science Foundation and other related. His research interests include data analytics, medical devices, recalls, healthcare, resource allocation, and voice of the customer. He is member of IEOM, INFORMS, SAVE Int'l, and IISE. He is also a peer reviewer for IISE, RIED and Frontiers for Psychology.

Sujeet Shrestha is a Ph.D. student, and Engineer at Fiat Chrysler Automotive for Jeep Gladiator. He earned B.Sc. in Industrial and Entrepreneurial Engineering from Western Michigan University, Michigan, USA and a Master's in Industrial Engineering from Wayne State University, Michigan, USA. His research interests include data analytics in Health care and automotive.

Suzan Arslanturk, PhD. is an Assistant Professor in the Department of Computer Science, Wayne State University. She has received her Ph.D. from Oakland University, Michigan. Her research interests and expertise falls in the broad area of health informatics and healthcare systems engineering. In particular, she is motivated with tackling challenging

theoretical and applied research problems that enable developing innovative machine learning and clinical solutions to enhance individual and population health outcomes, improve patient care, and optimize the operational performance of healthcare delivery systems. Prior to joining Wayne State, she has worked at a healthcare informatics and analytics services company in Michigan. She has authored several referred research papers and conference proceedings.