Data Mining Classification Algorithms for Heart Disease Prediction

Mirpouya Mirmozaffari[1], Alireza Alinezhad[2], and Azadeh Gilanpour[3]

***Abstract*—**Annually 17.3 million people approximately die from heart disease worldwide. A heart patient shows various symptoms and it is hard to attribute them to the heart disease in different steps of disease progress. Data mining, as a solution to extract hidden pattern from the clinical dataset are applied to a database in this research. The database consists of 209 instances and 8 attributes. All available algorithms in classification technique, are compared to achieve the highest accuracy. To further increase the accuracy of the solution, the dataset is preprocessed by different supervised and unsupervised algorithms. The system was implemented in WEKA and prediction accuracy in 9 stages, and 396 approaches, are compared. Random tree with an accuracy of 97.6077% and lowest errors is introduced as the highest performance algorithm.

***Keywords*—** Data mining, Classification, WEKA.

I. INTRODUCTION

AMONG all fatal disease, heart attacks diseases are considered as the most prevalent [1]. Medical practitioners conduct different surveys on heart diseases and gather information of heart patients, their symptoms and disease progression. Increasingly are reported about patients with common diseases who have typical symptoms. Thus, there is valuable information hidden in their dataset to be extracted.

Data mining is the technique of extracting hidden information from a large set of database [2]. It helps researchers gain both novel and profound insights of unprecedented understanding of large medical datasets. The principal goals of data mining are prediction and description of diseases.

To find the unknown trends in heart disease, all the available classification algorithms are applied to a unique dataset and their accuracy are compared. A dataset of 209 instances and 8 attributes (7 inputs and 1 output) are used to test and justify the differences between algorithms. To further enhance accuracy and achieve more reliable variables, the dataset is purified by supervised and unsupervised filters.

Mirpouya Mirmozaffari[1], Msc. student, Faculty of Industrial and Mechanical Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran (corresponding author's e-mail: m.mirmozaffari@gmail.com).

Alireza Alinezhad[2], Associate Professor, Faculty of Industrial and Mechanical Engineering, Qazvin Branch, Islamic Azad University, Qazvin,Iran (e-mail: alinezhad@qiau.ac.ir).

Azadeh Gilanpour[3], Islamic Azad University (IAU).

II. BACKGROUND AND LITERATURE REVIEW

Growing number of heart patients worldwide have motivated researchers to do comprehensive research to reveal hidden patterns in clinical datasets. This section provides an overview of previous computational studies on pattern recognition in heart disease. Not only are different techniques addressed, but also various heart disease datasets are covered to have a fair comparison. Finally, the gap in existing literature, which was the main motivation of this study is also provided. Some of the key studies are as follows:

- Das et al. introduced a neural network classifier for diagnosing of the valvular heart disease. The ensemble-based methods create new models by combining the posterior probabilities or the predicted values from multiple predecessor models. An effective model has been created and experimentally tested. A classification accuracy of 97.4% from the experiment on a dataset containing 215 samples is achieved [3].
- Pandey et al. proposed the performance of clustering algorithm using heart disease dataset. They evaluated the performance and prediction accuracy of some clustering algorithms. The performance of clusters will be calculated using the mode of classes to clusters evaluation. Finally, they proposed Make Density Based Cluster with the prediction accuracy of 85.8086%, as the most versatile algorithm for heart disease diagnosis [4].
- Karaolis et al. developed a data mining system using association analysis based on the Apriori algorithm for the assessment of heart-related risk factors with WEKA tools. A total of 369 cases were collected from the Paphos CHD Survey, most of them with more than one event. Selected rules were evaluated according to the importance of each rule. Each extracted rule was further evaluated by inspection of the number of cases within the database [5].

Therefore, pattern recognition in heart disease can be addressed through different computational techniques. In regard to classification algorithms, other respected works, focused on diverse aspects of heart disease on different datasets can be mentioned: Nahar et al., 2013 [6]; Tantimongcolwat et al., 2008 [7]; Jyoti et al., 2015 [8]; Manimekalai 2016 [9]; Durgadevi et al., 2016 [10]. Atkov 2012 [11]. Alizadehsani et al., 2013 [12]. Amin et al., 2013 [13]; Lakshmi et al., 2013 [14]. Also, different computational techniques for other health care issues have been reported in the literature [15-16].

It is observed various classifiers are frequently utilized in different studies to predict heart disease. Therefore, a comprehensive comparison of classification algorithms practically provides an insight into classifier performances. This comparison is of great importance to medical practitioners who desire to predict heart failure at a proper step of its progression. Furthermore, except for Ref. [17], which has evaluated 4 classification techniques, there is not any other study on the current dataset. Finally, a unique multilayer filtering in preprocessing step is applied which eventually results in increased accuracy within most of the classification algorithms, covered in this study.

III. Dataset Description

The standard dataset, compiled in this study contains 209 records, which is collected from a hospital in Iran, under the supervision of National Health Ministry. Data is gathered from a single resource, so it precludes any integration operations. Eight attributes are utilized, from them, 7 are considered as inputs which predict the future state of the attribute “Diagnosis”. All the attributes, along with their values and data types are discussed in Table I.

TABLE I
The Arrangement of Channels

Attributes	Descriptions	Encoding\Values	Feature
Age	Age in years	28-66	Numeric
Chest Pain Type	It signals heart attack and has four different conditions: Asymptotic, Atypical Angina, Typical Angina, and without Angina.	Asymptotic = 1 Atypical Angina = 2 Typical Angina = 3 Non-Angina = 4	Nominal
Rest Blood Pressure	Patient’s resting blood pressure in mm Hg at the time of admission to the hospital	94-200	Numeric
Blood Sugar	Below 120 mm Hg- Normal Above 120 mm Hg- High	High = 1 Normal = 0	Nominal Binary
Rest Electrocard iographic	Normal, Left Ventricular Hypertrophy (LVH) ST_T wave abnormality	Normal=1 Left Vent Hyper = 2 ST_T wave abnormality = 3	Nominal
Maximum Heart Rate	maximum heart rate attained in sport test	82-188	Numeric
Exercise Angina	It includes two conditions of positive and negative	Positive = 1 Negative = 0	Nominal Binary
Diagnosis	It includes two conditions of positive and negative	Positive = 1 Negative = 0	Nominal Binary

IV. Research methodology

The objective of this study is to effectively predict possible heart attacks, from the patient dataset. Using a prediction methodology, a model was developed to determine the characteristics of heart disease in terms of some attributes. Data mining in this research is utilized to build models for prediction of the class based on selected attributes. Waikato Environment for knowledge Analysis (WEKA) has been used for prediction due to its proficiency in discovering, analysis and predicting of patterns [18]. Generally, the whole process can be split into two steps as follows:

A. Multilayer filtering preprocess

The data in the real world is highly susceptible to noise, missing, and inconsistency. Therefore, pre-processing of data is very important. We apply a filter on datasets and purify them from dirty and redundant data present in the dataset. Both attribute (attribute manipulation), and instance (instance manipulation) filters in either case of supervised or unsupervised, can be applied in WEKA 2016 (version 3.9.0). In this study, a multilayer filtering process is applied to the dataset to make imbalanced data balanced. This process is implemented in three steps as follows:

- Step A: “Discretization” which is unsupervised attribute filter changes numeric data into nominal.
- Step B: The output of step A is applied to a “Resample” unsupervised instance filter.
- Step C: The output of step B is applied to a “Resample” supervised instance filter.

B. Evaluation in classification

To broaden our comparison, three different evaluation methods which are: 1- training set, 2- 10-Fold cross-validation, and 3- percentage split (66%) are considered to analyze each output of aforementioned steps.

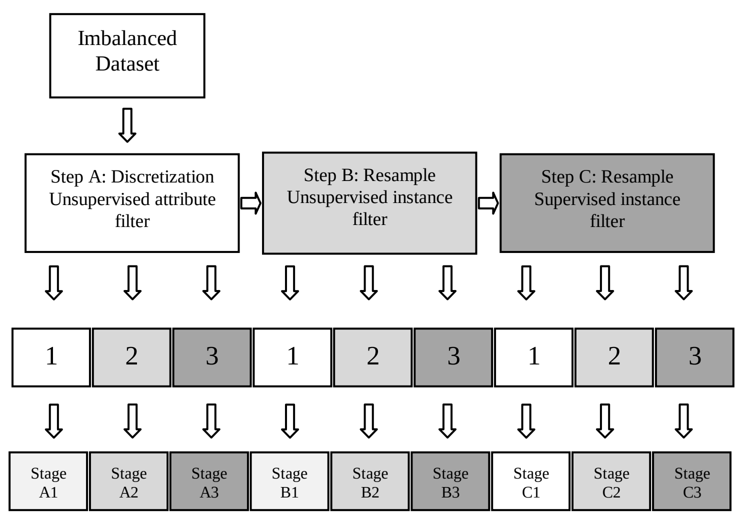

Fig 1: Implementation of classification Algorithms for accuracy analysis

Figure 1 elaborates the proposed model and different steps. The combination of each filtering step and each evaluation method results in a different stage. By applying 9 stages to 44 classifiers, 396 different approaches are yielded. The accuracy and average accuracy in each stage, are compared in Table II.

TABLE II
ACCURACY COMPARISON WITHIN CLASSIFICATION ALGORITHMS (ALL NUMBERS ARE IN PERCENT)

classifiers	StageA1	Stage B1	Stage C1	Stage A2	Stage B2	Stage C2	Stage A3	Stage B3	Stage C3
1.BayesNet(Bayes)	80.3828	81.3397	85.6459	78.4689	79.9043	84.6890	74.6479	78.8732	90.1408
2.NaiveBayes(Bayes)	80.3828	82.2967	85.6459	78.9474	78.9474	84.6890	74.6479	77.4648	90.1408
3.NaivveBayesMultinominalText(Bayes)	55.9809	56.4593	56.4593	55.9809	56.4593	56.4590	57.7465	43.6620	52.1127
4.NaiveBayesUpdatable(Bayes)	80.3828	82.2967	85.6459	78.9474	78.9474	84.6890	74.6479	77.4648	90.1408
5.Logistic(functions)	80.8612	85.1675	92.3445	75.5981	81.3397	87.0813	66.1972	84.5070	90.1408
6.MultyLayerPerceptron(functions)	89.9522	**95.6938**	96.6507	77.9904	**88.9952**	93.7799	73.2394	85.9155	95.7746
7.SGD(functions)	79.9043	83.2536	91.8660	78.4689	80.8612	88.0383	70.4225	74.6479	80.2817
8.SGDText (functions)	55.9809	56.4593	56.4593	55.9809	56.4593	56.4593	57.7465	43.6620	52.1127
9.SimpleLogistic(functions)	81.3397	84.6890	91.3876	77.0335	81.3397	87.5598	73.2394	84.5070	90.1408
10.SMO(functions)	79.9043	83.2536	92.3445	77.9904	80.3828	86.6029	73.2394	81.6906	90.1408
11.VotedPerceptron(functions)	78.9474	80.3828	85.1675	77.9904	80.3828	85.6459	69.0141	84.5070	85.9155
12.IBK(lazy)	**90.4306**	**95.6938**	**97.6077**	76.5550	88.5167	**94.2584**	67.6056	85.9155	95.7746
13.KStar(lazy)	88.9952	94.2584	97.1292	76.0766	87.5598	**94.2584**	69.0141	85.9155	95.7746
14.LWL(lazy)	81.8182	88.0383	87.5598	77.5120	81.3397	85.1675	**77.4648**	88.7324	83.0986
15.AdaBoost1M1(meta)	78.9474	83.7321	82.2967	78.4689	80.3828	80.3828	71.8310	80.2817	80.2817
16.AttributeSelectedClassifier(meta)	78.9474	86.1244	86.6029	77.9904	82.7751	84.6890	71.8310	81.6901	81.6901
17.Bagging(meta)	82.7751	88.5167	92.8230	78.4689	84.6890	88.5167	70.4225	88.7324	**97.1831**
18.ClassificationViaRegression(meta)	79.9043	89.4737	91.3876	78.4689	86.6029	91.3876	71.8310	83.0986	88.7324
19.CVparameterSelection(meta)	55.9809	56.4593	56.4593	55.9809	56.4593	56.4593	57.7465	43.6620	52.1127
20.FilteredClassifier(meta)	82.7751	89.9522	94.7368	77.9904	86.6029	88.9952	71.8310	87.3239	91.5493
21.IterativeClassifierOptimizer(meta)	80.3828	84.6890	91.3876	77.5120	82.2967	86.6029	73.2394	88.7324	84.5070
22.LogitBoost(meta)	81.3397	84.6890	91.3876	77.0355	83.7321	86.1244	67.6056	81.6901	84.5070
23.MultyClassClassifier(meta)	80.8612	85.1675	92.3445	75.5981	81.3397	87.0813	66.1972	84.5070	90.1408
24.MultyClassClassifierUpdatable(meta)	79.9043	83.2536	91.8660	78.4689	80.8612	88.0383	70.4225	74.6479	80.2817
25.MultyScheme(meta)	55.9809	56.4593	56.4593	55.9809	56.4593	56.4593	57.7465	43.6620	52.1127
26.RandomCommittee(meta)	**90.4306**	**95.6938**	**97.6077**	77.0335	87.0813	93.3014	66.1972	84.5070	95.7746
27.RandomizableFilteredClassifier(meta)	**90.4306**	**95.6938**	**97.6077**	71.7703	88.0383	92.8230	69.0141	78.8732	95.7746
28.RandomSubSpace(meta)	78.4689	88.0383	91.8660	77.5120	81.8182	85.6459	74.6479	**90.1408**	84.5070
29.Stacking(meta)	55.9809	56.4593	56.4593	55.9809	56.4593	56.4593	57.7465	43.6620	52.1127
30.VOTE(meta)	55.9809	56.4593	56.4593	55.9809	56.4593	56.4593	57.7465	43.6620	52.1127
31.WeightedInstancesHandlerWraper(meta)	55.9809	56.4593	56.4593	55.9809	56.4593	56.4593	57.7465	43.6620	52.1127
32.InputMappedClassifier(misc)	55.9809	56.4593	56.4593	55.9809	56.4593	56.4593	57.7465	43.6620	52.1127
33.DecisionTable(rules)	81.3397	88.5167	90.4306	**81.3397**	78.4689	78.9474	71.8310	78.8732	85.9155
34.JRep(rules)	82.2967	89.9522	93.7799	79.9043	83.2536	83.7321	71.8310	84.5070	87.3239
35.OneR(rules)	79.9043	81.8182	80.8612	79.9043	81.8182	80.8612	74.6479	85.9155	80.2817
36.PART(rules)	84.6890	93.3014	95.2153	75.5981	85.6459	90.9091	73.2394	88.7324	91.5493
37.ZeroR(rules)	55.9809	56.4593	56.4593	55.9809	56.4593	56.4593	57.7465	43.6620	52.1127
38.DecisionStump(trees)	78.9474	80.8612	79.9043	77.5120	80.8612	74.1627	71.8310	84.5070	80.2817
39.HoeffdingTree(trees)	80.3828	82.2967	85.6459	78.9474	78.9474	84.6890	74.6479	77.4648	90.1408
40.J48(trees)	82.7751	89.9522	94.7368	77.9904	86.6029	88.9952	71.8310	87.3239	91.5493
41.LMT(trees)	81.3397	84.6890	91.3876	77.0355	84.6890	87.0813	73.2394	84.5070	90.1408
42.RandomForest(trees)	**90.4306**	**95.6938**	**97.6077**	77.0355	88.0383	92.8230	66.1972	88.7324	94.3662
43.RandomTree(trees)	**90.4306**	**95.6938**	**97.6077**	75.5981	86.1244	92.8230	66.1972	88.7324	95.7746
44.RepTree(trees)	82.7751	88.5167	91.8660	79.4258	84.2105	84.6890	74.6479	85.9155	84.5070
Average of 44 classifiers	77.2184	81.1549	84.0474	73.2276	77.7620	80.1794	68.6889	75.5122	81.2099

V. RESULT AND DISCUSSION

It can be inferred from table II, as the layers of filtering increase:

- The maximum of accuracy within three evaluation methods is increased.
- The average accuracy of 44 classifiers, corresponds to each

filtering step is increased.

It also should be noted, in each filtering step, from stage A to stage C, the accuracy of most of the classifiers are increased. Therefore, to narrow down our study to the most accurate stages, a further comparison on other evaluators of the most accurate algorithms in stages C1, C2, and C3 are provided. The most accurate algorithm in stage C3, Bagging with 97.1831% accuracy, along with some evaluators are provided in table III.

TABLE III
EVALUATION OF THE BEST CLASSIFIERS IN STAGE C3

Classifier	Bagging(Meta)
TP Rate	0.972
FP Rate	0.028
precision	0.972
Recall	0.972
F-Measure	0.972
ROC	0.988
Kappa statistic	0.9436
MAE	0.1742
RMSE	0.2372
RAE	35.0971%
RRSE	47.0875%
Time	0 Sec

Table IV compares the best two classifiers in stage C2 with 94.2584 % accuracy. It is evident that IBK algorithm exhibits more appropriate performances in terms of many evaluators such as MAE, RMSE, RAE, and RRSE. Therefore, it is considered as the best algorithm in this stage.

TABLE IV
EVALUATION OF THE BEST CLASSIFIER IN STAGE C2

Classifiers	IBK(Lazy)	KStar (Lazy)
TP Rate	0.943	0.943
FP Rate	0.057	0.057
precision	0.943	0.943
Recall	0.943	0.943
F-Measure	0.943	0.943
ROC	0.954	0.968
Kappa statistic	0.8835	0.8835
MAE	0.0784	0.1442
RMSE	0.2292	0.2408
RAE	15.9366%	29.319%
RRSE	46.2296%	48.5564%
Time	0 Sec	0 Sec

Table V compares some other evaluators of five most accurate algorithms in stage C1. Comparing first two ones, Random Tree and Random Committee, it can be observed all the evaluators except for the time taken to build a model, are equal to each other. Therefore, Random Tree is considered as the superior algorithm in stage C1. The same can be inferred about next two algorithms in Table V.

TABLE V
EVALUATION OF THE BEST CLASSIFIER IN STAGE C1

Classifiers	Random Tree (Trees)	Random Committee (Meta)	IBK (Lazy)	Randomizable Filtered Classifier (Meta)	Random Forest (Trees)
TP Rate	0.976	0.976	0.976	0.976	0.976
FP Rate	0.021	0.021	0.021	0.021	0.021
precision	0.977	0.977	0.977	0.977	0.977
Recall	0.976	0.976	0.976	0.976	0.976
F-Measure	0.976	0.976	0.976	0.976	0.976
ROC	0.998	0.998	0.998	0.998	0.998
Kappa	0.9515	0.9515	0.9515	0.9515	0.9515
MAE	0.0311	0.0311	0.0325	0.0325	0.0546
RMSE	0.1247	0.1247	0.1247	0.1247	0.1381
RAE	6.3246%	6.3246%	6.6194%	6.6194%	11.1083%
RRSE	25.1509%	25.1509%	25.1538%	25.1538%	25.1509%
Time	0 Sec	0.02 Sec	0.03 Sec	0.05 Sec	0.02 Sec

Finally, in a more detailed discussion some other evaluators of five most accurate (97.6077%) algorithms within all approaches, are thoroughly discussed below:

- Random tree Random tree with the highest accuracy, TP Rate, precision (Sensitivity), Recall (Specificity), F-Measure, ROC area, Kappa Statistics and lowest FP Rate, MAE, RMSE, RAE, RRSE and Time taken to build the model, is considered as the best algorithm.
- Random committee (Meta) with all the same evaluators as Random Tree and just a little longer time to build the model (0.02 second), comes after Random Tree.
- The third best classifier is IBK (lazy) with the same evaluators as two aforementioned ones except for greater MAE, RAE, RRSE.
- Randomizable Filtered Classifier (Meta) is considered as the fourth best algorithm with 0.05 seconds building time.
- The fifth place is assigned to Random Forest (Trees) with significant different MAE, RMSE, and RAE with same evaluators of IBK and Randomizable classifiers.

VI. CONCLUSION

Various classification algorithms in data mining were compared to predict heart disease. A unique model consisting of different filters and evaluation methods are evolved. Multilayer filtering preprocess, as well as different evaluation methods, are applied to find the superior algorithm and more accurate clinical decision supports systems for diagnosis of diseases. Classifiers are compared regarding their accuracies, error functions, and building times. The high-performance algorithms within each stage were introduced. The experiment can serve as a practical tool for physicians to effectively predict uncertain cases and advise accordingly.

References

[1] A. K. Sen, S. B. Patel, and D. P. Shukla, "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level," International Journal of Engineering and Computer Science, Vol. 2, No. 9, pp. 1663–1671, 2013.

[2] G. Karraz, G. Magenes, "Automatic Classification of Heart beats using Neural Network Classifier based on a Bayesian Frame Work," IEEE, Vol 1, 2006.

[3] R. Das, I. Turkoglu, and A. Sengur, "Diagnosis of valvular heart disease through neural networks ensembles," Elsevier, 2009.

[4] A. K. Pandey, P. Pandey, K. L. Jaiswal, and A. K. Sen, "Data Mining Clustering Techniques in the Prediction of Heart Disease using Attribute Selection Method," International Journal of Science, Engineering and Technology Research (IJSETR), ISSN: 2277798, Vol 2, Issue10, October 2013.

[5] M. Karaolis, J. A. Moutiris, and C. S. Pattichis, "Association rule analysis for the assessment of the risk of coronary heart events," Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009.

https://doi.org/10.1109/iembs.2009.5334656

[6] J. Nahar, and T. Imam, "Computational intelligence for heart disease diagnosis: A medical knowledge driven Approach," Elsevier, 2013.

[7] T. Tantimongcolwat, and T. Naenna, "Identification of ischemic heart disease via machine learning analysis on Magnetocardiograms," Elsevier, 2008.

[8] R. Jyoti, G. Preeti, "Analysis of Data Mining Techniques for Diagnosing Heart Disease," International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol. 5, ISSUE. 7, July 2015.

[9] K. Manimekalai, "prediction of heart disease using data mining techniques," IJIRCCE, Vol.4, Issue 2, February 2016.

[10] A. Durgadevi and K. Saravanapriya, "comparative study of data mining classification algorithm in heart disease prediction," international journal of recent research in mathematics computer science and information technology, Vol.2, Issue 2, March 2016.

[11] O. Y. Atkov, "Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters," Elsevier, 2012.

[12] R. Alizadehsani, J. Habibi, M. Hosseini, R. Boghrati, A. Ghandeharioun, B. Bahadorian, Z. Alizadehsani, "A Data Mining Approach for Diagnosis of Coronary Artery Disease, " Elsevier, 2013.

[13] S. U. Amin, K. Agarwal, and R. Beg, "Genetic neural network based data mining in prediction of heart disease using risk factors," presented at the IEEE Conference on Information & Communication Technologies, 2013.

https://doi.org/10.1109/cict.2013.6558288

[14] K. R. Lakshmi, M. V. Krishna and S. P. Kumar, "Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability," International Journal of Scientific and Research Publications, ISSN 2250-3153, Vol.3, Issue.6, June 2013.

[15] M. Kumari, R. Vohra and A. Arora, "Prediction of Diabetes Using Bayesian Network," International Journal of Computer Science and Information Technologies, Vol. 5 (4), 5174-5178, 2014.

M. A. Banu, B. Gomathy, "Disease forecasting system using data mining methods," in IEEE International Conference on Intelligent Computing Applications (ICICA'14), pp. 130-133, 2014.

https://doi.org/10.1109/icica.2014.36

[16] B. Bahrami, and M. H. Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques," Journal of Multidisciplinary Engineering Science and Technology (JMEST), ISSN: 3159-0040, Vol. 2, Issue 2, February 2015.

[17] I. H. Witten and E. Frank, "Data Mining Practical Machine Learning Tools and Techniques," Morgan Kaufman Publishers, 2005.