Bagging Technique to Reduce Misclassification in Coronary Heart Disease Prediction Based on Random Forest

A Saifudin[1*], U U Nabillah[1], Yulianti[1] and T Desyani[1]
[1]Informatics Engineering, Pamulang University, Jalan Raya Puspitek 46, Banten 15310, Indonesia

*aries.saifudin@unpam.ac.id

Abstract. Knowing the existence of coronary heart disease is very important to reduce the risk caused. Coronary heart disease is influenced by many factors, in diagnose requires complex analysis. Many proposed the application of a machine-learning algorithm to diagnose/predict coronary heart disease, but have not given perfect results (excellent). The machine learning algorithm is used to classify someone affected by coronary heart disease or not based on factors that have been determined input. The results of diagnosis/prediction are not perfect due to misclassification that is still large.to reduce misclassification, bagging techniques are proposed. The classification algorithm used in the study is Random Forest. Experimental results show that bagging techniques can reduce misclassified predictions of coronary heart disease.

1. Introduction

The heart is one of the most important bodies that pumps blood throughout the body through coordinated contractions. While coronary heart disease is a compilation of blood vessels that causes blockages in blood vessels to the heart[1]. Coronary heart disease is influenced by several factors such as stress, high blood pressure, smoking habits, lack of exercise, cholesterol, etc [2]. The way to reduce the risk of coronary heart disease is to detect coronary heart disease early or describe the assessment of health risk manually. Namely by approaching to collect information from individuals then identify risk factors [3]. Early detection of coronary heart disease and by manually describing health risk assessments. Having a lack of comprehensive data about risk factors that are relevant and have limitations [4].

Recently, remote control based on the Internet of Things (IoT) has gained considerable success in the world of health which is a powerful feature that can increase the relevance of discovery and can make it easier to store large amounts of data. Observations made by IoT can produce the data is large enough and is a big provocation to achieve affordable, capacity-saving and high-quality portable devices for monitoring in a short time. [5]. Many ways are used to diagnose coronary heart diseases such as applying various classification methods such as SVM, Naïve Bayes, logistic regression and tree regression to predict coronary heart disease. However, this method is better than it should be, this technique does not affect disease changes from then until now to predict coronary heart disease [6].

Random Forest was first introduced by Breiman, Random Forest is a development of the Decision Tree used for integrated classification. and can be used to estimate the involvement of each feature in the classification with Gini and permutation calculations. [7]. Bagging is the Ensemble method in

Machine Learning which has demonstrated the ability to improve performance in classification in difficult practical settings which is a simple but effective Ensemble method [8]

To predict patients affected by coronary heart disease or not, this study will use the Random Forest method which is expected to increase the accuracy value using the Ensemble technique on Bagging.

2. Methodology

This research topic is the prediction of coronary heart disease. Where the dataset is tested using random forest technique classifiers and bagging techniques to get better accuracy values. In this study using secondary data, namely data collection contained in the UCI Learning Machine. The attributes of coronary heart disease dataset shown in table 1.

Table 1. Attribute of dataset

Attribute	Information	Score Information
Age	Patient age	29 – 80 years
Gender	Gender	1 = male 0 = female
Chest Pains	Angina stable, angina unstable, angina variant, asymptomatic	1 = Angina stable 2 = Angina unstable 3 = Angina variant 4 = Asymptomatic
Rest Blood Sugar	Mm/Hg	100 – 200
Cholestrol	mg/dl	135 - 420
Fasting Blood Sugar	>120 mg/dl	1 = true 0 = false
Electrocardiographic	Normal, ST-T abnormal, Hypertrophy	0 = Normal 1 = ST-T abnormal 0.25 - 1.8 2 = Hypertrophy 1.4 - 2.5
Heart Rate	Beats per minute	100 - 170
Exercise Induced	Yes or No	1 = Yes 0 = No

Old Peak Real	low, risk, bad	Low = < 2
		Risk = 1.5 – 4.2
		Bad = > 2.55

Slope	the slope of the peak exercise ST segment	- Value 1: upsloping - Value 2: flat - Value 3: downsloping
Major Vessels	number of major vessels colored by flourosopy	0 – 3
Thallium scan	Normal, fixed defect, reversable defect	3, 6, 7
Class	Predicted attribute choices	1-2

The purpose of this observation is how to predict coronary heart disease to be more precise and accurate. So that it can produce better accuracy. Figure 1 is the proposed framework of this study. Random Forest algorithm is implemented to conduct training on data samples. Meanwhile, to reduce misclassification prediction of coronary heart disease, the ensemble bagging algorithm is applied because it can improve classification accuracy [9].

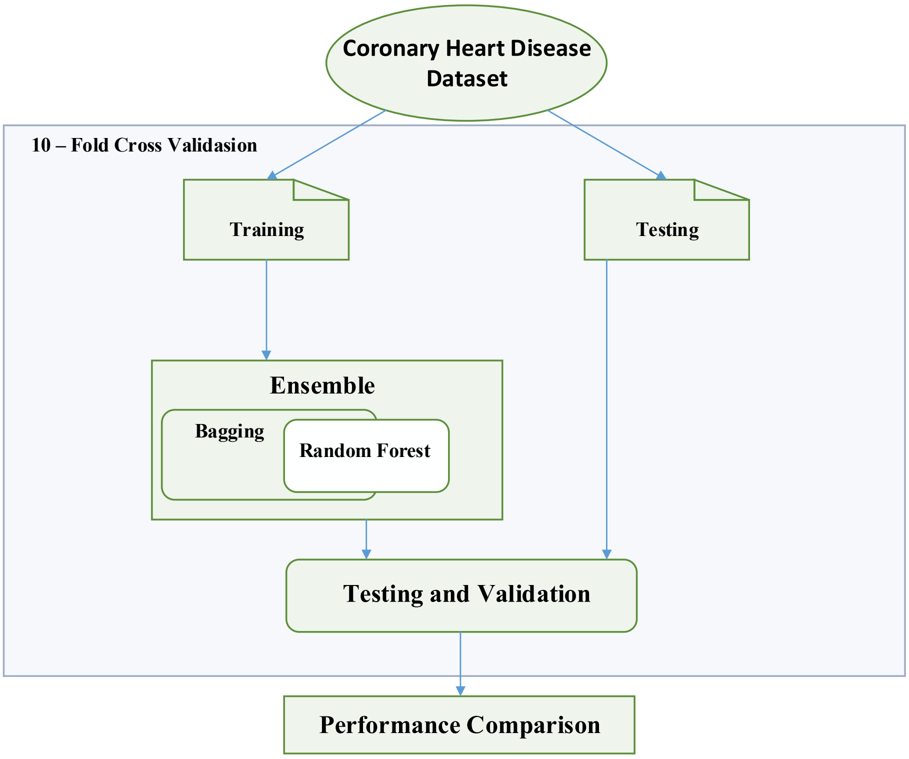

Figure 1. Coronary heart disease model

Random Forest is a method that combines trees with training on data samples owned [10]. The use of more trees will affect the accuracy that will get better. The Random Forest equation is written as follows:

$$\boldsymbol{m\infty, n(x)} = \frac{\sum_{i=1}^{n} Y_i \, K_n \; (x, X_i)}{\sum_{L=1}^{n} K_n \, (x, X_L)} \tag{1}$$

Random Forest is a method that uses decision trees as a basic classification, which consists of a collection of tree structures. Each tree grows with a random vector where k = 1, ... L is independent and statistically distributed [10]. How bagging works is to combine the average model to get an ensemble model with a lower variant then each model is simplified to get a relevant weight so. bagging equation can be written:

$$S_L\,(.) = \frac{\arg max}{k}\,[\text{ card }(l|\omega l\,(.) = k)] \tag{2}$$

The purpose of the model that has been applied is to predict coronary heart disease where the approved data is agreed to be divided into two parts, namely test data and training data. 10fold - cross validation is used to calculate the calculation time while still calculating the calculated accuracy and selecting the best model. which discussed in figure 2.

Validation	**Split**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	Testing	Training								
2	Training	Testing	Training							
3	Training		Testing	Training						
4	Training			Testing	Training					
5	Training				Testing	Training				
6	Training					Testing	Training			
7	Training						Testing	Training		
8	Training							Testing	Training	
9	Training								Testing	Training
10	Training									Testing

Figure 2. Dataset distribution for validation

In the first dataset is used as test data, while the second to a tenth dataset is training data. all datasets are repeated until all datasets are used as test data.

The results from the dataset have been used to measure the ability of the model using the confusion matrix. The confusion matrix is a tool to analyze how well the classifier recognizes the features of different classes. The confusion matrix is a 2-dimensional matrix shown in Table 2

Table 2. Confusion matrix

Class		Actual	
		True	False
Prediction	True	TP (True Positive)	FP (False Positive)
	False	FN (False Negative)	TN (True Negative)

Model capabilities that have been supported on Accuracy, AUC and memory values. shown as follows:[11]

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \tag{3}$$

$$\boldsymbol{TP_{rate}} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \tag{4}$$

$$\boldsymbol{FP_{rate}} = \frac{\boldsymbol{FP}}{\boldsymbol{FP+TN}} \tag{5}$$

AUC can be calculated based on the average estimates made by TP rate and FP rate. AUC is calculated as a measure of the area of the ROC (Receiver Operating Characteristics) curve using equations.

$$\boldsymbol{AUC} = \frac{\boldsymbol{1+TP_{rate}-FP_{rate}}}{\boldsymbol{2}} \tag{6}$$

3. Results and Discussion

Based on the model proposed in Figure 1. to see how the basic model applied by the Random Forest algorithm works as a classification without being optimized. The second model is integrating Random Forest with bagging techniques. The proposed model is applied using a coronary heart disease dataset that has been obtained and measured its performance. The results of measuring the performance of the proposed model are shown in Table 3.

Table 3 Average of model performance

Model	AVG	
	Accuracy	AUC
Random Forest	77,40 %	0.7561
Random Forest + Bagging	84,07 %	0.8317

Based on the graph in Figure 3, it can be seen about the prediction of coronary heart disease using the Random Forest algorithm with bagging techniques to have a better performance compared to the Random Forest algorithm without being optimized. The accuracy and AUC performance models have the same high values for the Random Forest algorithm with the bagging technique.

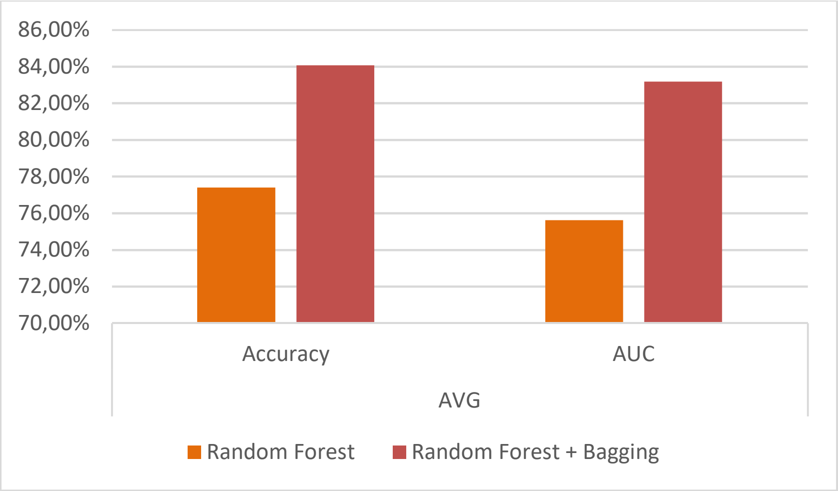

Figure 3. Model performance comparison

Based on the number of validation results, it can be seen that the performance of the model that implements bagging techniques with the Random Forest algorithm has high accuracy and AUC values compared to the Random Forest algorithm without optimization.

4. Conclusion

Prediction of coronary heart disease is an important disease topic because it is very necessary that the disease cannot be more severe because coronary heart disease is a dangerous disease and many causes it. Based on the proposed model that the results obtained have not gotten better results. The results of this study indicate that a model that integrates Random Forest with bagging techniques provides better performance values. The proposed model can help predict coronary heart disease better.

References

[1] Pareek V and Sharma R K 2016 Coronary heart disease detection from voice analysis *2016 IEEE Students' Conf. Electr. Electron. Comput. Sci. SCEECS 2016* 1–6
[2] Chauhan A, Jain A, Sharma P and Deep V 2018 Heart Disease Prediction using Evolutionary

Rule Learning *Int. Conf. &amp;amp;quot;Computational Intell. Commun. Technol. CICT 2018* 1–4

[3] Mohawish A, Rathi R, Abhishek V, Lauritzen T and Padman R 2015 Predicting Coronary Heart Disease risk using health risk assessment data *2015 17th Int. Conf. E-Health Networking, Appl. Serv. Heal. 2015* 91–6

[4] Murthy H S N and Meenakhi M 2014 Approach for Early Prediction of Coronary Heart Disease *Proc. Int. Conf. Circuits, Commun. Control Comput.* 21–2

[5] Venkatesan C, Karthigaikumar P and Satheeskumaran S 2018 Mobile cloud computing for ECG telemonitoring and real-time coronary heart disease risk detection *Biomed. Signal Process. Control* **44** 138–45

[6] Orphanou K, Stassopoulou A and Keravnou E 2016 DBN-extended: A dynamic Bayesian network model extended with temporal abstractions for coronary heart disease prognosis *IEEE J. Biomed. Heal. Informatics* **20** 944–52

[7] Wei G, Zhao J, Yu Z, Feng Y, Li G and Sun X 2018 An Effective Gas Sensor Array Optimization Method Based on Random Forest *Proc. IEEE Sensors* **2018-Octob** 1–4

[8] Liu L, Chin S P and Tran T D 2019 Reducing Sampling Ratios and Increasing Number of Estimates Improve Bagging in Sparse Regression *2019 53rd Annu. Conf. Inf. Sci. Syst. CISS 2019* 1–5

[9] Huda S, Liu K, Abdelrazek M, Ibrahim A, Alyahya S, Al-Dossari H and Ahmad S 2018 An ensemble oversampling model for class imbalance problem in software defect prediction *IEEE Access* **3536**

[10] Scornet E 2016 Random forests and kernel methods *IEEE Trans. Inf. Theory* **62** 1485–500

[11] Punitha K and Latha B 2016 Izbor neuravnoteženog niza podataka za predvidanje grešaka u racunalnom programu primjenom hibridnih neuro-fuzzy sustava s Naive Bayes klasifikatorom *Teh. Vjesn.* **23** 1795–804