Bagging Techniques to Reduce Misclassification of Breast Cancer Prediction Base on Gradient Boosted Trees (GBT) Algorithm

T Desyani[1*], Y Kasmayanti[1], A Saifudin[1] and Yulianti
Informatics Engineering, Universitas Pamulang, Jalan Raya Puspitek 46, Tangerang, Banten 15310, Indonesia

[*]dosen00839@unpam.ac.id

Abstract. Breast cancer is a malignant tumour that grows in breast cells and has a risk of death. Breast cancer has levels ranging from stage 0 to stage 4. The higher the stage of breast cancer, the higher the risk of death and is difficult to treat. The application of machine learning algorithms has been proposed to help predict breast cancer. Predictions made by classifying patients tend to have breast cancer or not. This research proposes to implement bagging techniques to reduce misclassification in the Gradient Boosting Trees (GBT) algorithm. The experimental results show that the application of bagging techniques can reduce misclassification and improve prediction accuracy.

1. Introduction

Breast cancer is the most common cause of death, especially women throughout the world [1]. Breast cancer is formed when cells in the breast grow abnormally and are out of control [2]. These cells generally form tumours that feel like benign or malignant lumps [3]. The collection of cells will spread throughout the body and grow in a network around the chest. Early symptoms of breast cancer are usually not detected or difficult to show. Therefore, it is very important to follow the screening guidelines recommended for detecting breast cancer early for women [2]. Current technological developments to diagnose or detect breast cancer continue to increase with the times and aim to provide less invasive choices for accurate and better diagnoses [1].

There have been many studies using machine learning technology specifically in predicting breast cancer from many sources[2]. To Diagnose Breast Cancer can use a Gradient Boosted Trees (GBT) algorithm.

The Gradient Boosting Trees (GBT) algorithm is an information-theoretical discriminative predictor for boosting regression 230 accuracy. GBT classification can work well and is also effective on a broad set of data and with a combination of many features that are not normalized.

Different hyperparameters used in the algorithm for each tree built (e.g., maximum tree depth) and others using the configuration of all models (e.g., numbers of trees to build) [3]. but the level of accuracy obtained from the GBT algorithm is still low at 0.58%. To increase the accuracy of prediction of the GBT algorithm by using bagging techniques. Whereas in the approach to combining or pairing

(ensemble) methods, there are two of the most popular ensemble-learning algorithms, namely boosting and bagging. so Bagging can improve model performance on breast cancer prediction[4].

Based on the prediction problem of breast cancer above, the Machine Learning algorithm will be implemented with a feature selection technique using a dataset downloaded from the UCI Machine Learning Repository. By using a bagging technique to reduce misclassification in the Gradient Boosting Trees (GBT) algorithm.

2. Methodology

The dataset used in this study uses a secondary dataset. Secondary data is data that is not obtained directly from the object of research, obtained has been collected by other parties. The secondary data used in this study is a collection of biomedical data taken directly from the UCI Machine Learning Repository which can be downloaded via the Dataset site used in this study. The Dataset collection of breast cancer in the UCI repository is divided into 4, namely breast cancer, Wisconsin (original) or WBCD breast cancer, Wisconsin (Prognostic) Breast Cancer or WPBC and Coimbra Breast Cancer. The specifications of the dataset used in this study are shown in table 1.

Table 1 Dataset specification

Attribut	Unit	Description	Value Range
Age	(years)	Patient Age	24-89
BMI	(kg/m2)	Body mass index	18,37-38,579
Glucose	(mg/dL)	Glucose Level	60-201
Insulin	(µU/mL)	Insulin	2,432-58,46
HOMA		Homeostasis Model Assessment	0,467-25,05
Leptin	(ng/mL)	Leptin Levels	4,311-90,28
Adiponectin	(µg/mL)	Protein Hormone Level	1,656-38,04
Resistin	(pg/dL)	Blood Sugar Level	3,21-82,1
MCP-1	(pg/dL)	Chemokine Monocyte Chemoattractant Protein 1	45,843-1698,44
Classification	class	1 for Healthy controls and 2 for Patients	1 or 2

The purpose of this study is to improve accuracy and AUC using bagging techniques on the GBT algorithm. This research is expected to be able to get a better diagnosis of breast cancer than using the GBT algorithm. The proposed framework of the prediction model for this work is shown in figure 1. The new GBT-based algorithm incorporates a novel component to the regular regression error component which is optimized through dataset testing [3]. While to reduce misclassification of software defect prediction, an ensemble algorithm (Bagging/AdaBoost) is applied because it can improve the classification accuracy [5].

Gradient Boosting Trees (GBT) algorithm is an information-theoretical discriminative predictor to improve accuracy and AUC.

It is described as follow:

$$L^{(t)} = \sum_{i=1}^{n} w\left(y_i \cdot y_i^{t-1} + f_t(x_i)\right) + \Omega(f_t) + c \quad (1)$$

$$\approx \sum_{i=1}^{n} \left[w(y_i \cdot y_i^{t-1} + \vartheta_{yi^{(t-1)}} w\left(y_i \cdot y_i^{t-1}\right) f_t(x_i) + \frac{1}{2} \vartheta_{yi^{(t-1)}} w\left(y_i \cdot y_i^{t-1}\right) {f_t}^2(x_i)) \right] \quad (2)$$

The $\vartheta_{yi^{(t-1)}} w\left(y_i \cdot y_i^{t-1}\right) f_t(x_i)$ is the first order of the target function and the $\frac{1}{2} \vartheta_{yi^{(t-1)}} w\left(y_i \cdot y_i^{t-1}\right) {f_t}^2(x_i)$ is the second order.

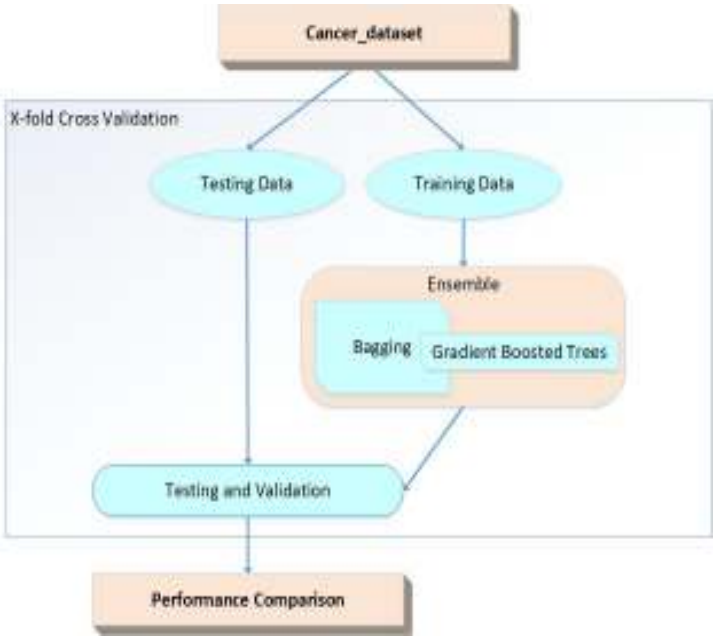

Figure 1. Software defect prediction model

Bagging or boosting has the same two, namely representative in the learning ensemble. To train the suboptimal model, then add and get approved in the final model [6]. Bagging (Bootstrap Aggregating), using the bootstrap sub-dataset to produce a training set of L (learning), L trains basic learning using unstable learning procedures and then during testing takes the average [7]. Bagging Algorithm:

Looping for b = 1, 2, . . ., B

1. Make a bootstrap sample $\{(X , Y_1 *) 1 *, (X , Y_2 *), \dots , 2 * (X_n * , Y_n *)\}$ by random replacement of training data $\{(X_1 , Y_1), (X_2 , Y_2), \dots , (X_n, Y_n)\}$ match the *Cb* classifier turned on in the appropriate bootstrap sample.

2. Output of final classifier:

$$(x) = B -1 \sum Cb\ (x) \qquad (3)$$

The purpose of this model is to increase accuracy accurately significantly greater than individual models, and stronger against the effects of noise and overfitting of the original training data. The proposed model is applied using 1 dataset from UCI Machine Leaning Repository. This dataset will be chosen alternately as testing data and others as training data until all datasets have tested the data. The distribution of the dataset is training data and testing data can be seen in Figure 2.

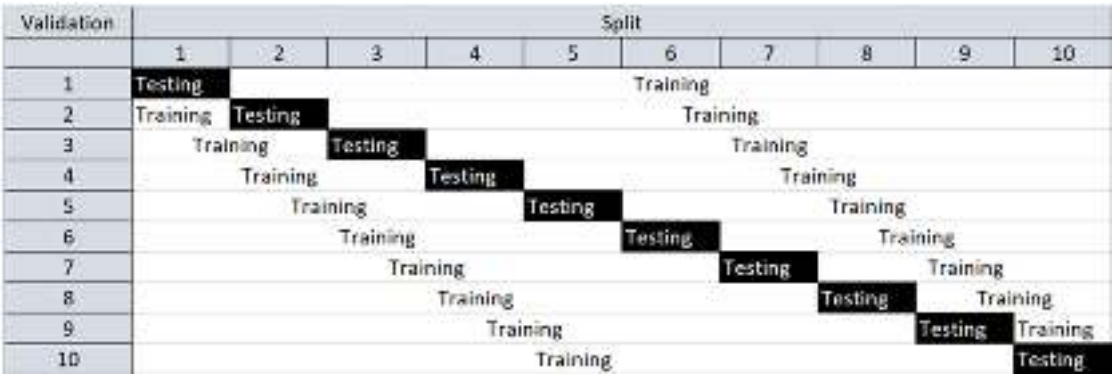

Figure 2. Dataset distribution for validation

In the first validation, the first dataset is used as the testing data, while the second until fifth datasets are training data. In the second validation, the second dataset is used as testing data, and the other as training data. Validation is repeated until all datasets have been used as testing data.

Validation results are used to measure model performance. Model performance is usually measured using a matrix. A confusion matrix is a useful tool for analyzing how well classifiers can recognize tuples/features of different classes [8]. Confusion matrix also provides performance appraisal of classification models based on the number of objects predicted correctly and incorrectly[9]. 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)

The performance of the model can be seen from the value of Accuracy or AUC. To calculate the performance of the model the following equation can be used [9]:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \tag{4}$$

$$\boldsymbol{TP_{rate}} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \tag{5}$$

$$\boldsymbol{FP_{rate}} = \frac{\boldsymbol{FP}}{\boldsymbol{FP+TN}} \tag{6}$$

The AUC can be calculated based on the approximate average trapezoidal plane for curves made by TP_{rate} and FP_{rate} [9].

$$\boldsymbol{AUC} = \frac{\mathbf{1}+\boldsymbol{TP_{rate}}-\boldsymbol{FP_{rate}}}{\mathbf{2}} \tag{7}$$

3. Results and Discussion

Based on the model proposed in figure 1, to find out the performance of the basic model applied by the Gradient Boosted Trees algorithm as a classification without being optimized. The second model is integrating Gradient Boosted Trees with bagging techniques.

Table 3 Average of Model Performance

Model	Performance	
	Accuracy	AUC
Gradient Boosted Trees	58%	0.55
Gradient Boosted Trees + Bagging	64%	0.63

Based on the graph in figure 3, it can be seen about the prediction of breast cancer using the Gradient Boosted Trees algorithm with bagging techniques to reduce misclassification and improve prediction accuracy to the Gradient Boosted Trees algorithm without being optimized. The accuracy and AUC performance models have the same high values for the Gradient Boosted Trees algorithm with the bagging technique.

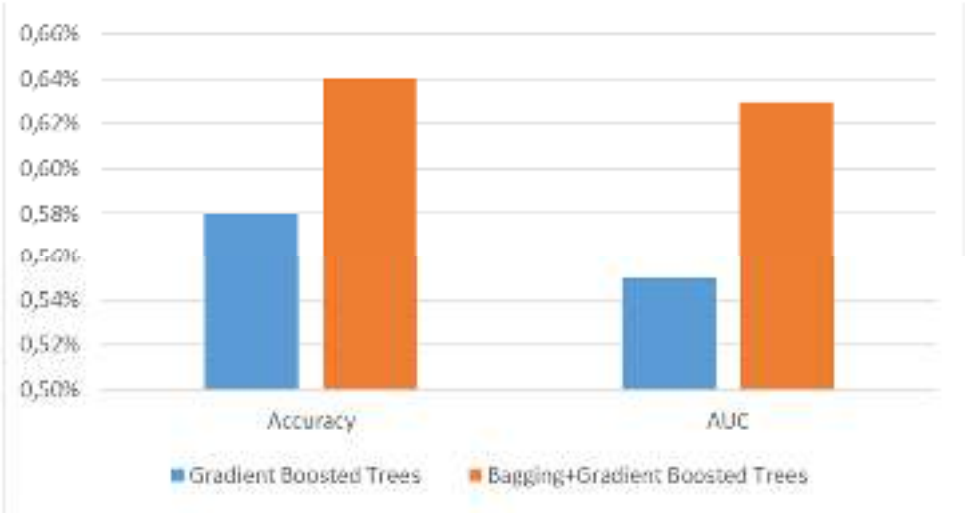

Figure 3. Model performance comparison

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

4. Conclusion

Breast cancer prediction is an important research topic to avoid because breast cancer can be deadly if new sufferers realize it after entering the final stage. The proposed model shows the results that there is no model that produces very good performance. The experimental results show that the application of bagging techniques can reduce misclassification and improve prediction accuracy. The proposed model can help predict breast cancer accurately.

References

[1] Chougrad H, Zouaki H and Alheyane O 2019 Multi-label transfer learning for the early diagnosis of breast cancer *Neurocomputing*

[2] Tseng Y J, Huang C E, Wen C N, Lai P Y, Wu M H, Sun Y C, Wang H Y and Lu J J 2019 Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies *Int. J. Med. Inform.* **128** 79–86

[3] Israeli A, Rokach L and Shabtai A 2019 Constraint learning based gradient boosting trees *Expert Syst. Appl.* **128** 287–300

[4] Wu Z, Li N, Peng J, Cui H, Liu P, Li H and Li X 2018 *Using an ensemble machine learning methodology-Bagging to predict occupants' thermal comfort in buildings* vol 173 (Elsevier B.V.)
[5] 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**
[6] Xia Y 2019 A Novel Reject Inference Model Using Outlier Detection and Gradient Boosting Technique in Peer-to-Peer Lending *IEEE Access* **7** 92893–907
[7] Breiman L 1996 Bagging Predictors, URL: https://link.springer.com/article/10.1007%2FBF00058655 *Mach. Learn.* **24** 123–40
[8] Han J 2011 *Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)*
[9] Gorunescu F 2011 *Data mining: concepts and techniques*