PROC. INTERNAT. CONF. SCI. ENGIN.
Volume 3, April 2020 | Pages: 455-459

ISSN 2597-5250
E-ISSN 2598-232X

Predicting Breast Cancer: A Comparative Analysis of Machine Learning Algorithms

Pulung Hendro Prastyo*, I GedeYudi Paramartha, Michael S. Moses Pakpahan***, Igi Ardiyanto******
Department of Electrical Engineering and Information Technology, Faculty of Engineering, Universitas Gadjah Mada
Jalan Grafika No. 2 Kampus UGM, Yogyakarta 55281, Indonesia.
Email: pulung.hendro@mail.ugm.ac.id*, gedeyudi@mail.ugm.ac.id**, msmpakpahan@mail.ugm.ac.id***, igi@ugm.ac.id******

Abstract. Breast cancer is the most common cancer among women (43.3 incidents per 100.000 women), with the highest mortality (14.3 incidents per 100.000 women). Early detection is critical for survival. Using machine learning approaches, the problem can be effectively classified, predicted, and analyzed. In this study, we compared eight machine learning algorithms: Gaussian Naïve Bayes (GNB), k-Nearest Neighbors (K-NN), Support Vector Machine(SVM), Random Forest (RF), AdaBoost, Gradient Boosting (GB), XGBoost, and Multi-Layer Perceptron (MLP). The experiment is conducted using Breast Cancer Wisconsin datasets, confusion matrix, and 5-folds cross-validation. Experimental results showed that XGBoost provides the best performance. XGBoost obtained accuracy (97,19%), recall (96,75%), precision (97,28%), F1-score (96,99%), and AUC (99,61%). Our result showed that XGBoost is the most effective method to predict breast cancer in the Breast Cancer Wisconsin dataset.

Keywords: XGBoost, MachineLearning, Breast Cancer, Classification

INTRODUCTION

Breast cancer is the most common cancer among women, with 43.3 incidents per 100.000 women. Breast cancer has a relatively low fatality rate with other types of cancer. However, with a large number of incidents, it has the highest mortality rate of any cancer among women (12.9 per 100 000). Early detection is critical for survival. Approximately 70% of deaths from cancer occur in low- and middle-income countries. Limited resources with underdeveloped health systems making it difficult for the patient to get access to the medical professional. Developing early diagnosis programs based on early signs and symptoms can improve the patient survival rate. (World Health Organization. 2019).

With a growing dataset of breast cancer patients, it is more feasible that machine learning methods are implemented to provide a quick, automated, and deeper understanding of cancer healthcare (Maity, G., and Das, S. 2017). Detection requires accurate prediction, and available large datasets give us the opportunity for an accurate prediction. However, the problem lies in what method will provide us with the best result.

Previous research has compared various machine learning methods to predict breast cancer. Using the Wisconsin Breast Cancer dataset, a performance comparison of Support Vector Machines or SVM, Decision Tree (C4.5), Naïve Bayes, and k-Nearest Network or kNN were conducted by Asri, H et al. (Asri, H. et al. 2016). SVM reached the highest accuracy by 97,13% and outperformed other algorithms. Bayrak, E. et al. (Bayrak, E. et al. 2019) compared SVM and Artificial Neural Network or ANN to predict breast cancer in early stages. The result showed that SVM has the best performance of 96,9957%. Gbenga, D. et al. (Gbenga, D. et al. 2017) compared eight machine learning algorithms to predict breast cancer using WEKA data mining and machine learning simulation environment. Algorithms compared in this research are SVM, Radial Based Function, Simple Linear Logistic Regression Model, Naïve Bayes, kNN, AdaBoost, Fuzzy Unordered Role Induction algorithm, and Decision Tree (J48). Their experimental result indicated that SVM has the best performance (97.07%).

Though previous research showed that SVM comparatively is the most accurate method, other comparative studies showed accurate methods. Amrane, M. et al. (Amrane, M. et al. 2018) compared two different classifiers: Naïve Bayes and kNN for breast cancer classification. They used cross-validation methods as a tool to evaluate accuracy. The result showed that K-NN gives better accuracy than Naïve Bayes (97,51%). Sharma, S. et al. (Sharma, S. et al. 2018) compared Random Forest or RF, K-NN, and Naïve Bayes to predict breast cancer using The Wisconsin Breast Cancer dataset. The research concluded that KNN has the highest accuracy of 94,20%. Liu, B. et al. (Liu, B. et al. 2018) compared several machine learning methods including SVM, AdaBoost, Decision Tree, and Random Forest or RF to predict the benign and malignant of breast cancer from a digitized image of a fine needle aspirate of a breast mass. Theresult showed that the random forest is the best method for prediction.

These results show that comparative research can be done on these methods. An objectively best method can

be identified using a common dataset. There are other methods that have never been compared to predict breast cancer. Therefore, we have evaluated the performance of the following machine learning algorithms: Gaussian Naïve Bayes (GNB), K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), Random Forest (RF), AdaBoost, Gradient Boosting (GB), XGBoost, and Multi-Layer Perceptron (MLP).

MATERIALS AND METHODS

Dataset

The Wisconsin Breast Cancer dataset from *UCI Machine Learning Repository* was used in this experiment. There are 30 numeric attributes of features in the dataset. Features are calculated from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe the characteristics of the cell nuclei present in the image. Dataset has 569 instances that divided into two classes: benign and malignant. Benign consists of 212 instances, and malignant consists of 357 instances (UCI, 2019).

Experiment Method

Design of the experiment is presented in Figure 1:

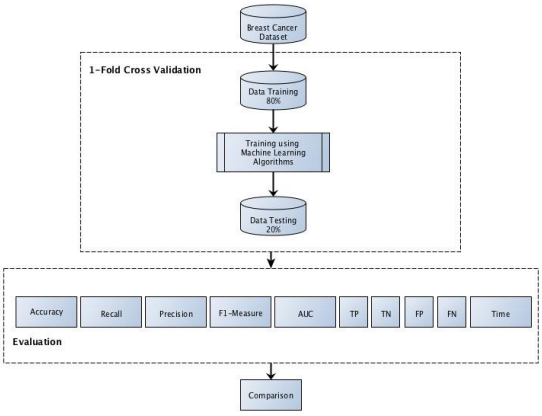

Figure 1. 1-Fold Cross Validation.

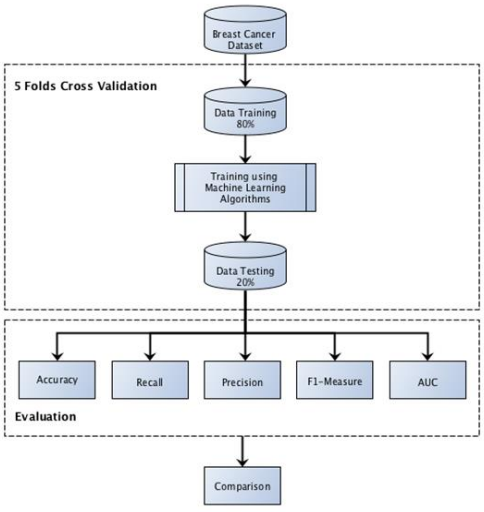

Figure 2. 5-Fold Cross Validation.

As we can see in Figure 1 and Figure 2, we used two ways to validate the performance of machine learning algorithms: (1) 1-Fold Cross-Validation and (2) 5-Folds Cross-Validation. Performance metrics that we used are accuracy, recall, precision, F1-Score, and Area Under Curve (AUC). In this experiment, we split the data into 80% data training and 20% data testing. Then, we train and test every algorithm for one iteration (1-fold cross-validation). After that, we run five iterations (5-folds cross-validation) using different, randomized data training and data testing from the dataset. Finally, we average every performance metric. All processes of the experiment are performed with Python 3.7 on a Laptop with 2.5GHz Intel Core i5 and 16 GB RAM, running on macOS High Sierra.

a. *Peformance Metrics/Confusion Matrix*

First, the definition of true positive, true negative, false positive, and false negative are defined in Table 1.

Table 1. Definitions of True Positive, True Negative, False Positive, and False Negative.

		Predicted class	
		Class = True	**Class = False**
Actual class	**Class = True**	True positive	False Negative
	Class = False	False Positive	True Negative

Performance metrics that we used are described as follows:

1) Accuracy
   Accuracy represents a degree of correctness in the training of the model. It is defined as the measurement of correct prediction compared to all predictions. The equation for accuracy is presented below:
   $$\text{Accuracy} = \frac{(True\ Positive + True\ Negative)}{(True\ Positive + False\ Positive + True\ Negative + False\ Negative)}$$
2) Recall
   Recall is a ratio to correctly determined positive instances to True Positive and False Negative. The equation for recall is presented below:
   $$\text{Recall} = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$
3) Precision
   Precision is a degree of correctness in determining the ratio between True Positives and all positive prediction. The equation for precision is presented below:
   $$\text{Precision} = \frac{True\ Positive}{(True\ Positive + False\ Positive)}$$
4) F1-Score/F-Measure
   F1-Score is a weighted average of Precision and Recall. The equation for F1-Scores is presented below:
   $$\text{F1-Score} = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

b. *AUC*
AUC measures accuracy by measuring the percentage of area under ROC curve. Wider the area under the ROC curve means that an algorithm has a better accuracy (Derisma, et al. 2018)

c. *K-Fold Cross Validation*
Cross-validation is a statistical technique used to check and evaluate learning algorithms or models. Cross-validation splits data randomly into a learning set to train and a testing set to evaluate the model. K-fold cross-validation is the basic form, one of the k partitions it is used as a validation set. We used k-fold cross-validation to validate performance metrics, has it represented the entire dataset (Suyanto. 2018).

Machine Learning Algorithms
In this section, we will describe every algorithm that we used in this experiment. Table 2 describes each basic algorithm concept, especially in classifying problem:

Table 2. Basic Concept of Each Algorithm.

GNB	**K-NN**	**SVM**	**RF**
Naive Bayes classifier is a simple probabilistic classifier based on the Bayes theorem. Bayes theorem considers every feature variable independent, as in features that do not depend on other features. It requires a small amount of training data, as it is one of the simplest learning methods. GNB is a type of naive Bayes classifier using continuous data(Gayathri, B. and Sumathi, C. 2016).	K-NN works by detecting input cases within the trained area by calculating the nearest neighbor with similar features of the case. There is no training period, as there is no need to build a decision model. Every new data can be added without impacting the system, as every learning period is done at the beginning of the prediction system. However, K-NN has a problem working on large datasets. As the larger dataset tends to have high dimensional data. K-NN also does not work well with noisy data. Outliners in training data may result in inaccurate predictions. (Kumar, A. et al. 2019)	SVM is a machine learning method that is conceptually similar to perceptron or ANN. SVM's goal is to find a hyperplane to separate the training data into two classes. This particular method is named binary class SVM, as there is only two class that we want to identify. Different from ANN, SVM's goal is to find the optimum hyperplane. Optimum hyperplane has equal distance with the most outside data of both classes. In other words, hyperplane has the maximum margin with both classes. (Suyanto. 2018)	RF is one the variant of the bagging machine learning method. RF is a combination of the decision tree that every tree depends on random vector value that has been sampled independently with the same distribution for every tree in the forest. Different from the bagging method, RF does not use every attribute to make an independent model. Instead, it uses only 20% of the features. Then, it results in the computational time reduction. (Suyanto. 2018)
AdaBoost	GB	XGBoost	MLP
The main idea of AdaBoost is to train different classifiers (weak classifiers) for the same training set. These weak classifiers then are grouped to form a robust classifier. AdaBoost determines the weight of each sample in each training, whether it is correct in the overall. The new dataset with modified weight is then sent to the lower classifier for training. Finally, classifier in each training is fused as final decision classifier. This whole process is done iteratively. (Liu, B. 2018) (Liu, B. 2018)	GB is a set of classification and regression trees that uses a gradient descent algorithm to minimize lost when adding new trees. GB can solve prediction and regression problems. (Punmiya, R. and Choe, S. 2019)	XGboost is a more regularized and improved version of GB. There are two objectives in XGBoost: (1) the sum of the specific loss functions evaluated on all predictions and (2) the sum of the regularization terms for all predictors (k-tree). XGBoost is based on a pre-sort-based algorithm(Punmiya, R. and Choe, S. 2019).	MLP is an ANN model that has multiple layers: (1) input layer, (2) hidden layer, and (3) output layers. Using multiple layers allows the hyperlink or decision boundaries to be flexible in classifying even the most random and complex data. (Pham, B. et al. 2017) (Suyanto. 2018)

RESULTS AND DISCUSSION

Comparison Amongthe Algorithms with 1-Fold Cross Validation

At first, we compared the algorithms with 1-fold cross validation method. The result are shown in Table 3:

As we can see in Table 1, excluding computing time, XGBoost outperforms other algorithms. In context, XGBoost correctly classifies two more cases than the second-best algorithm, KNN. Even though it is lower than GNB, XGBoost also has better computational time than other algorithms, because it takes 0.08 seconds to process. GNB performed better, only requires 0.008 seconds to process. Interestingly, the SVM method, one of the best algorithms in previous studies, has the longest computational time and is one of the lowest performance metric results as well.

Table 3. Performance Metrics with 1-Fold Cross Validation.

Algorithm	Recall	Precision	F1-Score	Accuracy	TP	TN	FP	FN	Time
GNB	94,03%	94,03%	94,03%	92,98%	63	43	4	4	0,008s
kNN	98,51%	95,65%	97,06%	96,49%	66	44	3	1	0,12s
SVM	94,03%	98,43%	96,18%	95,61%	63	46	1	4	1,57s
RF	94,03%	98,44%	96,18%	95,61%	63	46	1	4	0,23s
AdaBoost	97,01%	95,59%	96,30%	95,61%	65	44	3	2	0,14s
GB	97,01%	97,01%	97,01%	96,49%	65	45	2	2	0,23s
XGBoost	**98,51%**	**98,51%**	**98,51%**	**98,25%**	**66**	**46**	**1**	**1**	**0,08s**
MLP	95,52%	94,12%	94,81%	93,86%	64	43	4	3	0,91s

Comparison AmongtheAlgorithmswith5-fold Cross Validation Method

We implemented5-fold cross validation method and average performance metric was taken. Result are shown in Table 4:

Similar to the previous result in Table 3, XGBoost has the best result compared to other algorithms. Performance metrics of XGBoost are Recall 96,75%, Precision 97,28%, F1-Score 96,99%, and Accuracy 97,19%. XGBoost also has the highest AUC of 99,61%. From the experiment, we can conclude that XGBoost is the most accurate algorithm to classify breast cancer using the Wisconsin Breast Cancer dataset.

Table 4. Performance Metrics with Cross Validation.

Algorithm	Recall	Precision	F1-Score	Accuracy	AUC
GBN	93,30%	94,77%	93,82%	94,21%	98,78%
kNN	90,49%	92,55%	91,26%	91,93%	95,98%
SVM	94,33%	95,01%	94,60%	94,91%	99,50%
RF	95,45%	95,82%	95,05%	94,91%	99,11%
AdaBoost	95,34%	96,21%	95,71%	95,96%	99,14%
GB	95,28%	96,16%	96,29%	95,96%	99,45%
XGBoost	**96,75%**	**97,28%**	**96,99%**	**97,19%**	**99,61%**
MLP	90,45%	91,73%	90,97%	91,56%	97,82%

CONCUSIONS

In this research, we compared eight different algorithms on the Wisconsin Breast Cancer dataset to classify breast cancer. Using 1-fold and 5-fold cross-validation, we collected the performance metric to determine the best algorithm. The result showed that XGBoost has the best performance metric against other algorithms, and also has the highest AUC of 99,61%. We conclude that XGBoost is the most accurate algorithm to classify breast cancer using the Wisconsin Breast Cancer dataset.In the future, XGBoost can be compared with other algorithms that have not been tested in this experiment and canbe tested on different datasets as well.

ACKNOWLEDGMENTS

This work is partially supported by the IndonesiaEndowment Fund of Education (LPDP), Ministry of Finance, Indonesia.

REFERENCES

Amrane, M. et al. 2018. Breast cancer classification using machine learning. Electric Electronics, Computer Science, Biomedical Engineering's Meeting, pp. 1-4.

Asri, H. et al. 2016. Using machine learning algorithms for breast cancer risk prediction and diagnosis. International Symposium of Frontiers in Ambient and Mobile Systems, pp. 1064-1069.

Bayrak, E. et al. 2019. Comparison of machine learning methods for breast cancer diagnosis. IEEE Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science, pp. 4-6.

Center for Machine Learning and Intelligent Systems (UCI Machine Learning Repository). Breast Cancer Wisconsin (Diagnotic) Dataset. 1995. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) [6 November 2019]

Derisma, et al. 2018. Optimization of neural network with genetic algorithm for breast cancer classification. International Conference on Information Technology Systems and Innovation, pp: 398 – 403.

Gayathri, B. and Sumathi, C. 2016. An automated technique using gaussian naïve bayes classifier to classify breast cancer. International Journal of Computer Applications, pp. 16 – 21.

Gbenga, D. et al. 2017. Performance comparison of machine learning techniques for breast cancer detection. Nova Journal of Engineering and Applied Science, pp. 1-8.

Kumar A. et al. Machine learning based approaches for cancer prediction: a survey. International Conference on Advanced Computing and Software Engineering, pp. 326 – 330.

Liu, B. et al. 2018. Comparison of machine learning classifier for breast cancer diagnosis based on feature selection. Proceeding of IEEE International Conference on Systems, Man, and Cybernetics, pp. 4399 – 4404.

Maity, G., and Das, S. 2017. Machine learning for improved diagnosis and prognosis in healthcare. IEEE Aerospace Conference, pp. 1-9.

Pham, B. et al. 2017. Hybrid integration of multilayer perceptron neural networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. Catena, pp: 52 – 63.

Punmiya, R. and Choe, S. 2019. Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing, pp: 2326 – 2329.

Sharma, S. et al. 2018. Breast cancer detection using machine learning algorithms. International Conference on Computational Techniques, Electronics and Mechanical Systems, pp. 114 – 118.

Suyanto. 2018. Machine learning tingkat dasar dan lanjut. Informatika, Bandung.

World Health Organization. 2019. Early diagnosis and screening for breast cancer. https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ [13 November 2019]

World Health Organization. 2019. Key fact about breast cancer. https://www.who.int/en/news-room/fact-sheets/detail/cancer [13 November 2019]

THIS PAGE INTENTIONALLY LEFT BLANK