Original Article

Healthc Inform Res. 2019 October;25(4):283-288.
https://doi.org/10.4258/hir.2019.25.4.283
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Stacking Ensemble Technique for Classifying Breast Cancer

Hyunjin Kwon[1], Jinhyeok Park[1], Youngho Lee[2]

[1]Department of IT Convergence Engineering, Gachon University, Seongnam, Korea
[2]Department of Computer Engineering, Gachon University, Seongnam, Korea

Objectives: Breast cancer is the second most common cancer among Korean women. Because breast cancer is strongly associated with negative emotional and physical changes, early detection and treatment of breast cancer are very important. As a supporting tool for classifying breast cancer, we tried to identify the best meta-learner model in a stacking ensemble when the same machine learning models for the base learner and meta-learner are used. **Methods:** We used machine learning models, such as the gradient boosted model, distributed random forest, generalized linear model, and deep neural network in a stacking ensemble. These models were used to construct a base learner, and each of them was used as a meta-learner again. Then, we compared the performance of machine learning models in the meta-learner to determine the best meta-learner model in the stacking ensemble. **Results:** Experimental results showed that using the GBM as a meta-learner led to higher accuracy than that achieved with any other model for breast cancer data and using the GLM as a meta learner led to low root-mean-squared error for both sets of breast cancer data. **Conclusions:** We compared the performance of every meta-learner model in a stacking ensemble as a supporting tool for classifying breast cancer. The study showed that using specific models as a meta-learner resulted in better performance than single classifiers, and using GBM and GLM as a meta-learner is appropriate as a supporting tool for classifying breast cancer data.

Keywords: Breast Cancer, Machine Learning, Data Analysis, Medical Informatics, Classification

I. Introduction

According to statistics released by the National Statistical Office in 2016, cancer is the leading cause of death among Koreans. The mortality rates also increased from the previous year in the order of leukemia, breast cancer, and brain cancer [1]. In particular, breast cancer is the second leading cause of cancer among Korean women, and it is strongly associated with emotional and physical changes, such as postoperative depression, anxiety, loss of self-esteem, and low quality of life. Therefore, early detection and treatment of breast cancer is very important [2-4]. The most common method to diagnose breast cancer was preoperative mammography, but nowadays, the most frequently used method of diagnosing breast cancer is the fine-needle aspiration test [5]. The fine-needle aspiration test can be used to diagnose cancer easily without surgery by collecting a specimen from the lesion of the patient through a fine needle, observing its characteristics, and diagnosing the malignancy [6]. However,

Submitted: June 9, 2019
Revised: October 4, 2019
Accepted: October 6, 2019

Corresponding Author
Youngho Lee
Department of Computer Engineering, Gachon University, 1342, Seongnam-daero, Sujeong-gu, Seongnam 13120, Korea. Tel: +82-31-750-4759, E-mail: lyh@gachon.ac.kr (https://orcid.org/0000-0003-0720-0569)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ⓒ 2019 The Korean Society of Medical Informatics

there is a difficulty in accurate diagnosis since fine-cell aspiration is affected by various external variables, such as the small size of cancer, noninvasive breast cancer, and low stage [7]. In addition, diagnosis can be inaccurate due to various factors, such as the skills of the examiner and the similarity between organisms [8,9]. To overcome these limitations, it is necessary to provide an auxiliary tool to diagnose breast cancer accurately.

Machine learning, which is a field of artificial intelligence technology that makes machine judgments through data learning, has been widely used in industries such as medicine, manufacturing, and finance [10]. Machine learning is largely divided into supervised learning and unsupervised learning. Supervised learning is a method that enables a model to deduce correct labels for new data by learning data with labels. On the other hand, unsupervised learning refers to a learning technique in which data without labels is learned, and data are clustered according to the similarity of their characteristics. The common characteristics that supervised and unsupervised learning have are the importance of the data to use and the importance of choosing a prediction model for learning. There is a problem that if the label of the data is biased, or the selected model is over-fitted with the corresponding data, the model cannot show the correct performance. The machine running technology that emerged to solve these problems is ensemble techniques. Ensemble techniques are largely divided into three categories, namely, bagging, boosting, and stacking. Bagging a method of allowing redundancy when randomly classifying training data to a prediction model in the process of learning. Boosting is similar to bagging, but it reuses the data that the prior model cannot classify well in the learning process by providing weight. Stacking is an ensemble technique that is also known as super learning. It uses various models, such as random forest, k-nearest neighbors, and support vector machine as base learners to generate new data based on predicted results. Then this new data is used for another predictive model, which is called meta-learner, to finally derive the predicted value [11]. Studies on ensemble techniques have been steadily progressing, and many of them have been limited to the techniques of bagging and boosting ensembles [12-14]. If bagging and boosting ensemble techniques are used to generate strong predictive models using the same incomplete and weak prediction models, then staging ensemble techniques are available to combine various predictive models [15]. Many previous studies have introduced new stacking ensemble techniques for breast cancer prediction and to increase the performance. As prior research on stacking ensembles, research to determine the best meta-learner model is needed as when the same machine learning models are used in the base learner and meta-learner. Therefore, we compared the performance of meta-learners and identified the best meta-learner model in a stacking ensemble as a predictive aid for breast cancer patients.

II. Methods

1. Data Source

The Wisconsin Breast Cancer Diagnostic (WBCD) data provided by the University of Wisconsin consisted of data from 569 consecutive fine-needle aspiration tests, with 10 variables representing the characteristics of the nucleus, each with mean, standard deviation, and ideal values [16]. For example, each variable radius contains three variables, namely, the mean radius (mean radius of cell), radius SD (standard deviation of cell), and worst radius. Therefore, the WBCD consists of 30 variables, except for diagnostic results that indicate benign and malignant.

The Wisconsin Breast Cancer - Original (WBCO) data contained nine features and one diagnostic value for 699 breast biopsies, like the WBCD data.

For the WBCD data, each parameter value has different numerical measurement values. For this study, other numerical measurement values except for diagnosis results and ID were normalized by using the following equation:

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)}.$$

Unlike WBCD data, WBCO data has the same range of numerical measurement values and is not normalized. Only 10 rows with missing values are removed. In this study, the WBCO data and the WBCD data were randomly extracted at a ratio of 8:2 to train data and test data, so that the values were not duplicated. Training data was used to train the model, and the performance was evaluated using the test data.

2. Model Development

The algorithm used for the stacking ensemble in this study consists of four models.

1) Gradient Boosting Machine

Gradient Boosting Machine (GBM) is an ensemble model that creates a powerful model by grouping several decision trees; it can be used for both regression analysis and classification analysis. GBM creates a tree sequentially in a way that compensates for the error of the previous tree and is expressed as follows:

$$D(x) = d_1(x) + d_2(x) + d_3(x) + \cdots.$$

Here, $d_n(x)$ is generated by minimizing the error of $d_{n-1}(x)$ with each sequentially generated tree.

2) Distributed Random Forest

Distributed Random Forest (DRF) is an ensemble model used to reduce over-fitting while maintaining predictive performance by combining decision trees with slightly different variable values. DRF extracts features randomly with the bootstrap technique and generates several decision trees with corresponding values. The average of the predicted results in each decision tree is used as the final prediction result, and DRF has excellent performance even when the parameter is used as the default value.

3) Generalized Linear Model

Generalized Linear Model (GLM) is mainly used when the dependent variable is a categorical variable. The dependent variables used in this study are categorical variables (benign and malignant) and binomial variables at the same time, so it is appropriate to use the model that is expressed as follows:

$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n,$$

where E is the expected value, and $\beta_0, \beta_1 \cdots \beta_n$ are the regression coefficients. Here, y is the probability that an observation belongs to a group, x is the probability of not belonging to the group, and the higher the E value, the higher the probability of belonging to the group.

4) Deep Neural Network

Deep Neural Network (DNN) is a neural network with multiple hidden layers between the input layer and the output layer. DNN is a neural network structure that can produce different results depending on the number of layers and the number of nodes in each layer.

3. Experiment Environment and Model Architecture

This study was implemented using the H2O package of the R program (ver.3.4.3). The H2O package is an open-source library from H2O.ai, which provides a variety of supervised and unsupervised learning algorithms [17]. It also supports various languages, such as R, Python, and Java. Most of the prediction models in this study were implemented using the basic parameter values of the H2O package, and 5-fold cross-validation was performed to ensure the reliability of the data.

As shown in Figure 1, base learner models were implemented using h2o.gbm, h2o.randomForest, h2o.glm, and h2o.deeplearning functions, and a true value is given to determine whether to keep cross validation as the parameter value of the corresponding functions.

First, h2o.gbm has a default number of 50 trees with a maximum depth of 5. h2o.randomForest has a default number of 50 trees with a default depth of 20. h2o.glm has a family parameter Gaussian as a default value, but in this study, a binomial was used because the dependent variable is binomial distribution. By default, h2o.deeplearning has two hidden layers and 200 nodes, the epoch is 10, and the activation function is the ReLU function. The meta-learner was implemented using the h2o.stackedEnsemble function, and the functions used in the base-learner model were used once again.

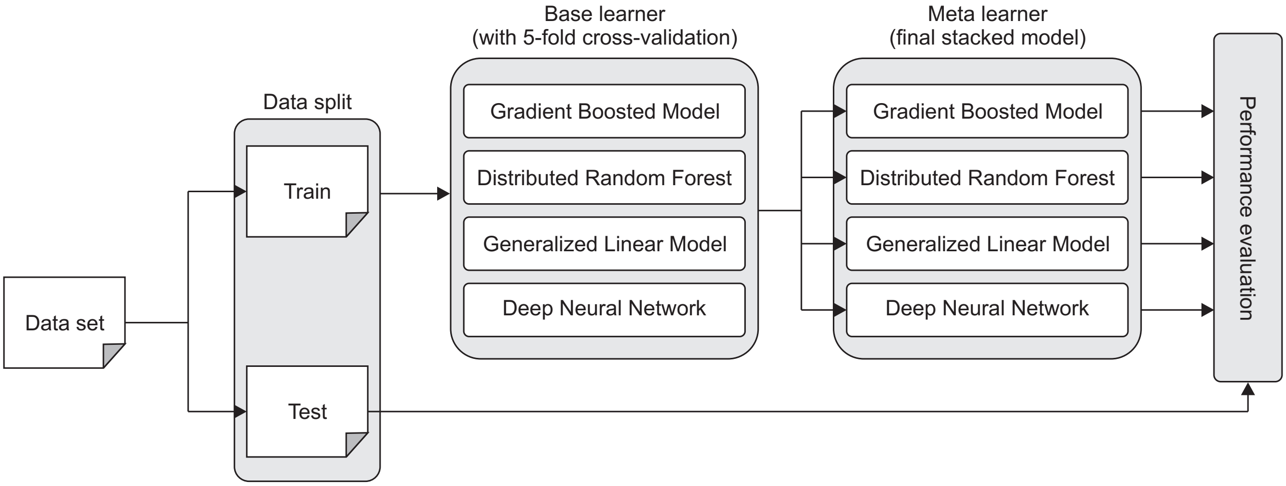

Figure 1. Architecture of stacking ensemble.

III. Results

In this study, we compared the accuracy and root-mean-square error (RMSE) of each prediction model for performance evaluation. In performance evaluation, accuracy indicates the accuracy of the prediction model, and the RMSE is a value calculated by comparing the predicted value with the observed value to determine how much error there is. Therefore, the lower the RMSE is, the better the performance is. To guarantee the reliability of the results, we randomly extracted the data and distributed it to the training data and the test data by randomly repeating this process five times and compared the average values of accuracy and RMSE. Figures 2 and 3 compare the accuracy and RMSE values of each model with the accuracy and RMSE values obtained when a stacking ensemble was used. In Figures 2, 3 and Tables 1, 2, 'single classifier' means a prediction model using single GBM, DRF, GLM and DNN. 'Ensemble classifier' means a prediction model using the GBM, DRF, GLM, and DNN models as the base learner, and each model is used as a meta-learner again. Figure 2 shows the results of each prediction model for WBCD data. Figure 2 shows that the ensemble model for WBCD data achieved better prediction accuracy than the single models used. Also, the RMSE value of the stacking ensemble using DRF, GLM, and DNN was lower than that of the conventional single model.

Table 1 shows the average prediction accuracy and average RMSE of each model used to draw the graph in Figure 2. Table 1 shows that the ensemble model achieved higher accuracy than the single model. In particular, the ensemble model using GLM as a meta-learner showed the highest average prediction accuracy of 97.37%. Also, the ensemble model using GLM as the meta-learner showed the lowest RMSE value, 0.1873. In contrast, the single GLM model

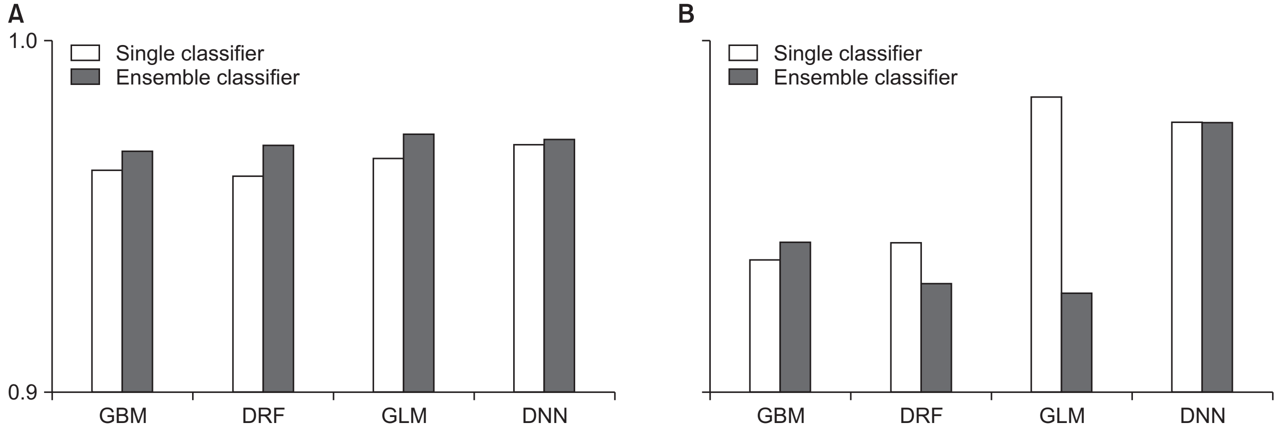

Figure 2. Performance comparison for the Wisconsin Breast Cancer Diagnostic: (A) accuracy and (B) root-mean-square error (RMSE). GBM: Gradient Boosted Model, DRF: Distributed Random Forest, GLM: Generalized Linear Model, DNN: Deep Neural Network.

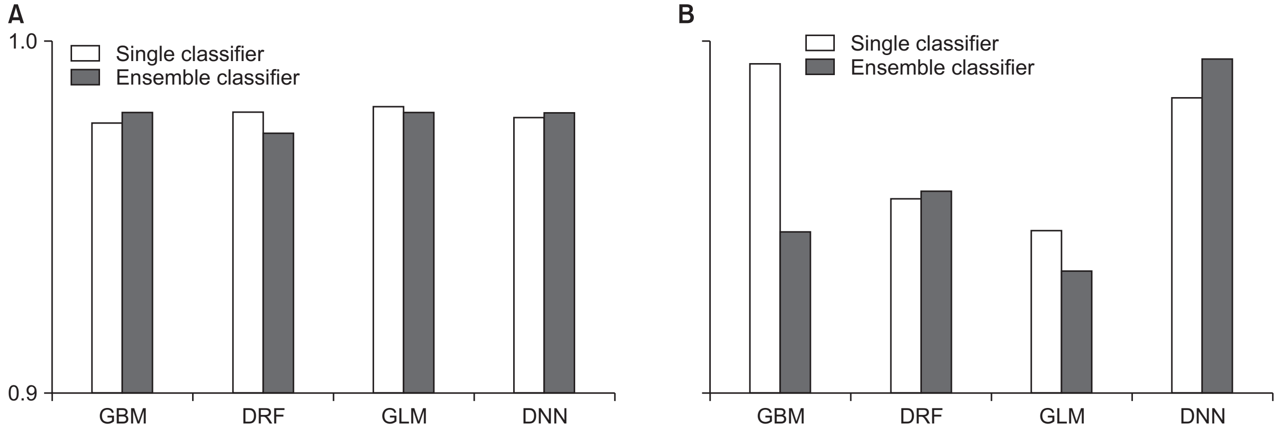

Figure 3. Performance comparison for the Wisconsin Breast Cancer - Original: (A) accuracy and (B) root-mean-square error (RMSE). GBM: Gradient Boosted Model, DRF: Distributed Random Forest, GLM: Generalized Linear Model, DNN: Deep Neural Network.

Table 1. Performance evaluation for the Wisconsin Breast Cancer Diagnostic

	Single classifier	Ensemble classifier
Accuracy		
GBM	0.9632	0.9684
DRF	0.9614	0.9702
GLM	0.9667	0.9737
DNN	0.9702	0.9719
RMSE		
GBM	0.1928	0.1959
DRF	0.1958	0.1888
GLM	0.2206	0.1873
DNN	0.2164	0.2163

RMSE: root-mean-square error, GBM: Gradient Boosted Model, DRF: Distributed Random Forest, GLM: Generalized Linear Model, DNN: Deep Neural Network.

Table 2. Performance evaluation for the Wisconsin Breast Cancer - Original

	Single classifier	Ensemble classifier
Accuracy		
GBM	0.9766	0.9796
DRF	0.9796	0.9737
GLM	0.9810	0.9796
DNN	0.9781	0.9796
RMSE		
GBM	0.1634	0.1515
DRF	0.1539	0.1544
GLM	0.1516	0.1487
DNN	0.1609	0.1637

RMSE: root-mean-square error, GBM: Gradient Boosted Model, DRF: Distributed Random Forest, GLM: Generalized Linear Model, DNN: Deep Neural Network.

showed the highest average RMSE value of 0.2206.

According to Figure 3, the WBCO data also shows that each model achieved high performance overall, but other ensemble models, except GBM and DNN, showed lower average predictive accuracy rates than those using a single model. The RMSE for the ensemble model using GBM and GLM as the meta-learner was lower, while the RMSE value for the ensemble model using DRF and DNN as the meta-learner was higher than that for the single model.

Table 2 shows the average prediction accuracy and average RMSE of each model used to draw the graph, as shown in Table 1. As seen in Table 2, only the ensemble model using GBM and DNN as the meta-learner showed an increase in average accuracy. On the contrary, the model with single GLM had the highest average prediction accuracy of 98.10%. In the case of average RMSE, the RMSE value was lower only when GBM and GLM were used as a meta-learner, and the model using GLM as a meta-learner showed the lowest RMSE value.

IV. Discussion

Unlike in the West, breast cancer in Korea has a very high incidence in people under 50 years old, rather than after 50 [18]. This means that the age which breast cancer can occur is lower than in the West. Therefore, early diagnosis and accurate prediction of breast cancer is needed through convergence with various fields. Recently, various studies have been conducted to improve the accuracy of cancer prediction using machine learning techniques [19]. In particular, research has focused on ensemble techniques that improve performance by combining various prediction models rather than one of the machine learning techniques [20]. In this study, we used a stacking ensemble technique that can combine various prediction models among ensemble techniques to generate prediction models and verified their performance. In this study, a stacking ensemble was constructed using GBM, DRF, GLM, and DNN, which are powerful prediction models, and each prediction model was used again as a meta-learner.

Since GBM, DRF, GLM, and DNN were predictive models that achieved high performance even if they were used alone, the same performance improvement could not be obtained in all models when the stacking technique was used. However, in this study, the stacking ensemble using GBM as a meta-learner showed better prediction accuracy than the existing single model for both sets of breast cancer data. In addition, the RMSE value was confirmed to be the meta-learner, and the stacking ensemble using the GLM decreased in different data. The results of this study demonstrated that a stacking ensemble using GBM and GLM as a meta-learner is suitable as an auxiliary tool for predicting breast cancer.

This research identified the best performance meta-learner model when the same models were used in the base learner and meta-learner. The result of this research could be a valuable reference for choosing machine learning models for further stacking ensemble research. However, this study had a limitation that it was based on the default parameter values provided by H2O; we expect that it would be possible to develop a stacking ensemble with higher performance by adjusting the detailed hyperparameters for each model.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was supported by Institute for Information & communications Technology Promotion grant funded by the Korea government (MSIT) (No. 2018-0-00861, Intelligent SW Technology Development for Medical Data Analysis).

ORCID

Hyunjin Kwon (http://orcid.org/0000-0003-3592-7431)
Jinhyeok Park (http://orcid.org/0000-0002-6839-4814)
Youngho Lee (http://orcid.org/0000-0003-0720-0569)

References

1. Statistics Korea. Causes of Death Statistics in 2016 [Internet]. Daejeon, Korea: Statistics Korea; c2019 [cited at 2018 Sep 10]. Available from: http://kostat.go.kr/portal/eng/pressReleases/1/index.board?bmode=read&aSeq=363695.
2. National Cancer Center. National Cancer Registration and Statistics in Korea 2015 [Internet]. Goyang, Korea: National Cancer Center; c2019 [cited at 2018 Sep 10]. Available from: http://ncc.re.kr/main.ncc?uri=english/sub04_Statistics.
3. Choi BJ, Park JH, Choe BM, Han SH, Kim SH. Factors influencing anxiety and depression in breast cancer patients treated with surgery. J Korean Soc Biol Ther Psychiatry 2011;17(1):87-95.
4. Nam SJ. Screening and diagnosis for breast cancers. J Korean Med Assoc 2009;52(10):946-51.
5. Jung SS, You YK, Park CH, Kim IC. Recent trends of breast cancer treatment in Korea. J Korean Surg Soc 1991;41(6):717-26.
6. Kim DD, Kim JH, Choi KW. The clinicopathologic factors affecting the false negativity of fine needle aspiration cytology (FNAC) in breast Cancer. J Korean Surg Soc 2002;62(5):403-7.
7. Layfield LJ, Glasgow BJ, Cramer H. Fine-needle aspiration in the management of breast masses. Pathol Annu 1989;24 Pt 2:23-62.
8. Oh BH, Park YS, Sung CW, Kim CS. Diagnostic value of ultrasound-guided fine needle aspiration cytology by a endocrine surgeon. Korean J Endocr Surg 2008;8(3):189-93.
9. Hong SW. Fine needle aspiration cytology of thyroid follicular proliferative lesions. Korean J Endocr Surg 2008;8(3):159-66.
10. Cho SJ, Kang SH. Industrial applications of machine learning (artificial intelligence). IE Mag 2016;23(2):34-8.
11. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol 2007;6:Article25.
12. Lim JS, Oh YS, Lim DH. Bagging support vector machine for improving breast cancer classification. J Health Info Stat 2014;39(1):15-24.
13. Krawczyk B, Galar M, Jelen L, Herrera F. Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. Appl Soft Comput 2016;38:714-26.
14. Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF. SVM and SVM Ensembles in Breast Cancer Prediction. PLoS One 2017;12(1):e0161501.
15. Nagi S, Bhattacharyya DK. Classification of microarray cancer data using ensemble approach. Netw Model Anal Health Inform Bioinform 2013;2(3):159-73.
16. UCI Machine Learning Repository [Internet]. Oakland (CA): University of California, Center for Machine Learning and Intelligent Systems; c2018 [cited at 2018 Aug 10]. Available from: http://archive.ics.uci.edu/ml.
17. H2O documentation overview [Internet]. [place unknown]: H2O.AI; 2019 [cited 2018 Aug 10]. Available from: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html.
18. Park JH, Jung YS, Jung Y. Factors influencing posttraumatic growth in survivors of breast cancer. J Korean Acad Nurs 2016;46(3):454-62.
19. Montazeri M, Montazeri M, Montazeri M, Beigzadeh A. Machine learning models in breast cancer survival prediction. Technol Health Care 2016;24(1):31-42.
20. Hsieh SL, Hsieh SH, Cheng PH, Chen CH, Hsu KP, Lee IS, et al. Design ensemble machine learning model for breast cancer diagnosis. J Med Syst 2012;36(5):2841-7.