EAI Endorsed Transactions
on Context-aware Systems and Applications

Research Article EAI.EU

# Predicting Breast Cancer with Ensemble Methods on Cloud

Au Van Pham[1][0000-0002-5244-2439], Tu Cam Thi Tran[2][0000-0001-5811-6952], Phuc Quang Tran[3][0000−0001−5271−2323], Hiep Xuan Huynh[4,*,][0000-0002-9213-131X]

[1]Cai Be Technical College, Tien Giang province, Vietnam
[2]Vinh Long University of Technology Education, Vinh Long province, Vietnam
[3] Department of Foreign Languages and Informatics People's Police College II HCM city, Vietnam
[4]Can Tho University, Can Tho city, Vietnam

## Abstract

There are many dangerous diseases and high mortality rates for women (including breast cancer). If the disease is detected early, correctly diagnosed and treated at the right time, the likelihood of illness and death is reduced. Previous disease prediction models have mainly focused on methods for building individual models. However, these predictive models do not yet have high accuracy and high generalization performance. In this paper, we focus on combining these individual models together to create a combined model, which is more generalizable than the individual models. Three ensemble techniques used in the experiment are: Bagging; Boosting and Stacking (Stacking include three models: Gradient Boost, Random Forest, Logistic Regression) to deploy and apply to breast cancer prediction problem. The experimental results show the combined model with the ensemble methods based on the Breast Cancer Wisconsin dataset; this combined model has a higher predictive performance than the commonly used individual prediction models.

**Keywords:** Bagging, Boosting, Stacking, Random Forest, Ensemble methods.

Received on 20 October 2022, accepted on 19 November 2022, published on 23 March 2023

doi: 10.4108/eetcasa.v8i2.2788

## 1. Introduction

Of all the cancers in the women, the breast cancer has the highest mortality rate, about 15%. Globally, an additional woman is diagnosed with the breast cancer every 15 seconds. Building an application to help the patients, they can detect early, and predict with high accuracy, that are limit the mortality rate of the breast cancer. In addition, the applications are developed on cloud computing to reduce the cost of infrastructure facilities, to storage devices, to process data, and it also helps users, who can be accessed anytime, anywhere on the applications.

To make the forecast with the highest (possible) accuracy, and to find the predictive model that is feasible, it is necessary. However, the current predictive models are only individual models that do not show generalization. Therefore, building a feasible, economical, highly generalizable predictive model is the desire not only of researchers but also of everyone in life. The method of combining many separate prediction models to give higher accuracy is called Ensemble methods.

In this study, we use ensemble methods including three models (Bagging, Boosting, Stacking). In it, we will implement the techniques of three models with R language with breast cancer dataset on Amazon cloud computing (AWS). The forecasting model by Ensemble method, specifically Stacking (Stacking model that stacks three sub-models: Random Forest, Gradient Boosting and

[*] Corresponding author. Email: hxhiep@ctu.edu.vn

Logistic Regression) has higher accuracy than previous individual models.

The structure of the article is divided into 6 Sections. Section 2 is Decision Forest (includes Decision tree; Bootrap; Bagging; Boosting; and Stacking). Section 3 showes the steps to connect the clouds. Section 4 is the Modeling. Section 5 is the Experiment with the Breast Cancer Wisconsin dataset; it presents the results of the proposed model with the comments and the evaluations. Section 6 is the Discussion and conclusion; it presents a summary of the results achieved.

## 2. Decision Forest

### 2.1. Decision tree

Decision tree [12] in **Figure 1** is a supervised learning algorithm that includes both classification and regression prediction algorithms. The use of decision trees is to create a training model by learning simple rules from the training data or previous data and to form the output of the model. Therefore, a decision tree has good predictive ability on the training data set and to use this decision tree to predict on the test data set, it is necessary to find this decision tree.

For example, decision tree with data is Beach Soccer, the data includes the attributes: Weather, Temperature, Humidity and Wind.

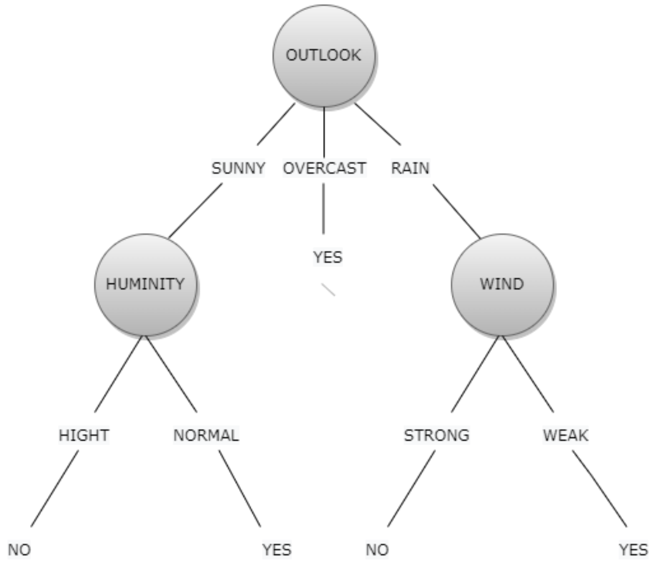


**Figure 1.** Decision tree

Based on the above model: If it is sunny, the humidity is normal, then the choice is to go out; And if it's sunny, the humidity is high, the choice is not to go out; If it's overcast, the choice is to hang out; If it rains, the wind is light, then the choice is to go out; When it rains, the wind is strong, the choice is not to go out.

There are many algorithms to build decision trees such as: ID3, C4.5, CART, CHAID, MARS… In which, the ID3 algorithm is a popular algorithm and it is widely used. The Iterative Dichotomiser 3(ID3) algorithm was proposed by Ross Quilan [27] in 1986.

At each node N, choose an attribute A (We can best classify N data based on attribute A).

Create sub-branches for A, then distribute the data into sub-branches, respectively.

Similarly, grow the tree until all training data is classified, or all attributes are used up.

Note that each attribute is only used once along the tree's path from the root to the leaf. The two factors to choose the root node for a decision tree are Entropy and Gain.

**Information gain:**

Given a set S, and class c, Entropy can be defined as follows:

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \quad (1)$$

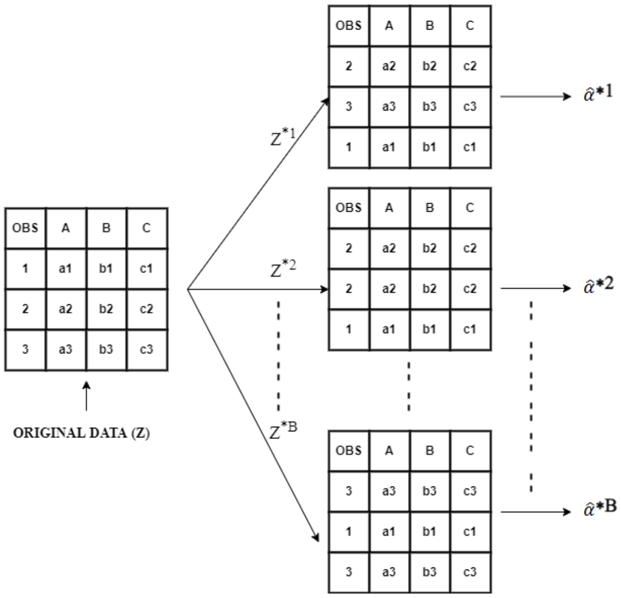
Information Gain measures the decrease in entropy if the set S is divided into subsets by attribute, it is calculated by the formula:

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (2)$$

### 2.2. Bootrap

The Bootstrap method [13] is a method of re-sampling or sampling with replacement, the Bootstrap method estimates parameters (parameters not resolved by other statistical methods).

Sampling with replacement (meaning taking a random sample from the samples, recording the value of the sample and returning it to the sample). For this method, an instance may appear more than once in a single sampling and continue to do so until the end of the sampling.



| OBS | A | B | C |
|---|---|---|---|
| 1 | a1 | b1 | c1 |
| 2 | a2 | b2 | c2 |
| 3 | a3 | b3 | c3 |

| OBS | A | B | C |
|---|---|---|---|
| 2 | a2 | b2 | c2 |
| 3 | a3 | b3 | c3 |
| 1 | a1 | b1 | c1 |

| OBS | A | B | C |
|---|---|---|---|
| 2 | a2 | b2 | c2 |
| 2 | a2 | b2 | c2 |
| 1 | a1 | b1 | c1 |

| OBS | A | B | C |
|---|---|---|---|
| 3 | a3 | b3 | c3 |
| 1 | a1 | b1 | c1 |
| 3 | a3 | b3 | c3 |

**Figure 2.** Boostrap

For example on **Figure 2**, with initial samples of 1, 2, 3

**Sample 1** has the values of A, it is a1, the value of B is b1, the value of C is c1.

**Sample 2** has the values of A, it is a2, the value of B is b2, the value of C is c2.

**Sample 3** has the values of A, it is a3, the value of B is b3, the value of C is c3.

Applying the bootstrap method, the first result is 2; 3; 1 corresponds with the values: a2, b2, c2; a3, b3, c3; a1, b1, c1. And return the sample, continue to take the second sample, the second result: 2; 2; 1 corresponds with the values: a2, b2, c2; a2, b2, c2; a1, b1, c1. And return the sample, continue sampling for the Bth time, the result of the B time: 3; 1; 3 corresponds with the values a3, b3, c3; a1, b1, c1; a3, b3, c3.

## 2.3. Bagging

Bagging method was proposed by Breiman [25], the aim of this method is to improve the efficiency with unbalanced data with single algorithm as decision tree. From the initial data set, the incense method is used to sample boostrap, and divide it into several subsets, and then train the same algorithm in parallel. The model results are the mean values of the models.

### Random Forest

Random Forest [9]: A forest is a collection of many trees; a decision forest is a collection of many decision trees. The bagging method builds a large collection of uncorrelated trees to improve prediction performance, it is shown in **Figure 3**.

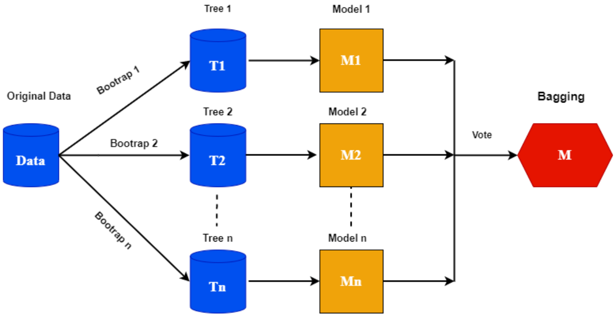


**Figure 3.** Bagging

## 2.4. Boosting

Boosting [6] was built with the desire to improve some limitations of Bagging (since the models in Bagging learn individually, it does not affect and relate to each other, leading to bad results when the results of the models are overlapped). Therefore, the weak models in Boosting will be learned from each other to limit the error of the previous model. The training process in this method takes place sequentially in sequence. Boosting's powerful algorithms include: Adaptive, Gradient and Extreme gradient.

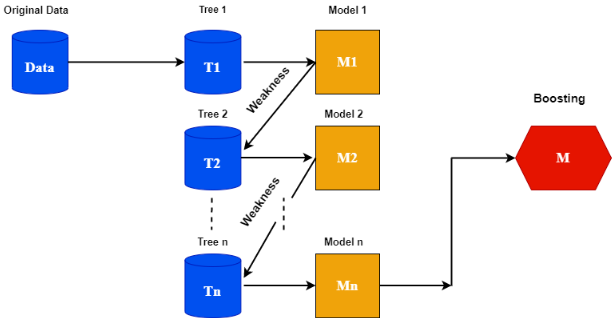


**Figure 4.** Boosting

Based on **Figure 4**, Boosting includes the following basic steps:

Step 1: Creating multiple data sets through random sampling with replacement over weighted data;

Step 2: Building learners sequentially;

Step 3: Combining all learners using a weighted averaging strategy.

### Adaptive Boosting (AdaBoost)

AdaBoost [14] is a form of Boosting algorithm that is widely used for binary classification model. AdaBoost is implemented by starting from the training data, successive models are generated from the error correction of the first model. This method is used with decision trees (the dicision trees have a short depth). When the first decision tree is created, the performance of the tree per training sample is used as information to decide the next tree. This next tree will focus on selecting the training sample. The training data is difficult to predict, they will be assigned more weight than other samples. In turn, the model is created, the performance of the former will impact the later models.

After building all the models, the prediction process will be performed on the new data. At this point, each node in the decision tree will be weighted depending on its accuracy on the training data.

### Gradient Boosting

Gradient Boosting [28] is an algorithm in the Boosting group. Step 1: train the decision tree model to predict the outcomes. Step 2: calculate and compare the difference error (difference between the 1st error of the first model and the actual value). Step 3: train the decision tree model with the same previous feature, but add the error of the previous model. Step 4: The predicted value of the second model is added to the prediction error of the first model. Step 5: The value combined by step 3 is considered as the new predicted value. Finally, calculate the error of error (2nd error) based on the error between this value and the actual value. Repeat until the required quantity or error value remains constant.

### Extreme gradient boosting (XGBoost)

Extreme Gradient or XGBoost [15] is a specific implementation of the Gradient Boost method, which uses approximations to find the best tree model. It uses a number of nifty techniques to implement with the structured data. Advantage of XGBoost: Training is very fast and it can be parallelized with distribution between clusters.

## 2.5. Stacking

Stacking [16] is a variant of the Ensemble model, which uses weak models. It then superimposes these weak models to create a predictive model, which has higher accuracy from the separate models (this method is intended to increase the robustness of the model). Like Boosting, Stacking uses more complex weighting schemes than Bagging (Bagging only uses simple uniform weight schemes). Stacking will combine the forecast results of several models. The Stacking model is shown in **Figure 5**.

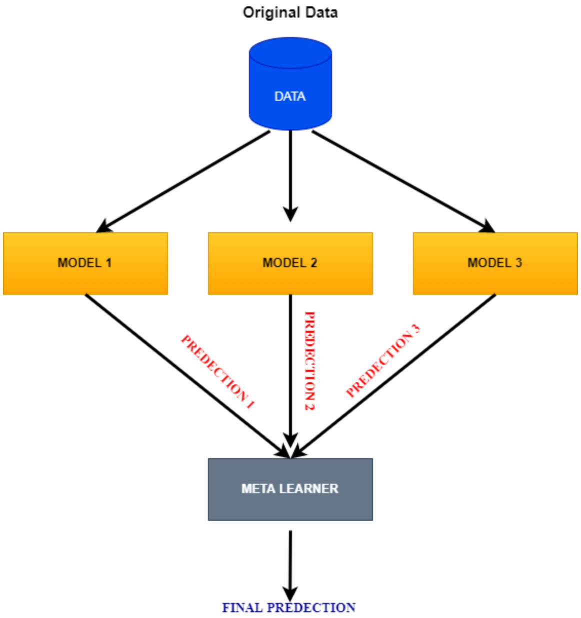


**Figure 5.** Stacking

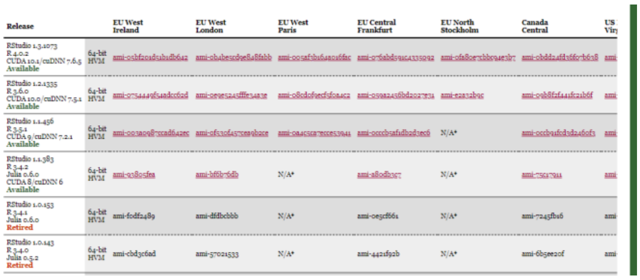
## 3. Connecting the clouds

The cloud computing services reduce the cost of infrastructure facilities, storage and data processing equipment. In addition, these services help users, who can access anywhere.

There are many ways to use cloud technology such as: IBM, Microsoft, Amazon, Google, VMware. In which, Amazon cloud technology (AWS) is selected to use with the R language.

Step 1: Accessing the page "https://aws.amazon.com/" to create an account.

Step 2: Using Amazon Machine Image (AMI), it is shown **Figure 6**.

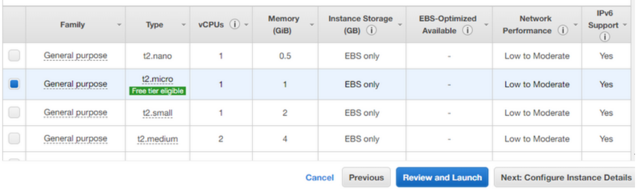


| Release | | EU West<br>Ireland | EU West<br>London | EU West<br>Paris | EU Central<br>Frankfurt | EU North<br>Stockholm | Canada<br>Central | US<br>Vir |
|---|---|---|---|---|---|---|---|---|
| RStudio 1.3.1073<br>R 4.0.2<br>CUDA 10.1/cuDNN 7.6.5<br>Available | 64-bit<br>HVM | ami-05bf201d51b1db642 | ami-0b4be5cd9e848fabb | ami-005af3b164a016fac | ami-076abd591c4335092 | ami-0fa80e7cbb094e3b7 | ami-0bdd24fd36f07b638 | ami- |
| RStudio 1.2.1335<br>R 3.6.0<br>CUDA 10.0/cuDNN 7.5.1<br>Available | 64-bit<br>HVM | ami-0754449f54adc062d | ami-0e9e5245fffe3483e | ami-08cd0f9ecf5f0a4c2 | ami-0592a2456bd2027e31 | ami-e2832b9c | ami-09b8f2f441fc21b6f | ami- |
| RStudio 1.1.456<br>R 3.5.1<br>CUDA 9/cuDNN 7.2.1<br>Available | 64-bit<br>HVM | ami-003a0987ccad642ec | ami-0f530f457cea9b0ce | ami-0a4c5ca7ecce53941 | ami-0cccb5af1db2d3ec6 | N/A* | ami-0ccb91fcd3d2460f3 | ami- |
| RStudio 1.1.383<br>R 3.4.2<br>Julia 0.6.0<br>CUDA 8/cuDNN 6<br>Available | 64-bit<br>HVM | ami-93805fea | ami-bf6b76db | N/A* | ami-a80db3c7 | N/A* | ami-75c17911 | ami- |
| RStudio 1.0.153<br>R 3.4.1<br>Julia 0.6.0<br>Retired | 64-bit<br>HVM | ami-f0df2489 | ami-dfdbcbbb | N/A* | ami-0e5cf661 | N/A* | ami-7245fb16 | ami- |
| RStudio 1.0.143<br>R 3.4.0<br>Julia 0.5.2<br>Retired | 64-bit<br>HVM | ami-cbd3c6ad | ami-57021533 | N/A* | ami-4421f92b | N/A* | ami-6b5ee20f | ami- |

**Figure 6**. Amazon Machine Image (AMI)

Step 3: Choosing the server to match.

Step 4: Selecting the machine configuration, which you want to use, it is shown in **Figure 7**.



|  | Family | Type | vCPUs ⓘ | Memory (GiB) | Instance Storage (GB) ⓘ | EBS-Optimized Available ⓘ | Network Performance ⓘ | IPv6 Support ⓘ |
|---|---|---|---|---|---|---|---|---|
| □ | General purpose | t2.nano | 1 | 0.5 | EBS only | - | Low to Moderate | Yes |
| ■ | General purpose | t2.micro<br>Free tier eligible | 1 | 1 | EBS only | - | Low to Moderate | Yes |
| □ | General purpose | t2.small | 1 | 2 | EBS only | - | Low to Moderate | Yes |
| □ | General purpose | t2.medium | 2 | 4 | EBS only | - | Low to Moderate | Yes |

**Figure 7**. Choose an Instance

Step 5: Selecting Configure Security Group

Step 6: Finding the computer's Public IP address, this IP has just been created and accessed

Step 7: Accessing and using the Rstudio on the AWS.

## 4. Modeling

### 4.1. Bagging model (random forest) in Rstudio

Bagging model (random forest) in Rstudio is shown in **Figure 8**, it includes the following steps:
Step 1: Using the Boostrap method to make 150 bags. Based on the variable "Diagnosis" (Benant(B) and Malignant(M)) to predict.
Step 2: Splitting the data into training data (80%) and test data (20%) to train 150 bags in parallel, which was built in step 1.
Step 3: 150 bags are recorded in this step.
Step 4: the final result of the model is given by selecting a majority of votes out of 150 bags (150 bags are recorded in step 4).
Step 5: Evaluating the performance of the decision tree model and the bagging model (random forest).

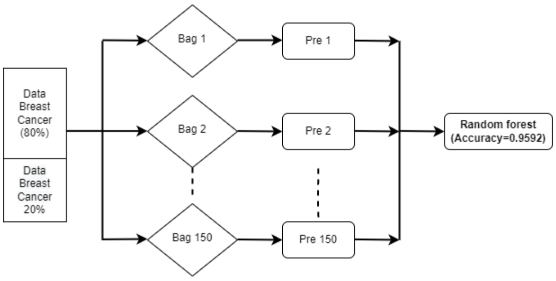


**Figure 8.** Bagging model

### 4.2. Boosting model in Rstudio

Boosting model in Rstudio is shown in **Figure 9**, it includes the following steps:
Step 1: Installing and using library (gbm), library (caret). Based on the variable "Diagnosis" Benign(B) and Malignant(M) to predict.
Step 2: Splitting the data into training data (80%) and test data (20%).
Step 3: Training the ada Boost model and the Gradient Boosting model.
Step 4: Making predictions with the test data set.
Step 5: Evaluating the performance of each model.

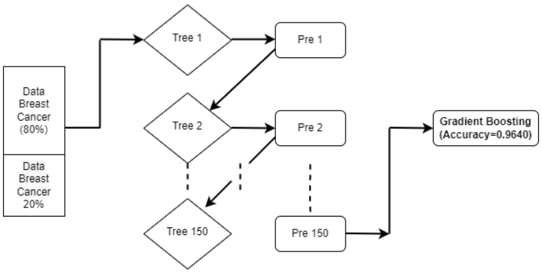


**Figure 9.** Boosting model

### 4.3. Stacking model with Random Forest, Gradient Boosting and Logistic Regression

Stacking model with Random Forest, Gradient Boosting and Logistic Regression is shown in **Figure 10**, it includes the following steps:
Step 1: Installing and using library (tidyverse), library (h20). Based on the variable "Diagnosis" Benign(B) and Malignant(M) to predict
Step 2: Splitting the data into training data (80%) and test data (20%)
Step 3: Training and cross-validate for three models: Random Forest, Gradient Boosting and Logistic Regression.
Step 4: Training the Stacking model.
Step 5: Evaluating the performance of each individual model and Stacking model.

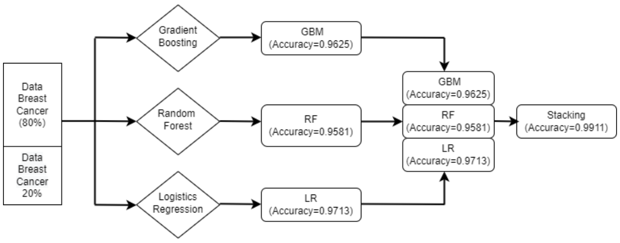


**Figure 10.** Stacking model

## 5. Experiment

### 5.1. Dataset

In this paper, we used the Breast Cancer Wisconsin dataset. The features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Ten real-valued features are computed for each cell nucleus: radius (mean of distances from center to points on the perimeter); texture (standard deviation of gray-scale values); perimeter; area; smoothness (local variation in radius lengths); compactness; concavity (severity of concave portions of the contour); concave points (number of concave portions of the contour); symmetry; fractal dimension ("coastline approximation" - 1).

## 5.2. Tools

We have programmed in R language for predictive model, the models used are: Bagging model, Boosting model and Stacking model. In addition, we also use libraries: library(dplyr), library(ggplot2), library(e1071), library(caret), library(rpart), library(ipred), library(tidyverse), and package h2o, library(h2o) to integrate models.

## 5.3. Scenario 1: The Bagging model

Based on the variable “Diagnosis” Benign (B) and Malignant (M) to make predictions about performance M, B. In this experiment, we chose Bagging algorithm with Random Forest to split the data into 150 bags (or 150 decision trees), that's to build the Bagging model. The results are presented in **Table 1.**

Table 1. The accuracy of Decision Tree and Random Forest

| Model | Decision Tree | Random Forest |
|---|---|---|
| Accuracy | 0.9276 | 0.9592 |

The average accuracy of the Bagging model (random forest) is 0.9592, which is higher than the average accuracy of Decision Tree Bagged of 0.9276. The comparison results are shown in **Figure 11**.

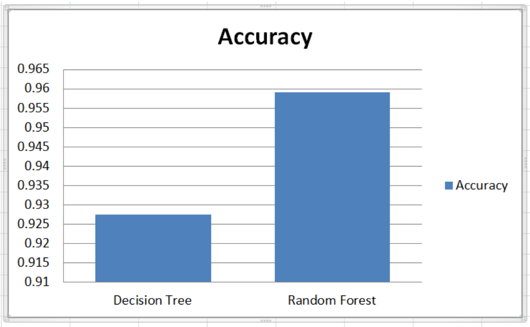


**Figure 11.** Evaluate accuracy for the Bagging model

## 5.4. Scenario 2: The Boosting model

In this experiment, we used the Gradient boosting method with library(tidyverse), library(h2o). Split the dataset into Train (80%) and Test (20%). The results are presented in **Table 2.**

Table 2. The accuracy of Ada Boost and Gradient Boosting.

| Model | Ada Boost | Gradient Boosting |
|---|---|---|
| Accuracy | 0.9598 | 0.9640 |

Accuracy of Gradient Boosting model is higher than AdaBoost's Model Accuracy. The comparison results are shown in **Figure 12.**

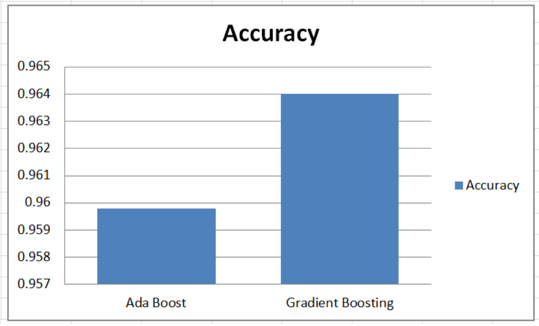


**Figure 12.** Evaluate accuracy for the Boosting model

## 5.5. Scenario 3: The Stacking model

We built the Stacking model by applying the following steps:

- Splitting dataset into 2 parts: 1 Train set (80%) and 1 Test set (20%).
- Running three models: Gradient Boosting, Random Forest and Logistic Regression with Cross-Validation method (k-folds = 5).
- Applying Random Forest to stack three 3 child models by.
- Comparing the accuracy of each model with the Stacking model on the test data set.

Table 3. The accuracy of Gradient Boosting, Random Forest, Logistic Regression and Stacking.

| Model | Gradient Boosting | Random Forest | Logistic Regression | Stacking |
|---|---|---|---|---|
| Accuracy | 0.9625 | 0.9581 | 0.9713 | 0.9911 |

With the results shown in **Table 3**, we can see the performance for each individual model. In which, the accuracy of Random Forest is 0.9581, it is the lowest. The accuracy of Gradient Boosting is 0.9625. The accuracy of Logistic Regression is 0.9713. And the Accuracy of Stacking model is 0.9911, it is the highest. The comparison results are shown in **Figure 13**.

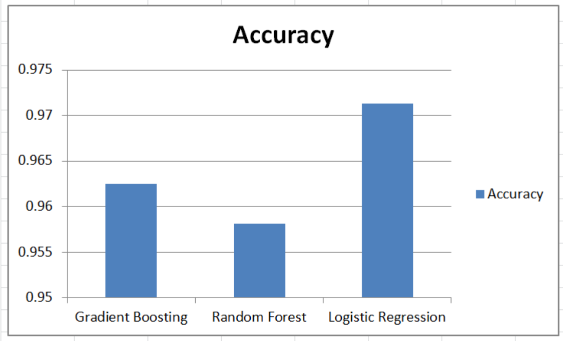


**Figure 13.** Evaluate accuracy for the Stacking model

## 6. Discussion and conclusion

In this article, we've explored and built three standard approaches in machine learning, and running the experiments in the AWS cloud using Rstudio. Apply them to the problem of predicting the likelihood of breast cancer with the variable “Diagnosis” (Benant(B) and Malignant(M)). The experimental results show that Ensemble methods are better than others. In which, the Stacking model based on three 3 child models give the best results in our test. The combination of the Ensemble method on the cloud platform will reduce the cost of facilities and storage devices.

## References

[1] Saleh H, Abdelghany FS, Alyami H, Alosaimi W. Predicting Breast Cancer Based on Optimized Deep Learning Approach. Hindavi. 2022; Article ID 1820777:11 pages.

[2] Asri H, Mousannif H, Al HM, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science. 2016; vol 83: pp 1064–1069.

[3] Yang R. Enterprise Network Marketing Prediction Using the Optimized GA-BP Neural Network. Complexity Article. 2020; ID 6682296.

[4] Zang C, Ma Y. Ensemble Machine Learning Methods and Applications. Springer Science+Business Media. 2012.

[5] Rosly R, Makhtar M, Awang M H. Rahman N D, Deris M H. Comparison of Ensemble Classifiersfor Water Quality Dataset. Proceedings of the UniSZA Research Conference 2015 (URC ’15). 2015; Universiti Sultan Zainal Abidin.

[6] Drucker H, Cortes C, Jackel L, LeCun Y. Boosting and Other Ensemble Methods. Neural Computation. 1994; vol 6: 1289-130.

[7] Todorovski L, Dzeroski S. Combining classifiers with meta decision trees. Researchgate. 2003; 50(3): 223-249.

[8] Wolpert DH. Stacked generalization. Researchgate. 1992; vol5(2): 241-259.

[9] Adele C, David R, John R. Random Forests. Springer. 2011; vol 45(1): pp 157-176.

[10] Pintelas P, Livieris E I. Ensemble Algorithms and Their Applications. Mdpi AG. 2020; ISBN 978-3-03936-959-1

[11] Aldhyani HHT, AI-Yaari M, Hasan Alkahtanni, Mashael Maashi. Water Quality Prediction Using Artificial Intelligence Algorithms. Hindawi. 2020; vol. 2020: Article ID 6659314: 12 pages.

[12] Rokach L, Maimon O. Decision Tree. researchGate, (2005).

[13] SOCIAL-SCIENCES https://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/bootstrap-method, (2022).

[14] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. December 19, 1996.

[15] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. August 2016.

[16] Nakano FK, Mastelini SM, Barbon S, Cerri R. Stacking Methods for Hierarchical Classification. IEEE 2017; vol 2017: 289-296.

[17] Robert E. Schapire. The strength of weak learnability. Manufactured in The Netherlands; 2017; vol 5 (2) :197-227

[18] Sultana J. Predicting Breast Cancer using Logistic Regression and Multi-Class Classifiers. Researchgate . 2018; vol 7.

[19] Cheng X, Whan W, Liang Y, Lin X, Luo J, Zhong W, Chen D. Risk Prediction of Coronary Artery Stenosis in Patients with Coronary Heart Disease Based on Logistic Regression and Artificial Neural Network. Computational and Mathematical Methods in Medicine. 2022; Article ID 3684700.

[20] Asri H, Mousannif H, Al Moatassime H, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Sciencedirect. 2016; vol: 83: 1064-1069.

[21] Chen H, Du M, Zhang Y, Yang C. Research on Disease Prediction Method Based on R-Lookahead-LSTM. Computational Intelligence and Neuroscience. 2022; vol: 2022, Article ID 8431912.

[22] Islam M Md, Haque Md R, Iqbal H, Hasan Md M, Hasan M, Kabir MN. Breast cancer prediction: a comparative study using machine learning techniques. Original research. 2020; vol: 1; no: 5; pp: 1–14.

[23] Prananda AR, Nugroho HA, Frannita EL. Rapid assessment of breast cancer malignancy using deep neural network. Springer, Surabaya, Indonesia Cairo, Egypt, October 2021; pp. 639–649.

[24] Alickovic E, Subasi A. Breast cancer diagnosis using ga feature selection and rotation forest. Researchgate. 2017; vol: 28; no. 4; pp: 753–763.

[25] Leo Breiman. Bagging predictors. Machine learning. 1996; 24(2):123–140.

[26] Sahran S, Qasem A, Omar K, Albashih D, Adam A, Abdullah SNHS, Abdullah A, Hussain RI, Ismail F, Abdullah N, Pauzi Md HS, Shukor Adb N. Machine Learning Methods for Breast Cancer Diagnostic. 2018,

Avialable: http://dx.doi.org/10.5772/intechopen. 79446, retrieved on 13th September, 2020.

[27] Quinlan J R. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81-106, 1986.

[28] Jerome H. Friedman. Stochastic Gradient Boosting. Jscimedcentral. 29 October 2018.