

# Breast Cancer Diagnosis Using Imbalanced Learning and Ensemble Method

Tongan Cai<sup>1,\*</sup>, Hongliang He<sup>2</sup>, Wenyu Zhang<sup>2</sup>

<sup>1</sup>Department of Electrical Engineering & Computer Science, University of Michigan, Ann Arbor, USA

<sup>2</sup>School of Information, Zhejiang University of Finance and Economics, Hangzhou, China

## Email address:

johncta@umich.edu (Tongan Cai), hhl@zufe.edu.cn (Hongliang He), wyzhang@e.ntu.edu.sg (Wenyu Zhang)

\*Corresponding author

## To cite this article:

Tongan Cai, Hongliang He, Wenyu Zhang. Breast Cancer Diagnosis Using Imbalanced Learning and Ensemble Method. *Applied and Computational Mathematics*. Vol. 7, No. 3, 2018, pp. 146-154. doi: 10.11648/j.acm.20180703.20

**Received:** May 25, 2018; **Accepted:** July 21, 2018; **Published:** August 3, 2018

---

**Abstract:** Worldwide, breast cancer is one of the most threatening killers to mid-aged women. The diagnosis of breast cancer aims to classify spotted breast tumor to be Benign or Malignant. With recent developments in data mining technique, new model structures and algorithms are helping medical workers greatly in improving classification accuracy. In this study, a model is proposed combining ensemble method and imbalanced learning technique for the classification of breast cancer data. First, Synthetic Minority Over-Sampling Technique (SMOTE), an imbalanced learning algorithm is applied to selected datasets and second, multiple baseline classifiers are tuned by Bayesian Optimization. Finally, a stacking ensemble method combines the optimized classifiers for final decision. Comparative analysis shows the proposed model can achieve better performance and adaptivity than conventional methods, in terms of classification accuracy, specificity and AuROC on two mostly-used breast cancer datasets, validating the clinical value of this model.

**Keywords:** Data Mining, Breast Cancer, Ensemble Method, Imbalanced Learning

---

## 1. Introduction

Accuracy matters most in clinical diagnosis. Such is especially the case when it comes to cancer diagnosis, where failure to detect fatal medical condition may result in death of the patient. Breast cancer, with highest patient death rate among all cancers (U.S. Breast Cancer Statistics), is one of the most threatening killer of women over 45. According to statistics from Breastcancer.org, 1 in 8 US women (about 12.4%) will develop invasive breast cancer over the course of her lifetime. To confirm the presence of abnormality early is to save patients' lives. When breast tumor is spotted, medical workers will need to classify it to be Benign (non-invasive) or Malignant (invasive cancer). With the help of information technology, computer-aided diagnosis (CAD), first proposed by Johnston (1994), has been bringing great changes to clinical decision making. Recent years, machine learning and data mining models have been well used in clinical field, various high-performance models are going to help medical workers on the detection and prediction of medical situations,

further improving the accuracy of cancer diagnosis. Breast cancer is one of the diseases that benefit from CAD, as well as many new data mining techniques.

In this study, an ensemble machine learning model is proposed for accurate diagnosis of breast cancer. This model deploys ensemble algorithm on multiple baseline classifiers including Random Forest (RF), Extra Tree (EXT), Gradient Boosting Decision Tree (GBDT), XGBoost (XGB), Support Vector Machine (SVM), Multilayer Perceptron (MLP) Neural Network and Logistic Regression (LR) after imbalanced learning algorithm is applied to data. Two datasets from Wisconsin Breast Cancer Database (WBCD) are used to validate and demonstrate the performance of the model, and the performance of the overall model is compared to the best performance of single baseline classifiers with optimal parameters, as well as some typical works of other scholars.

The following parts of this paper are organized as: 2) reviews literatures on the methods and algorithms used in this study, then evaluates several representative works on the

same topic; 3) elaborates the technical details of the proposed model, including data exploration & preprocessing, imbalanced learning approach, parameter optimization and ensemble structure; 4) presents the performance of the proposed model on the two datasets, and validates balancing method and ensemble structure by comparative analysis; finally, 5) draws the conclusion and 6) discusses the possibility of future works.

## 2. Related Works

### 2.1. Data Mining in Clinical & Medical Field

Data mining techniques are now well used in all works of life. In terms of data mining in clinical and medical field, great importance has been attached to the classification, diagnosis and prediction of the kind and course of diseases. Ever since the concept of “Data Mining” came out in the 1990s, clinical and medical data have been widely collected and learned through computational and statistical methods in search of useful and indicative information. In 1997, “Knowledge Discovery” was proposed by Prather et al. (1997) on obstetrical patient data, which was evolutionary at that time. New perspectives of “Knowledge Discovery” inspired scholars and a Bayesian method with neural network structure on adverse drug reactions database was proposed as early as Coulter et al. (2001).

It was not until 2005 that the number of works in clinical data mining boosts, and various data mining work has been proposed in fields like breast cancer (Peña-Reyes and Sipper, 1999), diabetes (Wang et al., 2005), pharmacovigilance (Wilson et al., 2003), readmission (Strack et al., 2014) and even gene association (Perez-Iratxeta et al., 2002). Up to now, with novel learning methods like random forest, deep neural network, more and more works have been proposed with ever higher target accuracy and automaticity. Niemeijer et al. (2010) developed an automated diabetic retinopathy detection system on clinical images; Xu et al. (2014) constructed deep neural network for colon cancer histopathology images classification. However, most data mining works only consider the performance of one single method, without considering the adaptivity and robustness of their method on different datasets.

### 2.2. Ensemble Methods & Breast Cancer Diagnosis

Ensemble method in machine learning describes the process of training multiple classifiers and make final decision based on the combination of classifiers rather than rely on a single one. Among all ensemble methods, three basic ideas, boosting, bagging(bootstrap) and stacking, are widely used to improve the adaptivity and robustness of the overall model, and potentially the classification accuracy. Boosting was proposed by Schapire (1990). It involves training multiple weak classifiers and weighs each classifier based on its classification performance. Breiman (1994) proposed Bagging, which generates multiple resampled training sets of the same size with repetition and the final

decision is made by the voting of all classifiers. The concept of stacking came from “Stacked Generalization” proposed by Wolpert (1992), that takes the prediction result of baseline classifiers on training sets to train a combinator classifier, often with cross validation. He et al. (2018) applied stacking algorithm for credit scoring, and a Stacking-based Approach was used by Han and Cook. (2013) to predict Twitter user geolocation. Ensemble learning methods are also widely developed for clinical needs. Eom et al. (2008) constructed a decision support system for Cardiovascular disease level prediction with ensemble method. Sarwar et al. (2015) ensembles Naïve Bayes, PART and decision table by hybrid ensemble algorithm for cervical cancer screening. Emamjomeh et al. (2014) chose RF, NB, SVM MLP as base classifiers and ensemble by a MLP classifier in their study of hepatitis C protein-reaction prediction.

Numerous works are seen in breast cancer diagnosis with machine learning. Wolberg and Mangasarian (1990) proposed a Multi-surface method of pattern separation as one of the earliest analysis of breast cancer data; Peña-Reyes and Sipper (1999) proposed a fuzzy-genetic approach that takes advantage of fuzzy logic and genetic algorithm. Akay (2009) proposed a SVM method for this diagnosis; Karabatak and Ince (2009) constructed neural networks with association rule. Osareh and Shadgar (2010) combines SVM, k-NN methods with feature engineering. Zheng et al. (2014) hybrids K-means method and SVM with feature extraction and Asri et al. (2016) compared the performance of C4.5, SVM, Naïve Bayes and k-NN method with the conclusion that SVM achieves best performance for this classification task. Ensemble methods also help with this task. Hsieh et al. (2011) alternatively chose Neural Fuzzy (NF), k-NN and Quadratic classifier (QC) and combine them with majority voting. Yavuz et al. (2017) combines Radial Basis Function Network (RBFN), Generalized Regression Neural Network (GRNN) and Feed Forward Neural Network (FFNN) by a solid voting and weighed sum method. However, few researches took imbalance learning algorithms and ensemble methods together into consideration, and related works are seldom found to have validated their models on both two datasets.

## 3. Method

In this research, the proposed model is validated on two datasets from Wisconsin Breast Cancer Database (WBCD). The WBCD was created by Dr. William H. Wolberg, at University of Wisconsin Hospitals, aiming to do accurate classification (Benign/Malignant) for the diagnosis of breast cancer. These two datasets are also available on Machine Learning Repository at the University of California – Irvine (UCI) <http://archive.ics.uci.edu/ml/index.php>.

### 3.1. Data Exploration

The first dataset (Original) contains 699 clinical cases reported by July 15, 1992, with an ID and 10 attributes for each case. The first 9 attributes are: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape,

Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses. Attributes are independent from each other and are presented by integer 1 to 10, categorically informing the measure of each attribute the case holds. The last attribute, Class, is the classification goal of the dataset and takes two categorical integers, i.e., 2 for Benign and 4 for Malignant. Here, 458 of cases are Benign and 241 are Malignant, the imbalance ratio is 1.90. 16 instances have missing value “?” for Bare Nuclei attribute.

The second dataset (Diagnostic) contains 569 instances, each corresponds to a digitized image of a fine needle aspirate (FNA) of a breast mass, from which the features are extracted. There are 32 attributes in the dataset, including ID, diagnosis (classification goal) and 30 real-valued features computed for each cell nucleus. 10 attributes, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension are considered, and the mean, best and worst values of measurements are calculated for each attribute. The diagnosis takes “M” for Malignant and “B” for “Benign” and in this dataset, there are 357 Benign and 212 Malignant, the imbalance ratio is 1.68. No missing value is spotted.

### 3.2. Data Preprocessing

For both datasets, attribute “ID” is dropped and the instances with missing value are omitted. After this data cleaning the first dataset (Original) has 683 instances, including 444 Benign and 239 Malignant. In the Class attribute of Original dataset, 2 was replaced by 0 and 4 by 1 for

easy manipulation, and in Diagnostic dataset, B was replaced by 0 and M by 1 for Diagnosis attribute. As the data in both datasets are not normalized and vary greatly in range, Min-Max-Scale is used to normalize the data. The idea of this method is described by the following formula:

$$X_{normalized} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

This scaling method allows each of the features to be normalized to a range of 0 to 1, which is suitable for this classification task. Note that the data in binary attribute manipulated above (Class and Diagnosis) will remain the same after scaling.

### 3.3. Modeling

The flow chart of the model proposed in this study is shown in Figure 1. As the graph indicates, the modelling process can be decomposed into 3 parts:

- 1) Dealing with training set with heavy imbalance ratio, by applying Synthetic Minority Over-Sampling Technique (SMOTE) algorithm
- 2) Using Bayesian Optimization to train baseline classifiers, including RF, EXT, GBDT, XGB, SVM, MLP and LR.
- 3) Deploying stacking method on optimized baseline classifiers, then training a combiner classifier

The optimized ensemble model is then used on the classification & validation of testing set.

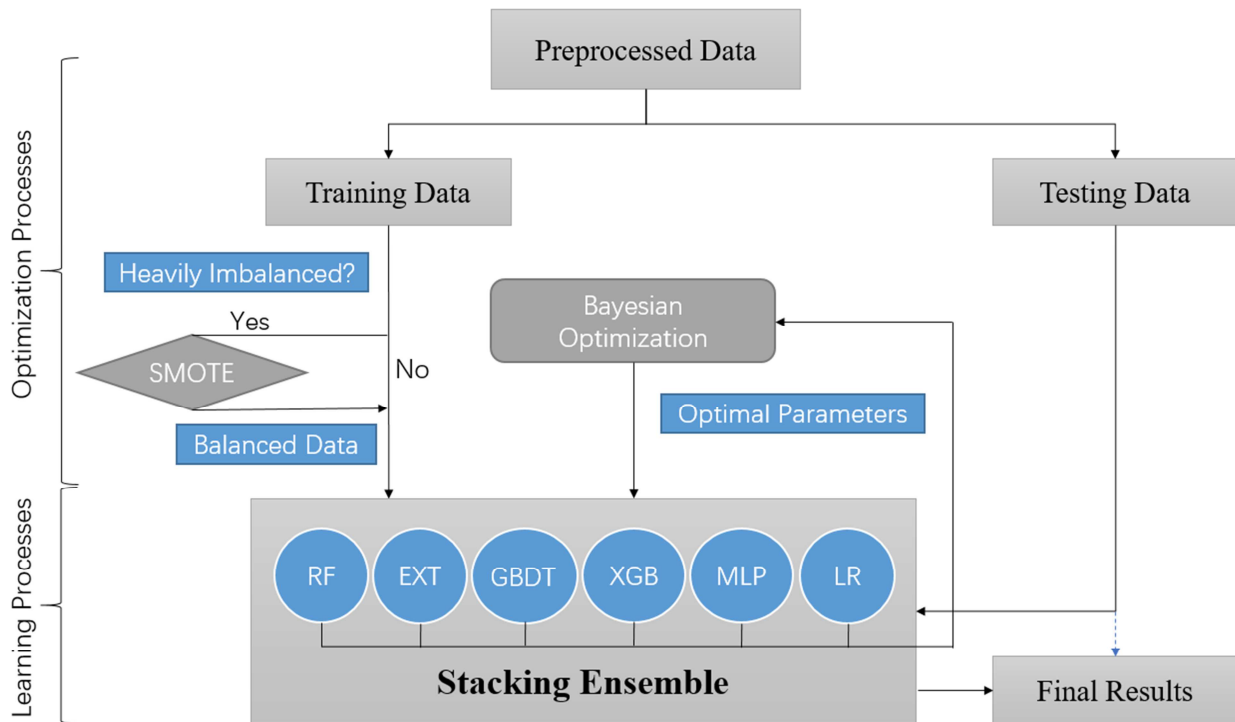


Figure 1. Flow of the proposed model.

### 3.3.1. SMOTE Imbalanced Learning Algorithm

When the training set is heavily imbalanced, minority class is harder to detect and will result in low prediction accuracy. SMOTE is one of the algorithms that deal with such imbalance.

SMOTE was proposed by Chawla et al. (2002). It does over-sampling on the minority class and helps achieve a better classification performance on imbalanced datasets. Figure 2 shows the flow of SMOTE algorithm.

---

#### Algorithm SMOTE algorithm

---

**Input:** Dataset, # minority samples T, Ratio N%, # nearest neighbors k

**Output:**  $(N/100) * T$  synthetic minority class samples

```

1: for Points in minority class do
2:   for i from 1 to N do
3:     Compute the k-nearest neighbors of the selected point
4:     Select one of the neighbors as point 2
5:     for each attribute of data do
6:       Choose a random place between two points selected
7:       record as the attribute of a synthetic point
8:     Add the synthetic point to original dataset as minority class
9: return Synthesized dataset

```

---

Figure 2. SMOTE algorithm.

The novelty inside this algorithm that makes it different from traditional oversampling method—which uses replica of instances—is that the SMOTE algorithm considers “synthetic points”. After applying SMOTE algorithm on the training set, the addition of synthetic points ensures the quantities of original majority and minority dataset are almost the same. This is of great importance especially when the minority class indicates abnormality: failure to detect abnormal cases can result in huge cost. This is often the case in terms of clinical data, including the data used in this study, where low classification accuracy on minority class Malignant may cause patients’ deaths. Therefore, SMOTE algorithm can be a useful method to eliminating imbalance in training classifier in this

study.

### 3.3.2. Bayesian Optimization

Bayesian optimization is a function optimizing process, in which, a posterior distribution—a guess—what the functions are expected to be—is constructed to describe the properties of the “black-box” function we’re interested in. This is called a Gaussian Process. Then, the posterior distribution is used to determine the next point to explore or exploit by an acquisition function. Increasing the number of observation will offer us a description of the target function with higher confidence. The following pseudo code in Figure 3 illustrates the algorithm:

---

#### Algorithm Bayesian Optimization

---

```

1: for t = 1,2,... do
2:   Define cumulative observation  $D_{1:k} = \{x_{1:k}, f(x_{1:k})\}$ 
3:   Construct posterior distribution  $P(f|D_{1:t})$ 
4:   Find  $x_t$  by optimizing the acquisition function over the GP:
5:    $x_t = \operatorname{argmax}_x u(x|D_{1:t-1})$ 
6:   Sample the objective function:  $y_t = f(x_t) + \text{error}$ 
7:   Add the t observation to  $D_{1:t-1}$ , construct  $D_{1:t} = \{D_{1:t-1}, (x_t, y_t)\}$ 
8:   Update GP with  $D_{1:t}$ 

```

---

Figure 3. Algorithm for Bayesian Optimization.

This method was first introduced to machine learning by Snoek et al. (2012) and is now widely used in the process of tuning the hyperparameters of machine learning classifiers. Unlike the mostly used grid search or random search method, Bayesian Optimization in hyperparameter tuning does not require a parameter map since it decides for itself the next point to try. Moreover, using Bayesian Optimization in parameter tuning is more likely to give us the real optimal solution with minimum number of iterations. The automaticity of Bayesian Optimization tuning method handles the manipulation difficulty in grid search and random search,

allowing same model to be applied to different datasets.

### 3.3.3. Stacking Ensemble Model

Stacking is a combinative learning algorithm for training multiple classifiers into a “blended” classifier. The basic idea of stacking is to train separate baseline classifiers on dataset first, then train a combinator classifier which takes the prediction result of baseline classifier as input. The final prediction is obtained using this combinator classifier. Figure 4 shows the core idea of stacking ensemble method:

**Algorithm** Stacking Algorithm**Input:** Training data, classifiers, #folds N**Output:** Combinator Classifier

```

1: for Classifier in classifiers do
2:   Splitting training & testing subsets by N folds
3:   //just like cross-validation
4:   for i from 1 to N do
5:     Train Classifier on a training set, predict the testing set
6:     Record the testing data and prediction results
7:   Record the mean of predictions as the prediction of classifier on the
   training subset
8: Take the prediction and labels, use Logistic regression as combiner
9: return the overall classifier

```

**Figure 4.** Stacking Algorithm.

Comparing to boosting and bagging, stacking is more adaptive for classifiers, and with classifier of different organism, the overall classifier will be more robust. In this study, 7 optimized baseline classifiers are used and trained using 5-fold cross validation, they are: RF, EXT, GBDT, XGB, SVM, MLP and LR. One extra LR model is selected as the combiner classifier.

## 4. Experiment

This section focuses on the experiment setup, procedures and results for the proposed ensemble model on two breast cancer datasets. The split of training and testing set follows the ratio of 0.85:0.15 and is randomly chosen. Training sets from both datasets are processed and balanced by SMOTE algorithm. Hyperparameters of baseline classifiers are automatically tuned using Bayesian Optimization. All processes of the experiment are performed with Python 3.6 on a Laptop with 2.4GHz Intel Core i7 and 8GB RAM, running Ubuntu 16.04 LTS OS.

### 4.1. Metrics of Model Performance

For classification tasks, four possible prediction cases would be spotted: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). This study follows the convention and define 1, the Malignant cases to be negative and 0, the Benign cases to be positive. A confusion matrix (shown as Table 1) is used to demonstrate the four cases:

**Table 1.** Confusion Matrix

Prediction \ Real	Positive (Predict Benign)	Negative (Predict Malignant)
Positive (Real Benign)	TP	FN
Negative (Real Malignant)	FP	TN

The metrics used in this study include:

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

$$Specificity = \frac{TN}{TN + FP}$$

Accuracy (Precision) focuses on the overall performance of classifiers on target attribute. This metric may face with a serious problem when the dataset is extremely imbalanced, i.e. the positive class takes up 90% and the negative class 10%. Classifiers achieve good accuracy (90%) if they predict every instance to be positive. Hence, different metrics will potentially be needed for performance analysis. Specificity (Negative Recall) focuses on illustrating the performance of classifier in predicting the negative class. In this study, the minority class is Malignant, and the prediction precision of this abnormality is the overall goal as the accuracy indicates the clinical value of the proposed model. Meanwhile, this study utilizes the area under ROC (Receiver Operation Characteristic) curve as metric in tuning classifier parameters as AuROC considers the prediction performance for positive and negative class alike and is less likely to be affected by data imbalance.

### 4.2. Experimental Results & Analysis

#### 4.2.1. Baseline Classifiers

The performance of 7 baseline classifiers on two datasets are presented in Table 2. For comparison, the performance of the same classifiers is appended to the table when no balancing technique is used. These results are noted with “NON-SMOTE” label. 10-fold cross validations are applied to test the chosen classifiers to reduce the effect of random choice of training & testing set. Due to the randomness in splitting data, the optimal parameters of each classifier from Bayesian Optimization are not recorded.

From Table 2, it's noticed that RF, GBDT and XGB have good classification performance on both datasets, with and without SMOTE algorithm applied.

Rather than only using these 3 tree-based classifiers, this study still adds non-tree-based classifiers like MLP, LR and SVM to the proposed ensemble model so that the model has better flexibility and may performs better robustness on different situations.



Table 2. Baseline Performance.

Dataset	Classifier	SMOTE			NON-SMOTE		
		Accuracy	Specificity	AuROC	Accuracy	Specificity	AuROC
Original	EXT	0.9775	0.9685	0.9775	0.9649	0.9498	0.9650
	RF	0.9797	0.9707	0.9797	0.9723	0.9665	0.9710
	GBDT	0.9764	0.9730	0.9764	0.9619	0.9414	0.9572
	XGB	0.9741	0.9707	0.9741	0.9634	0.9498	0.9604
	MLP	0.9741	0.9685	0.9741	0.9635	0.9540	0.9635
	LR	0.9718	0.9685	0.9718	0.9619	0.9331	0.9555
	SVM	0.9611	0.9302	0.9639	0.9577	0.9575	0.9656
Diagnostic	EXT	0.9679	0.9552	0.9679	0.9510	0.9292	0.9614
	RF	0.9623	0.9603	0.9623	0.9243	0.9292	0.9235
	GBDT	0.9749	0.9692	0.9749	0.9269	0.9151	0.9302
	XGB	0.9706	0.9664	0.9706	0.9578	0.9434	0.9616
	MLP	0.8004	0.7451	0.8004	0.7033	0.4104	0.7770
	LR	0.9638	0.9608	0.9638	0.9495	0.9245	0.9562
	SVM	0.4067	0.6611	0.4067	0.6680	0.3129	0.6786

#### 4.2.2. The Analysis on the Proposed Mode

The performance of the proposed ensemble model is then tested by performing prediction on different randomly split training/testing dataset 10 times. The testing set labels and the prediction result labels are recorded and concatenated to give

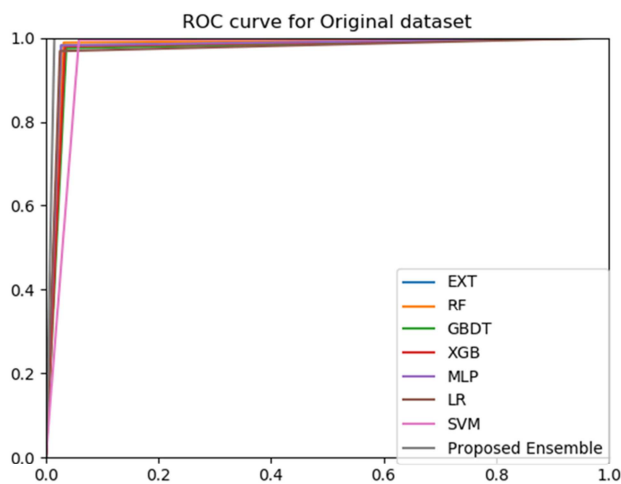
a final performance measurement in terms of accuracy and specificity. In the same manner, this procedure is performed without balancing the data for comparison. The results are included in Table 3:

Table 3. Ensemble Performance.

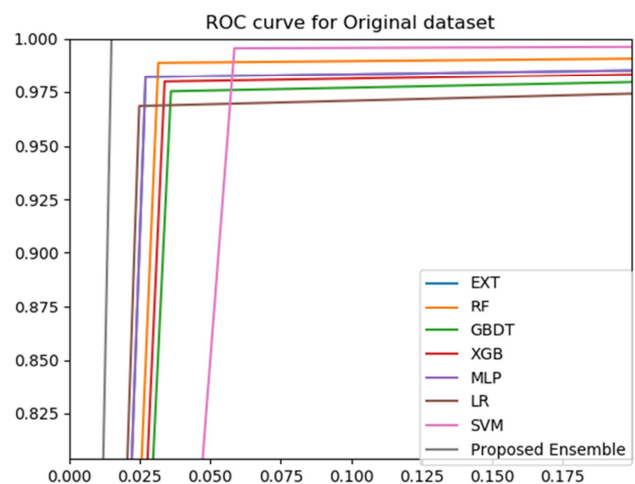
Dataset	Classifier	SMOTE			NON-SMOTE		
		Accuracy	Specificity	AuROC	Accuracy	Specificity	AuROC
Original	Best for baseline	0.9797 (RF)	0.9730 (GBDT)	0.9797 (RF)	0.9723 (RF)	0.9665 (RF)	0.9710 (RF)
	Proposed Ensemble	0.9814	0.9750	0.9800	0.9767	0.9688	0.9810
Diagnostic	Best for baseline	0.9749 (GBDT)	0.9692 (GBDT)	0.9749 (GBDT)	0.9578 (XGB)	0.9434 (XGB)	0.9616 (XGB)
	Proposed Ensemble	0.9745	0.9833	0.9760	0.9709	0.9444	0.9712

Then the ROC curves for baseline classifiers and the proposed model are compared. The ROC curve takes False Positive Rate (indicating the percentage of wrong classification on positive class) as x-axis and True Positive rate (indicating the percentage of correct classification on positive class) as y-axis. The area under the ROC curve (AuROC) indicates the probability that the classifier will give a higher prediction on a true positive instance than a true negative instance. As the performance of every classifier is

good, the ROC curve for each classifier is difficult to identify in the figures. Each of the two plots are zoomed in for better visual effect. Figure 5(a) shows the whole ROC curves for the baseline classifiers together with the proposed ensemble model on WBCD Original dataset, and Figure 6(a) shows that of the Diagnostic dataset. Figure 5(b) and Figure 6(b) are the zoomed parts of the upper-left corner regions of Figure 5(a) and Figure 6(a) respectively:



(a). Whole ROC curves for Original dataset



(b). Zoomed ROC curves for Original dataset

Figure 5. ROC curves for Original dataset.

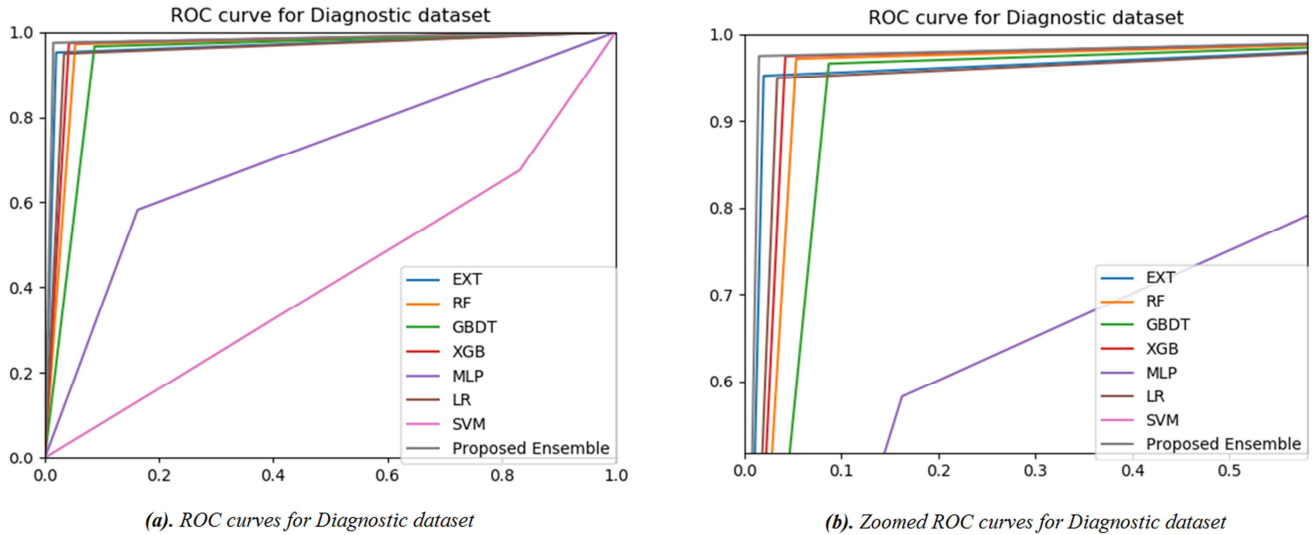


Figure 6. ROC curves for Diagnostic dataset.

The tables and plots indicate that the proposed ensemble model improves the performance metrics comparing to work solely on single classifiers in terms of accuracy, specificity and AuROC, indicating that the ensemble model outperforms other baseline classifiers. Also, from Figure 6, it's noticed that even if MLP and SVM achieve rather poor classification performance, the overall ensemble model accuracy is still improved comparing to the best baseline classifier. This may be due to the improvement of model adaptivity through including non-tree-based classifiers.

From the results above, one other conclusion may also be drawn that a balancing method can make a difference when dealing with classification tasks on imbalanced dataset. The improvement of metric statistics is relatively small but at least, demonstrates that the SMOTE algorithm is useful to some

extent. As the negative class (Malignant) is the minority class, without balancing practice the classifiers are more likely to give prediction as the majority class. When SMOTE algorithm is applied, the specificity, which indicates the prediction accuracy of the negative class, increases significantly. This indicates that the prediction accuracy will increase when the idea of this ensemble model is adopted in real clinical cases.

#### 4.2.3. Comparative Analysis with Related Works

For demonstration purpose, several representative works by other scholars about breast cancer are chosen and compared with the proposed model. Most researches only involve one dataset, and only the corresponding performance metrics reported in those works are compared, and the results are presented in Table 4:

Table 4. Performance Compare with Related Works.

Model Information	Original		Diagnostic	
	Accuracy	Specificity	Accuracy	Specificity
SVM (Asri et al., 2016)	0.9700	0.9700	-	-
fuzzy-genetic (Peña-Reyes and Sipper, 1999)	0.9780	-	-	-
AR+NN (Karabatak and Ince, 2009)	0.9740	-	-	-
K-SVM (Zheng et al., 2014)	-	-	0.9738	-
SVM+feature engineering (Akay, 2009)	0.9955	0.9664	-	-
SVM+KNN+feature engineering (Osareh and Shadgar, 2010)	0.9880	-	0.9633	-
RBFN+GRNN+FFNN+solid voting (ensemble) (Yavuz et al., 2017)	-	-	0.9643	0.9589
NF+k-NN+QC+majority voting (ensemble) (Hsieh et al., 2011)	0.9714	-	-	-
Proposed ensemble model	0.9814	0.9750	0.9745	0.9833

For Original dataset, Akay's SVM method with feature engineering achieves the highest classification accuracy 99.55%, and Osareh's work also achieves higher accuracy than the proposed model, while the proposed model outperforms other models for Specificity of Original dataset and both metrics for Diagnostic dataset. Feature engineering may help achieving a higher accuracy for Original dataset.

## 5. Conclusion

Breast cancer, as the second most diagnosed cancer among

women in America (U.S. Breast Cancer Statistics), risks every mid-aged female. The performance of classifier to classify breast cancer is of great significance, especially for the detection of Malignant cases. This study proposed an ensemble model with SMOTE algorithm to classify instances to be Benign or Malignant on two public datasets and achieved an overall accuracy of 98.14% on WBCD Original dataset and 97.45% on Diagnostic dataset. By comparing with baseline classifiers, the validity of stacking ensemble model is validated in terms of accuracy, specificity and AuROC. Furthermore, the result of the overall algorithm with and without SMOTE

balancing algorithm are compared to demonstrate the necessity of balancing algorithm for heavily imbalanced data, confirming the impact of imbalanced learning method on this problem. The validity and clinical value of the ensemble model proposed in this study is therefore confirmed.

## 6. Discussion

The main idea of imbalanced learning and ensemble method in this study can be applied to similar situations, like predicting or classifying diabetes type, cervical cancer survival rate and even other fields like credit scoring or spam detection, where datasets are more likely to be imbalanced and the minority class indicates abnormality. Furthermore, by deploying stacking ensemble method with Bayesian Optimization after SMOTE algorithm, it's actually allowing modularization of the entire model. After necessary data preprocessing, datasets with imbalance and binary classification goal may directly use the program of this study.

Meanwhile, there are still several drawbacks of the proposed model. First, as the two datasets are relatively small in terms of numbers of instances and features. Clinical and medical data are more likely to be less dedicated for classification, containing more missing values and outliers, together with a more information that may potentially influence the classification performance. When dealing with high-dimensional datasets, feature selection techniques like Principle Component Analysis and feature importance should be considered. Second, the random choices of initializing range in Bayesian Optimization gives this method a possibility for a false optimal solution to be generated, and a small number of experiments face abnormally low performance. Such issues prevent the proposed model from being directly applied to clinical use. Also, the choice of imbalance learning method, the choice of type and number of baseline classifiers may further influence classification performance, as well as the task's time efficiency. Future works may include procedures to check if the baseline classifiers are truly optimal and try feature engineering if needed. With higher dimensionality and more instances contained, deep learning method may also help to achieve better classification performance.

## Acknowledgements

This study is supported by the National Natural Science Foundation of China (No. 51375429).

## References

- [1] Akay, M. F., "Support vector machines combined with feature selection for breast cancer diagnosis." *Expert Systems with Applications*, vol. 36, no. 2, 2009, pp. 3240-3247.
- [2] Asri, H., Mousannif, H., Moatassime, H. A., and Noel, T., "Using machine learning algorithms for breast cancer risk prediction and diagnosis." *Procedia Computer Science*, vol. 83, 2016, pp. 1064-1069.
- [3] Breiman, L., "Bagging predictors." *Machine Learning*, vol. 24, no. 2, 1996, pp. 123-140.
- [4] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., "SMOTE: synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research*, vol. 16, no. 2002, pp. 321-357.
- [5] Coulter, D. M., Bate, A., Meyboom, R. H., Lindquist, M., and Edwards, I. R., "Antipsychotic drugs and heart muscle disorder in international pharmacovigilance: data mining study." *BMJ*, vol. 322, no. 7296, 2001, pp. 1207-1209.
- [6] Emamjomeh, A., Goliaei, B., Zahiri, J., and Ebrahimpour, R., "Predicting protein-protein interactions between human and Hepatitis C virus via an ensemble learning method." *Molecular Biosystems*, vol. 10, no. 12, 2014, pp. 3147-3154.
- [7] Eom, J., Kim, S., and Zhang, B., "AptaCDSS-E: a classifier ensemble-based clinical decision support system for cardiovascular disease level prediction." *Expert Systems with Applications*, vol. 34, no. 4, 2008, pp. 2465-2479.
- [8] Han, B., and Cook, P., "A stacking-based approach to twitter user geolocation prediction." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Sonifa, Bulgaria, August 4-9, 2013, pp. 7-12.
- [9] He, H. L., Zhang, W. Y., and Zhang, S., "A novel ensemble method for credit scoring: adaption of different imbalance ratios." *Expert Systems with Applications*, vol. 98, 2018, pp. 105-117.
- [10] Hsieh, S. L., Hsieh, S. H., Cheng, P. H., Chen, C. H., Hsu, K. P., Lee, I. S., Wang, Z., and Lai, F., "Design ensemble machine learning model for breast cancer diagnosis." *Journal of Medical Systems*, vol. 36, no. 5, 2011, pp. 2841-2847.
- [11] Johnston, M. E., Langton, K. B., Haynes, R. B., and Mathieu, A., "Effects of computer-based clinical decision support systems on clinician performance and patient outcome: a critical appraisal of research." *Annals of Internal Medicine*, vol. 120, no. 2, 1994, pp. 135-142.
- [12] Karabatak, M., and Ince, M. C., "An expert system for detection of breast cancer based on association rules and neural network." *Expert Systems with Applications*, vol. 36, no. 2, 2009, pp. 3465-3469.
- [13] Niemeijer, M., Ginneken, B. V., Russell, S. R., Suttrop-Schulten, M. S., and Abramoff, M. D., "Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis." *Investigative Ophthalmology & Visual Science*, vol. 48, no. 5, Jan. 2007, pp. 2260-2267.
- [14] Osareh, A., and Shadgar, B., "Machine learning techniques to diagnose breast cancer." In *Proceedings of the 5th International Symposium on Health Informatics and Bioinformatics*, Antalya, Turkey, April 20-22, 2010, pp. 114-120.
- [15] Peña-Reyes, C. A., and Sipper, M., "A fuzzy-genetic approach to breast cancer diagnosis." *Artificial Intelligence in Medicine*, vol. 17, no. 2, 1999, pp. 131-155.
- [16] Perez-Iratxeta, C., Bork, P., and Andrade, M. A., "Association of genes to genetically inherited diseases using data mining." *Nature Genetics*, vol. 31, no. 3, 2002, pp. 316-319.



- [17] Prather, J. C., Lobach, D. F., Goodwin, L. K., Hales, J. W., Hage, M. L., and Hammond, W. E., "Medical data mining: knowledge discovery in a clinical data warehouse." In Proceedings of the 1997 American Medical Informatics Association Annual Fall Symposium, Nashville, USA, Oct. 25-29, 1997, pp. 101-105.
- [18] Sarwar, A., Sharma, V., and Gupta, R., "Hybrid ensemble learning technique for screening of cervical cancer using papanicolaou smear image analysis." *Personalized Medicine Universe*, vol. 4, 2015, pp. 54-62.
- [19] Schapire, R. E., "The strength of weak learnability." *Machine Learning*, vol. 5, no. 2, 1990, pp. 197-227.
- [20] Snoek, J., Larochelle, H., and Adams, R. P., "Practical Bayesian optimization of machine learning algorithms." *Neural Information Processing Systems*, 2012, pp. 2951-2959.
- [21] Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., and Clore, J. N., "Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records." *Biomed Research International*, 2014, pp. 1-11.
- [22] "U.S. Breast Cancer Statistics." *Breastcancer.org*, Jan. 9, 2018, [www.breastcancer.org/symptoms/understand\\_bc/statistics](http://www.breastcancer.org/symptoms/understand_bc/statistics).
- [23] Wang, Y., Rimm, E. B., Stampfer, M. J., Willett, W. C., and Hu, F. B., "Comparison of abdominal adiposity and overall obesity in predicting risk of Type 2 diabetes among men." *The American Journal of Clinical Nutrition*, vol. 81, no. 3, 2005, pp. 555-563.
- [24] Wilson, A. M., Thabane, L., and Holbrook, A., "Application of data mining techniques in pharmacovigilance." *British Journal of Clinical Pharmacology*, vol. 57, no. 2, 2003, pp. 127-134.
- [25] Wolberg, W. H., and Mangasarian, O. L., "Multisurface method of pattern separation for medical diagnosis applied to breast cytology." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 23, 1990, pp. 9193-9196.
- [26] Wolpert, D. H., "Stacked generalization." *Neural Networks*, vol. 5, no. 2, 1992, pp. 241-259.
- [27] Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M., and Chang, I. C., "Deep learning of feature representation with multiple instance learning for medical image analysis." In Proceedings of 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing, Florence, Italy, May 4-9, 2014, pp. 1626-1630.
- [28] Yavuz, E., Eyupoglu, C., and Sanver, U., "An ensemble of neural networks for breast cancer diagnosis." In Proceedings of International Conference on Computer Science and Engineering, Antalya, Turkey, Oct 5-8, 2017, pp. 538-543.
- [29] Zheng, B., Yoon, S. W., and Lam, S. S., "Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms." *Expert Systems with Applications*, vol. 41, no. 4, 2014, pp. 1476-1482.