Classification of Breast Cancer Risk Factors Using Several Resampling Approaches

Md Faisal Kabir
Department of Computer Science
North Dakota State University
Fargo, USA
mdfaisal.kabir@ndsu.edu

Simone A. Ludwig
Department of Computer Science
North Dakota State University
Fargo, USA
simone.ludwig@ndsu.edu

***Abstract*—Breast cancer is the most common cancer in women worldwide and the second most common cancer overall. Predicting the risk of breast cancer occurrence is an important challenge for clinical oncologists as it has direct influence in daily practice and clinical service. Classification is one of the supervised learning models that is applied in medical domains. Achieving better performance on real data that contains imbalance characteristics is a very challenging task. Machine learning researchers have been using various techniques to obtain higher accuracy, generally by correctly identifying majority class samples while ignoring the instances of the minority class. However, in most of the cases the concept of the minority class instances usually is of higher interest than the majority class. In this research, we applied three different classification techniques on a real world breast cancer risk factors data set. First, we applied specified classification techniques on breast cancer data without applying any resampling technique. Second, since the data is imbalanced meaning data has an unequal distribution between the classes, we applied several resampling methods to get better performance before applying the classifiers. The experimental results show significant improvement on using a resampling method as compared to applying no resampling technique, particularly for the minority class.**

***Index Terms*—classification; class imbalance; breast cancer; risk factors.**

I. Introduction and Related Work

Cancer has become one of the most devastating diseases worldwide, with more than 10 million new cases every year, according to WHO (World Health Organization) [1]. The causes and types of cancer vary in different geographical regions, however, nearly every family in the world is touched by cancer. The disease burden is enormous, not only for affected individuals but also for their relatives and friends. Breast cancer is the most common cancer in women worldwide, with nearly 1.7 million new cases diagnosed in 2012 [1]. Breast cancer makes up 25 percent of all new cancer diagnoses in women globally, according to the American Cancer Society (ACS) [1]. Predicting the risk of breast cancer occurrence is an important challenge for clinical oncologists as this has direct influence in their daily practice and clinical service. A reliable prediction will help oncologists and other clinicians in their decision-making process and allow clinicians in choosing the most reliable and evidence-based treatment and prevention strategies for their patients.

Researchers have developed different models for breast cancer risk prediction, and association between risk factors [2]–[5]. In [2], the authors applied statistical methods to show a positive association between Hormone Replacement Therapy (HRT) and breast cancer risk, although this relationship varies according to race/ethnicity, BMI (Body Mass Index), and breast density. The Gali model is used to estimate the number of expected breast cancers for white females who are examined annually [3]. In [4], the authors used commonly identified risk factors such as race/ethnicity, breast density, BMI, and the use of hormone therapy, type of menopause, and previous mammographic results to improve the model using logistic regression. In [5], the Breast cancer risk score is determined using k-nearest-neighbor (KNN) to improve readability for physician and patients.

Machine Learning (ML) or Data Mining (DM) algorithms are applied in the medical domain in order to assist with the decision-making process, for example, for the prediction of cancer risk. ML and DM algorithms [1],[2],[6] can be classified into supervised or unsupervised learning depending on the goal of the data mining task. Classification is a supervised learning techniques and the goal of the classification model is to predict qualitative or categorical outputs which assume values in a finite set of classes (e.g. Yes/No or Benign-cancer/Malignant-cancer, etc.) without an explicit order [4]. The primary objective of traditional classifiers is to get higher accuracy by reducing the overall classification error [5]. However, the overall classification error is biased towards the majority class for imbalanced data problems.

The problem of class imbalance is common that affects ML or classification models due to having a disproportionate number of different class instances in practice [7]. There are many approaches that deal with this problem such as cost function based, and sampling based solutions. In this research, we focused on sampling based approaches that can be classified into three major categories - random under-sampling, random over-sampling, and hybrid of over-sampling and under-sampling.

Sampling methods modify the data set to balance the class distribution before using the data set to train the classifier. Random under-sampling is the process of removing some of instances of the majority class whereas over-sampling is the

process of adding more samples of the minority class so it has a larger effect on the ML algorithm. Although the methods are simple, however, both of these techniques have some shortcomings. The random under-sampling technique has the potential to lose information as it removes instances from the major class. On the other hand, over-sampling generates instances from the minority class that creates the potential risk of over-fitting. The hybrid method is a mix of the oversampling and under-sampling technique.

The Synthetic Minority Over-sampling Technique (SMOTE) generates synthetic minority instances to balance the class distribution [8] and has been widely used. SMOTE produces synthetic minority instances by linear interpolation between neighbors in the input space. ENN (Edited Nearest Neighbor) is a technique of under-sampling of the majority class. It removes points or instances whose class labels differ from the majority of its k nearest neighbors [8]. Tomek Link [9] is a method of under-sampling which is used as a method of guided under-sampling where the observations from the majority class are removed. The combinations of these techniques are also applied in the literature to achieve better performance.

In this research, we applied three different classification algorithms on breast cancer risk factors data, and calculated the predicted performance on a test set. Since the data is imbalanced, we also applied various resampling techniques on the training data and applied classifiers on the 'modified' training data. Performance comparisons on the test data based on the all classification models were also conducted.

The remainder of the paper is organized as follows. Section II describes the different resampling and classification techniques that were used in this research. Our proposed approach is also discussed in this section. Section III shows the experimental results; the proposed techniques were evaluated using breast cancer risk factors data and their results are presented. Section IV is the summary section of this paper; we conclude our paper and suggest possible future research directions.

II. Methodology

A. Classification Phase

We used three different classifiers namely Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) to train the breast cancer data set of imbalanced data (original data) as well as modified training data obtained by using different resampling methods. These trained models were used to predict the target class for the test data set. The three classifiers that are used in this research are briefly described below.

1) Decision Tree (DT): DT is a supervised learning approach that learns from class-labeled instances. It works very well with different types of data and results are easy to interpret. In addition, building a model using decision tree is comparatively easy, and data can be represented in a visualizing form. The decision tree model generation is however sensitive to overfitting and may get stuck in local minima. When the number of dimensions gets too high, the decision tree model generation may fail. The decision tree classifier has been widely applied to solving many real world problems including in areas of healthcare, medicine, business, education, and so on [10]–[12].

2) Random Forest (RF): RF is a powerful classification and regression tool that generates a forest of classification trees, rather than a single classification tree [13]. RF creates decision trees on randomly selected data samples, obtains the prediction from each tree and selects the best solution by means of voting. There are two stages in the RF algorithm, the first one is RF building, and the second stage is to make a prediction from the RF classifier created during the first stage. RF is considered as a highly accurate and robust method because of the number of decision trees participating in the process. In addition, if there are more trees in the forest, the RF classifier will avoid the over-fitting problem.

3) Extreme Gradient Boosting (XGBoost): XGBoost [14] is an implementation of gradient boosted decision trees designed for speed and performance. XGBoost provides a wrapper class to allow models to be treated like a classifier or a regressor in the scikit-learn framework. The XGBoost model for classification is called XGBClassifier. XGBoost is a scalable and accurate implementation of the gradient boosting machines and it has proven to push the limits of computing power for boosted trees algorithms as it was built and developed for the sole purpose of model performance and computational speed. Specifically, it was engineered to exploit every bit of memory and hardware resources available for the tree boosting algorithm.

XGBoost has been widely used in a number of machine learning and data mining challenges. For example, in Kaggle, which is a ML competetion site; among the 29 challenge winning solutions published on the Kaggle site during 2015, 17 solutions used XGBoost. The second most popular method was deep neural network and was used in 11 solutions [15].

B. Resampling Phase

The data set that we used in this research is imbalanced data, meaning there are significantly more samples for one category than the other. For that reason, different resampling techniques were applied to the training data set (imbalanced) and thus the training data is modified accordingly. The resampling techniques that were used in this work are briefly discussed below.

1) Random under-sampling (RUS) of majority class: is a form of data sampling that randomly picks majority class instances and removes them from the dataset until the desired class distribution is achieved [16]. This means that for a dataset containing 100 positive and 500 negative instances, RUS removes 400 negative instances in order to achieve a 50:50 post-sampling positive:negative class ratio.

2) Random over-sampling (ROS) of minority class: is a form of data sampling that randomly picks minority class instances with replacement until the desired class distribution is achieved [16]. This means that for a dataset containing 100 positive and 500 negative instances, ROS adds 400

positive instances in order to achieve a 50:50 post-sampling positive:negative class ratio.

3) SMOTE: works by creating synthetic observations based upon the existing minority instances [8],[17]. For each minority instance, SMOTE calculates the k nearest neighbors. Depending upon the amount of oversampling needed, one or more of the k-nearest neighbors are selected to create synthetic examples.

4) Edited Nearest Neighbor (ENN): is the technique of under-sampling of the majority class [8]. It removes points or instances whose class label differs from a majority of its k-nearest neighbors.

5) SMOTE + ENN: combines the over-sampling and under-sampling techniques [8]. It performs over-sampling using SMOTE and under-sampling or cleaning using ENN. Thus, instead of removing only the majority class examples, instances from both classes are removed. ENN tends to remove more instances than Tomek links do, so it is expected that it will provide more in-depth data cleaning.

6) SMOTE + Tomek Link: also combines over-sampling and under-sampling techniques. It performs over-sampling using SMOTE and under-sampling or cleaning using Tomek links [8],[9]. Thus, instead of removing only the majority class examples, instances from both classes are removed. Tomek links remove less instances compared to ENN.

C. Proposed Approach

To obtain a better classification performance, we used specified classifiers to train the model using the original training data. We also used various types of resampling methods on the training data to train the model using specified classifiers with the modified training data. We then used all the trained models to obtain class information on the test data. The diagram of our proposed approach is shown in Fig. 1 consisting of the following main steps:

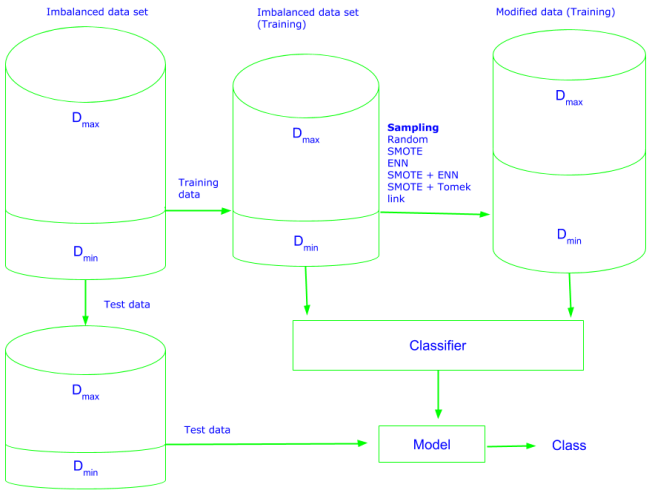

Fig. 1. Proposed model to handle imbalanced data.

1) Step 1: This step includes obtaining the classification model data and test data for classification; we constructed the classification model data, or training data, and sample, or test data, for classification. The training set contains 80% of the data while the test set contains the remaining 20%. The Stratified shuffle split technique available in scikit-learn (sklearn), a machine learning library for the Python programming language, was used since it preserves the percentage of samples for each class which is important for imbalance data. $Dmax$ is the number of instances belonging to the negative class, or majority class, while $Dmin$ is the number of instances of the positive, or minority class.

2) Step 2: This step resamples the training data. Several resampling techniques were used on training data that changed the number of instances of the training data. Based on the techniques of the resampling methods, the instances of the majority class were removed and/or instances of the minority class were added. The test data was kept unchanged.

3) Step 3: In this step, the classification model data was trained with the specified classifiers. First, we used the original training data without using any sampling methods, and built models using the specified classifiers. Second, for the training we used the modified training data obtained by applying the different sampling techniques. Each of these training data sets were used to train all three classifiers. All of the above models were saved for the prediction on the test data.

4) Step 4: The last step was to apply test data on the saved models obtained in Step 3 to generate predictions on the test data.

III. Experiments and Results

Detailed data description and pre-processing is discussed in this section. This section also presents the experimental results and performance evaluation of the different models.

A. Data Description and Pre-processing

The dataset includes information from 6,318,638 mammography examinations obtained from the Breast Cancer Surveillance Consortium (BCSC) database collected from January 2000 to December 2009 [18]. Data for this study was obtained from the BCSC Data Resource and more information is available at http://www.bcsc-research.org.

The data is aggregated such that the total number of instances or records is 1,144,565, with 13 attributes or columns. The dataset also contains missing or unknown values denoted by 9. To build a reliable model, we discarded the records containing at least one missing or unknown value. We also removed the attribute year that represents the calendar year of the observation. After discarding these records and one attribute, there are 219,524 available records with 12 attributes. In the dataset, there is an attribute named count, representing the number of records that have the combination of variable-values shown in the row. For instance, the value of the count column for the particular row is 12. It indicates that there were 12 similar records, the same as that particular row in the original data. For that reason, we created the number of rows

or records the same as the count value in the original dataset, and discarded the count column after that. Finally, there are a total of 1,015,583 records with 11 attributes for building the model. Among 1,015,583 records, 60,800 individuals have prior breast cancer, and 954,783 are non-breast cancer individuals. Among the 11 attributes, "prior breast cancer" values yes or no is considered as the response or dependent variable and the remaining 10 attributes are considered as explanatory or predictors or independent variables.

The summary of the BCSC data along with train/test split are shown in Table I.

TABLE I
SUMMARY OF BCSC DATA WITH TRAIN/TEST SPILT.

Types	Class = yes	Class = no	Total
BCSC data	60,800	954,783	1,015,583
Training (80 %)	48,640	763,826	812,466
Test (20%)	12,160	190,957	203,117

We used different resampling methods on the training data. The distribution of the training data after applying different resampling techniques is shown in Table II.

TABLE II
DISTRIBUTION OF MODIFIED TRAINING DATA AFTER APPLYING DIFFERENT RESAMPLING METHODS.

Resampling type	Class = yes	Class = no	Total
Random under-sample	48,640	48,640	97,280
Random over-sample	763,826	763,826	1,527,652
SMOTE	763,826	763,826	1,527,652
ENN	48,640	685,963	734,603
SMOTE + ENN	437,256	658,167	1,095,423
SMOTE + Tomek link	763,825	763,825	1,527,650

B. Evaluation Measures

To measure the performance of our model, several evaluation measures were used such as accuracy, recall, precision, area under the Receiver Operating Characteristic curve (ROC) or AUC, and F-measure [19]. These were derived from the confusion matrix, and applied to the classifier evaluation.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \quad (1)$$

$$recall = TP/(TP + FN) \quad (2)$$

$$precision = TP/(TP + FP) \quad (3)$$

where Y is the binary response or class variable; α is the intercept to be calculated; β_i is the estimated vector of parameters, and X_i is the vector of independent variables.

Here, TP denotes the number of positive examples correctly classified, TN denotes the number of negative samples correctly classified, FN represents the number of positive observations incorrectly classified, and FP indicates the number of negative samples incorrectly classified by the estimator.

The ROC curve is a representation of the best decision boundaries for the cost between the True Positive Rate (TPR), and the False Positive Rate (FPR). The ROC curve plots TPR against FPR. TPR, and FPR are defined as follows.

$$TPR = TP/(TP + FN) \quad (4)$$

$$FPR = FP/(FP + TN) \quad (5)$$

The area below the ROC curve is called AUC and is widely utilized for weighing classifier performance. The value of AUC ranges from 0.0 to 1.0, where a value of AUC equals 1.0 means perfect prediction, a value of 0.5 means random prediction, and a value less than 0.5 is considered as a poor prediction.

If only the performance of the positive class in this case the minority class is considered, two measures namely recall, and precision are important. Recall or true positive rate denoting the percentage of retrieved objects that are relevant, while precision or positive predictive value denoting the percentage of relevant objects that are identified for retrieval. The F-measure or F1 score is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test, which is defined as follows:

$$F - measure = 2 * \frac{precision * recall}{precision + recall} \quad (6)$$

It is to be noted that for balanced class F1 score can effectively be ignored, the accuracy is key. For the imbalance class, if the class distribution is highly skewed, then the classifier can have a higher accuracy simply by choosing the majority class. In such a situation, the classifier that gets a high F1 score on both classes, as well as high accuracy should be selected. However, if a particular class generally the minority class is more important than the other then it is more important to correctly classify instances for the minority or important class as opposed to the majority class. In this case, the classifier that has a good F1 score only on the important class should be considered.

C. Results

In this paper, we applied three different classifier models on the original training data (imbalanced), and the modified training data sets. We compared the performance of the trained models on the test data. The overall performance of the classification models (built based on the original training data) on test data are shown in Table III whereas the performance of the minority class is shown in Table IV, respectively.

TABLE III
OVERALL PERFORMANCE OF SPECIFIED CLASSIFIERS ON TEST DATA (TRAINED WITH THE ORIGINAL TRAINING DATA).

Methods	Precision	Recall	F1-score	Accuracy	AUC
DT	0.92	0.94	0.92	0.9406	0.9272
RF	0.92	0.94	0.92	0.9399	0.9164
XGBoost	0.92	0.94	0.91	0.9404	0.9287

Although, the performance of these classifiers seem very good (according to Table III) when no resampling techniques were used, however, the performance of classifying the instances of the minority class was very low. For the minority

TABLE IV
PERFORMANCE OF MINORITY CLASS ON TEST DATA.

Methods	Precision	Recall	F1-score
DT without sampling	0.54	0.05	0.10
RF without sampling	0.49	0.11	0.18
XGBoost without sampling	0.54	0.03	0.05

class, maximum recall, and F1-score were reported as 0.11 and 0.18, respectively for the RF classifier.

Different resampling methods on the training data were used to modify the training data accordingly. The modified training data sets were used for the training of the specified classifiers. Results were obtained from the models applied to the test data. Table V shows the overall performance of the DT classification models (built based on the modified training data) on test data for all the different training data sets, whereas Table VI shows the performance of the minority class for the DT classifier, respectively.

TABLE V
OVERALL PERFORMANCE OF DT CLASSIFIER (MODEL BUILT ON MODIFIED TRAINING DATA) ON TEST DATA.

Methods	Precision	Recall	F1-score	Accuracy	AUC
DT with RUS	0.95	0.82	0.86	0.8171	0.9255
DT with ROS	0.95	0.82	0.86	0.8157	0.9263
DT with SMOTE	0.95	0.82	0.87	0.8244	0.9266
DT with ENN	0.93	0.91	0.92	0.9069	0.9249
DT with SMOTE + ENN	0.94	0.87	0.90	0.8722	0.9207
DT with SMOTE + Tomek link	0.95	0.82	0.86	0.8208	0.9270

TABLE VI
PERFORMANCE OF MINORITY CLASS ON TEST DATA BASED ON DT CLASSIFIER.

Methods	Precision	Recall	F1-score
DT with RUS	0.24	0.96	0.39
DT with ROS	0.24	0.97	0.39
DT with SMOTE	0.25	0.95	0.39
DT with ENN	0.33	0.56	0.42
DT with SMOTE + ENN	0.29	0.80	0.43
DT with SMOTE + Tomek link	0.24	0.96	0.39

For DT, the best accuracy obtained was 90.69% when sampling method ENN was applied, but the AUC value was little (0.0021) less than the highest AUC value of 0.9270. For the minority class, The best recall (0.80) and the best F1-score (0.43) values were obtained when the resampling technique SMOTE and ENN were applied.

Table VII shows the overall performance of the RF classification models (built based on the modified training data) on test data for all the different training data sets whereas Table VIII shows the performance of the minority class, respectively.

TABLE VII
OVERALL PERFORMANCE OF RF CLASSIFIER ON TEST DATA.

Methods	Precision	Recall	F1-score	Accuracy	AUC
RF with RUS	0.95	0.82	0.87	0.8219	0.9180
RF with ROS	0.95	0.84	0.87	0.8356	0.9145
RF with SMOTE	0.95	0.85	0.89	0.8540	0.9140
RF with ENN	0.93	0.88	0.90	0.8820	0.9039
RF with SMOTE + ENN	0.94	0.88	0.91	0.8855	0.8606
RF with SMOTE + Tomek link	0.95	0.85	0.89	0.8532	0.9135

TABLE VIII
PERFORMANCE OF MINORITY CLASS ON TEST DATA BASED ON RF CLASSIFIER.

Methods	Precision	Recall	F1-score
RF with RUS	0.24	0.94	0.39
RF with ROS	0.26	0.91	0.40
RF with SMOTE	0.27	0.85	0.41
RF with ENN	0.28	0.63	0.39
RF with SMOTE + ENN	0.31	0.74	0.44
RF with SMOTE + Tomek link	0.27	0.85	0.41

For RF, the best accuracy obtained was 88.55% when sampling method SMOTE followed by ENN was applied. But in case of SMOTE followed by ENN, the AUC value (0.8606) was the lowest among all other sampling methods. The maximum AUC (0.9180) for RF was reported when RUS used. For the minority class, the best recall (0.94) was found when RUS was applied and the highest F1-score (0.44) was obtained when resampling technique SMOTE and ENN were applied.

Table IX shows the overall performance of XGBoost classification models (built based on the modified training data) on test data for all the different training data sets whereas Table X shows the performance of the minority class, respectively.

TABLE IX
OVERALL PERFORMANCE OF XGBOOST CLASSIFIER ON TEST DATA.

Methods	Precision	Recall	F1-score	Accuracy	AUC
XGBoost with RUS	0.95	0.81	0.86	0.8118	0.9287
XGBoost with ROS	0.95	0.81	0.86	0.8128	0.9288
XGBoost with SMOTE	0.95	0.82	0.87	0.8218	0.9284
XGBoost with ENN	0.93	0.91	0.92	0.9149	0.9281
XGBoost with SMOTE + ENN	0.95	0.86	0.89	0.8626	0.9270
XGBoost with SMOTE + Tomek link	0.95	0.82	0.86	0.8210	0.9282

For XGBoost, the best accuracy obtained was 91.49% when the sampling method ENN was used. Surprisingly, the

TABLE X
PERFORMANCE OF MINORITY CLASS ON TEST DATA BASED ON XGBOOST CLASSIFIER.

Methods	Precision	Recall	F1-score
XGBoost with RUS	0.24	0.97	0.38
XGBoost with ROS	0.24	0.97	0.38
XGBoost with SMOTE	0.25	0.96	0.39
XGBoost with ENN	0.35	0.52	0.42
XGBoost with SMOTE + ENN	0.29	0.87	0.43
XGBoost with SMOTE + Tomek	0.25	0.96	0.39

AUC value (close to 0.93) remained almost same for all the resampling techniques. For the minority class, the best recall (0.97) was found when both RUS and ROS were applied, and the highest F1-score (0.43) was obtained when the resampling technique SMOTE and ENN were applied.

D. Performance Comparison

Although we obtained the best overall performance for all the classifiers when no resampling methods were used for the training phase, however, for minority class performance was very low. The accuracy for all three classifiers were about 94% when no resampling methods were applied which is about 3% more than the best accuracy obtained when the resampling techniques were used.

However, for the minority class, the performance was not better when no resampling methods were used. For instance, the best recall and F1 score for the minority class for RF were reported as 0.11 and 0.18, respectively when no resampling was used on the training data. Yet, the best recall and F1 score for the minority class were reported as 0.87 and 0.43, respectively for the XGBoost classifier when the resampling method SMOTE and ENN was used. It is also worth to mention that the overall performance for the same combination was also good (not best). For example, the accuracy and AUC score for this combination were reported as 86.26% and 92.70%, respectively. The performance for the minority class was far better when applying all the specified resampling methods as compared to not applying any resampling method. Thus, it is important to consider all the factors when dealing with imbalanced data such as if both classes are important or only the minority class is significant. Therefore, the appropriate model should be selected based on the objective.

IV. CONCLUSION AND FUTURE WORK

Predicting the risk of breast cancer occurrence is an important challenge for clinical oncologists as this has direct influence in their daily practice as well as their clinical service. In this research, we explored breast cancer risk factors data and applied different resampling techniques before applying machine learning methods. The data that we used in this research was severely imbalanced (60,800 versus 954,7834). Our main objective was to improve the classification performance of the standard machine learning algorithms towards the prediction of the important or minority class. We compared the impact of using several resampling techniques on the training data before using the specified classifiers in terms of the overall performance and the performance of the minority class. Experimental results show that the performance improves particularly for the minority class when the resampling techniques were used as compared to applying the classification techniques without using any resampling techniques.

We intend to extend this research by considering more risk factors not only for breast cancer but also for other cancer types. Furthermore, we plan to build more accurate predictive models that could provide better performance for both the minority and the majority class.

REFERENCES

[1] J. Ferlay, et al. "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012." International journal of cancer 136.5:E359-E386, 2015.

[2] N. Hou, et al. "Hormone replacement therapy and breast cancer: heterogeneous risks by race, weight, and breast density." Journal of the National Cancer Institute 105.18:1365-1372, 2013.

[3] M. H. Gail et al. "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually." JNCI: Journal of the National Cancer Institute 81.24: 1879-1886, 1989.

[4] W. E. Barlow, et al. "Prospective breast cancer risk prediction model for women undergoing screening mammography." Journal of the National Cancer Institute 98.17: 1204-1214, 2006.

[5] E. Gauthier et al. "Breast cancer risk score: a data mining approach to improve readability." The International Conference on Data Mining. CSREA Press, 2011.

[6] J. Han, M. Kamber. "Data mining concept and technology." Publishing House of Mechanism Industry: 70-72, 2001.

[7] J. Mathew, et al. "Classification of Imbalanced Data by Oversampling in Kernel Space of Support Vector Machines." IEEE transactions on neural networks and learning systems, 2017.

[8] A. More, "Survey of resampling techniques for improving classification performance in unbalanced datasets." arXiv preprint arXiv:1608.06048 (2016).

[9] T. Elhassan, et al. "Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method." Journal of Informatics and Data Mining 1 (2016): 1-12.

[10] J. R. Quinlan, "Constructing decision tree." C4 5, 17-26, 1993.

[11] Md F. Kabir, et al. "Information theoretic SOP expression minimization technique." Computer and information technology, 2007. iccit 2007. 10th international conference on. IEEE, 2007.

[12] Y. Zheng, et al. "R-C4. 5 Decision tree model and its applications to health care dataset." Services Systems and Services Management, 005. Proceedings of ICSSSM'05. 2005 International Conference on. Vol. 2. IEEE, 2005.

[13] J. Vanerio, and P. Casas, "Ensemble-learning approaches for network security and anomaly detection." Proceedings of theWorkshop on Big Data Analytics and Machine Learning for Data Communication Networks. ACM, 2017.

[14] Daz-Uriarte, Ramn, and Sara Alvarez De Andres. "Gene selection and classification of microarray data using random forest." BMC bioinformatics 7.1 (2006): 3.

[15] T. Chen, and C. Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016.

[16] G. Batista, R. C. Prati, and M. C. Monard. "A study of the behavior of several methods for balancing machine learning training data." ACM SIGKDD explorations newsletter 6.1 (2004): 20-29.

[17] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research, 16: 321-357, 2002.

[18] Data collection and sharing was supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (HHSN261201100031C), A list of the BCSC investigators and procedures for requesting BCSC data for research purposes, last retrieved July 2018 from http://www.bcsc-research.org.

[19] T. Fawcett, "An introduction to ROC analysis." Pattern recognition letters 27.8 (2006): 861-874.