2020 10th IEEE International Conference on Control System, Computing and Engineering (ICCSCE2020), 21–22 August 2020, Penang, Malaysia

Cost-Sensitive Extreme Gradient Boosting for Imbalanced Classification of Breast Cancer Diagnosis

Manop Phankokkruad
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand 10520
Email: manop@it.kmitl.ac.th

***Abstract*—The clinical information can enhance the doctors for predicting and diagnosing the diseases also making the right decisions. Breast cancer is the most dangerous disease, early diagnosis can improve a chance of survival and can support clinical treatment. Detecting breast cancer takes a lot of time and it is hard to classification. However, the problem of the classification occurs when there is an unequal distribution of classes the dataset. This is caused by the low performance in the traditional machine learning models. For this reason, this work proposed the cost-sensitive XGBoost model, which is an improved version of the XGBoost model in conjunction with cost-sensitive learning. The models were applied to classify the four breast cancer datasets that contained the imbalanced data. In the experiment, this work determined the best parameters on each dataset by the hyperparameters optimization technique before configuring the models. The results indicated that the cost-sensitive XGBoost model had been skillful, and could improve classification accuracy in four datasets. In addition, this work evaluated the model performance by accuracy, ROC AUC, and k-Fold cross-validation to ensure that the new models is accurate.**

***Index Terms*—Extreme Gradient Boosting, Classification, Cost-Sensitive, Imbalanced Classification, Breast Cancer, Deep Learning, Boosting Tree, Medical Information**

I. Introduction

The benefits of clinical information and data analysis of healthcare will enhance the medical staff and doctors for predicting diseases and diagnosing them at their early stages. Breast cancer is the most dangerous disease and a frequent cause of death in women. The early diagnosis of breast cancer can improve a chance of survival and can support clinical treatment. Detecting breast cancer is a time-consuming process and hard to diagnose. For this reason, automatic diagnosis techniques are developed by using many algorithms. Thus, the correct diagnosis of breast cancer and the classification of patients into malignant or benign groups is the subject of much research. Because of the critical feature detection in the complex breast cancer dataset is unique. Therefore deep learning is recognized as the best choice in pattern classification in much research. For instance, Khuriwal and Mishra [1] proposed an adaptive ensemble voting technique for diagnosing breast cancer. Moreover, they compared the proposed method to provide a better solution by using machine learning algorithms.

A lot of studies have been done in the cancer classification with serveral techniques include linear discriminant analysis (LDA), k-nearest neighbors (KNN), Naive Bayes, and convolutional neural network (CNN), support vector machine (SVM), and so on. The most-used machine learning algorithms such as the k-nearest neighbors algorithm [2], and KNN can be applied to measure the performance of the various classification techniques [3], especially for cancer diagnosis. The accuracy of many of them was related to the attribute of the dataset.

Classification is the most important technique. The classification model learns a piece of information from training data, then it makes a minimal error as possible. The study of classification modeling algorithms has been conducted, and successfully used in many application areas, including the medical datasets [4]. The problem of the classification occurs when there is an unequal distribution of classes in the training dataset. This distribution can vary from a slight bias to a severe imbalance. This is caused by the low performance because of the traditional machine learning models, and evaluation metrics assumes based on the balanced class distribution. An imbalanced in medical data problem can obstruct classifier learning include an impact of noisy data and disjunction due to nonequivalent instances. [5]. For this reason, the imbalance classification is not only a challenge for making predictive modeling, but also it is very hard when creates the model is implemented on the imbalanced datasets. An extreme gradient boosting (XGBoost) [6] is one of the most popular deep learning technique which is frequently used for the classification problem. XGBoost is the most effective technique for creating the predictive model. It is built on the principles of boosting decision trees which is designed for the computation speed [7]. For this reason, the objectives of this work are improving the quality of XGBoost in conjunction with cost-sensitive learning for imbalanced classification problems by enhancing the procedure with insights from XGBoost and introducing a correction of weighted sampling for boosting algorithms.

This paper is organized as follows. Section 2 explains

978-1-7281-7243-9/20/$31.00 ©2020 IEEE

the problem definition, and the background of imbalanced classification, cost-sensitive learning, XGBoost, and the proposed technique details. The experiment steps, data preparation, model tuning and model configurations are presented in Section 3. Section 4 discusses the results, and evaluation. Section 5 represents the research conclusion and contribution.

II. Background and Proposed Techniques

This section describes the problem background, the theoretical basis of extreme gradient boosting, imbalanced classification, cost-sensitive learning, and the proposed technique details.

A. Problem Background

The inaccurate medical diagnosis classification of a sample could have fatal results in any decisions, especially in the case of suffering patients. In addition, the performed classification model can provide a precise diagnosis for the diseases. Basically, the machine learning algorithms are trained on a dataset and search for minimize error. Later, this method fit a model on the trained data to solves an optimization problem. The functions can be used to evaluate the error of a model on training data, which is referred as loss. In this way, the machine learning method is applied to overcome tasks such as classification and clustering. Note that the development of a machine learning algorithm needs the appropriate structured of data, from diverse sources, and in diverse formats. Most machine learning algorithms are designed for optimizing the classification accuracy. In medical data classification, they often face the imbalanced number of data instances, at least one of the classes have a very small minority of the data. Consequently, it is a hard problem in most of the machine learning algorithms. [8]. Inaccurate medical diagnoses on the minority class have a more probable chance than the majority class. These problems are mostly caused by data imbalanced classification. This case, the most distinct characteristic is the skewness of data distribution between the classes. Much research indicates that the skewness of data distribution is not only affected by the parameters of the model but also a classifier in identifying events. Moreover, the other impact includes small sample size and reparability [9].

In order to achieve higher accuracy for cancer classification, this work applies XGBoost algorithms for imbalanced classification in conjunction with cost-sensitive learning. At the data level, this work will re-balance the class distribution by resampling the data frame. At the algorithm level, this work will adjust the classifier, parameters, the weighted score of learning algorithms to bias towards the small class by the XGBoost algorithm. Therefore, the objective of this work on imbalanced classification problem is to improve diagnosis in the minority class. The methods developed to overcome the imbalance data problem.

B. Extreme Gradient Boosting

The extreme gradient boosting (XGBoost) proposed by Chen and Guestrin [10], is a scalable machine learning technique based on the gradient boosting algorithm [11]. XGBoost operates with speed and accurate prediction. It does by merging a set of weak and strong learner iteratively in order to meet the best prediction accuracy [6].

$$\bar{L}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{1}$$

As depicted in Eq. (2) denotes the scoring function for determing the quality of tree structure q.

$$Gain = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{2}$$

Differ from a common gradient boosting algorithms, that based on the only first derivative γ_i, whereas XGBoost uses both the first g_i, and second derivative h_i. XGBoost performs on the right cost function, and create a prediction model.

C. Cost-Sensitive Imbalanced Classification

Cost-sensitive learning [12] is a technique of machine learning. This technique focus on the data that have uneven costs when making prediction and classification. The theory of cost-sensitive learning [13] describes the process of misclassification cost in to consideration. The goal of technique is to minimize the total cost. An imbalanced classification problem occurs when the distribution of multi-class data is not equal. The classes with a small number of data are referred to as the minority class whereas the other data are merged into the majority class. The distribution can vary from a slight bias to a severe imbalance. In the case of imbalanced classification, cost-sensitive learning is focused on early assigning different costs to the types of misclassification errors, then using specific methods to take those costs. By considering the identical table with the same rows and columns and assign a cost to each of the cells. This is called a cost matrix that assigns a cost to each into the confusion matrix.

The costs of false positive(FP), false negative (FN), true positive (TP), and true negative (TN) can be given in a cost matrix as depiected in Table I.

Table I
An example of cost matrix for classification

#	Actual negative	Actual Positive
Predict negative	TP	FN
Predict positive	FP	TN

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{3}$$

where $Accuracy$ is a ratio of correctly predicted observation to the total observations.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

where $Precision$ is the ratio between the correctly predicted positive and the total predicted positive observations.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

where $Recall$ is calculation of correctly predicted positive by the all observations in actual class.

$$F1Score = \frac{2 \times Precison \times Recall}{Precision + Recall} \tag{6}$$

where $F1Score$ is the weighted average of $Precision$ in Eq.(4) and $Recall$ in Eq.(5).

A classification performance is evaluated by using the confusion matrix. This matrix contains information about the actual and the predicted class which presents the values of TP, FP, FN and TN. In order to achieve best quality results for both classes, this work also apply the Receiver Operating Characteristic (ROC) [14]. The Area Under the ROC Curve (AUC) [15] is used to measure a classifier for evaluating the better model performance which the calculation formula is depicted in Eq.(7).

$$\text{AUC} = \frac{1 + TP_{rate} - FP_{rate}}{2} \tag{7}$$

$$IR = \frac{T_{max}}{T_{min}} \tag{8}$$

In order to determine the degree of uneven class in a dataset, it is described in terms of the imbalance ratio (IR). The foumulation of IR is depicted in Eq.(8). where, T_{max} is a number of instances of majority class and T_{min} is number of instances of minority class. In this case, the classifiers provide the imbalanced degree of predictive accuracy. Normally, the majority class gives better accuracy than the minority class. [16]. Since the class distribution is not balanced, most machine learning algorithms require modification to avoid poorly prediction in the majority class. Therefore, the ROC Area Under Curve is needed for evaluating the prediction accuracy.

Imbalanced classification makes a challenge for building the predictive model because most machine learning algorithms were performed on the presumption in an equal number of instances for each class. Consequently, the results of models may show nominal predictive performance, especially for the minority class because the minority class is more important and more sensitive to classification errors than the majority class.

III. Experiments

This work conducted the experiments by separating into four steps as follows. Firstly, the data will be pre-processed, and do the data analysis. Then, the experiment will create the common XGBoost model, configure the optimal model parameters, and make the feature selection. In step 3, the new model will be created by combining the cost-sensitive learning in conjunction with the XGBoost model. This new version of XGBoost is referred to as cost-sensitive XGBoost. Finally, the experiment will tune the cost-sensitive XGBoost model by class weighting hyperparameter.

A. Data Preparation and Description

This work collected the cancer data from four different sources. The Wisconsin Breast Cancer dataset and Coimbra Breast cancer dataset are selected from the UCI Machine Learning repository. Likewise, Biopsy breast cancer dataset are collected from Rdatasets. In addition, this work makes the randomly generated dataset in order to control the imbalance ratio. The Wisconsin Breast Cancer, Coimbra Breast cancer, Biopsy breast cancer, and randomly generated dataset include 569, 116, 699, and 1,186 instances, respectively. The details of each dataset are shown in Table II. Most of data are preprocessed by data cleaning, and data transformation.

Table II
A SUMMARY OF DATASETS USED IN THE EXPERIMENTS.

#	Datasets	In-stances	Minority	Majority	IR
1	Breast Can-cer(Wisconsin)	569	212	257	1.684
2	Breast Cancer(Coimbra)	116	52	64	1.231
3	Breast Cancer(Biopsy)	699	241	458	1.900
4	Randomly Generated	1,186	267	919	3.442

The distribution of all data is represented by the scatter plot. The scatter plots are created to display values for two classes for each dataset that belongs to the minority class and majority class. Furthermore, we can see some measure of overlap between the two main classes. The scatter plots of each dataset are depicted in Figure 1.

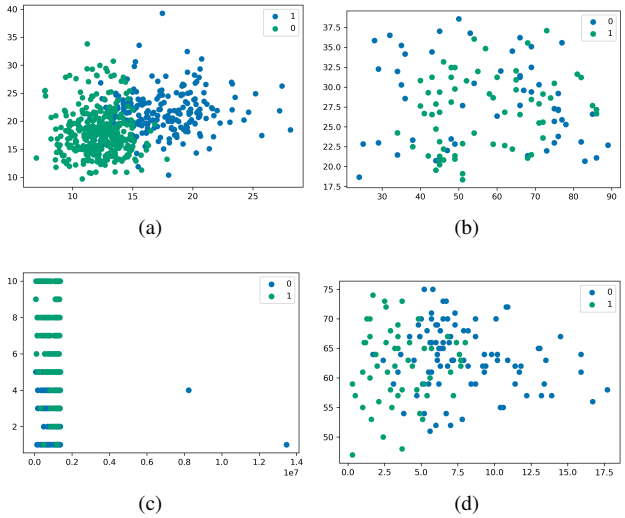

Figure 1. The scatter plot of different breast cancer dataset (a) Wisconsin, (b) Coimbra, (c) Biopsy, and (d) randomly generated

B. XGBoost Modeling and Feature Seclection

1) Feature Seclection: Feature importance is a technique that assigns a score to input features. This technique base on

the degree of usefulness, that is at predicting a target variable. The feature importance scores can be calculated from many ways such as statistical correlation scores, coefficients of linear models, and so on. These scores are very important value in the predictive model because it provides insight into the data model, some basis for dimensionality reduction and feature selection, etc., Furthermore, feature importance is used for improving the efficiency predictive the predictive model.

2) XGBoost Configuration: After XGBoost models are built, the configuring and tuning the model is the difficult step because the XGBoost model requires carefully parameters tuning to make the optimal model performance. There are many essential parameters such as the number of rounds for performing, the number of trees, maximum depth, learning rate, importance score, etc. Therefore, this work determines by using hyperparameter optimization technique. This technique is conducted extensively using a random search, by assigning a collection of essential used parameters. A benefit of using the hyperparameter optimization technique can automatically provide the best major parameters from a trained predictive model.

3) Cost-sensitive for XGBoost: Cost-sensitive learning for imbalanced classification is focused on assigning the different costs to the types of misclassification errors. The experiments tune the behavior of the XGBoost algorithm for imbalanced classification adjusts the ratio of the number of negative classes to the positive class. This work vary this ratio from 1 through 100 by using hyperparameter optimization technique. This ratio has the effect of weighing the balance of positive examples, relative to negative examples when boosting decision trees. For this imbalanced binary classification dataset, the negative class refers to the majority class is 0 and the positive class refers to the minority class is 1. Furthermore, the method starts by calculating and reviewing the confusion matrix. This matrix summarizes the number of predictions for each class, separated by the actual class to which each instance belongs.

IV. Results and Evaluations

This section is organized into two main parts. The results from the experiments will be represented, and also conducted the evaluation of the proposed model.

A. Results

The performance of XGBoost model relates to the number of selected features. Therefore, this work determines the importance scores for indicating the value of each feature. The right feature importance score can improve performance, simplify the models, and reduce over-fitting. The feature important score of each dataset is depicted in Figure 3.

The model accuracy of the prediction is illustrated in Table III. The experiments, it was found classification using cost-sensitive XGBoost model performed better when compared with classification from the original XGBoost model. For instance, the classification accuracy for breast cancer dataset is 96.40% while the accuracy for cost-sensitive XGBoost is 99.12%. The cancer classification was done using cost-sensitive XGBoost and the accuracy was measured using confusion matrix. Figure 2 show the confusion matrix for the classification on four datasets. From Eq.(3), Eq.(4), Eq.(5), and Eq.(6), we can calculate the accuracy, precision, recall, and F1 score of earch data as follows. For Wisconsin dataset as depicted in Figure 2(a), we have got 0.965 which means the proposed model is 96% accurate, 0.961 precision, recall of 2.0, and F1 score is 1.298. For Wisconsin dataset as depicted in Figure 2(b), we have got accuracy of 0.875, 0.961 precision, recall of 2.0, and F1 score is 1.298. For Coimbra dataset as shown in Figure 2(b), we have got accuracy of 0.875, 0.900 precision, recall of 0.643, and F1 score is 0.750. In the case of Biopsy dataset as shown in Figure 2(c), we have got accuracy of 0.964, 0.977 precision, recall of 1.556, and F1 score is 1.200. Finally, arndomly generated dataset as shown in Figure 2(d), we have got accuracy of 0.692, 0.806 precision, recall of 2.544, and F1 score is 1.224.

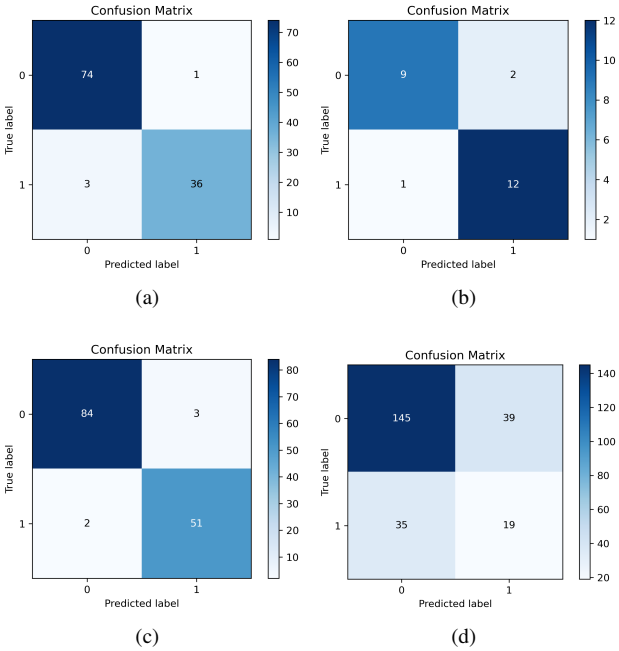

Figure 2. The confusion matrix of different breast cancer dataset (a) Wisconsin, (b) Coimbra, (c) Biopsy, and (d) randomly generated

In addition, this work keeps track of performance of the training data on each dataset. Figure 4 shows the model classification error on each iteration on each dataset. Figure 5 depicts the plot of logarithmic loss, the model could stop the learning of the different datasets around epoch 30, 30, 80, and 40, respectively.

B. Evaluations

To evaluate the proposed model, this work has done by creating the training dataset and testing dataset. Later, this work evaluates the proposed model by using k-Fold cross-validation. Furthermore, this work also measured the Area Under the Curve (AUC) of the Receiver Operating Curve

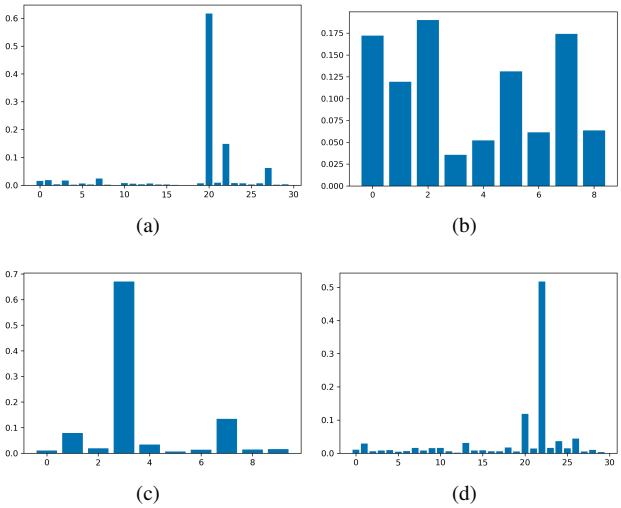

Figure 3. The feature importances of different breast cancer dataset (a) Wisconsin, (b) Coimbra, (c) Biopsy, and (d) randomly generated

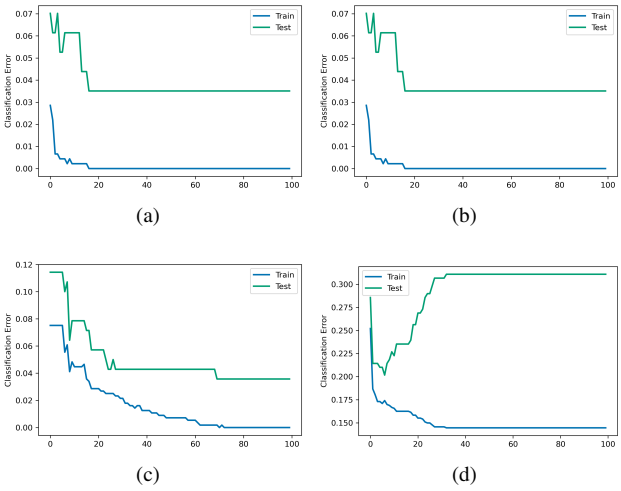

Figure 4. A classification error plot of different breast cancer dataset (a) Wisconsin, (b) Coimbra, (c) Biopsy, and (d) randomly generated

Table III
A SUMMARY OF ACCURACY COMPARISON BETWEEN COMMON XGBOOST AND COST-SENTSITIVE XGBOOST MODELS.

Datasets	ROC AUC	Accuracy	
		None	Cost-Sensitive
Breast Cancer (Wisconsin)	0.9933	96.491	99.123
Breast Cancer (Coimbra)	0.8074	57.348	87.500
Breast Cancer (Biopsy)	0.9928	95.996	96.429
Randomly Generated	0.7674	66.802	76.365

(ROC) as shown in Table III. The results indicate that cost-sensitive XGBoost model can improve the prediction accuracy.

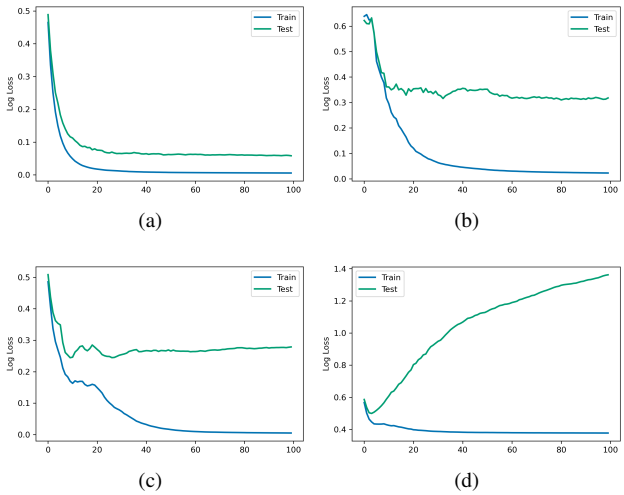

Figure 5. A logarithmic loss plot of different breast cancer dataset (a) Wisconsin, (b) Coimbra, (c) Biopsy, and (d) randomly generated

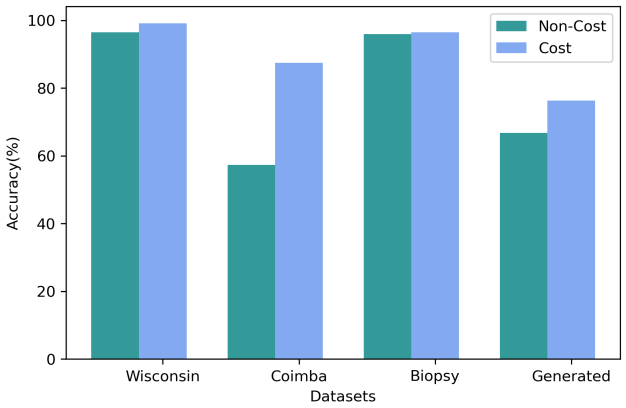

Figure 6. An accuracy comparison of XGBoost and cost-sensitive XGBoost on each dataset

In the case of Wisconsin breast cancer dataset, the model could make the prediction accuracy from 96.49% to 99.12%. The model could improve the prediction accuracy from 57.35% to 87.50% in the Coimbra breast cancer dataset. In the Biopsy breast cancer dataset the model could make the prediction accuracy from 95.99% to 96.43%. Finally, the model could make the prediction accuracy from 66.80% to 76.36% in the case of the randomly generated dataset. The model accuracy of classification is compared as shown in Figure 6.

V. CONCLUSION

This work proposed the cost-sensitive XGBoost model, which is an improved version of the XGBoost model in conjunction with cost-sensitive learning. The models were applied to classify the four breast cancer datasets that contained the imbalanced data. In the experiment, this work determined the best parameters on each dataset by the hyperparameters optimization technique before configuring the models. The results indicated that the cost-sensitive XGBoost model had

been skillful, and could improve classification accuracy in four datasets. In addition, this work evaluated the model performance by accuracy, ROC AUC, and k-Fold cross-validation to ensure that the new models is accurate.

The contribution of this work is to create a cost-sensitive XGBoost model for imbalanced data classification, which can be applied to another field data classification and analysis methods.

REFERENCES

[1] N. Khuriwal and N. Mishra, "Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm," in *2018 IEEMA Engineer Infinite Conference (eTechNxT)*, March 2018, pp. 1–5.

[2] J. Sánchez, R. Mollineda, and J. Sotoca, "An analysis of how training data complexity affects the nearest neighbor classifiers," *Pattern Analysis and Applications*, vol. 10, pp. 189–201, 08 2007.

[3] P. Bhuvaneswari and B. Therese, "Detection of cancer in lung with k-nn classification using genetic algorithm," *Procedia Materials Science*, vol. 10, pp. 433–440, 12 2015.

[4] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Science*, vol. 83, pp. 1064 – 1069, 2016, the 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops.

[5] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, July 2012.

[6] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 10 2000.

[7] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119 – 139, 1997.

[8] S. Belarouci and M. Chikh, "Medical imbalanced data classification," *Advances in Science, Technology and Engineering Systems Journal*, vol. 2, pp. 116–124, 04 2017.

[9] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, pp. 429–449, 2002.

[10] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794.

[11] L. Breiman, "Arcing the edge," Statistics Department, University of California at Berkeley, Tech. Rep., 1997.

[12] M. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume, and C. Brunk, "Reducing misclassification costs," in *Machine Learning Proceedings 1994*, W. W. Cohen and H. Hirsh, Eds. San Francisco (CA): Morgan Kaufmann, 1994, pp. 217 – 225.

[13] C. Elkan, "The foundations of cost-sensitive learning," in *In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973–978.

[14] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[15] Jin Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, March 2005.

[16] B. Pal and M. K. Paul, "A gaussian mixture based boosted classification scheme for imbalanced and oversampled data," in *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2017, pp. 401–405.