Universal Journal of Public Health 12(3): 434-440, 2024
DOI: 10.13189/ujph.2024.120302

Ensemble of XGBoost Classifiers Based on LDA Dimensionality Reduction for Predicting Breast Cancer

Mai Nhu Uyen Le[1,2], Jianlin Zhou[1], Dinh Phu Cuong Le[3,4,*], Dong Wang[3]

[1]State Key Laboratory of Developmental Biology of Freshwater Fish & Key Laboratory of Protein Chemistry and Developmental Biology of the Ministry of Education, College of Life Science, Hunan Normal University, China
[2]Faculty of Medicine and Pharmacy, Yersin University of Dalat, Vietnam
[3]College of Computer Science and Electronic Engineering, Hunan University, China
[4]Yersin University of Dalat, Vietnam

Received November 16, 2023; Revised April 12, 2024; Accepted May 20, 2024

Cite This Paper in the Following Citation Styles

(a): *[1] Mai Nhu Uyen Le, Jianlin Zhou, Dinh Phu Cuong Le, Dong Wang , "Ensemble of XGBoost Classifiers Based on LDA Dimensionality Reduction for Predicting Breast Cancer," Universal Journal of Public Health, Vol. 12, No. 3, pp. 434-440, 2024. DOI: 10.13189/ujph.2024.120302.*

(b): *Mai Nhu Uyen Le, Jianlin Zhou, Dinh Phu Cuong Le, Dong Wang (2024). Ensemble of XGBoost Classifiers Based on LDA Dimensionality Reduction for Predicting Breast Cancer. Universal Journal of Public Health, 12(3), 434-440. DOI: 10.13189/ujph.2024.120302.*

Abstract As reported by the World Health Organization, breast cancer is recognized as the most popular disease in women. Thus, the need for early and accurate detection of this cancer for effective treatment is highly demanded. In this paper, a novel machine learning-based method is proposed to improve the success of breast cancer prediction. To be specific, Extreme Gradient Boosting (XGBoost), which is an efficient machine learning algorithm to deal with large datasets, is applied with the help of the Linear Discriminant Analysis (LDA) algorithm, which is often used for dimensionality reduction by fusing the original multidimensional data features, to create the cancer predictive model. From the experimental results, with the LDA, it is shown that the XGBoost classifier can help to improve the classification accuracy by 2.7 % compared to the classifier without using LDA. Moreover, when compared to other machine learning methods, the proposed method also shows a better classification result with the root mean squared error of 0.115, which means that its error is at least 2.6 % lower than others. The proposed method aims to support doctors in enhancing clinical application as well as improving medical quality, especially when detecting the very first moment of breast cancer.

Keywords Breast Cancer, Machine Learning, XGBoost, Dimensionality Reduction, LDA

1. Introduction

Recently, from the reports of the World Health Organization about the latest global cancer data, breast cancer has become the number one cancer in the world [1], especially in women. It is obviously noticed that the patients with breast cancer symptoms cannot be underestimated. The seriousness of breast cancer prevention and treatment has received lots of attention from the entire society, therefore, early detection, diagnosis, and treatment are really important. Nowadays, a large number of data analysis technologies are used to work with breast cancer.

As machine learning algorithms are paid attention and are being applied in various fields of human lives, these algorithms are also utilized to predict breast cancer. Using the support vector machine (SVM), Chenwen et al. [2] proposed an effective algorithm based on a weighted eigenvector kernel function to classify and predict breast cancer. With the same idea of utilizing SVM, Yu et al. [3] introduced the model with the help of a systematic clustering technique to diagnose different tumor characteristics of breast cancer. For breast cancer diagnosis, Jun et al. [4] applied the BP-GamysBoost algorithm while

the authors in [5] attempted to use the convolutional neural networks. Moreover, Feng and Fan [6] proposed a novel method to predict the recurrence rate of breast cancer patients based on a double-weighted Naive Bayes algorithm. Although the above algorithms have achieved good results, the accuracy of classification prediction still needs to be improved. The reason is that the mechanism of breast cancer is uncertain and the physical and mental quality of different patients varies widely, thus, it is necessary to develop a personalized treatment plan for each individual to improve treatment effectiveness [7].

In the process of personalized treatment of breast cancer, patients need to test many indicators. Therefore, finding meaningful indicators among the whole collected information has an important impact on the patients and it can be the key to preventing them from the effects of cancer. Initial breast cancer screening indicators often contain redundant information and noisy data. If the size of the data is getting larger, then the problem complexity and time complexity will increase. Hence, reducing the dimensionality of data features can reduce the amount of computing time in the system. Moreover, utilizing machine learning algorithms to build prediction models can help doctors detect the appearance of breast cancer early and treat the patients in the most effective way. The contributions of this paper are as follows.

- The LDA algorithm is utilized to merge object features, reduce errors due to data redundancy, quickly reduce data size, and reduce system complexity.
- The XGBoost algorithm is applied to deal with large a mount of data when diagnosing the very first signs of breast cancer.
- The performance of the proposed method with the two algorithms is evaluated in the public breast cancer diagnosis dataset and compared to other machine learning methods.

2. Materials and Methods

2.1. Dataset

In this paper, the proposed method with XGBoost and LDA algorithms is evaluated in the breast cancer diagnosis dataset named Breast Cancer Wisconsin, which is provided by the Kaggle database [9]. To be specific, there are 357 benign samples (marked as 0) and 212 malignant samples (marked as 1), each sample has 30 features which are calculated from the digitized images, one type label, and one sample id. The type label indicates whether the patient's breast cell nucleus is benign or malignant to diagnose breast cancer. Figure 1 displays the difference between benign and malignant samples. Moreover, Figure 2 presents the distribution of the benign (i.e., the orange dots) and malignant samples (i.e., the blue dots). The figure points out that these two samples have overlapping parts, thus, the dimensionality reduction is required. The descriptions of the 30 features used for classification are shown in Table 1. Furthermore, Figure 3 shows the distribution of the features. From the figure, it can be seen that the *concave points_mean*, *radius_worst*, and *concave points_worst* are the three features that obtain the largest distributions among other features, respectively. To evaluate the proposed method, these properties are divided into the training samples and testing samples with percentages of 70 % and 30 %, respectively.

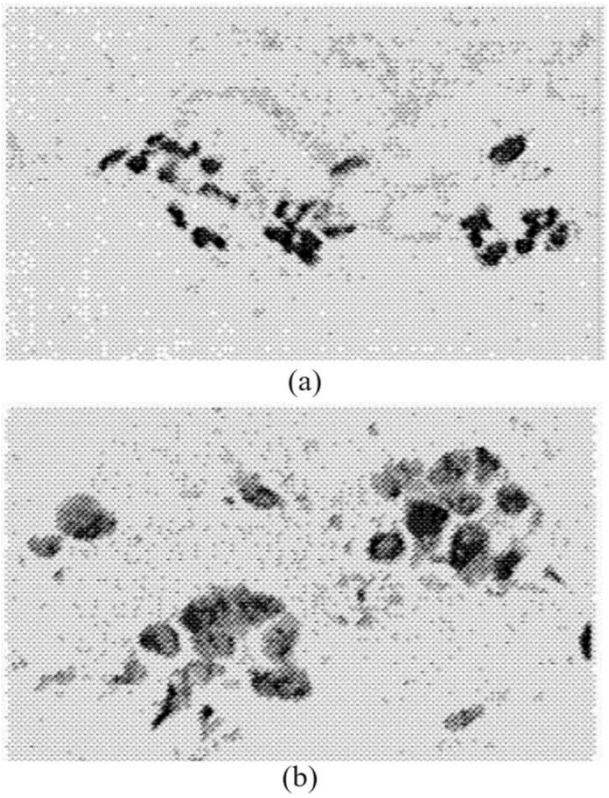

Figure 1. Images of (a) benign and (b) malignant samples [8]

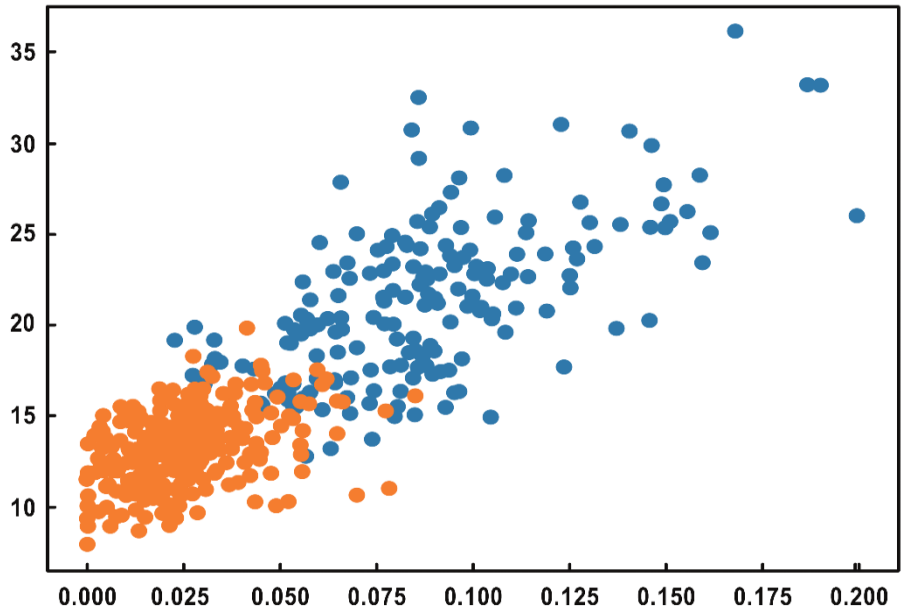

Figure 2. Distribution of benign (orange dots) and malignant (blue dots) samples

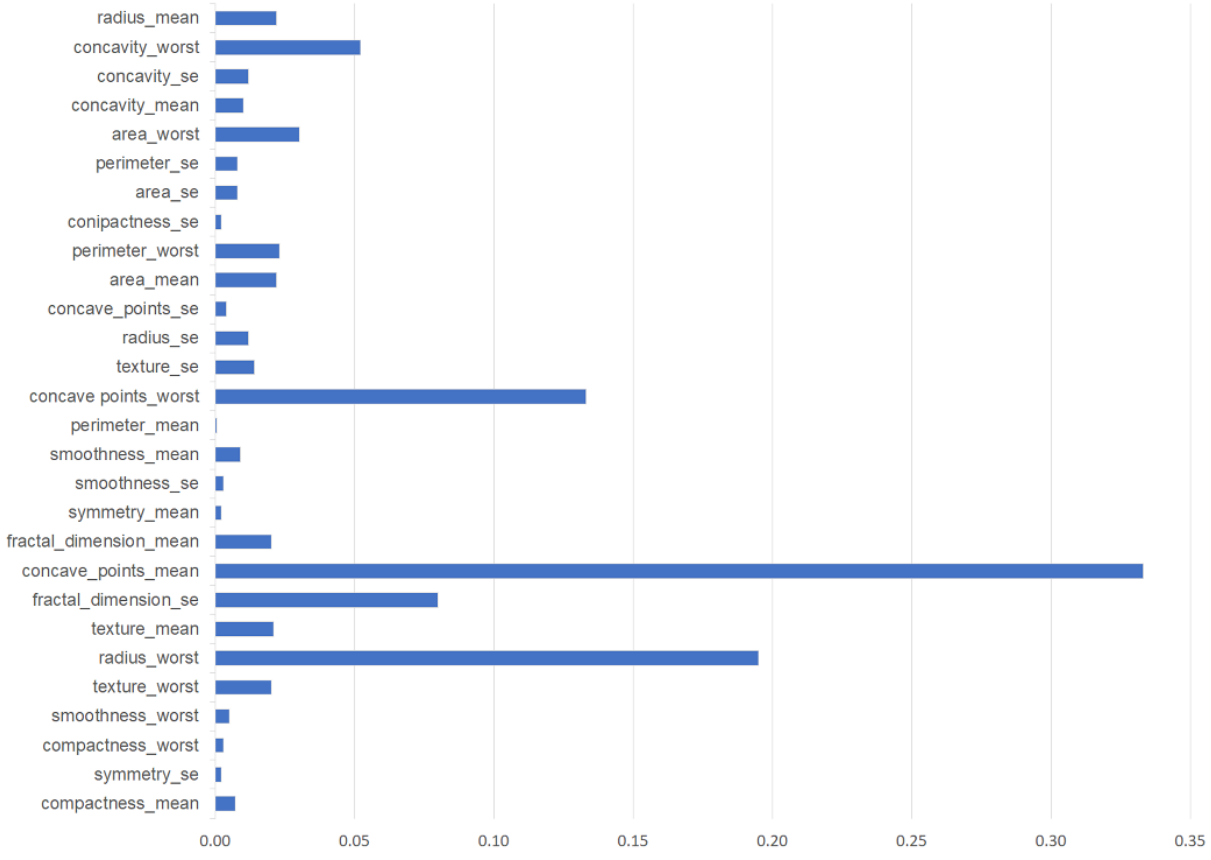

Figure 3. Distribution of the whole feature

Table 1. Summary of the 30 features

Features
Average value from the center to points around the lesion
Gray value standard deviation
Average tumor core area size
Average area
Smoothness (local variation in radius length) mean value
Compression (mean squared perimeter/ area -1.0)
Average degree of borderline depression
Average number of contour indentations
Symmetrical meaning
Fractal dimension
Standard error of the average distance from the center to the perimeter points
Standard error of gray value standard deviation
Standard error of circumference
Regional standard error
Smoothness (local variation in radius length) standard error
Compression (mean squared perimeter/ area -1.0) standard error
Standard error of contour indentation severity
Standard error of the number of contour indentations
Standard errors are symmetrical
Fractal dimension standard error
Worst radius (large)
Worst standard deviation grayscale value standard deviation (large)
Circumference worst (large) sense of symmetry
Worst region (large) fractal dimension
Worst smoothness (large) (local variation in radius length)
Worst (large) compactness (mean squared circumference/area -1.0)
Worst degree of borderline depression (large)
Worst number of contour dents (large)
Worst symmetry (large)
Worst fractal dimension (large)

After the dataset preprocessing process to avoid missing data, abnormal data, duplicate data, etc., the benign and malignant data are relatively balanced. The diagnostic categories of this dataset are formatted as strings. To facilitate comparison, the diagnostic results are classified into two categories, where 0 represents benign breast cell nuclei and 1 represents malignant breast cell nuclei. When the data size is getting bigger, to extract useful features and facilitate computation, the data size reduction process is required. The original dataset contains 30 features which often bring redundant and meaningless information, thus, it leads to the increased computational load and time complexity of the whole classification system.

2.2. Methods

2.2.1. LDA method

LDA is a method of merging object features based on a classification model and is a supervised dimensionality reduction method [10]. The basic principle is to project data with categorized labels into a lower dimensional space to distinguish projected points by category. After the projection process, the distance between similar points in space will be closer and the variance will be smaller while the variance between different categories will be larger [11]. Dimensionality reduction can reduce errors caused by redundant information and improve classification accuracy.

Provided that dataset $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ has n-vector sample dimension $x_i, N_j (j = 0, 1)$ is the number of samples of the j^{th} type, $X_j \ (j = 0, 1)$ is the set of samples of the j^{th} type. The following steps describe how the LDA helps to reduce the dimensionality.

As the first step, the mean vector μ_j and the covariance matrix $\sum j$ of each sample are calculated.

$$\mu_i = \frac{1}{N_j}\sum_{x \in X_j} x \quad (j = 0,1) \tag{1}$$

$$\sum_j = \sum_{x \in X_j} (x - \mu_j)\left(x - \mu_j\right)^T (j = 0,1) \tag{2}$$

Next, the intra-class divergence matrix S_W and the intra-class divergence matrix S_b are calculated.

$$S_W = \sum_0 + \sum_1 = \sum_{x \in X_0} (x - \mu_0)(x - \mu_0)^T + \sum_{x \in X_1} (x - \mu_1)(x - \mu_1)^T \tag{3}$$

$$S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T \tag{4}$$

The given projection line is a vector w, then for any pattern x_i, its projection onto the given w is defined as $W^T X_i$. In the case there are two center points μ_0 and μ_1 of the two input classes, i.e. benign and malignant, the above projection w becomes $w^T\mu_0$ and $w^T\mu_1$. According to the LDA principle, it is necessary to maximize the distance between different categories, i.e., making $\|w^T\mu_0 - w^T\mu_1\|^2$ as big as possible, while minimizing the distance among the same categories, i.e., minimizing $w^T \sum_o w + w^T \sum_1 w$. Therefore, the optimization goal to reduce the data dimensionality is:

$$argmax_W J(W) = \frac{w^T S_b w}{w^T S_w w} \tag{5}$$

To sum up, the main idea of LDA is to make the optimal projection straight line w, transform each sample feature xi in the sample set to get new output samples $w^T x_i$, and finally, obtain the output dataset with completed dimensionality reduction.

According to the LDA principle, the object size after dimensionality reduction for binary classification problems needs to be no larger than 1. When using LDA to reduce the data dimensionality of the aforementioned dataset, it reduces the data object size from 30 dimensions (i.e., features) to 1 dimension, as shown in Figure 4. All data are projected onto the same line, the distance between projection points of data of the same type is reduced, and the distance between data centers of different types is increased.

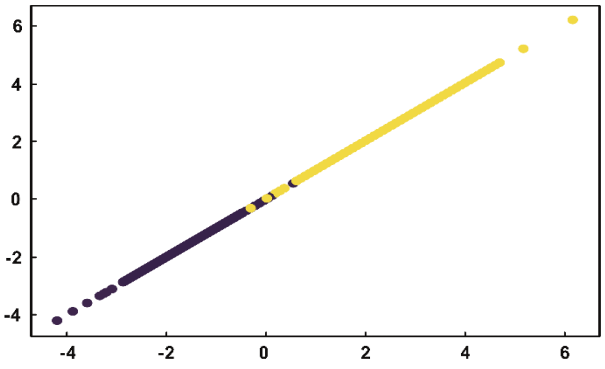

Figure 4. The output after using LDA for dimensionality reduction

2.2.2. XGBoost Method

XGBoost [12] is a kind of ensemble learning method that can be used as an ensemble of multiple linear classification decision trees. It is implemented with the gradient boosting algorithm to increase the classification speed and performance, thus, it is sufficient to deal with a large dataset. The basic idea of this method is to integrate multiple tree models, which are considered weak classifiers, into a stronger classifier with higher classification accuracy. XGBoost supports feature sampling, which can effectively prevent oversampling installation problems and reduce the amount of system computation [13]. The training process of the XGBoost method is calculated as in Equation (6).

$$\gamma_i = \varphi(x_i) = \sum_{k=1}^{K} f_k(x_i) \quad (6)$$

where γ_i is the predicted value, K represents the number of trees, xi represents the i^{th} sample and the weighted sum of the outputs of the K trees is the final predicted value of the XGBoost model. Its model training functions are as Equation (7).

$$L(\varphi) = \sum_i l(\gamma_i, h_i) + \sum_k \Omega(f_k) \quad (7)$$

where the first item is the loss function, which represents the prediction error of the i^{th} sample. The second item is the regularization item, which adjusts the model complexity to prevent over-fitting. The original dataset outputs the results after training the first weak classifier, i.e., the decision tree classifier, then continuously adjusts the next weak classifier based on the returned balance until the specified number of classifiers has been trained.

3. Experimental Results

3.1. Evaluation Metrics

In this paper, the components of the confusion matrix are used to compare the classification results with the actual sample information, each row in the matrix represents the predicted class and each column represents the actual class, as shown in Table 2. The accuracy is the first evaluation metric that is calculated from the confusion matrix as in Equation (8).

$$accuracy = \frac{(TP+TN)}{TP+TN+FP+FN} \quad (8)$$

The second metric is the receiver operating characteristic (ROC) curve which can objectively measure the performance of the model. If the ROC curve is near the upper left corner, then it means that the model has better classification performance of the classification model. If the ROC curve is smooth, it can be determined that there is no over-fitting phenomenon. The Area Under Curve (AUC) is another related metric, which is the area under the ROC curve and the coordinate axis. AUC is usually in the range of 0.5 to 1. The larger the AUC value, the better the model performance and the higher accuracy.

The root mean square error (RMSE) is the last evaluation metric. It is also known as a standard error which reflects the deviation between the predicted sample and the actual sample, the smaller the value the higher the accuracy and stability of the model.

Table 2. Confusion matrix

	Positive (Actual samples)	Negative (Actual samples)
Positive (Predicted samples)	True Positive (TP)	False Positive (FP)
Negative (Predicted samples)	False Negative (FN)	True Negative (FN)

3.2. Evaluation Results

In this paper, the grid search is used for performing cross-validation, which aims to obtain optimal parameters to build the XGBoost model. The parameters of XGBoost are shown in Table 3.

Table 3. The chosen parameters of XGBoost

Parameter	Value
Learning rate	0.1
Maximum tree depth	2
Minimum total leaf node sample weight	5
Number of weak classifiers	66
Model penalty gamma	0
Proportion of columns	0.3
Proportion of random sampling of each tree in the subsample	0.9

To evaluate the performance of the XGBoost method, the other three methods named Adaboost, Random forest, and GaussianNB are implemented as the classifiers for

performance comparison. The LDA method is applied to the above methods for a fair comparison. The results between the two cases, i.e., with and without using the LDA, before the classification process are shown in Figure 5. According to the figure, it can be seen that the classification accuracies of the models with dimensionality reduction by LDA (the orange bars) are on average 2.7% higher than the models that do not use the LDA (the blue bars).

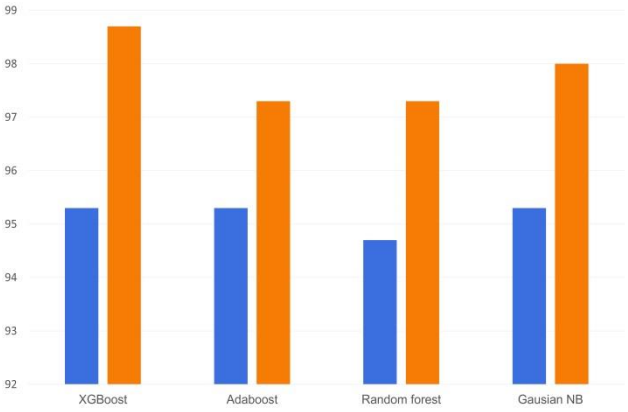

Figure 5. Accuracy comparison of the models with and without using the LDA method

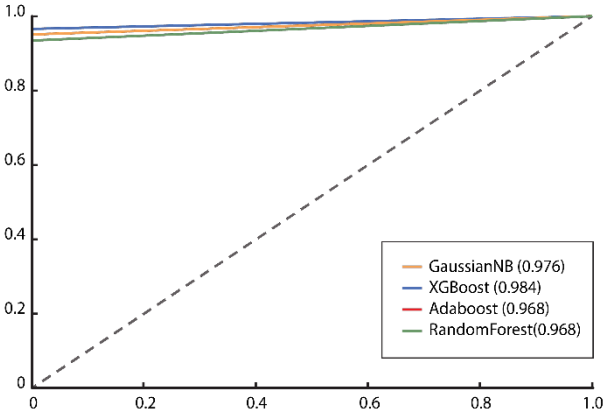

Figure 6. ROC curve of each algorithm model after dimensionality reduction by using LDA

Table 4. Model implementation results

Model	Accuracy	AUC	RMSE
XGBoost	0.984	0.980	0.115
Adaboost	0.968	0.964	0.153
Random Forest	0.968	0.964	0.153
GaussianNB	0.976	0.972	0.131

Next, the ROC curve of each method after using LDA is presented in Figure 6. The dashed line represents the ROC curve of the random classifier. If one method is further away from the dashed line, then this method is considered to achieve better classification performance. From the figure, it can be seen that the performance of the XGBoost methods is the best compared to others, in which the average classification accuracy of the XGBoost reaches 0.987 and the average AUC index reaches 0.984. The RMSE of XGBoost is only 0.115, which is 4.8%, 4.8%, and 2.6% lower than the errors of Adaboost, Random Forest, and GaussianNB, respectively. The detailed results are summarized in Table 4.

4. Conclusions

The need for early detection and treatment of breast cancer is severe, thus, improving the detection accuracy of cancer is practical and significant. In this paper, the LDA method is used to merge the multi-dimensional data features to reduce the data dimensionality and system complexity while the XGBoost method is used to build a breast cancer prediction model. The experimental results show that the classification performance of the prediction model when using dimensionality reduction from LDA is much better than the models without using the LDA. Moreover, the breast cancer prediction model constructed by the XGBoost method has good classification accuracy, which supports doctors in enhancing clinical application and improving medical level. In future works, different machine learning methods will be combined to utilize the strength of each method, and then finally can help to improve the classification accuracy.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61502162, Grant 61702175, and Grant 61772184, in part by the Fund of the State Key Laboratory of Geoinformation Engineering under Grant SKLGIE2016-M-4-2, in part by the Hunan Natural Science Foundation of China under Grant 2018JJ2059, in part by the Key R & D Project of Hunan Province of China under Grant 2018GK2014, and in part by the Open Fund of the State Key Laboratory of Integrated Services Networks under Grant ISN17-14. Chinese Scholarship Council (CSC) through College of Computer Science and Electronic Engineering, Changsha 410082, Hunan University with grant CSC No. 2018GXZ020784.

REFERENCES

[1] WHO, Breast Cancer Now Most Common Form of Cancer: WHO taking action, Online available from https://www.who.int/news/item/03-02-2021-breast-cancer-now-most-common-form-of-cancer-who-taking-action

[2] W. Chenwen, L. Changsheng, W. Wei, L. Jinghan, Y. Guanghui. Application of an Improved SVM Algorithm in

Breast Cancer Diagnosis, Computer Engineering and Science, Vol. 39, No. 03, 562-566.

[3] Y. Yu, Z. Fan, R. Zhu, H. Xiong. Research on Breast Cancer Diagnosis Based on Hierarchical Clustering and SVM Model, Intelligent Computers and Applications, Vol. 10, No. 11, 97-100.

[4] L. Jun, P. Hui-xian, H. Bin, T. Shay. Breast Cancer Diagnosis Model Based on BP-GamysBoost, Computer and Modernization, No. 04, 1-7.

[5] X. I. A. Yongqiang, Y. A. N. Yusheng. Breast Cancer Pathological Image Classification Based on a Convolutional Neural Network, Journal of Harbin Engineering University, Vol. 42, No. 4, 567-573, 2021.

[6] Z. Feng, Q. Fan. Doubly Weighted Naïve Bayesian Algorithm Predicts Breast Cancer Recurrence Rate, Journal of Mudanjiang Normal University, Vol. 2, 11-15.

[7] C. Bing-guo, L. Yu-qin, F. Zhi-chao, Y. Shan-Hu. New Intelligent Prediction of Chronic Liver Disease Based on Principal Component Machine Learning Algorithm, Computer Science, Vol. 44, No. 11A, 65-72.

[8] A. Agarap. On breast cancer detection: an application of machine learning algorithms on the Wisconsin diagnostic dataset, The 2nd international conference on machine learning and soft computing, 5-9, 2018.

[9] W. William, M. Olvi, S. Nick, and W. Street. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.

[10] H. Dong. Comparative Study of Two Data Dimensionality Reduction Algorithms, Principal Component Analysis and Linear Discriminant Analysis, Modern Computing, Vol. 29, 36-40.

[11] P. Belhumeur, J. Hespanha, D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7, 711-720.

[12] T. Chen, C. Guestrin. Xgboost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, 785-794, 2016.

[13] Z. Li, Z. Liu, G. Ding. Feature Selection Algorithm Based on XGBOOST, Journal on Communications, Vol. 40, No. 10, 101-108.