IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 10, No. 1, March 2021, pp. 184~190
ISSN: 2252-8938, DOI: 10.11591/ijai.v10.i1.pp184-190

Breast cancer prediction model with decision tree and adaptive boosting

Tsehay Admassu Assegie[1], R. Lakshmi Tulasi[2], N. Komal Kumar[3]
[1]Department of Computer Science, Faculty of Computing Technology, AIT, Aksum University, Aksum, Ethiopia
[2]Department of Computer Science and Engineering, R.V.R & J.C College of Engineering, Guntur, India.
[3]Department of Computer Science and Engineering, St. Peter's Institute of Higher Education and Research, Avadi, Chennai, India

Article Info

Article history:

Received Dec 25, 2019
Revised Oct 10, 2020
Accepted Jan 4, 2021

Keywords:

Adaboost
Breast cancer
Breast cancer prediction
Decision tree
Machine learning

ABSTRACT

In this study, breast cancer prediction model is proposed with decision tree and adaptive boosting (Adboost). Furthermore, an extensive experimental evaluation of the predictive performance of the proposed model is conducted. The study is conducted on breast cancer dataset collected form the kaggle data repository. The dataset consists of 569 observations of which the 212 or 37.25% are benign or breast cancer negative and 62.74% are malignant or breast cancer positive. The class distribution shows that, the dataset is highly imbalanced and a learning algorithm such as decision tree is biased to the benign observation and results in poor performance on predicting the malignant observation. To improve the performance of the decision tree on the malignant observation, boosting algorithm namely, the adaptive boosting is employed. Finally, the predictive performance of the decision tree and adaptive boosting is analyzed. The analysis on predictive performance of the model on the kaggle breast cancer data repository shows that, adaptive boosting has 92.53% accuracy and the accuracy of decision tree is 88.80%, Overall, the adaboost algorithm performed better than decision tree.

Corresponding Author:

Tsehay Admassu Assegie
Department of Computer Science, Aksum University
1010 Aksum, Ethiopia
Email: tsehayadmassu2006@gmail.com

1. INTRODUCTION

Breast cancer is caused by an abnormal growth and cell division in the breast tissues without control. The abnormal growth of the cells is called a tumor and results in either benign (non-cancerous) or malignant (cancerous). In recent years, breast cancer has become one of the deadliest and epidemic diseases in the world [1-5]. A literature review on the breast cancer shows that, breast cancer has become common in women [1] and cancer disease cases are expected to be 27 million by 2030 [2]. In the literature, different machine learning models are proposed as a solution in the reduction of death rate caused by breast cancer with computer assisted breast cancer diagnosis system.

Breast cancer is the second major cancer disease in women in the world [3]. The disease is common in developed countries in the past but is rapidly increasing in middle-income and low-income countries too. This shows that, the cancer disease cases are increasing rapidly and machine-learning algorithms are required for decision support to reduce the epidemic cases by predicting breast cancer as early as possible. The major problem in breast cancer prediction with machine learning is the imbalance between the benign and malignant observations in breast cancer dataset [4]. Breast cancer prediction involves a binary classification problem

where an observation belongs to either malignant or benign class. However, the number of benign observations is always greater than the number of malignant observations in the dataset as the numbers of non-cancerous people are greater than the number of cancerous people in the real world. The imbalance of observation in the dataset creates a problem to machine learning algorithm which results in incorrect predictions on the class of interest which is the malignant (minority class). As machine learning algorithm more frequently learns the majority class, the model also predicts the benign (majority class) with better accuracy than the minority class. Hence, a standard machine-learning model makes biased prediction towards the majority class.

In this research, we have proposed breast cancer prediction model with adaptive boosting algorithm to optimize the prediction performance of decision tree algorithm due to biased prediction towards benign observation. Furthermore, this study, investigates the answers to the following research questions:

1. How to optimize predictive performance of decision tree for classification of imbalanced breast cancer?
2. What is the performance of decision tree and adaptive boosting algorithm for predicting breast cancer?
3. Which feature (s) in the breast cancer dataset has strong relationship to the class feature?

2. LITREATURE REVIEW

Many research works have been conducted on breast cancer classification. The research works applied different machine learning algorithms for developing predictive model for classification of breast cancer. Some of the previous research works on breast cancer classification [5-25] are discussed in this section. In [5], naïve bayes, RBF and J48 algorithms are applied to Wisconsin breast cancer dataset. The dataset consists of 699 observations and two classes (malignant and benign) and 9 features. The experimental result of the study shows that naïve bayes algorithm performed better than RBF and J48- decision tree algorithm.

In [6], deep neural network and support vector machine is applied to an online breast cancer data repository collected from broad GDAC firehouse available online at https://gdac.broadinstitute.org/. The algorithms are evaluated against their predictive accuracy and result shows that the highest accuracy achieved by the support vector machine is 69.8%. The deep neural network performed lower than the support vector machine. In [7], the authors applied support vector machine (SVM), naïve bayes (NB), decision tree (DT) and k-nearest neighbor (KNN) on Wisconsin breast cancer dataset and proposed a breast cancer prediction model with SVM, NB, DT and KNN. The data repository contains 699 observations of which 459 are benign and 241 are malignant. The comparative performance analysis on the efficiency of the prediction models shows that SVM has better accuracy than the other algorithms.

In another study [8], on breast cancer prediction model is proposed by employing three machine-learning algorithms namely, linear regression, decision tree and random forest. In the study, the authors applied these machine-learning algorithms on the Wisconsin breast cancer data repository. The predictive performance of the proposed model is analyzed and the result of analysis shows an accuracy of 84.14%. The regression algorithm is used to analyze the relationship between the attributes in the data repository. In [9], support vector machine algorithm is applied to 573 observations collected from medical repository. The authors compared the performance of linear and non-linear support vector machine. The result of performance analysis shows that linear support vector machine outperformed than the non-linear support vector machine.

In another study [10], NB and logistic regression is applied to the Wisconsin breast cancer data repository. The data repository contains 697 observations and 11 features. The authors compared the performance of the proposed model and the result of performance analysis shows that the naïve bayes algorithms outperformed than the logistic regression algorithm. In [11], breast cancer prediction model is proposed by employing the support vector machine algorithm Wisconsin data repository. The number of observations used in the dataset is 569 and the number of features is 10. The predictive performance of the proposed breast cancer prediction model is evaluated and the accuracy of the algorithm is 90.86%. The accuracy result shows that support vector machine performed well on the prediction of breast cancer.

In [12], a support vector machine and convolutional neural network (CNN) based breast cancer classification model is proposed. In the study, CNN is used for feature extraction and the support vector machine is employed for prediction of the breast cancer. In [13], KNN based breast cancer prediction model is proposed. The dataset consists of 209 observations collected manually by the authors. The predictive performance of the proposed model is accepTable with prediction accuracy of 93%. In another study [14], a decision tree algorithm is applied to Wisconsin breast cancer prognosis dataset and a breast cancer prediction model is proposed.

In [15], the authors compared the accuracy of naïve bayes algorithm with decision tree and support vector machine algorithm on breast cancer data collected from Wisconsin data repository. The dataset consists of 699 observations and among the observations, 458 are malignant and 248 are benign. The result of performance analysis shows that the support vector machine outperformed the KNN and naïve bayes algorithm having a better accuracy score on breast cancer prediction.

In [16], SVM and KNN is applied to Wisconsin breast cancer and a predictive model is proposed using these algorithms. The dataset contains 699 observations and 11 features. The authors compared the performance of the algorithms and result shows the support vector machine as a better algorithm with higher accuracy than the KNN algorithm. Another study [17], employed the Wisconsin breast cancer data repository to analyze the predictive performance of KNN algorithm on prediction of breast cancer. The predictive performance of the proposed KNN based breast cancer prediction model has an average accuracy of 76%.

3. RESEARCH METHOD

In this research, breast cancer dataset collected from the kaggle repository is employed in training and testing the proposed model. In the implementation and experimental testing, Python programming language is employed. A statistical method that is Pearson's correlation analysis and data visualization as well as feature relationship measures are employed for identification and interpretation of breast cancer data repository to discover the relationship between the class and the features in observations. Decision tree and adaptive boosting algorithms are employed for developing the prediction model. The data repository consists of a list observations that belong to malignant (cancerous) and benign (non-cancerous) class. The percentage of the malignant and benign observations in the data repository is demonstrated in Figure 1.

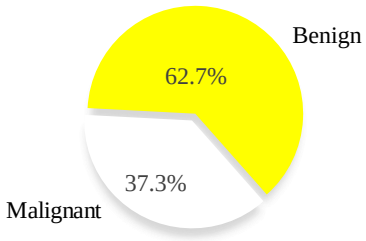


Figure 1. Percentage of malignant and benign observations in the kaggle breast cancer data repository

3.1. Dataset description

The kaggle breast cancer data repository used in this study consists of 569 observations and 31 features. Among a total of the 569 observations and 212 observations are benign or breast cancer negative and 357 are malignant or breast cancer positive. This shows 37.25% of the observation consists of breast cancer negative and 62.74% of the observation is breast cancer positive. The dataset has no missing feature values. The features of the breast cancer data repository are summarized in Table 1. The dataset observations used in training is 75% and in testing 25% of the observations is used.

Table 1. The kaggle cervical cancer data repository features description

Observations	Feature	Description
1	Mean radius	The mean of distances from center to points on the perimeter, integer
2	Mean-texture	Standard deviation of gray-scale values, integer
3	Mean-perimeter	mean size of the core tumor, integer
4	Mean-area	Mean of area, integer
5	Mean-smoothness	the local variation in radius lengths, integer
6	Diagnosis	Class label (1=Malignant, 0=Benign)

The breast cancer dataset features are demonstrated in Figure 2. As demonstrated in Figure 2, the number of malignant observations is more than the benign observations.

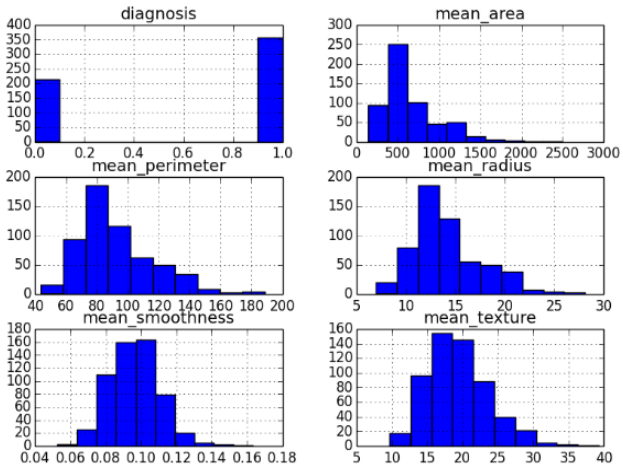


Figure 2. The breast cancer data repository features

3.2. Correlation analysis

We have employed Pearson's correlation analysis for visualization of the relationship between each feature. This helps to identify the feature that is strongly related to the class feature in the data repository. The Pearson's correlation matrix for each features of the breast cancer dataset is shown in Figure 3. As shown in Figure 3 the class is perfectly related to mean radius and mean perimeter features. This shows that breast cancer prediction is highly influenced by those features.

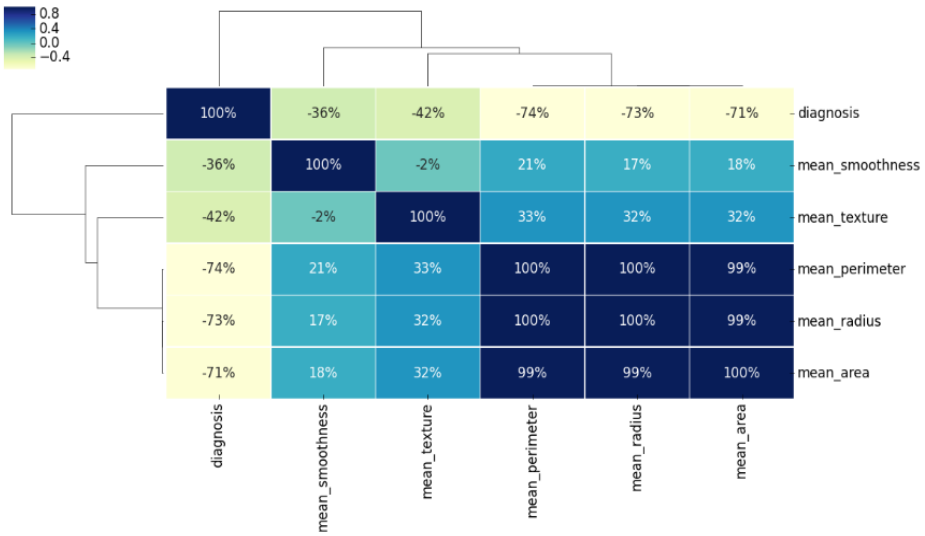


Figure 3. The relationship between breast cancer features

4. RESULTS AND DISCUSSION

In this section, the experimental test results on the proposed model is explained. The predictive performance of decision tree and adaptive boosting algorithm is analyzed by employing the performance metrics such as accuracy and confusion matrix along with learning curve of the algorithms.

4.1. Predictive accuracy analysis

The predictive performance of the proposed model is experimented on the training set. The predictive accuracy of the proposed model is shown in Figure 4. Moreover, the accuracy for decision tree and adaptive boosting for breast cancer classification on random test is given in Table 2.

Table 2. Accuracy of adaptive boosting and decision tree

Learning algorithm	Accuracy in % on experimental test
Adaptive boosting	90.20
	90.90
	96.50
Decision tree	88.81
	87.41
	90.20

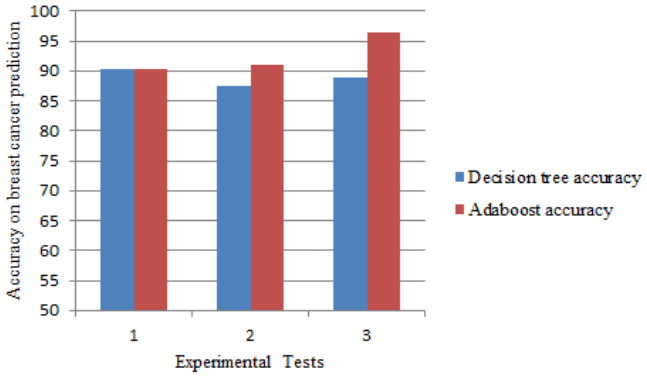


Figure 4. Accuracy of decision tree and adaptive boosting algorithm

4.2. Confusion matrix analysis

A confusion matrix is a measure the predictive performance of the proposed models in terms of the number of correct and incorrect predictions on the test set by the decision tree and adaptive boosting algorithm. The confusion matrix of the decision tree and adaptive boosting algorithm is shown in Figure 5(a) and Figure 5(b) respectively.

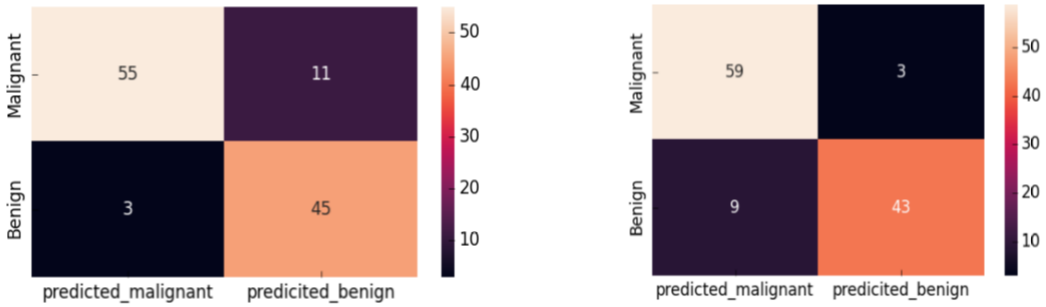


Figure 5. Confusion matrix for the decision tree and adaptive boosting, (a) Decision tree confusion matrix, (b) Adaptive boosting confusion matrix

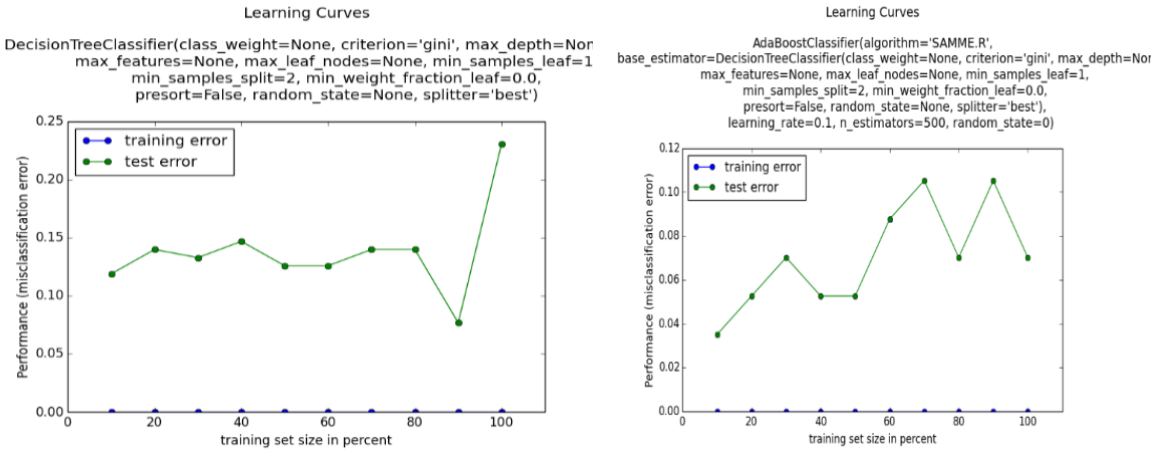
As shown in Figure 5(a) and Figure (b) the accuracy of the adaptive boosting algorithms is better than the accuracy of the decision tree algorithm. The accuracy of the models can be calculated form the confusion matrix using (1).

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN)*100 \quad (1)$$

The accuracy of the decision tree model is calculated as using the (1). Accuracy=(55+45)/(55+45)/(55+45+11+3)*100=87.71%, likewise, the accuracy of the adaptive boosting algorithm is calculated as, Accuracy=(59+43)/(59+43+3+9)*100=89.47%. This result shows that the adaptive boosting algorithm outperformed than the decision tree algorithm.

4.3. Learning curves

Learning curves of the proposed model shows the performance of the model on training set as demonstrated in Figure 6. As demonstrated in Figure 6, the learning curve for the proposed model's testing error is higher for the decision tree model than the adaptive boosting model. The testing error for decision tree model falls in the range 12.5% to 25%, which shows that the accuracy of the model falls in the interval 75% to 87.5%. The testing error for the adaptive boosting algorithm falls in the range 0.03% to 0.11% and this shows that the accuracy of the adaptive boosting algorithm falls in the range 89% to 97%.



(a) (b)

Figure 6. The learning curve for Adaboost and decision tree, (a) Decision tree learning curve, (b) Adboost learning curve

5. CONCLUSION

In this research, we have proposed a breast cancer prediction model with adaptive boosting and decision tree algorithm on breast cancer dataset collected form kaggle data repository. The proposed model solves the problem of biased classification on imbalanced observation by non-ensemble algorithm through ensemble classifier namely the adaptive boosting. The predictive performance of the proposed model is evaluated by employing different performance metrics such as accuracy and confusion matrix on the test set. The result of performance analysis reveals that the adaptive boosting algorithm has better performance than the decision tree. Hence, the adaptive boosting algorithm is a better classifier for imbalanced dataset where the use of non-ensemble algorithm such as decision tree, results in biased prediction towards the majority class yielding better performance on prediction of the majority class and poor performance on the minority class.

REFERENCES

[1] R. Chand, D. K. Rao, T. B. Tekabu and M.G.M Khan, "Modeling Breast Cancer Cases in Fiji," *Asia-Pacific World Congress on Computer Science and Engineering*, 2018.

[2] Mohammed Abdulrazaq Kahya, "Classification enhancement of breast cancer histopathological image using penalized logistic regression," *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 13, No. 1, 2019.

[3] Vikas Chaurasia, Saurabh Pal, BB Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," *Journal of Algorithms & Computational Technology*, 2018.

[4] May Phu Paing, C. Pintavirooj, Kazuhiko Hamamoto, "Comparison of Sampling Methods for Imbalanced Data Classification in Random Forest," *Biomedical Engineering International Conference, IEEE*, 2018.

[5] Dongdong Sun, Minghui Wang, Huanqing Feng, Ao Li, "Prognosis Prediction of Human Breast Cancer by integrating Deep Neural Network and Support Vector Machine, Supervised Feature Extraction and Classification for Breast Cancer Prognosis Prediction," *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, 2017.

[6] Ebru Aydındag Bayrak, Pınar Kırcı, Tolga Ensari, "Comparison of Machine Learning Methods for Breast Cancer Diagnosis," *IEEE*, 2019.

[7] S. Murugan, Muthu Kumar, S. Amudha, "Classification and Prediction of Breast Cancer using Linear Regression," *Decision Tree and Random Forest, International Conference on Current Trends in Computer, Electrical, Electronics and Communication*, 2017.

[8] Shahrbanoo Goli, Hossein Mahjub, Javad Faradmal, Hoda Mashayekhi, Ali-Reza Soltanian, "Survival Prediction and Feature Selection in Patients with Breast Cancer Using Support Vector Regression," *Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine*, 2016. https://doi.org/10.1155/2016/2157984.
[9] P. Sathiyanarayanan, Pavithra, Sai Saranya.M, Makeswari.M, "Identification of Breast Cancer Using The Decision Tree Algorithm," *Proceedings of international conference on systems computation automation and networking, IEEE*, 2019. DOI: 10.1109/ICSCAN.2019.8878757.
[10] Ahmet Mert, Niyazi KJlJç, Erdem Bilgili, Aydin Akan, "Breast Cancer Detection with Reduced Feature Set," *Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine*, Volume 2015.
[11] Abdullah-Al Nahid, Yinan Kong, "Involvement of Machine Learning for Breast Cancer Image Classification: A Survey," *Hindawi Computational and Mathematical Methods in Medicine*, Volume 2017.
[12] Mohd Rasoul Al-hadidi, Abdulsalam Alarabeyyat, Mohannad Alhanahna, "Breast Cancer Detection using K-nearest Neighbor Machine Learning Algorithm," *International Conference on Developments in eSystems Engineering, IEEE*, 2017. DOI: 10.1109/DeSE.2016.8.
[13] Shabina Sayed, Shoeb Ahmed, Rakesh Poonia, "Holo Entropy Enabled Decision Tree Classifier For Breast Cancer Diagnosis Using Wisconsin (Prognostic) Data Set," *International Conference on Communication Systems and Network Technologies, IEEE*, 2017.
[14] Wan Nor Liyana Wan Hassan Ibeni, *et al*., "Comparative analysis on Bayesian classification for breast cancer problem," *Bulletin of Electrical Engineering and Informatics (BEEI)*, Vol. 8, No. 4, 2019. DOI: https://doi.org/10.11591/eei.v8i4.1628.
[15] Md. Milon Islam, Hasib Iqbal, Md. Rezwanul Haque, Md. Kamrul Hasan, "Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors," *Humanitarian Technology Conference, IEEE*, 2017.
[16] Alberto Palacios Pawlovsky, Mai Nagahashi, "A Method to Select a Good Setting for the kNN Algorithm when Using it for Breast Cancer Prognosis," *IEEE*, 2014. DOI: 10.1109/BHI.2014.6864336.
[17] Assegie, Tsehay Admassu, Sushma S J., Prasanna Kumar S C, "Weighted Decision Tree Model for Breast Cancer Detection," Technology Reports of Kansai University, Volume 62, Issue 03, 2020.
[18] Assegie Tsehay Admassu, Sushma S. J., "A Support Vector Machine and Decision Tree Based Breast Cancer Prediction," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 9, no. 3, pp. 2972-2976, 2020, DOI: 10.35940/ijeat.A1752.029320.
[19] Assegie Tsehay Admassu, "An optimized K-Nearest Neighbor based breast cancer detection," *Journal of Robotics and Control (JRC)*, vol. 2, no. 3, pp. 115-118, 2020, DOI: 10.18196/jrc.2363.
[20] Sushma S. J, S. C. Prasanna Kumar, A novel approach to jointly address localization and classification of breast cancer using bio-inspired approach, *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 2, pp. 992-1001, 2019, DOI: 10.11591/ijece.v9i2.pp992-1001.
[21] Amandeep Kaur, Prabhjeet Kaur, "Breast Cancer Detection and Classification using Analysis and Gene-Back Proportional Neural Network Algorithm," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 8, pp. 2789-2803, 2019, available: https://www.ijitee.org/wp-content/uploads/papers/v8i8/H6992068819.pdf.
[22] P. Suryachandra and P. V. S. Reddy, "Comparison of machine learning algorithms for breast cancer," *2016 International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, pp. 1-6, 2016, doi: 10.1109/INVENTIVE.2016.7830090.
[23] Achmad Ridok, Nashi Widodo, Wayan Firdaus Mahmudy, Muhaimin Rifa, "A hybrid feature selection on AIRS method for identifying breast cancer diseases," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 1, pp. 728, 735, 2021, DOI: 10.11591/ijece.v11i1.pp728-735.
[24] Mohammed Y. Kamil, "Computer-aided diagnosis system for breast cancer based on the Gabor filter technique," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 5, pp. 5235-5242, 2020, DOI: 10.11591/ijece.v10i5.pp5235-5242.
[25] Susama Bagchi, Kim Gaik Tay, Audrey Huong, Sanjoy Kumar Debnath, "Image processing and machine learning techniques used in computer-aided detection system for mammogram screening-A review," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 3, pp. 2336-2348, 2020, DOI: 10.11591/ijece.v10i5.pp5235-5242.