Chemotherapy: Open Access
ISSN: 2167-7700
OPEN ACCESS Freely available online
Research Article

Use of Machine Learning to Diagnose Benign and Malignant Breast Tissues with the Best Degree of Accuracy and in the Shortest Amount of Time

Seyed Matin Malakouti*

Department Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran

ABSTRACT

A tumor is an abnormal lump or growth of cells. Sometimes a tumor is made up of cells that aren't a threat to invade other tissues, this is considered benign. When the cells are abnormal and can grow uncontrollably and spread to other body parts, they are cancerous cells that mean the tumor is malignant. This spreading process is called metastasis. If the cells are not cancerous, the tumor is benign. A benign tumor is less problematic. Doctors may need to remove benign tumors through surgery. These tumors can grow very large, sometimes weighing pounds. They can also be dangerous. They can press on vital organs or block channels. Some benign tumors, such as intestinal polyps, are considered precancerous. They are removed to prevent them from becoming malignant. Benign tumors usually don't come back once removed. But if they do, they return to the same place. In this research, using the features recorded for tumor tissues such as mean radius and mean texture and other available features that are fully mentioned in the article, with the help of Light Gradient Boosting Machine (LGBM), random forest, extra tree, ada boost, and the ensemble method classified benign and malignant tumors with perfect accuracy.

Keywords: Tumor; Malignant; Benign; Surgery; Ensemble method

INTRODUCTION

Breast cancer is a kind of cancer that develops from a malignant tumor in the breast tissue's cells. A malignant tumor is a mass of cancer cells that may invade nearby tissues or travel throughout the body [1]. Breast cancer is an unchecked cell proliferation in the breast tissue. A lump or structural deformities may emerge when the number of cells divides quickly. Breast cancer, which occurs immediately after lung cancer, is women's second most common cause of mortality. Breast cancer is a fatal condition, and early identification may undoubtedly lower the fatality rate. The survival rate is 88 percent after five years of treatment and 80 percent after ten days of therapy, according to an examination of the most current data [2]. In addition to 60,290 new instances of non-invasive (in situ) breast cancer, there are expected to be 231,840 cases reported of invasive breast cancer among women in the United States in 2015 [3]. Other than lung cancer, breast cancer has the highest mortality rate for women in the United States. With 25% of all occurrences, breast cancer is the most common kind of cancer among women globally. It is more prevalent in industrialized nations and affects women almost 100 times more often than it does males [4]. The kind of breast cancer, the severity of the condition, and the patient's age all affect the outcome [5]. The industrialized world has reasonable survival rates, with 80% to 90% of english people in United states are alive for at least five years [6,7]. Males are more likely than females to get lung cancer, which accounts for 23% of all cancer fatalities and 17% of all new cancer cases. Among financially developing nations, breast cancer is currently one of women's leading causes of cancer-related deaths [8]. There has been a change from the previous ten years when cervical cancer was the leading cause of cancer-related mortality.

However, lung and cervical cancer both account for 11% of all female cancer fatalities in emerging nations, placing them at comparable levels in terms of mortality load. The overall cancer death rates are typically similar, even though overall cancer incidence rates in the developing world are half as high as those in the developed world for both sexes. Each year, 4500 new instances of breast cancer are identified in Portugal, and it is expected that 1600 women will pass away from the illness [9]. Mammography is the most efficient tool for detecting early breast cancer [10]. The creation of a classifier is a critical phase in the design of a Computer Aided Design (CAD)

Correspondence to: Seyed Matin Malakouti, Department Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran, E-mail: m.malakoti98@ms.tabrizu.ac.ir

Received: 01-Nov-2022, Manuscript No. CMT-22-19874; **Editor assigned:** 03-Nov-2022, Pre QC No. CMT-22-19874 (PQ); **Reviewed:** 17-Nov-2022, QC No. CMT-22-19874; **Revised:** 24-Nov-2022, Manuscript No CMT-22-19874 (R); **Published:** 02-Dec-2022, DOI: 10.35248/ 2167-7700.22.10.166.

Citation: Malakouti SM (2022) Using Machine Learning to Diagnose Benign and Malignant Breast Tissues with the Best Degree of Accuracy and in the Shortest Amount of Time. Chemo Open Access.10:166.

system. A classifier must be able to combine the supplied input feature data and provide an accurate assessment. Back Propagation Neural (BPN) networks and Linear Discriminants (LDA) are two popular classifiers for CAD that is effective in lesion classification tasks [11-28].

MATERIALS AND METHODS

Research algorithms

Random forest classifier: Ensemble learning methods (such as random forest, bagging, and boosting) are gaining popularity as they're more reliable and resistant to noise than single classifiers. Classifier ensembles are built on the fundamental idea that a group of classifiers performs better classifications than a single classifier. Breiman [29] proposed a novel and effective classifier called a random forest, which has numerous benefits for remote sensing applications includes:

a) It performs well on substantial data sets.

b) It can manage hundreds of inputs without deleting any of them.

c) It assesses which factors are relevant in categorization.

Ada boost classifier: The ada boost method produces robust classifiers from poor ones. The ada boost ensemble classifier includes the weak classifiers as a member. Ada boost develops a committee of member weak classifiers by adaptively modifying the weights at each cycle. Training samples with incorrect classifications by a current weak classifier have their consequences raised, while training samples with accurate classifications by a current weak classifier have their weights dropped. Ada boost is an excellent approach for creating ensemble classifiers, although it doesn't always produce classifiers with the most minor generalization average error. The ada boost algorithm performs well due to its capacity to provide growing variety. It comprises a variety of weak classifiers to enhance the performance of the final ensemble. The ada boost method is used by Viola and Jones [30] to choose a small number of crucial visual characteristics from a vast pool of alternative features. Ada boost offers a strong constraint on generalization performance and an efficient learning technique. They employed the Ada boost method to find a limited number of high-quality features with substantial variation. The Ada boost approach improves performance by limiting the weak learner to a collection of classification functions that each rely on a single element. They use single threshold features, weak classifiers that may be compared to single node decision trees.

Light gradient boosting machine classifier: Think again if you believed XG-boost to be the most significant algorithm available. Another boosting algorithm known as Light Glioblastoma Multiforme (GBM) has shown to be quicker and sometimes more accurate than XG-boost. Gradient-based One-Side Sampling (GOSS) is a particular method that Light GBM employs to filter out the data instances and determine a split value. This is distinct from XG-boost, which determines the optimal split using pre-sorted and histogram-based methods.

Extra tree classifier: Similar to Random forest classifier, it is a form of ensemble learning approach that combines the outcomes of many de-correlated decision trees. The extra tree often performs as well as or better than the random forest. Between random forest and extra tree classifier, the following is the main distinction in contrast to the random forest; extra tree classifier does not use bootstrap aggregation. To put it simply, it takes a random selection of data without replacing it. So, nodes are divided at random rather than using the optimal splits. Randomness in the extra tree classifier thus derives from the data's random splits rather than bootstrap aggregation.

Ensemble: Using a variety of modeling techniques or training data sets, ensemble modeling is the process of building numerous varied models to predict a result. The ensemble model then combines each base model's forecast into a single overall prediction for the unobserved data.

Data preparation: Breast tissues [31] of 569 women were examined. The tissues had different characteristics, which determined whether breast cancer was benign or malignant based on the characteristics of the examined tissues. 'mean radius', 'mean texture', 'mean perimeter', 'mean area', 'mean smoothness', 'mean compactness', 'mean concavity', 'mean concave points', 'mean symmetry', 'mean fractal dimension', 'radius error', 'texture error', 'perimeter error', 'area error', 'smoothness error', 'compactness error', 'concavity error', 'concave points error', 'symmetry error', 'fractal dimension error', 'worst radius', 'worst texture', 'worst perimeter', 'worst area', 'worst smoothness', 'worst compactness', 'worst concavity', 'worst concave points', 'worst symmetry', 'worst fractal dimension', 'Benign and malignant tissue, benign tissue is marked with 0 and malignant tissue with 1'

Evolution process: A resampling technique called 10-fold cross-validation [32,33] evaluates and trains a model utilizing a variety of input bits throughout many rounds. Use it to test how well a forecasting model will perform in the actual world. The remaining 455 people are employed as training sets, leaving just 114 for validation and testing.

RESULTS AND DISCUSSION

Classification report

A classification report is used to evaluate the algorithm's accuracy in making predictions. Which of the predictions were accurate and which were incorrect? True-Positive (TP), False-Positive (FP), True-Negative (TN), and False-Negative (FN) are all measures that may be used to predict the outcome of a test.

Precision, F1, and recall need an understanding of TP, FP, TN, and FN by giving a simple example and we explain what they are:

- True-Positive (TP): People that are sick are accurately labeled as such.
- False-Positive (FP): People who are well are mistakenly labeled as sick.
- True-Negative (TN): Ones in good health are rightly labeled "healthy."
- False-Negative (FN): Those who are unwell are mistaken for those who are healthy.

We now define precision, F1, and recall as follows:

$$Precision=TP/(TP+FP)$$ (Equation 1)

$$Recall=TP/(TP+FN)$$ (Equation 2)

$$F1=2*\frac{(Precision*Recall)}{(Precision+Recall)}$$ (Equation 3)

A total of 569 tissues were used, 445 were used to train the algorithms, and 114 were tested and validated. Out of these

114 tissues, 71 were malignant, and 43 were benign. Out of 71 malignant tissues, 70 were diagnosed correctly. Out of 43 benign tissues, 42 were diagnosed correctly (Figure 1).

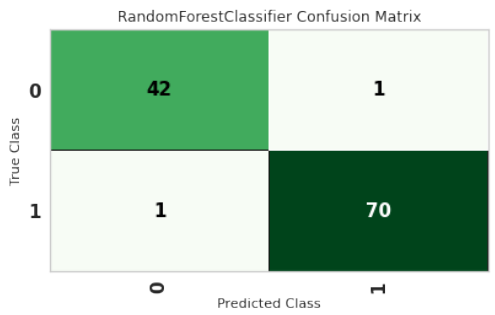

Figure 1: Confusion matrix for random forest classifier for predicting malignant and benign tissue.

The recall is more critical if missed instances (False-Negatives) cost more than false alarms (False-Positive). The fundamental goal is to identify answers to these problems. When false positives (false alarms) cost more than missed instances, precision becomes more critical (False-Negatives). Figure 2 displays the classification report for random forest classifier for predicting malignant and benign tissue. The precision and F1 evaluation criteria are less critical in medical issues than the recall assessment criterion. The diagnosis of malignant is 98.6% accurate and benign 97.7% correct for tissues (Figure 2).

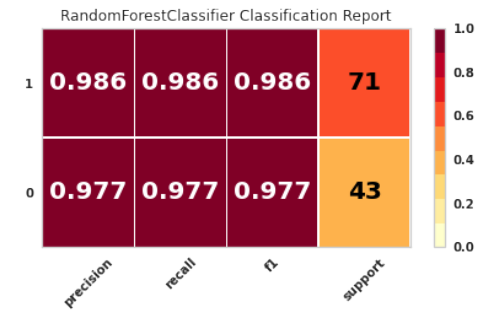

Figure 2: Classification report for random forest classifier for predicting malignant and benign tissue.

A total of 569 tissues were used, 445 were used to train the algorithms, and 114 were tested and validated. Out of these 114 tissues, 71 were malignant, and 43 were benign. 71 malignant tissues, 71 were diagnosed correctly. Out of 43 benign tissues, 40 were diagnosed correctly (Figure 3).

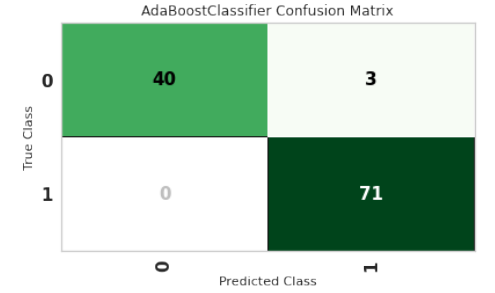

Figure 3: Confusion matrix for ada boost classifier for predicting malignant and benign tissue.

Figure 4 displays the Classification Report for Ada Boost Classifier for predicting malignant and benign tissue. The Precision and F1 evaluation criteria are less critical in medical issues than the recall assessment criterion. The diagnosis of malignant is 100% accurate and benign 93% correct for tissues (Figure 4).

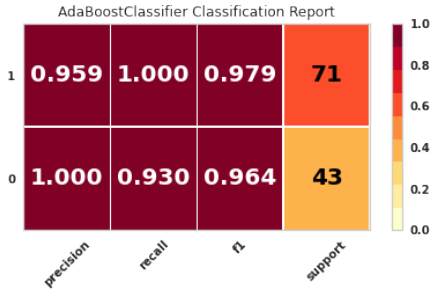

Figure 4: Classification report for ada boost classifier for predicting malignant and benign tissue.

A total of 569 tissues were used, 445 were used to train the algorithms, and 114 were tested and validated. Out of these 114 tissues, 71 were malignant, and 43 were benign. Out of the 71 malignant tissues, 69 were diagnosed correctly (Figure 5). Out of 43 benign tissues, 41 were diagnosed correctly.

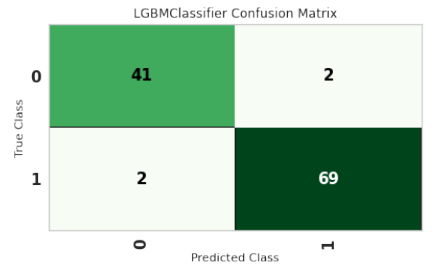

Figure 5: Confusion matrix for light-GBM classifier for predicting malignant and benign tissue.

Figure 6 displays the classification report for light GBM classifier for predicting malignant and benign tissue. The Precision and F1 evaluation criteria are less critical in medical issues than the recall assessment criterion. The diagnosis of malignant is 97.2% accurate and benign 95.3% correct for tissues (Figure 6).

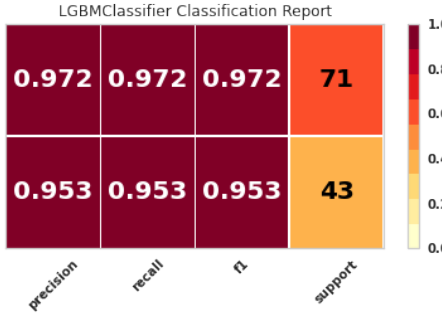

Figure 6: Classification report for light-GBM classifier for predicting malignant and benign tissue.

A total of 569 tissues were used, 445 were used to train the algorithms, and 114 were tested and validated. Out of these 114 tissues, 71 were malignant, and 43 were benign. Out of the 71 malignant tissues, 69 were diagnosed correctly. Out of 43 benign

tissues, 42 were diagnosed correctly (Figure 7).

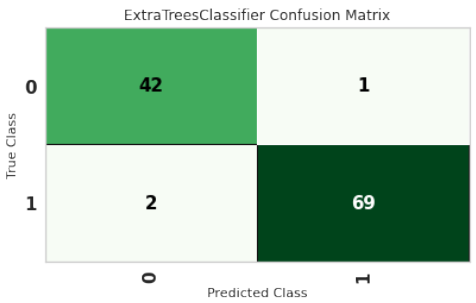

Figure 7: Confusion matrix for extra tree classifier for predicting malignant and benign tissue.

Figure 8 displays the classification report for extra tree classifier for predicting malignant and benign tissue. The Precision and F1 evaluation criteria are less critical in medical issues than the recall assessment criterion. The diagnosis of malignant is 97.2% accurate and benign 97.7% correct for tissues (Figure 8).

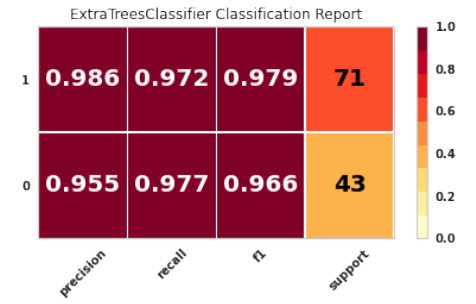

Figure 8: Classification report for extra tree classifier for predicting malignant and benign tissue.

A total of 569 tissues were used, 445 were used to train the algorithms, and 114 were tested and validated. Out of these 114 tissues, 71 were malignant, and 43 were benign. 71 malignant tissues, 71 were diagnosed correctly. Out of 43 benign tissues, 42 were diagnosed correctly. Only our proposed method could accurately detect all malignant tissues. This is an outstanding achievement in medical science, and human lives can be saved in a good way (Figure 9).

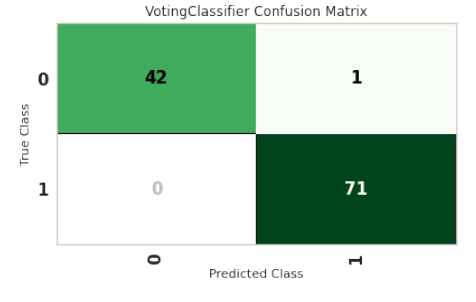

Figure 9: Confusion matrix for ensemble (Lightgbm,rf,et,ada) classifier for predicting malignant and benign tissue.

Figure 10 displays classification report for ensemble (Light GBM, rf, et, ada) classifier for predicting malignant and benign tissue. The precision and F1 evaluation criteria are less important in medical issues than the recall assessment criterion. The diagnosis of malignant is 100% accurate and benign 97.7% accurate for tissues (Figure 10).

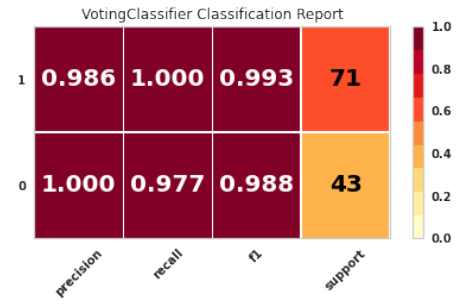

Figure 10: Classification report for ensemble (Lightgbm,rf,et,ada) Classifier for predicting malignant and benign tissue.

Figure 11 shows that the algorithm proposed by Ensemble (Light GBM, rf, et, ada) is able to detect benign and malignant breast tissue within 35 seconds. Also, Figures 10 and 11 show the high accuracy of the proposed method Ensemble (Light GBM, rf, et, ada) in the diagnosis of benign and malignant breast tissue (Figure 11).

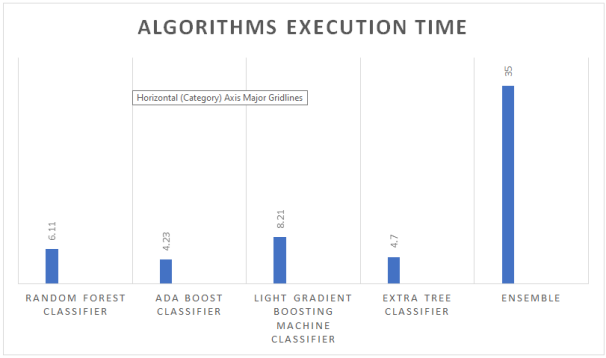

Figure 11: Algorithm execution time for predicting malignant and benign tissue.

CONCLUSION

This research looked back at 569 pieces of cancerous and healthy tissue. All cancerous and healthy tissue was predicted using algorithms called Extra tree, random forest classifier, ada boost classifier, and light gradient boosting machine, ensemble (Light GBM, rf, et, ada). Ensemble (Light GBM, rf, et, ada) was able to correctly identify 71 out of 71 cancerous tissues. The Ensemble (Light GBM, rf, et, ada) algorithm was the only one able to diagnose all malignant tumors correctly. In medical science, it's essential to figure out what's wrong with a sick person as a patient. If a healthy person is told they are sick, it doesn't matter as much as if a sick person is told they are healthy. Ensemble (Light GBM, rf, et, ada) predict a healthy person for a patient. This is the algorithm's weakness, and neither machine learning algorithms nor the best doctors have ever been able to diagnose diseases with 100% accuracy.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin. 2015;65(1):5-29.

2. Heymach J, Krilov L, Alberg A, Baxter N, Chang SM, Corcoran RB, et al. Clinical cancer advances 2018: annual report on progress against cancer from the American Society of Clinical Oncology. J Clin Oncol. 2018;36(10):1020-1044.

3. Breastcancer.org - Breast Cancer Information and Support.

4. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, et al. Cancer statistics in China, 2015. CA Cancer J Clin. 2016;66(2):115-132.

5. PDQ Adult Treatment Editorial Board. Breast cancer treatment (PDQ®): health professional version. PDQ cancer information summaries. Bethesda (M.D.): National Cancer Institute (U.S.); 2002.

6. Mustafa M, Nornazirah A, Salih F, Illzam E, Suleiman M, Sharifa A. Breast cancer: detection markers, prognosis, and prevention. IOSR-JDM. 2016;15(08):73-80.

7. B.P., L.B. World cancer report. 2008

8. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011;61(2):69-90.

9. Pérez NP. Improving Variable Selection and Mammography-based Machine Learning Classifiers for Breast Cancer CADx. Doctoral dissertation, Universidade do Porto (Portugal).185-185.

10. Zuckerman HC. The role of mammography in the diagnosis of breast cancer. Breast cancer, diagnosis and treatment. 1987:152-72.

11. Lachenbruch PA. Discriminant Analysis, New York: Hafner. Lachenbruch Discriminant Analysis 1975. 1975.

12. Duda RO, Hart PE. Pattern classification. John Wiley & Sons. 2006.

13. Werbos P. Beyond regression: new tools for prediction and analysis in the behavioral sciences. Ph. D. dissertation, Harvard University. 1974.

14. Rumelhart D, Hinton GE, Williams RJ, Rumelhart DE. Parallel and Distributed Processing. Cambridge, MA: MIT Press. 1986; 1:318-318.

15. Hertz J, Krogh A, Palmer RG. Introduction to the theory of neural computation. CRC Press; 2018.

16. Chan HP, Wei D, Helvie MA, Sahiner B, Adler DD, Goodsitt MM, et al. Computer-aided classification of mammographic masses and normal tissue: linear discriminant analysis in texture feature space. Phys Med Biol. 1995;40(5):857-857.

17. Wei D, Chan HP, Helvie MA, Sahiner B, Petrick N, Adler DD, et al. Classification of mass and normal breast tissue on digital mammograms: multiresolution texture analysis. Med Phys. 1995;22(9):1501-1513.

18. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM. Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis. Med Phys. 1998;25(4):516-526.

19. Sahiner B, Chan HP, Wei D, Petrick N, Helvie MA, Adler DD, et al. Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue. Med Phys. 1996;23(10):1671-1684.

20. Chan HP, Sahiner B, Petrick N, Helvie MA, Lam KL, Adler DD, et al. Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network. Phys Med Biol. 1997;42(3):549-567.

21. Rangayyan RM, El-Faramawy NM, Desautels JL, Alim OA. Measures of acutance and shape for classification of breast tumors. IEEE Trans Med Imaging. 1997;16(6):799-810.

22. Fogel DB, Wasson EC, Boughton EM, Porto VW, Angeline PJ. Linear and neural models for classifying breast masses. IEEE Trans Med Imaging. 1998;17(3):485-488.

23. Highnam RP, Brady JM, Shepstone BJ. A quantitative feature to aid diagnosis in mammography. Proc Digital Mammography'96. 1996;17:201-206.

24. Wu Y, Giger ML, Doi K, Vyborny CJ, Schmidt RA, Metz CE. Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. Radiology. 1993;187(1):81-87.

25. Goldberg V, Manduca A, Ewert DL, Gisvold JJ, Greenleaf JF. Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence. Med Phys. 1992;19(6):1475-1481.

26. Kilday J, Palmieri F, Fox MD. Classifying mammographic lesions using computerized image analysis. IEEE Trans Med Imaging. 1993;12(4):664-669.

27. McNitt-Gray MF, Huang HK, Sayre JW. Feature selection in the pattern classification problem of digital chest radiograph segmentation. IEEE Trans Med Imaging. 1995;14(3):537-547.

28. Huo Z, Giger ML, Vyborny CJ, Wolverton DE, Schmidt RA, Doi K. Automated computerized classification of malignant and benign masses on digitized mammograms. Acad Radiol. 1998;5(3):155-168.

29. Breiman L. Random forests. Machine learning. 2001;45(1):5-32. [Google Scholar]

30. Viola P, Jones MJ. Robust real-time face detection. Int J Comput. Vis. 2004;57(2):137-154.

31. Nyandw JD. Breast Cancer Dataset UCI ML. Kaggle.

32. Malakouti SM, Ghiasi AR, Ghavifekr AA, Emami P. Predicting wind power generation using machine learning and CNN-LSTM approaches. Wind Eng. 2022;46(6).

33. Malakouti SM, Ghiasi AR. Evaluation of the application of computational model machine learning methods to simulate wind speed in predicting the production capacity of the Swiss basel wind farm. IEEE. 2022;31-36.