Frontiers in Health Informatics
2023; 12: 132
www.ijmi.ir
Open Access

Improvement of the performance of machine learning algorithms in predicting breast cancer

Maryam Poornajaf[1], Sajad Yousefi[2]*

[1]Faculty Member, Department of Computer Engineering, Technical and Vocational University (TVU), Tehran, Iran
[2]Faculty Member, Department of Electrical Engineering, Technical and Vocational University (TVU), Tehran, Iran

Article Info

Article type:
Research

Article History:
Received: 2022-12-15
Accepted: 2023-01-18
Published: 2023-03-18

**** Corresponding author:***
Sajad Yousefi

Department of Electrical Engineering, Technical and Vocational University (TVU), Tehran, Iran

Email: syosefi_1980@yahoo.com

Keywords:
Machine Learning
Dataset
Importance Score
Accuracy
Breast Cancer
Logistic Regression

A B S T R A C T

Introduction: Breast cancer is one of the most common cancers among women compared to all other ones. Machine learning (ML) techniques can bring a large contribute on the process of prediction and early diagnosis of breast cancer, became a research hotspot and has been proved as a strong technique. Using ML models performed on multidimensional dataset, this article aims to find the most efficient and accurate ML models for tumor classification prediction.

Material and Methods: Several supervised ML algorithms were utilized to diagnosis and prediction of cancer tumor such as Logistic Regression Decision Tree, Random Forest and KNN. The algorithms are applied to a dataset taken from the UCI repository including 699 samples. The dataset includes Breast cancer features. To enhance the algorithms' performance, these features are analyzed, the feature importance score and cross validation are considered. In this research, ML algorithms improved coupled by limited and selective features to produce high classification accuracy in tumor classification.

Results: As a result of evaluation, Logistic Regression algorithm with accuracy value equal to 99.14%, AUC ROC equal to 99.6%, Extra Tree algorithm with accuracy value equal to 99.14% and AUC ROC equal to 99.1% have better performance than other algorithms. Therefore, these techniques can be useful for diagnosis and prediction of cancer tumor and prescribe it correctly.

Conclusion: The technique of ML can be used in medicine for analyzing the related data collections to a disease and its prediction. The area under the ROC curve and evaluating criteria related to a number of classifying algorithms of ML to evaluate breast cancer and indeed, the diagnosis and prediction of breast cancer is compared to determine the most appropriate classifier.

Cite this paper as:
Poornajaf M, Yousefi S. Improvement of the performance of machine learning algorithms in predicting breast cancer. Front Health Inform. 2023; 12: 132. DOI: 10.30699/fhi.v12i0.400

INTRODUCTION

Breast cancer is one of the most common types of cancer in women worldwide. Breast cancer is associated with a high fatality rate. Breast cancer affects more than 1.5 million women worldwide each year, according to the World Health Organization [1]. Tumors can be used to detect breast malignancy. Tumors are classified as either malignant or benign. To detect malignant cancers, doctors need to use an active determination approach. However, even for specialists, identifying malignancies is extremely difficult [2]. As a result, in order to detect cancer, an automatic approach is needed.

Diagnostic mammography can assess abnormal breast cancer tissue in patients with subtle and inconspicuous malignancy signs. Due to a large number of images, this method cannot effectively be used in assessing cancer suspected areas. According to a report, approximately 50% of breast cancers were not detected in screenings of women with very dense breast tissue. However, about a quarter of women with breast cancer are diagnosed negatively within two years of screening [3, 4]. Therefore, the early and timely diagnosis of breast cancer is crucial.

Most mammography-based breast cancer screening is performed at regular intervals - usually annually or every two years - for all women. This "A fix screening program for everyone" is not effective in diagnosing cancer at the individual level and may impair the effectiveness of screening programs. On the other hand, experts suggest that considering other risk factors along with mammography screening can help a more accurate diagnosis of women at risk. Moreover, effective risk prediction through modeling can not only help radiologists in setting up a personal screening for patients and encouraging them to participate in the program for early detection but also help identify high-risk patients [5].

Data mining algorithms applied in healthcare industry play a significant role due to their high performance in predicting, diagnosis of the diseases, reducing costs of medicine, making real time decision to save people's lives. The most Common data mining modeling goals are classification and prediction, which uses several algorithms for the prediction of breast cancer.

Machine learning (ML), as a modeling approach, represents the process of extracting knowledge from data and discovering hidden relationships, widely used in healthcare in recent years to predict different diseases [6-8]. Some studies only used demographic risk factors (lifestyle and laboratory data) in predicting breast cancer, and several studies predicted based on mammographic stereotypes or used data from patient biopsy. Others showed the application of genetic data in predicting breast cancer [5].

Many studies have attempted to use ML approaches to determine the survivability of carcinoma in people, and they have shown that these algorithms are more effective in diagnosing carcinoma diagnosis [9, 10]. ML is well known for its use in the categorization and modeling of breast cancer. A large number of ML algorithms are available for prediction and diagnosis of breast cancer. The current study aimed to predict breast cancer using different ML approaches considering various factors in modeling.

The objective is to predict and diagnosis breast cancer, using ML algorithms, and find out the most effective based on the performance of each classifier in terms of confusion matrix, accuracy, precision and sensitivity. This paper is organized as follows: introduces methods and results of previous research on breast cancer diagnosis, describes the proposed methodology for their work, presents and explains in detail the experiments results. All the work is done in the Anaconda environment based on python programming language and Scikit-learn library.

MATERIAL AND METHODS

Classification

Classification is a method of ML and is used to learn how to assign a class tag to an input instance. For example, classification can determine whether a person is ill. Class labels here are malignant and benign that must be converted to numeric values. Two classes are considered: class zero (benign) and class one (malignant). Classification is actually a predictive issue that predicts class labels. The data set records under analysis are divided into two categories: Training Set and Test Set. The individual records that make up the test dataset are randomly sampled from the r set under analysis. Test data set records are independent of training records.

Performance Measures

It used confusion matrix, accuracy, precision, sensitivity, F1 score, area under the curve (AUC) as performance metrics to evaluate and compare the models and identify the best algorithm for the breast cancer prediction. A confusion matrix is a table with two dimensions as "Actual" and "Predicted" and furthermore, both the dimensions have true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Accuracy is most common performance metric for classification algorithms. It defined as the number of correct predictions made as a ratio of all predictions made. Precision, used in document retrievals, defined as the number of correct documents returned by our ML model. Sensitivity defined as the number of positives returned by your ML model. F1 score gives us the harmonic mean of precision and sensitivity. Mathematically, F1 score is the weighted average of the precision and sensitivity [11].

Receiver operating characteristic (ROC) is a graphical way to show how good the performance of a classifier is. It is a plot of a true positive rate against a false positive rate.

That is, the number of correct predictions divided by the number of actual positive results, and the rate of positive predictions is calculated. FPR, on the other hand, indicates the number of positive identifications among negative observations. This ratio is also used as a false positive rate in the ROC. AUC is used as a criterion for evaluating the performance of a classifier. Therefore, the closer the area under the graph to the number one, the better the classifier performance [12].

Dataset

The Wisconsin Breast Cancer Diagnostic (WBCD) dataset from the UCI repository is used in this paper. This dataset was obtained from the University of Wisconsin's hospitals, Madison from Dr. William H. Walberg, Frank and Asuncion [13]. The data set worked on is related to the information of 699 people

(sample) which includes a series of characteristics such as clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, which each characteristic contains number from 1-10 and 2 classes (malignant or benign). Since the class variable is a categorical response variable, it assumes 0 for benign and 1 for malignant. Table 1 lists the range of values for each feature. All the work is done in the Anaconda environment based on python programming language and Scikit-learn library.

Table 1: Details of features

SN	Attribute name	Domain
1	Id Number	Int
2	Clump Thickness	1-10
3	Uniformity of Cell Size	1-10
4	Uniformity of Cell Shape	1-10
5	Marginal Adhesion	1-10
6	Single Epithelial Cell Size	1-10
7	Bare Nuclei	1-10
8	Bland Chromatin	1-10
9	Normal Nucleoli	1-10
10	Mitoses	1-10
11	Class	0 for benign, 1 for malignant

Data Pre-Processing

In order to build up a more accurate ML model, data preprocessing is required. Data pre-processing is the process of cleaning the data. It will remove all the NAN values from our data. This process is also known as Data Wrangling. This includes the identification of missing data, noisy data and inconsistent data.

K-fold cross-validation

The approach of K-fold cross-validation is used to train and test the model. In this approach, the data set is divided into a number of groups. K refers to the number of groups, also known as 'fold'. At the same time, cross-validation is an approach to evaluate a ML model. K-fold cross-validation is such a technique, where the data set is split into k number of groups and the model is trained by (k-1) groups and the other group participates to test or evaluate the trained model. In this approach, the model is trained k number of times and each time, different fold participates to evaluate the model. It indicates that each fold participates to train and test a model in K-fold cross-validation [14, 15].

RESULTS

Fig 1 shows the dependency values between all attributes in the dataset. Lower values indicate a low dependence and upper values indicate a high dependence. According to Fig 1, the properties of uniformity of cell size, uniformity of cell shape and bare nuclei have a high relationship with target variable. For example, the dependency coefficient of target and uniformity of cell size is equal to 0.82, the dependence coefficient of target and uniformity of cell shape is equal to 0.82 and the dependence coefficient of target and bare nuclei is equal to 0.82. In the data set used, the Mitoses property have the least amount of dependence on the target variable. The dependence coefficient of target and mitoses is equal to 0.42.

In this study after identifying the features with high correlation, the data was splitting into two portions for train and for testing purposes. Feature Selection to decrease the number of features in the classifiers and avoid overfitting, coefficients were used as a feature ranking method on the training set. using the classification algorithms in the test phase, the value of the target variable is predicted according to other features. After obtaining and predicting the variable target, this predicted value is compared with its actual value in the test set and the degree of proximity of the prediction values to the real values in the test set is calculated. model validation was conducted through k-fold cross-validation. In cross-validation, trained data was randomly partitioned into k folds of similar sizes, the k-1 folds are used for model training, and the rest one-fold was used for testing. K-Fold Cross Validation is used to prevent overfitting in forecasting models. In this paper, k is considered equal to 10. Feature Importance Score is also used in classification algorithms.

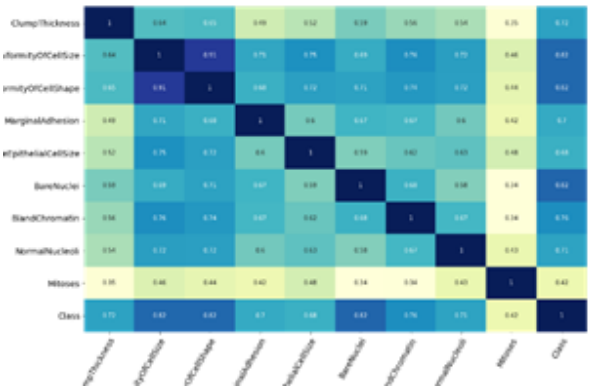

Fig 1: Correlation between all available features

Taking into account the significance score of the features in the algorithms (in algorithms such as KNN and MLP the feature importance score cannot be calculated), the values of the evaluation criteria are calculated.

Fig 2 to 5 show the performance outcome parameters of the classification algorithms employed, namely accuracy, precision, recall and f1 score. The pre and post dimensionality reduction accuracy of algorithms is compared in Fig 2 to 5, after the attributes were reduced, the outcome of measures also got affected.

The logistic regression, extra tree and SVM algorithms outperformed others by producing 99.14% accuracy. The Decision Tree delivered 96.28% and Random Forest made it with 98.42%. The value of this parameter is equal to 97.71% for MLP and 99.57% for bagging algorithm.

The value of the precision criterion for the Logistic Regression algorithm is equal to 99.82%, which has the highest value compared to the rest. In addition, the value of this parameter is equal to 98.07% for ddecision tree, 99.72% for bagging, 96.54% for random forest, 94.85% for SVM, 96.44% for extra tree and 96.44% for MLP algorithm. The highest recall criteria for SVM and extra tree algorithms are 1.00% and 99.14%, respectively.

The value of F1 score parameter is equal to 94.45% for decision tree, 97.41% for random forest, 98.29% for SVM, 98.36% for extra tree, 97.41% for KNN, 97.74% for logistic regression, 98.62% for bagging and 96.44% for MLP algorithm.

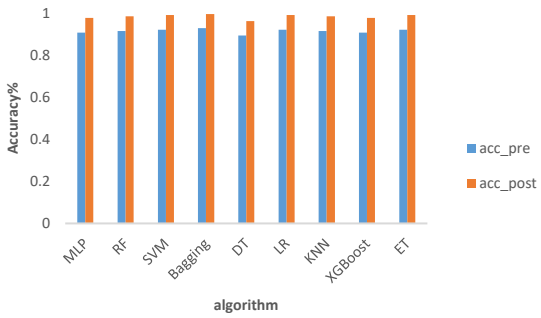

Fig 2: Accuracy of learning techniques

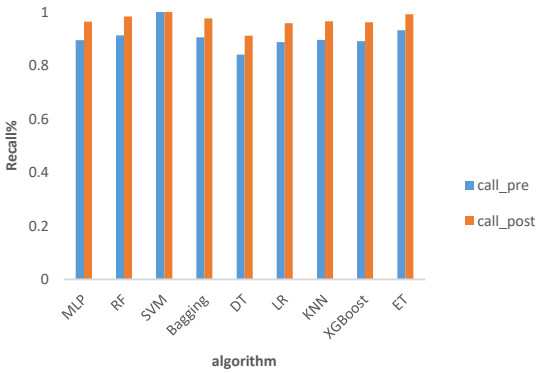

Fig 3: Recall of learning techniques

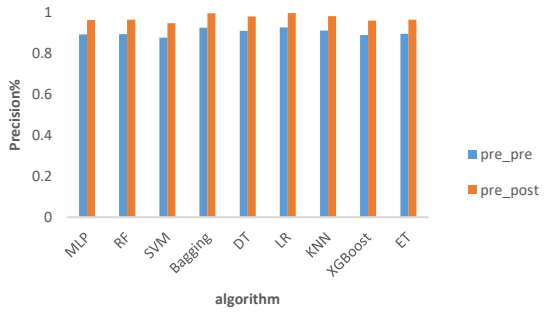

Fig 4: Precision of learning techniques

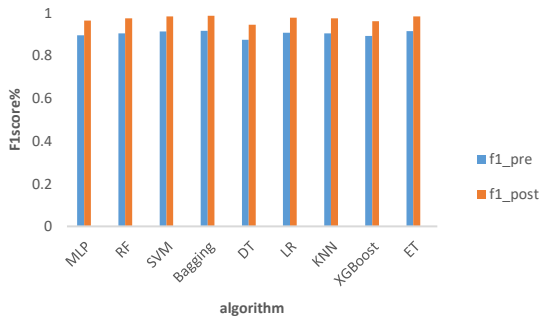

Fig 5: F1 score of learning techniques

Fig 6 to 11 are for a better understanding of feature ranking and importance according to classification algorithms. Each of these figures is a graphical representation of Table 2. These figures show the feature ranking based on feature importance and coefficient scores for all the applied classification algorithms except MLP and KNN. These figures also tend to represent the highly responsible attributes for breast cancer. Table 2 shows the five most significant features according to feature importance and correlation value. According to the table, it is found that UniformityOfCellSize is the significant feature or factor for identification and prediction. Besides BareNuclei, UniformityOfCellShape, ClumpThickness and BlandChromatin also significant factors predicting breast cancer.

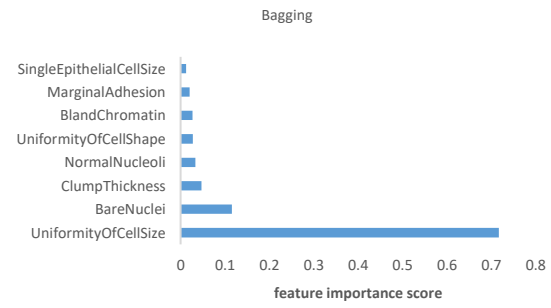

Fig 6: feature importance

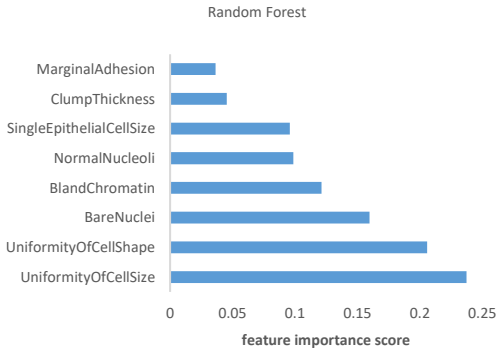

Fig 7: Feature importance

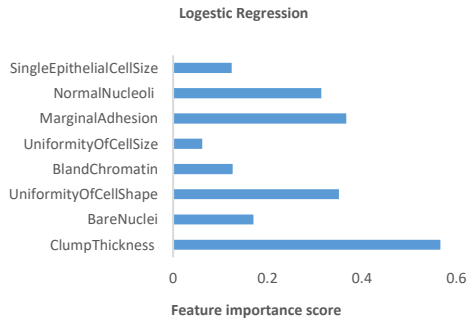

Fig 8: Feature importance

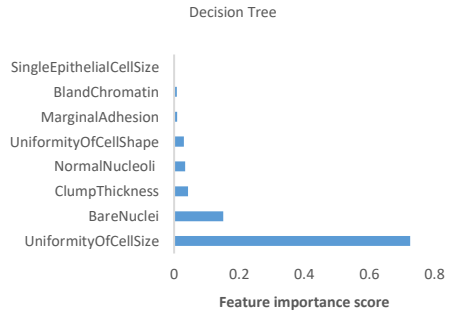

Fig 9: Feature importance

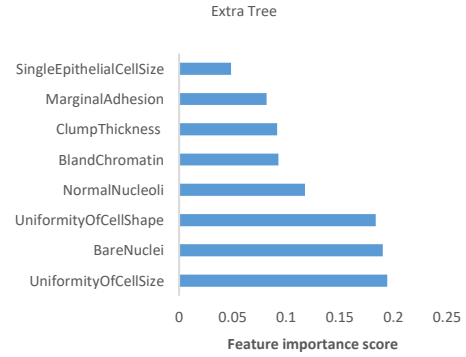

Fig 10: Feature importance

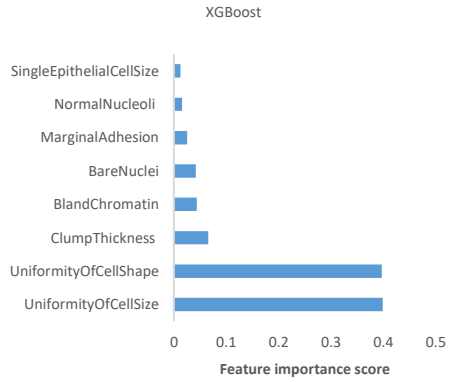

Fig 11: Feature importance

Table 2: Feature ranking for Breast Cancer

Ranking	ET	LR
First	UniformityOfCellSize	ClumpThickness
Second	BareNuclei	BareNuclei
Third	UniformityOfCellShape	UniformityOfCellShape
Forth	NormalNucleoli	BlandChromatin
Fifth	BlandChromatin	UniformityOfCellSize
Ranking	**DT**	**RF**
First	UniformityOfCellSize	UniformityOfCellSize
Second	BareNuclei	UniformityOfCellShape
Third	ClumpThickness	BareNuclei
Forth	NormalNucleoli	BlandChromatin
Fifth	UniformityOfCellShape	NormalNucleoli
Ranking	**Bagging**	**XGBoost**
First	UniformityOfCellSize	UniformityOfCellSize
Second	BareNuclei	UniformityOfCellShape
Third	ClumpThickness	ClumpThickness
Forth	NormalNucleoli	BlandChromatin
Fifth	UniformityOfCellShape	BareNuclei

Table 3 shows the value of area under ROC for all the applied classification algorithms. AUC represents a common area of true positive rate and false positive rate. Logistic regression shows higher performance than they do. On the other hand, XGBoost provided the good result. Fig 12 represents the ROC curve, which is built by the value of the true positive rate and false positive rate. It is a graphical representation of the AUC.

Table 3: Value of area under ROC

Algorithms	AUC
MLP	0.991
Random Forest	0.990
SVM	0.978
Bagging	0.987
Decision Tree	0.992
LR	0.996
KNN	0.984
Extra Tree	0.991
XGBOOST	0.995

The ROC curves of each ML algorithms are presented on Fig 12 ROC curve is an important metric for the

performance of classifiers. The area under ROC curve (AUC) is computed. The area is bigger than others are, the performance of the classifier is better. The Logistic Regression has the highest AUC score 0.996% while the AUC score of SVM 0.978% is the lowest as shown in table 3.

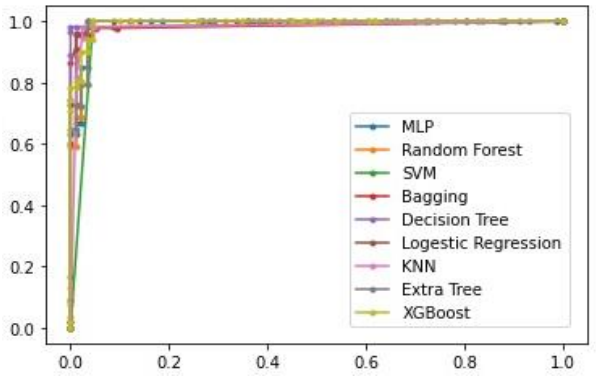

Fig 12: ROC curves of ten classifiers

DISCUSSION

In this research, we investigated the application of several ML methods to predict breast cancer. The methods used were SVM, LR, KNN, decision trees etc.

In order to Improvement of the performance of ML algorithms in predicting breast cancer, this paper highlight the importance of applying k-fold cross validation and feature importance score to compare the trained models generated from dataset with characteristics.

Selecting good features and reducing dimension has been effective in improving the results of algorithms, as shown in Fig 2 to 5. The pre and post dimensionality reduction accuracy of algorithms is compared in Fig 2 to 5, After the attributes were reduced, the outcome of measures also got affected. Because of evaluation, logistic regression algorithm with accuracy value equal to 99.14% and AUC equal to 99.6% have better performance than other algorithms, which compared to the work of others, a good result has been obtained.

Some ML studies reported higher accuracy (100%) and sensitivity (100%) for breast cancer prediction compared to the present study, which is likely due to using different databases. Similar to the database used in the current study, some studies used databases from specific medical or research centers.

A lot of researcher have realized research in breast cancer by using several datasets [16]. The [17] observed that SVM outperformed all other classifiers and achieved the highest accuracy 97.2%. They used Breast Cancer Wisconsin Diagnostic dataset from University of Wisconsin Hospitals Madison Breast Cancer Database. The [18], demonstrates the use of various supervised ML algorithms in classification of breast cancer from using 3D images and find out that SVM is the best based on his overall performance. The [19], worked on comparative study of relevance vector machine, which provides Low computational cost while comparing with other ML techniques, which are used for breast cancer detection, and explain how RVM is better than other ML algorithms for diagnosing breast cancer even the variables are reduced and achieved 97% accuracy. Asri [20] demonstrated that SVM proves its efficiency in breast cancer prediction and diagnosis and achieves the best performance in terms of precision and low error rate with an accuracy of 97.13%. Khoudfi and Bahaj [21] similarly proposed a comparison between ML algorithms and they found the SVM is the best classifier with an accuracy of 97.9% compared with K-NN, RF and NB, they are based on multilayer perception with 5 layers and 10 times cross validation using MLP. Latchoumiet [22] found a classification value of 98.4% proposing an optimization weighting of the particle swarm (WPSO) based on the SSVM for the classification. Osman [23] proposed a solution for the diagnosis of Wisconsin breast cancer (WBCD) with a prediction of 99.10% found by the SVM algorithm by combining a clustering algorithm with an efficient probabilistic vector support machine.

CONCLUSION

The early detection and classification of breast cancer help to prevent the disease's spread. The use of ML reduces the cost and time of diagnosis of tumor type. ML techniques in the field of medicine can be used to analyze a set of data related to a disease and to predict the disease. In this paper, ML algorithms are used to predict and diagnose of breast cancer. The data set is related to breast cancer collected from the UCI repository. This study presents ML algorithms to identify high correlated features that are closely associated with malignant identification. In this research, ML algorithms improved coupled by limited and selective features to produce high classification accuracy in tumor classification. In algorithms, the importance of features and cross-validation are the result of effective performance and increase accuracy. The value under the ROC curve and evaluation criteria such as accuracy, sensitivity, accuracy and F1 score are compared to a number of ML classification algorithms to assess breast cancer risk and actually predict breast cancer to identify the best appropriate classifier.

AUTHOR'S CONTRIBUTION

All authors contributed to the literature review, design, data collection and analysis, drafting the manuscript, read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding

the publication of this study.

FINANCIAL DISCLOSURE

No financial interests related to the material of this manuscript have been declared.

REFERENCES

1. Mohammed MA, Al-Khateeb B, Rashid AN, Ibrahim DA, Abd Ghani MK, Mostafa SA. Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images. Computers & Electrical Engineering. 2018; 70: 871-82.
2. Al-Hashimi MM, Wang XJ. Breast cancer in Iraq, incidence trends from 2000-2009. Asian Pac J Cancer Prev. 2014; 15(1): 281-6. PMID: 24528040 DOI: 10.7314/apjcp.2014.15.1.281 [PubMed]
3. Langarizadeh M, Mahmud R, Ramli AR, Napis S, Beikzadeh MR, Rahman WEZWA. Improvement of digital mammogram images using histogram equalization, histogram stretching and median filter. J Med Eng Technol. 2011; 35(2): 103-8. PMID: 21204610 DOI: 10.3109/03091902.2010.542271 [PubMed]
4. Langarizadeh M, Mahmud R, Ramli AR, Napis S, Beikzadeh MR, Wan Abdul Rahman WEZ. Effects of enhancement methods on diagnostic quality of digital mammogram images. Iranian Journal of Cancer Prevention. 2010; 3(1): 36–41.
5. Rabiei R, Ayyoubzadeh SM, Sohrabei S, Esmaeili M, Atashi A. Prediction of breast cancer using machine learning approaches. J Biomed Phys Eng. 2022; 12(3): 297-308. PMID: 35698545 DOI: 10.31661/jbpe.v0i0.2109-1403 [PubMed]
6. Maghooli K, Langarizadeh M, Shahmoradi L, Habibi-Koolaee M, Jebraeily M, Bouraghi H. Differential diagnosis of Erythmato-Squamous Diseases using classification and regression tree. Acta Inform Med. 2016; 24(5): 338-42. PMID: 28077889 DOI: 10.5455/aim.2016.24.338-342 [PubMed]
7. Shahmoradi L, Langarizadeh M, Pourmand G, Aghsaei Fard Z, Borhani A. Comparing three data mining methods to predict kidney transplant survival. Acta Inform Med. 2016; 24(5): 322-7. PMID: 28163356 DOI: 10.5455/aim.2016.24.322-327 [PubMed]
8. Tahmasebian S, Ghazisaeedi M, Langarizadeh M, Mokhtaran M, Mahdavi-Mazdeh M, Javadian P. Applying data mining techniques to determine important parameters in chronic kidney disease and the relations of these parameters to each other. J Renal Inj Prev. 2016; 6(2): 83-7. PMID: 28497080 DOI: 10.15171/jrip.2017.16 [PubMed]
9. Shaikh FJ, Rao DS. Prediction of cancer disease using machine learning approach. Materials Today: Proceedings. 2022; 50 :40-7.
10. Gayathri BM, Sumathi CP, Santhanam T. Breast cancer diagnosis using machine learning algorithms: A survey. International Journal of Distributed and Parallel Systems. 2013; 4(3): 105-12.
11. Saleh H, Abd-El Ghany SF, Alyami H, Alosaimi W. Predicting breast cancer based on optimized deep learning approach. Comput Intell Neurosci. 2022; 2022: 1820777. PMID: 35345799 DOI: 10.1155/2022/1820777 [PubMed]
12. Yousefi S. Comparison of the performance of machine learning algorithms in predicting heart disease. Frontiers in Health Informatics. 2021; 10(1): 99.
13. Frank A, Asuncion A. UCI machine learning repository [Internet]. 2010 [cited: 8 Jul 2022]. Available from: http://archive.ics.uci.edu/ml
14. Ali MM, Paul BK, Ahmed K, Bui FM, Quinn JM, Moni MA. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Comput Biol Med. 2021; 136: 104672. PMID: 34315030 DOI: 10.1016/j.compbiomed.2021.104672 [PubMed]
15. Battineni G, Chintalapudi N, Amenta F. Performance analysis of different machine learning algorithms in breast cancer predictions. EAI Endorsed Transactions on Pervasive Health and Technology. 2020; 6(23): e4.
16. World Health Organization. Preventive cancer [Internet]. 2008 [cited: 18 Feb 2020]. Available from: http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/
17. Naji MA, El Filali S, Aarika K, Benlahmar EH, Abdelouhahid RA, Debauche O. Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Computer Science. 2021; 191: 487-92.
18. Nayak S, Gope D. Comparison of supervised learning algorithms for RF-based breast cancer detection. Computing and Electromagnetics International Workshop. IEEE; 2017.
19. Gayathri BM, Sumathi CP. Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer. IEEE International Conference on Computational Intelligence and Computing Research. IEEE; 2016.
20. Asri H, Mousannif H, Al Moatassime H, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science. 2016; 83: 1064-9.
21. Khourdifi Y, Bahaj M. Applying best machine learning algorithms for breast cancer prediction and classification. International Conference on Electronics, Control, Optimization and Computer Science. IEEE; 2018.
22. Latchoumi TP, Parthiban L. Abnormality detection using weighed particle swarm optimization and smooth support vector machine. Biomedical Research. 2017; 28(11): 4749-51.
23. Osman AH. An enhanced breast cancer diagnosis scheme based on two-step-SVM technique. International Journal of Advanced Computer Science and Applications. 2017; 8(4): 158-65