Prediction of Breast Cancer using Machine Learning Approaches

Reza Rabiei[1], Seyed Mohammad Ayyoubzadeh[2], Solmaz Sohrabei[3]*, Marzieh Esmaeili[2], Alireza Atashi[4]

ABSTRACT

Background: Breast cancer is considered one of the most common cancers in women caused by various clinical, lifestyle, social, and economic factors. Machine learning has the potential to predict breast cancer based on features hidden in data.

Objective: This study aimed to predict breast cancer using different machine-learning approaches applying demographic, laboratory, and mammographic data.

Material and Methods: In this analytical study, the database, including 5,178 independent records, 25% of which belonged to breast cancer patients with 24 attributes in each record was obtained from Motamed cancer institute (ACECR), Tehran, Iran. The database contained 5,178 independent records, 25% of which belonged to breast cancer patients containing 24 attributes in each record. The random forest (RF), neural network (MLP), gradient boosting trees (GBT), and genetic algorithms (GA) were used in this study. Models were initially trained with demographic and laboratory features (20 features). The models were then trained with all demographic, laboratory, and mammographic features (24 features) to measure the effectiveness of mammography features in predicting breast cancer.

Results: RF presented higher performance compared to other techniques (accuracy 80%, sensitivity 95%, specificity 80%, and the area under the curve (AUC) 0.56). Gradient boosting (AUC=0.59) showed a stronger performance compared to the neural network.

Conclusion: Combining multiple risk factors in modeling for breast cancer prediction could help the early diagnosis of the disease with necessary care plans. Collection, storage, and management of different data and intelligent systems based on multiple factors for predicting breast cancer are effective in disease management.

Citation: Rabiei R, Ayyoubzadeh SM, Sohrabei S, Esmaeili M, Atashi AR. Prediction of Breast Cancer using Machine Learning Approaches. *J Biomed Phys Eng.* 2022;12(3):297-308. doi: 10.31661/jbpe.v0i0.2109-1403.

Keywords

Artificial Intelligence; Breast Cancer; Computing Methodologies; Genetic Algorithm; Machine Learning

[1]PhD, Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[2]PhD, Department of Health Information Technology and Management, School of Allied Medical Sciences, Tehran University of Medical Science, Tehran, Iran
[3]MSc, Department Deputy of Development, Management and Resources, Office of Statistic and Information Technology Management, Zanjan University of Medical Sciences, Zanjan, Iran
[4]PhD, Department of E-Health, Virtual School, Tehran University of Medical Sciences, Medical Informatics Research Group, Clinical Research Department, Breast Cancer Research Center, Motamed Cancer Institute, ACECR, Tehran, Iran

*Corresponding author:
Solmaz Sohrabei
Department Deputy of Development, Management and Resources, Office of Statistic and Information Technology Management, Zanjan University of Medical Sciences, Zanjan, Iran
E-mail: solmazsohrabee1@gmail.com

Received: 25 September 2021
Accepted: 5 March 2022

Introduction

Breast cancer is considered a multifactorial disease and the most common cancer in women worldwide [1,2] with approximately 30% of all female cancers [3, 4] (i.e. 1.5 million women are diagnosed with breast cancer each year, and 500,000 women die from this disease in the world). Over the past 30 years, this disease has increased, while the death rate has decreased. However, the reduction in mortality due to mammography screening is estimated at 20% and improvement in cancer treatment is estimated at 60% [5,6].

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 [7]. However, about a quarter of women with breast cancer are diagnosed negatively within two years of screening. Therefore, the early and timely diagnosis of breast cancer is crucial [8].

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 [9]. 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 [9-11]. 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 [12,13].

Machine learning, as a modeling approach, represents the process of extracting knowledge from data and discovering hidden relationships [14], widely used in healthcare in recent years [15] to predict different diseases [16-18]. Some studies only used demographic risk factors (lifestyle and laboratory data) in predicting breast cancer [19,20], and several studies predicted based on mammographic stereotypes [21] or used data from patient biopsy [22]. Others showed the application of genetic data in predicting breast cancer [23].

A major challenge in predicting breast cancer is the creation of a model for addressing all known risk factors [24-26]. Current prediction models might only focus on the analysis of mammographic images or demographic risk factors without other critical factors. In addition, these models, which are accurate enough for identifying high-risk women, could result in multiple screening and invasive sampling with magnetic resonance imaging (MRI) and ultrasound. The financial and psychological burden could be experienced by patients [27-29].

The effective prediction of breast cancer risk requires different factors, including demographic, laboratory, and mammographic risk factors [24,25,30,31]. Therefore, multifactorial models with many risk factors in their analysis can be effective in assessing the risk of breast cancer through more accurate analysis [32,33]. The current study aimed to predict breast cancer using different machine learning approaches considering various factors in modeling.

Material and Methods

In this analytical study, the database was obtained from a clinical breast cancer research center (Motamed cancer institute) in Tehran, Iran. The research was conducted in 4 stages: data collection, data pre-processing, modeling, and model evaluation.

Data Collection

In the first stage, 5178 records of people, referred to the research center over the past 10 years (2011-2021), were prepared retrospectively. Each record covered 24 features (11 demographic features, 9 laboratory features, and 4 mammography features) (Table 1), all labeled to indicate the presence or absence of breast cancer, of which 1,295 records (25%) were identified as breast cancer.

Data preprocessing

The second step was associated with data preprocessing in which five records related to men were removed, and a total of 1290 records remained. Some of the patients' laboratory features that were outside the considered range were repositioned in the central registry as their laboratory results were available. In

Table 1: The relevant features of breast cancer

Feature name	Description	Type	Values
Age	age at diagnosis	Demographic	<100 Years
Age.menop	age of menopause	Demographic	38-65 Years
First pregnancy	age at first pregnancy	Demographic	13-42 Years
Age.menarch	age of menarche	Demographic	11-18 Years
BMI	Body mass index	Demographic	Underweight (Below 18.5) =0, Normal (18.5 - 24.9) =1, Overweight (25.0 - 29.9) =2, Obese (30.0 and Above) =3
Lactation	Breastfeeding status	Demographic	0-96 Mount
Physical Activity	Have a regular Physical Activity	Demographic	Yes=1 No=0
Education	Academic education	Demographic	Illiterate=1, primary=2, high school=3, university=4
Life event stress	life event statues	Demographic	No=0, death of father=1, family problems=2, death of mother=3, death of child=4, death of husband=5, divorced=6
Smoking	Smoking status	Demographic	Yes=1, No=0
Marital	marital status	Demographic	Single=0 other=1
Duration Ocp. used	Mount of used Oral Contraceptive Pills	Laboratory	0-120 Mount
Duration HRT used	mount of Hormone replacement therapy use	Laboratory	0-120 Mount
Personal. Other. Cancer	Personal. Other. Cancer	Laboratory	No=0, ovary=1, endometrium=2, colon=3, meningioma=4, lymphoma=5
Family.BC	FAMILY Breast Cancer	Laboratory	Yes=1 No=0
Exposure X-ray	Exposure X-ray to chest	Laboratory	Negative=0 positive=1
Vitamin D3	Amount vitamin D in body	Laboratory	>10 mg=0 deficiency 10-30 mg=1 insufficiency 30-100 mg=2 sufficient >100 mg=3 Overdose
Biopsy	pathology of biopsy	Laboratory	no malignancy detected= 0 lobular carcinoma insitu=1 ductal carcinoma insitu=2 ductal carcinoma insitu=3 invasive lobular carcinoma=4 medullary=5 microinvasion=6
Hysterectomy	history of hysterectomy	Laboratory	Yes=1 No=0
Personal.BC	Personal Breast Cancer history	Laboratory	Yes=1 No=0, surgery=2, RT (Radio Therapy) =3
Breast density	screening	Mammography	Fatty tissue=0, glandular and fibrous tissue=1, dense =2, heterogeneously dense extremely dense=3
Micro lobulated	screening	Mammography	None=0, Fibroadenoma=1, Papilloma=2, Phyllodes tumor=3, DCIS=4, IDC=5, ILC=6, Lactating and tubular adenomas =7
Circumscribed	screening	Mammography	None=0 cysts=1, complicated cyst=2, clustered microcyst=3, solid mass=4
Micro calcification, Macro calcification	screening	Mammography	Probably benign Punctate Intermediate=1 concern Coarse heterogeneous Amorphous =2 Higher probability of malignancy Fine pleomorphic Fine linear/ branching=3
Class	**Breast Cancer**		**malignant=1 benign=0**

DCIS: Ductal carcinoma in situ, IDC: Invasive ductal carcinoma, ILC: Invasive lobular carcinoma

addition, for records with missing values, the method of maximum frequency or the same mod was used. Finally, the Synthetic Minority Oversampling Technique (SMOTE) was used to balance the training data due to the difference in the number of study class records.

Modeling for breast cancer prediction

In the third step, the Scikit-Learn 0.18.2 library, NumPy v1.20, TPOT, and Python open-source programming were used for modeling. Three leaners, i.e. Random forest (RF), Gradient Boosting trees (GBT), and Multi-layer Perceptron (MLP) were applied to the dataset. In addition, the K-Fold (K=3) validation was used to gain the optimized hyper-parameter of each model in the genetic algorithm step. In the final evaluation, the train-test split method (75% for training and 25% for testing) was used to more accurately estimate the performance of the model. In this study, a genetic algorithm (GA) with a population of 5, the number of children 50, and the number of 10 generations with the criterion of the highest accuracy in model selection were used to optimize values for variables. Further, these models were then trained with demographic and laboratory features (20 features). Finally, the model was trained with all demographic, laboratory, and mammography features (24 features) to measure the effect of mammography features in predicting breast cancer. In the current study, MLP hidden layers numbers were considered 10, and the alpha value for the training rate was 0.01-0.2. The sigmoid and hyperbolic tangent functions were selected for activation function. The value of the solver optimizer function was set to a gradient-based optimizer method, such as Adam and Stochastic Gradient Descent (SGD) to find the optimal weights. In the GBT model, the learning rate was considered 0.01-0.2, and the maximum depth was regarded as 3, 5, and 8. The buoyancy level learning was 0.1 and the estimator value for the gradient boosting was 10. In the random forest (RF) model, the minimum number of sheets required to split an external node was considered 4 and 12. The estimator value was 151, and the node evaluation parameter to prevent splitting (min_samples_split) was considered 5 and 10. The block diagram for the methods is shown in Figure 1.

Random Forest (RF)

As a non-parametric approach, the RF uses the classification method. For each set of data, the RF performs categorization at high speed and applies a large number of decision trees [34]. In each tree, there is a random number of input variables, then all the trees are combined for a better inference from the variables [35].

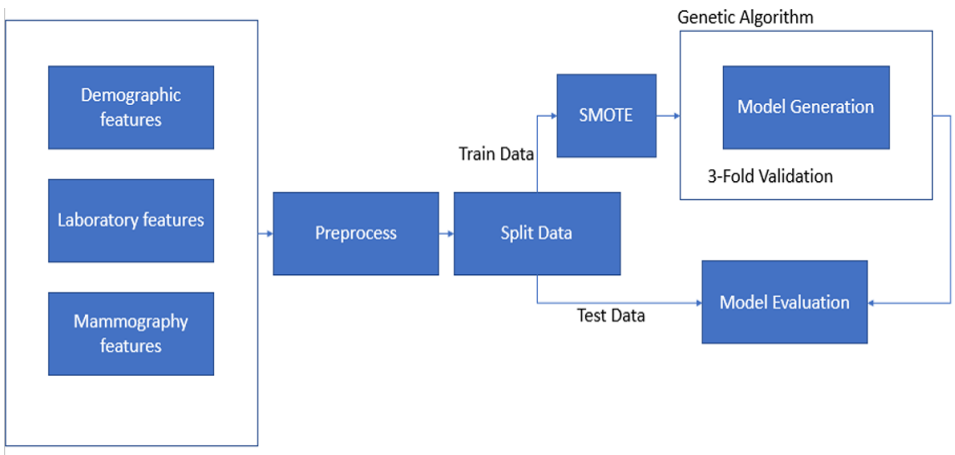

Figure 1: Block diagram of methods

Gradient Boosting Trees (GBT)

This algorithm is one of the reinforcement gradient algorithms with a very good performance in classification and performs the best classification for each of the data [36]. In this method, the trees are trained one after another; each subset tree is taught primarily with data erroneously predicted by the previous tree. This process continuously reduces the model error since each model is sequentially improved against the weaknesses of the previous model [37,38].

Multi-Layer Perceptron (MLP)

As a deep artificial neural network, the MLP is composed of an input layer for receiving the signal, an output layer used for prediction, and in between those two, some hidden layers are acting as the computation engine. The MLP is trained by a backpropagation algorithm, which is part of the supervised networks. In this network, data are driven from input nodes to output nodes. If there is an error in the output, this error must be somehow returned from the output to the input, and this corrects the weights. The most commonly used method for this is the post-diffusion algorithm [39,40].

Genetic Algorithm (GA)

As a subset of the evolutionary computing algorithm, GA is directly associated with artificial intelligence and used for solving optimization problems through the evolution process [41,42]. To obtain the best answer, the GA applies the best survival rule to a series of problems for patterning the best solution for problems [43,44]. In each generation, the optimal solution is achieved based on a natural biological process and by selecting the best chromosomes for creating the subsequent generation to solve the problem optimally [45].

Model Evaluation

The test results of the database samples (confusion matrix) are shown in Table 2. In the final stage, the performance of the created models was measured by different criteria. The classification of samples is one of the common criteria in evaluating and measuring the ability of classifiers, the degree of separation or accuracy, and the separation of classes [46]. In this study, accuracy, sensitivity, specificity, and the area under the receiver operating characteristic (ROC) curve were used to measure the overall performance of the classifiers.

Results

A total of 1290 records containing 24 demographic, laboratory, and mammographic features related to breast cancer were used in the study; the weight of the features based on their degree of importance is shown in (the weights are between (0.0 - 1) (Figure 2). Family history of breast cancer, personal history of breast cancer, breast density, and age of diagnosis is 5 important factors in the diagnosis of this disease.

The performance of the models shown based on the ROC area under the curve demonstrated the Gradient Boosting Trees (GBT) as the model with the highest performance. The modeling results using RF, GBT, and MLP are shown in Table 3, and the comparison of their ROC curve is demonstrated in Figure 3 and Table 4.

Discussion

According to the findings of the current study, the mammographic features along with other features could improve the performance of models. The RF model showed the highest

Table 2: Confusion matrix of a binominal classifier

		Predicted	
		Negative	Positive
Actual	Negative	TN	FP
	Positive	FN	TP

TN: True Negative, FN: False Negative, FP: False Positive, TP: True Positive

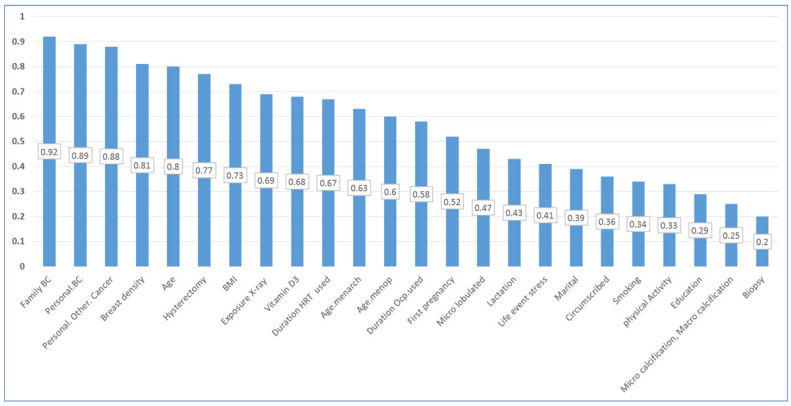

Figure 2: The weight of the features in breast cancer prediction

Table 3: Performance comparison of the breast cancer prediction models

Models	Features	AUC	Sensitivity (%)	Specificity (%)	Accuracy (%)
Random Forest	Demographics	0.53	93	83	79
	Demographics + Mammography	0.53	95	83	80
Gradient Boosting	Demographics	0.59	63	87	62
	Demographics + Mammography	0.59	82	86	74
Multi-Layer Perceptron	Demographics	0.56	78	85	71
	Demographics + Mammography	0.56	82	84	73

AUC: Area under the ROC curve, ROC: Receiver operating characteristic

sensitivity (95%), but was more efficient due to the sensitivity of breast cancer diagnosis, models, such as gradient boosting with higher specificity (86%).

In a study by Rosner et al. [47,48], the findings showed that family and personal history of breast cancer were two of the key influential factors in breast cancer, which are consistent with the findings of the current study as these two factors demonstrated the highest weight (0.92 and 0.89) compared to other factors. Breast density and age are influential in tumor appearance and increase the proportion of breast cancers [49] with the weights (0.80, 0.80), respectively. However, the hysterectomy feature was used along with other risk factors that could influence the performance of models. The study by Chow et al. assessed the risk of breast cancer after hysterectomy and showed a statistical significance between hysterectomy and breast cancer [50].

The use of optimization algorithms with feature weighting and proper adjustment of classification parameters could improve the performance of classification algorithms [51]. Studies reported that the classifiers that used GA in feature selection demonstrated better performance compared to those that did not use the GA. For the prediction of breast cancer, Bhattacharya et al. [52] approached three

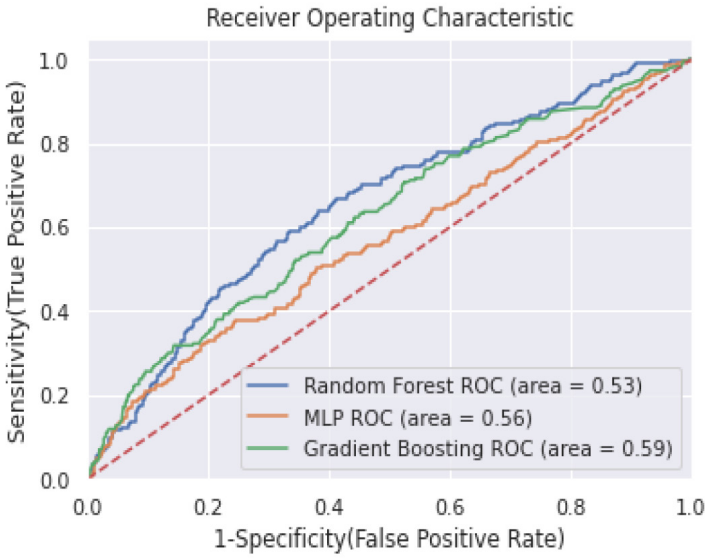

Figure 3: Receiver operating characteristic (ROC) curve of models

Table 4: Area under the Receiver operating characteristic (ROC) curve

Test Result Model(s)	Area
GBT	0.59
MLP	0.56
RF	0.53

GBT: Gradient Boosting Tree, MLP: Multi-Layer-Perceptron, RF: Random Forest

machine learning algorithms and used GA for feature selection; the findings of this study showed that the GA led to an improved performance for models created. In a study by Sakri et al. [53] to predict breast cancer recurrence in 198 instances with 34 clinical attributes, the GA was used for optimization. The Naive Bayes accuracy, sensitivity, specificity, and area under the ROC curve were reported at 70%, 81%, 79%, and 0.82, respectively in this study. Kumar et al. [54] used GA on a breast cancer dataset containing 611 records with 10 features to predict breast cancer survival and the reported accuracy, and ROC were 88% and 0.966 for GA, showing a better performance compared to Naive Bayes, DT, and K-nearest neighbor (KNN); in their study conducted to classify the masses observed in mammographic stereotypes, Thawkar and Ingolikar [55] used a dataset composed of 651 records with 25 mammography features. In the current study, the models were optimized by GA, and the ROC, accuracy, sensitivity, and specificity were 0.974, 95%, 96.14%, and 93.94% for RF, respectively. In the studies noted above, the modeling was performed using one set of influencing factors.

Some machine-learning studies [56-62] reported higher accuracy (100%) and sensitivity (100%) for breast cancer prediction compared to the present study, which is likely due to using different databases, such as "Wisconsin" and "SEER". Similar to the database used in the current study, some studies used databases from specific medical or research centers. Behravan and Hartikainen [33] predicted breast cancer using a database containing 695 records, including demographic risk factors and genetic data; their findings suggested that the XGBoost model with different factors showed improved performance (AUC= 0.788) compared to a model with just one set of factors (AUC= 0.678). In a study by Feld

et al. [10] to predict breast cancer, the modeling was performed on 738 records, including demographic, genetic, and abnormal mammographic data, and the reported AUC was 0.75. Other studies suggest that considering different factors in modeling would improve modeling performance. For example, by Ayvaci MU et al. [63], the analysis of demographic, mammography, and biopsy data using logistic regression resulted in an AUC of 0.84. Rajendran k et al. [64] analyzed 2.4 million records of mammography screening and demographic risk factors associated with breast cancer to predict breast cancer using the Naïve Bayes, RF, and C4.5 techniques; the findings indicated the highest AUC (0.993) for Naïve Bayes.

The findings of a study by Atashi et al. [65] conducted on a database with 4004 records, including demographic risk factors showed the higher performance of the neural network (sensitivity= %80.9, specificity= %99.8, accuracy= %62.8) compared to other approaches, such as C5.0. Mosayebi et al. study [66] was conducted on a database with 5471 records, including demographic and laboratory features reported for C.50 (accuracy 82%, sensitivity 86%. and specificity 77%). In a study by Jalali et al. [67] performed on 644 records (with 10 clinical features), the support vector machine (SVM) was reported with the highest sensitivity (94.33%), accuracy (93.72%), and specificity (92.26%). Afshar et al. [68] studied the survival of breast cancer patients using a dataset with 856 records and 15 clinical features using machine learning models. In this study, C5.0 showed the highest sensitivity (92.21%) and accuracy (84%). In addition, in a similar study by Nourelahi et al. [69] to predict patient survival on a database consisting of 5673 cases and 41 clinical features, logistic regression presented a sensitivity of 71.85%, specificity of 72.83%, and accuracy of 72.49%. In addition, Tapak et al. [70] performed a study on a database with 550 records to predict the survival and metastasis of breast cancer and also reported the sensitivity and specificity of 99% for AdaBoost, the findings of the current study suggest that modeling with a variety of related risk factors from different sources could improve the performance of models in breast cancer prediction.

In the current study, limitations are considered as follows: modeling based on records of only one database, and the lack of access to genetic data that could influence the findings of the study. However, different machine learning approaches were used considering demographic, laboratory, and mammography features, resulting in comparing the performance of different approaches in predicting breast cancer.

Conclusion

The proposed machine-learning approaches could predict breast cancer as the early detection of this disease could help slow down the progress of the disease and reduce the mortality rate through appropriate therapeutic interventions at the right time. Applying different machine learning approaches, accessibility to bigger datasets from different institutions (multi-center study), and considering key features from a variety of relevant data sources could improve the performance of modeling.

Authors' Contribution

R. Rabiei proposed conceptualization and design, supervision of modeling, manuscript drafting, editing, and critical review. Deta modeling, interpretation, and manuscript drafting was done by SM. Ayyoubzadeh. S. Sohrabei provided conceptualization and design, data modeling and interpretation, manuscript drafting, and editing. M. Esmaeili presented data interpretation and manuscript drafting. A. Atashi collected data and manuscript drafting. All the authors read, modified, and approved the final version of the manuscript.

Ethical Approval

This study was approved by Clinical Research Department, Breast Cancer Research Center, Motamed Cancer Institute (ACECR), Tehran, Iran, with Approval ID IR, ACECR, IBCRC, REC.1394.68.

Informed consent

We used anonymous data for modeling and no consent was required for conducting this study.

Funding

There was no funding for conducting this study.

Conflict of Interest

None

References

1. Zhang X, Shengli SU, Hongchao WA. Intelligent diagnosis model and method of palpation imaging breast cancer based on data mining. *Big Data Research.* 2019;**5**(1):2019005. doi: 10.11959/j.issn.2096-0271.2019005.
2. Chen SI, Tseng HT, Hsieh CC. Evaluating the impact of soy compounds on breast cancer using the data mining approach. *Food & function.* 2020;**11**(5):4561-70. doi: 10.1039/C9FO00976K. PubMed PMID: 32400770.
3. Aavula R, Bhramaramba R, Ramula US. A Comprehensive Study on Data Mining Techniques used in Bioinformatics for Breast Cancer Prognosis. *Journal of Innovation in Computer Science and Engineering*. 2019;**9**(1):34-9.
4. Kaushik D, Kaur K. Application of Data Mining for high accuracy prediction of breast tissue biopsy results. 2016 Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC); Moscow, Russia: IEEE; 2016. p. 40-5. doi: 10.1109/DIPDMWC.2016.7529361.
5. Mokhtar SA, Elsayad A. Predicting the severity of breast masses with data mining methods. *ArXiv preprint arXiv:1305.7057*. 2013. doi: 10.48550/ARxIV.1305.7057.
6. Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology.* 2018;**12**(2):119-26. doi: 10.1177/1748301818756225.
7. Fan J, Wu Y, Yuan M, Page D, Liu J, Ong IM, Peissig P, Burnside E. Structure-leveraged methods in breast cancer risk prediction. *The Journal of Machine Learning Research.* 2016;**17**(1):2956-70.
8. Burnside ES, Liu J, Wu Y, Onitilo AA, McCarty CA, Page CD, et al. Comparing Mammography Abnormality Features to Genetic Variants in the Prediction of Breast Cancer in Women Recommended for Breast Biopsy. *Acad Radiol.* 2016;**23**(1):62-9. doi: 10.1016/j.acra.2015.09.007. PubMed PMID: 26514439. PubMed PMCID: PMC4684977.
9. Stephens K. New Mammogram Measures of Breast Cancer Risk Could Revolutionize Screening. *AXIS Imaging News.* 2020.
10. Feld SI, Fan J, Yuan M, Wu Y, Woo KM, Alexandridis R, Burnside ES. Utility of Genetic Testing in Addition to Mammography for Determining Risk of Breast Cancer Depends on Patient Age. *AMIA Jt Summits Transl Sci Proc.* 2018;**2017**:81-90. PubMed PMID: 29888046. PubMed PMCID: PMC5961791.
11. Guan Y, Nehl E, Pencea I, Condit CM, Escoffery C, Bellcross CA, McBride CM. Willingness to decrease mammogram frequency among women at low risk for hereditary breast cancer. *Sci Rep.* 2019;**9**(1):9599. doi: 10.1038/s41598-019-45967-6. PubMed PMID: 31270367. PubMed PMCID: PMC6610104.
12. American Cancer Society. Cancer facts & figures 2018. Atlanta: American Cancer Society; 2018.
13. Blandin Knight S, Crosbie PA, Balata H, Chudziak J, Hussell T, Dive C. Progress and prospects of early detection in lung cancer. *Open Biol.* 2017;**7**(9):170070. doi: 10.1098/rsob.170070. PubMed PMID: 28878044. PubMed PMCID: PMC5627048.
14. Jothi N, Husain W. Data mining in healthcare-a review. *Procedia Computer Science.* 2015;**72**:306-13. doi: 10.1016/j.procs.2015.12.145.
15. Maxwell K, Nathanson K. Common breast cancer risk variants in the post-COGS era: a comprehensive review. *Breast Cancer Res.* 2013;**15**(6):212. doi: 10.1186/bcr3591. PubMed PMID: 24359602. PubMed PMCID: PMC3978855.
16. McCarthy AM, Keller B, Kontos D, Boghossian L, McGuire E, Bristol M, et al. The use of the Gail model, body mass index and SNPs to predict breast cancer among women with abnormal (BI-RADS 4) mammograms. *Breast Cancer Res.* 2015;**17**(1). doi: 10.1186/s13058-014-0509-4. PubMed PMID: 25567532. PubMed PMCID: PMC4311477.
17. Bent CK, Bassett LW, D'Orsi CJ, Sayre JW. The positive predictive value of BI-RADS microcalcification descriptors and final assessment categories. *AJR Am J Roentgenol.* 2010;**194**(5):1378-83. doi: 10.2214/AJR.09.3423. PubMed PMID: 20410428.
18. Burnside ES, Rubin DL, Fine JP, Shachter RD, Sisney GA, Leung WK. Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: initial experience. *Radiology.* 2006;**240**(3):666-73.

doi: 10.1148/radiol.2403051096. PubMed PMID: 16926323.

19. Ghani MU, Alam TM, Jaskani FH. Comparison of classification models for early prediction of breast cancer. 2019 International Conference on Innovative Computing (ICIC); Lahore, Pakistan: IEEE; 2019. p. 1-6. doi: 10.1109/ICIC48496.2019.8966691.
20. Williams K, Idowu PA, Balogun JA, Oluwaranti AI. Breast cancer risk prediction using data mining classification techniques. *Transactions on Networks and Communications*. 2015;**3**(2):1-11. doi: 10.14738/tnc.32.662.
21. Ferreira P, Fonseca NA, Dutra I, Woods R, Burnside E. Predicting malignancy from mammography findings and image-guided core biopsies. *International Journal of Data Mining and Bioinformatics.* 2015;**11**(3):257-76. doi: 10.1504/IJDMB.2015.067319. PubMed PMID: 26333262. PubMed PMCID: PMC4764253.
22. Oyewola D, Hakimi D, Adeboye K, Shehu MD. Using five machine learning for breast cancer biopsy predictions based on mammographic diagnosis. *International Journal of Engineering Technologies*. 2016;**2**(4):142-5. doi: 10.19072/ijet.280563.
23. Hajiloo M, Damavandi B, Hooshsadat M, Sangi F, et al. Breast cancer prediction using genome wide single nucleotide polymorphism data. *BMC Bioinformatics.* 2013;**14**(Suppl 13):S3. doi: 10.1186/1471-2105-14-S13-S3. PubMed PMID: 24266904. PubMed PMCID: PMC3891310.
24. Brédart A, Kop JL, Antoniou AC, Cunningham AP, De Pauw A, et al. Clinicians' use of breast cancer risk assessment tools according to their perceived importance of breast cancer risk factors: an international survey. *J Community Genet*. 2019;**10**(1):61-71. doi: 10.1007/s12687-018-0362-8. PubMed PMID: 29508368. PubMed PMCID: PMC6325038.
25. Hou C, Zhong X, He P, Xu B, Diao S, Yi F, Zheng H, Li J. Predicting Breast Cancer in Chinese Women Using Machine Learning Techniques: Algorithm Development. *JMIR Med Inform*. 2020;**8**(6):e17364. doi: 10.2196/17364. PubMed PMID: 32510459. PubMed PMCID: PMC7308891.
26. Maas P, Barrdahl M, Joshi AD, Auer PL, et al. Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. JAMA Oncology. 2016;**2**(10):1295-302. doi: 10.1001/jamaoncol.2016.1025.
27. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology*. 2019;**292**(1):60-6. doi: 10.1148/radiol.2019182716. PubMed PMID: 31063083.
28. Koopmann BDM, Harinck F, Kroep S, Konings ICAW, Naber SK, et al. Identifying key factors for the effectiveness of pancreatic cancer screening: A model-based analysis. *Int J Cancer.* 2021;**149**(2):337-46. doi: 10.1002/ijc.33540. PubMed PMID: 33644856. PubMed PMCID: PMC8251934.
29. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA Cancer J Clin.* 2019;**69**(2):127-57. doi: 10.3322/caac.21552. PubMed PMID: 30720861. PubMed PMCID: PMC6403009.
30. Arefan D, Mohamed AA, Berg WA, Zuley ML, Sumkin JH, Wu S. Deep learning modeling using normal mammograms for predicting breast cancer risk. *Med Phys.* 2020;**47**(1):110-8. doi: 10.1002/mp.13886. PubMed PMID: 31667873. PubMed PMCID: PMC6980268.
31. Yanes T, Young MA, Meiser B, James PA. Clinical applications of polygenic breast cancer risk: a critical review and perspectives of an emerging field. *Breast Cancer Res.* 2020;**22**(1):21. doi: 10.1186/s13058-020-01260-3. PubMed PMID: 32066492. PubMed PMCID: PMC7026946.
32. Feld SI, Woo KM, Alexandridis R, Wu Y, Liu J, et al. Improving breast cancer risk prediction by using demographic risk factors, abnormality features on mammograms and genetic variants. *AMIA Annu Symp Proc*. 2018;**2018**:1253-62. PubMed PMID: 30815167. PubMed PMCID: PMC6371301.
33. Behravan H, Hartikainen JM, Tengström M, Kosma VM, Mannermaa A. Predicting breast cancer risk using interacting genetic and demographic factors and machine learning. *Sci Rep*. 2020;**10**(1):11044. doi: 10.1038/s41598-020-66907-9. PubMed PMID: 32632202. PubMed PMCID: PMC7338351.
34. Dai B, Chen RC, Zhu SZ, Zhang WW. Using random forest algorithm for breast cancer diagnosis. 2018 International Symposium on Computer, Consumer and Control (IS3C); Taichung, Taiwan: IEEE; 2018. p. 449-52. doi: 10.1109/IS3C.2018.00119.
35. He T, Puppala M, Ogunti R, Mancuso JJ, Yu X, Chen S, Chang JC, Patel TA, Wong ST. Deep learning analytics for diagnostic support of breast cancer disease management. 2017 IEEE EMBS international conference on biomedical & health informatics (BHI); Orlando, FL, USA: IEEE; 2017. p. 365-8. doi: 10.1109/BHI.2017.7897281.
36. Shrivastav LK, Jha SK. A gradient boosting machine learning approach in modeling the impact

of temperature and humidity on the transmission rate of COVID-19 in India. *Appl Intell (Dordr)*. 2021;**51**(5):2727-39. doi: 10.1007/s10489-020-01997-6. PubMed PMID: 34764559. PubMed PMCID: PMC7609380.

37. Xenochristou M, Hutton C, Hofman J, Kapelan Z. Water demand forecasting accuracy and influencing factors at different spatial scales using a Gradient Boosting Machine. *Water Resources Research*. 2020;**56**(8):e2019WR026304. doi: 10.1029/2019WR026304.
38. Wang S, Wang Y, Wang D, Yin Y, Wang Y, Jin Y. An improved random forest-based rule extraction method for breast cancer diagnosis. *Applied Soft Computing*. 2020;**86**:105941. doi: 10.1016/j.asoc.2019.105941.
39. Verma D, Mishra N. Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques. 2017 International Conference on Intelligent Sustainable Systems (ICISS); Palladam, India: IEEE; 2017. p. 533-8. doi: 10.1109/ISS1.2017.8389229.
40. Janghel RR, Shukla A, Tiwari R, Kala R. Breast cancer diagnosis using artificial neural network models. The 3rd International Conference on Information Sciences and Interaction Sciences; Chengdu, China: IEEE; 2010. p. 89-94. doi: 10.1109/ICICIS.2010.5534716.
41. Venkatesan E, Velmurugan T. Performance analysis of decision tree algorithms for breast cancer classification. *Indian Journal of Science and Technology.* 2015;**8**(29):1-8. doi: 10.17485/ijst/2015/v8i29/84646.
42. Devarriya D, Gulati C, Mansharamani V, Sakalle A, Bhardwaj A. Unbalanced breast cancer data classification using novel fitness functions in genetic programming. *Expert Systems with Applications.* 2020;**140**:112866. doi: 10.1016/j.eswa.2019.112866.
43. Chiesa M, Maioli G, Colombo GI, Piacentini L. GARS: Genetic Algorithm for the identification of a Robust Subset of features in high-dimensional datasets. *BMC Bioinformatics.* 2020;**21**(1):54. doi: 10.1186/s12859-020-3400-6. PubMed PMID: 32046651. PubMed PMCID: PMC7014945.
44. Kim G, Kim S, Turbo Tek SK. Feature selection using genetic algorithms for handwritten character recognition. Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition; Amsterdam, Nijmegen: International Unipen Foundation; 2002. p. 103-12.
45. Guo J, White J, Wang G, Li J, Wang Y. A genetic algorithm for optimized feature selection with resource constraints in software product lines. *Journal of Systems and Software.* 2011;**84**(12):2208-21.
46. Lavanya D, Rani KU. Ensemble decision tree classifier for breast cancer data. *International Journal of Information Technology Convergence and Services.* 2012;**2**(1):17-24.
47. Rosner B, Tamimi RM, Kraft P, Gao C, et al. Simplified Breast Risk Tool Integrating Questionnaire Risk Factors, Mammographic Density, and Polygenic Risk Score: Development and Validation. *Cancer Epidemiol Biomarkers Prev.* 2021;**30**(4):600-7. doi: 10.1158/1055-9965.EPI-20-0900. PubMed PMID: 33277321. PubMed PMCID: PMC8026588.
48. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. *Stat Med.* 2004;**23**(7):1111-30. doi: 10.1002/sim.1668. PubMed PMID: 15057881.
49. Conant EF, Barlow WE, Herschorn SD, Weaver DL, Beaber EF, et al. Association of Digital Breast Tomosynthesis vs Digital Mammography With Cancer Detection and Recall Rates by Age and Breast Density. *JAMA Oncol.* 2019;**5**(5):635-42. doi: 10.1001/jamaoncol.2018.7078. PubMed PMID: 30816931. PubMed PMCID: PMC6512257.
50. Chow S, Raine-Bennett T, Samant ND, Postlethwaite DA, Holzapfel M. Breast cancer risk after hysterectomy with and without salpingo-oophorectomy for benign indications. *Am J Obstet Gynecol.* 2020;**223**(6):900.e1-7. doi: 10.1016/j.ajog.2020.06.040. PubMed PMID: 32585221.
51. Raiesdana S. Breast Cancer Detection Using Optimization-Based Feature Pruning and Classification Algorithms. *Middle East Journal of Cancer*. 2021;**12**(1):48-68. doi: 10.30476/MEJC.2020.85601.1294.
52. Mohan S, Bhattacharya S, Kaluri R, Feng G, Tariq U. Multi-modal prediction of breast cancer using particle swarm optimization with non-dominating sorting. *International Journal of Distributed Sensor Networks.* 2020;**16**(11). doi: 10.1177/1550147720971505.
53. Sakri SB, Rashid NB, Zain ZM. Particle swarm optimization feature selection for breast cancer recurrence prediction. *IEEE Access*. 2018;**6**:29637-47. doi: 10.1109/ACCESS.2018.2843443.
54. Kumar K, Singh VV, Ramaswamy R. Different Perspective of Machine Learning Technique to Better Predict Breast Cancer Survival. *BioRxiv*. 2020. doi: 10.1101/2020.07.03.186890.
55. Thawkar S, Ingolikar R. Classification of masses in digital mammograms using the genetic en-

semble method. *Journal of Intelligent Systems.* 2020;**29**(1):831-45. doi: 10.1515/jisys-2018-0091.

56. Bayrak EA, Kırcı P, Ensari T. Comparison of machine learning methods for breast cancer diagnosis. 2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT); Istanbul, Turkey: IEEE; 2019. p. 1-3. doi: 10.1109/EBBT.2019.8741990.
57. Alghunaim S, Al-Baity HH. On the scalability of machine-learning algorithms for breast cancer prediction in big data context. *IEEE Access*. 2019;**7**:91535-46. doi: 10.1109/ACCESS.2019.2927080.
58. Kumar GR, Ramachandra GA, Nagamani K. An efficient prediction of breast cancer data using data mining techniques. *International Journal of Innovations in Engineering and Technology (IJIET)*. 2013;2(4):139.
59. Aruna S, Rajagopalan SP. A novel SVM based CSSFFS feature selection algorithm for detecting breast cancer. *International Journal of Computer Applications.* 2011;**31**(8):14-20.
60. Memon MH, Li JP, Haq AU, Memon MH, Zhou W. Breast cancer detection in the IOT health environment using modified recursive feature selection. *Wireless Communications and Mobile Computing*. 2019;**2019**:1-19. doi: 10.1155/2019/5176705.
61. Sun YS, Zhao Z, Yang ZN, Xu F, Lu HJ, et al. Risk Factors and Preventions of Breast Cancer. *Int J Biol Sci.* 2017;**13**(11):1387. doi: 10.7150/ijbs.21635. PubMed PMID: 29209143. PubMed PMCID: PMC5715522.
62. 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. doi: 10.1016/j.procs.2016.04.224.
63. Ayvaci MU, Alagoz O, Chhatwal J, Munoz del Rio A, Sickles EA, Nassif H, Kerlikowske K, Burnside ES. Predicting invasive breast cancer versus DCIS in different age groups. *BMC Cancer.* 2014;**14**:584. doi: 10.1186/1471-2407-14-584. PubMed PMID: 25112586. PubMed PMCID: PMC4138370.
64. Rajendran K, Jayabalan M, Thiruchelvam V. Predicting breast cancer via supervised machine learning methods on class imbalanced data. *Int J Adv Comput Sci Appl.* 2020;**11**(8):54-63. doi: 10.14569/IJACSA.2020.0110808.
65. Atashi A, Sohrabi S, Dadashi A. Applying two computational classification methods to predict the risk of breast cancer: A comparative study. *Multidisciplinary Cancer Investigation.* 2018;**2**(2):8-13. doi: 10.30699/acadpub.mci.2.2.8.
66. Mosayebi A, Mojaradi B, Bonyadi Naeini A, Khodadad Hosseini SH. Modeling and comparing data mining algorithms for prediction of recurrence of breast cancer. *PLoS One.* 2020;**15**(10):e0237658. doi: 10.1371/journal.pone.0237658. PubMed PMID: 33057328. PubMed PMCID: PMC7561198.
67. Jalali SM, Moro S, Mahmoudi MR, Ghaffary KA, Maleki M, Alidoostan A. A comparative analysis of classifiers in cancer prediction using multiple data mining techniques. *International Journal of Business Intelligence and Systems Engineering.* 2017;**1**(2):166-78. doi: 10.1504/IJBISE.2017.10009655.
68. Lotfnezhad Afshar H, Jabbari N, Khalkhali HR, Esnaashari O. Prediction of Breast Cancer Survival by Machine Learning Methods: An Application of Multiple Imputation. *Iran J Public Health*. 2021;**50**(3):598-605. doi: 10.18502/ijph.v50i3.5606. PubMed PMID: 34178808. PubMed PMCID: PMC8214598.
69. Nourelahi M, Zamani A, Talei A, Tahmasebi S. A model to predict breast cancer survivability using logistic regression. *Middle East Journal of Cancer.* 2019;**10**(2):132-8. doi: 10.30476/MEJC.2019.78569.
70. Tapak L, Shirmohammadi-Khorram N, Amini P, Alafchi B, Hamidi O, Poorolajal J. Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. *Clinical Epidemiology and Global Health*. 2019;**7**(3):293-9. doi: 10.1016/j.cegh.2018.10.003.