Using Data Mining Techniques for Early Diagnosis of Breast Cancer

Sorayya Rezayi
Ph.D. candidate in Medical Informatics, Department of Health Information Management and Medical Informatics, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran.

Keivan Maghooli
Department of Biomedical Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.

Soheila Saeedi ✉
Ph.D. candidate in Medical Informatics, Department of Health Information Management and Medical Informatics, School of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran. Email: saeedi_s@razi.tums.ac.ir. ORCID: https://orcid.org/ 0000-0003-1315-794X

Citation: Rezayi S, Maghooli K, Saeedi S. **Using Data Mining Techniques for Early Diagnosis of Breast Cancer.** Applied Health Information Technology 2022; 3(1): 14-24.

Received: 2022-04-05
Accepted: 2022-07-19

Abstract

Aim: The present study aimed to compare six data mining approaches and find the best methods for predicting breast cancer.

Method: In this study, six classification methods, including Random Forest (RF), Neural Network (NN), Support Vector Machine (SVM), Auto Multilayer Perceptron (AutoMLP), Naïve Bayes (NB), and Deep Learning (DL) were applied for breast cancer detection. Data related to 116 patients and healthy people from the UCI repository with nine predictors were used for training and testing. To develop the model, data were first divided into two parts: training and testing. The data of the training set (70%) produced the models, and the data of the test set (30%) was applied to validate the models.

Results: To compare the performance of the techniques used to diagnose breast cancer, accuracy, recall, precision, AUC (Area Under the ROC Curve), sensitivity, and specificity were calculated and reported for all approaches. Evaluation of data mining algorithms revealed that deep learning with 81.89% accuracy performed better than other techniques. The results of one-way ANOVA for performance in six modeling methods showed no statistically significant difference between the methods (P-value <0.05).

Conclusion: Choosing the most effective computer diagnostic methods can provide a comprehensive system for the early detection of breast cancer. By reducing the cost of treating patients and increasing the quality of services offered, these intelligent methods take practical steps to improve medicine and lead to a systematic diagnosis.

Keywords: Data Mining; Breast Cancer; Neural Network; Deep Learning

Breast cancer is a type of cancer that begins in the breast tissue and occurs due to the abnormal growth of abnormal cells in the breast. It is safe to say that breast cancer is the second most common cause of death among women today after skin cancer. It is the most common cancer diagnosed in women worldwide. Over the period from 2012-2016, the breast cancer incidence rate increased slightly by 0.3% per year, mainly because of ascending rates of local stage and hormone receptor-positive disease (1). Breast cancer exhibits symptoms, but most people are unaware of some of these symptoms in the body and ignore the risk factors. Although there is no need to worry unnecessarily and obsessively about breast cancer, the symptoms of breast cancer should not be overlooked (1). This is because breast cancer is the deadliest cancer in women. The outcome of breast cancer treatment varies according to the severity of the disease.

If the disease is diagnosed in the early stages, the percentage of treatment will typically increase (2). However, early detection of breast cancer before metastasis to various parts of the body and timely treatment will increase the life expectancy of patients and improve the quality of life of women with cancer (3).

It is generally believed that hidden and valuable knowledge is concealed within health data (4). The volume of medical data increases daily, and physicians are often required to obtain valuable information about diseases in particular and their relationship to other disease-causing agents. However, this raw data set possesses no value in itself. To make sense of these data, experts must analyze them, turn them into information, and better know them. (5). Due to the prevalence of various diseases worldwide, utilizing new artificial intelligence-oriented biomedical research methods has received much attention. One of the most popular artificial intelligence-based algorithms that can detect hidden patterns in health-related data is known as data mining; using data mining algorithms to develop predictive models is extremely helpful in identifying high-risk individuals to reduce the complications of the disease (5, 6). Data mining, in general, is the process of selecting, analyzing, exploring, and modeling a massive amount of data to discover unknown patterns and create prediction models and automated decision systems (7). However, the specific purpose of data mining is to discover valid, new, and traceable patterns in a huge amount of data using statistical tools and artificial intelligence (8). Over the previous decades, in many studies, data mining techniques were applied in different fields of medicine to better understand the disease causes, discover new treatment methods, and devise prediction models (9). Employing various data mining techniques provides high-precision models which can be used as decision-making systems. Numerous data mining methods like artificial neural networks, Bayesian networks, support vector machines, decision trees, deep learning, and bagging algorithm have been utilized actively in clinical decision support systems for early diagnosis of breast neoplasms (10). Many studies have focused on the confirmed diagnosis of breast cancer using data mining techniques. These studies have used different data mining approaches and achieved different results. Details of several studies are provided in this section.

Chaurasia et al. used data mining techniques to diagnose benign and malignant breast cancer. In this study, three widely used data mining techniques (NB, Radial Basis Function Network (RBFN, and J48) have been selected to develop prediction models. The Breast Cancer Wisconsin dataset was used with 683 samples and nine features. WEKA version 3.6.9 was chosen as a tool for dataset analysis and evaluation of data mining techniques. A 10-fold cross-validation method has been applied to compare the performance of prediction models. The results of evaluating data mining techniques illustrated that NB had the best performance (97.36% accuracy) compared to other methods (11).

Alshammari and Mezher, in their study, compared the performance of various data mining algorithms including NB, Logistic regression (LR), Lazy Instance-Bases learning with parameter K (IBK), Lazy Kstar, Lazy Locally weighted learner, Rules ZeroR, Decision stump, Decision tree (DT: J48), RF, and Random tree (RT). In this study, WEKA tool was used to perform data mining on an academic experimental breast cancer dataset. The results showed the Lazy IBK classifier KNN had the best accuracy (98%) among other classifiers (12).

Nalini and Meera proposed two common data mining algorithms for diagnosing breast cancer. In this study, Bayesian network and J48 algorithms as supervised learning methods have been used. Seven hundred and sixty-eight

instances with eight features have been used in this comparative analysis. This comparison was performed in the WEKA tool environment. Comparison of J48 and NB prediction models have been performed based on classification accuracy and execution time. As a result, Naïve Bayes performed better than the J48 algorithm (13).

Asri et al. used the Wisconsin Breast Cancer (original) dataset from the UCI Machine Learning Repository to assign breast cancer patients into benign and malignant groups using data mining approaches. The Wisconsin Breast Cancer dataset includes 699 instances, two classes, and 11 integer-valued features. Four data mining approaches were selected to categorize patients, including SVM, DT: C4.5, NB, and K-Nearest Neighbors (K-NN).The results indicated that SVM had the best performance in classifying patients into two groups (14).

In this study, the main aim was to determine the performance of data mining techniques in breast cancer diagnosis. The novelty of this study is that it uses six modeling techniques to diagnose breast cancer, including RF, NN, SVM, AutoMLP, NB, and DL. Because clinicians cannot use all these methods in the clinical environment and perform trial and error, authors can introduce the best algorithm and help clinicians choose the proper method by comparing the performance of these six techniques. In this paper, a proposed model based on Cross-Industry Standard Process for Data Mining (CRISP) is presented, which consists of five steps. Each of these steps includes sub-sections. Moving back and forth between different steps is required because the input of each step depends on the output of the previous step. Each of these five steps is addressed in Figure 1.

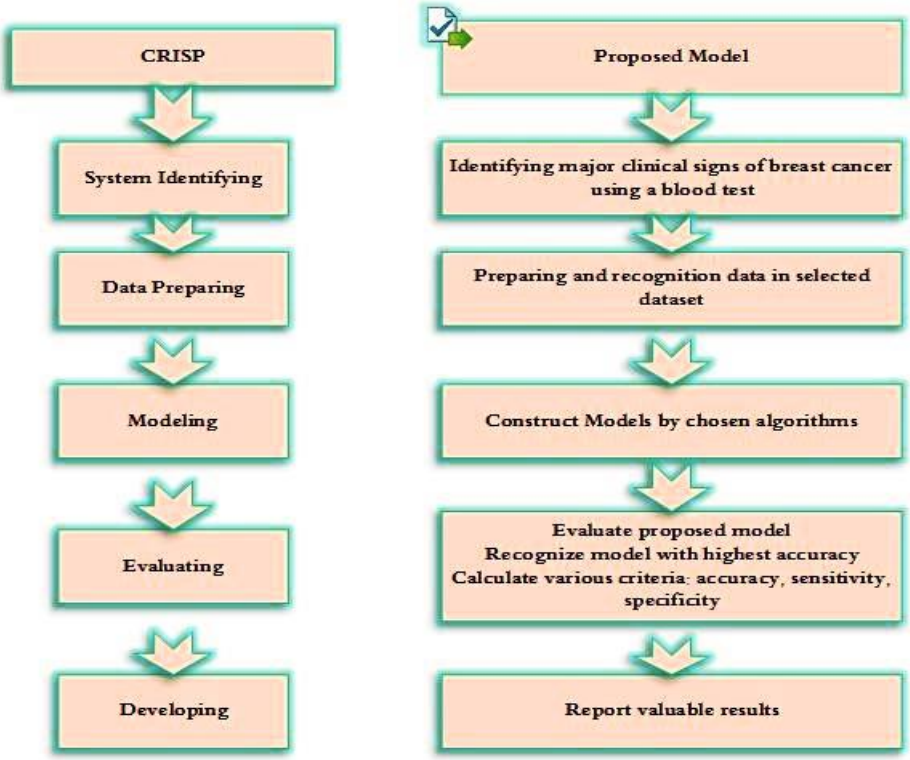

Figure 1: CRISP method steps and proposed model

Method

- ***Dataset***

Due to researchers' lack of access to patients' data in clinical settings,, they used the data available in the UCI repository in this study. These data will help recognize whether a person has breast cancer or not based on several pre-defined standard signs. There are nine quantitative predictors and binary dependent variables that indicate the presence or absence of breast neoplasm. These independent predictors were anthropometric data and specific parameters which can be gathered in routine blood analysis. Classification models based on these predictors can be used as a potential biomarker of breast cancer if accurately identified and categorized. The chosen dataset typically involved ten critical variables; the detailed description of quantitative attributes is as follows:

1) Age (years old), 2) BMI (kg/m2), 3) Glucose (mg/dL), 4) Insulin (μU/mL), 5) HOMA (The homeostatic model assessment), 6) Leptin (ng/mL), 7) Adiponectin (μg/mL), 8) Resistin (ng/mL), 9) MCP-1 (Monocyte Chemoattractant protein-1: pg/dL), 10) Classifications (1=healthy controls, 2=patients**)**

Based on all these categorical features, each label's unique combination would be efficiently generated in the independent variable; therefore, in total, 116 instances or unique combinations were produced.

This dataset has some missing values; the authors have replaced them with the average for numerical features.

- ***Applied algorithms***

There are many data mining methods for modeling. In this step, through using different data mining techniques, researchers produced the optimal model and pattern. Classification as one of the most well-known data mining methods typically consists of two stages. In the first stage, which is called inference, the goal is to discover a model for defining predefined data categories. The model is created based on training examples provided to the system. The inference algorithm creates a definition for that particular category by using the characteristic values of the samples belonging to each category (15-17). In the second stage, which is called prediction, for samples that do not belong to a specific category, their label can be predicted based on the inferred model. In this study, to investigate and analyze the occurrence of signs in people with breast cancer, the authors used supervised classification methods such as RF, NN, SVM AutoMLP, NB, and DL.

- ***Support vector machines***

SVMs are one of the supervised learning methods used for classification and regression. This method is one of the relatively recent methods with good performance in recent years compared to older methods for classification (18). The basis of SVM classifier is a linear classification of data. For linear segmentation of data, the authors tried to choose the line that had the most reliable margin. Solving the equation and finding the optimal line for data was done by Quadratic Programming (QP) methods, which are known methods for solving constrained problems (19).

- ***Naive Bayes classifier***

NB classifier in machine learning, a simple classifier based on probabilities, is based on the assumption of the independence of random variables according to the Bayesian theorem (20). The Bayesian method is simply a method of classifying phenomena based on the probability of a phenomenon occurring or not. This method is one of the simplest forecasting algorithms with acceptable accuracy (21).

- ***Random forest***

RF is an easy-to-use machine learning algorithm that often delivers great results even without adjusting its meta-parameters (22). Due to its simplicity and usability, this algorithm is

one of the most widely used machine learning algorithms for both "classification" and "regression". The leading idea of the bagging method is that a combination of learning models enhances the overall results of the model. Simply put, an RF makes several decision trees and merges them to provide more accurate and consistent predictions (23).

- ***Neural Net, AutoMLP, and Deep learning***

The NN methods are used for classification, clustering, mining characteristics, prediction, and pattern recognition**.** An NN includes hidden layers of interconnected nodes or neurons (24). Each node is a perceptron; the perceptron feeds the signal produced by a multiple linear regression into an activation. Remarkably, the perceptron is a machine learning algorithm that falls into the category of supervised learning techniques. The perceptron algorithm is a binary classification algorithm (a classifier that can recognize whether an input belongs to a specific category based on the input vector). These algorithms are linear classifiers, meaning that they make their predictions according to the weighted linear composition of the algorithm input (25). Besides, they are online algorithms because they examine their inputs one at a time. In an MLP, perceptrons are arranged in interconnected layers or hidden layers. The input layer can assemble input patterns; the output layer has output signals or classifications to map input patterns. Hidden layers set the input weightings until the margin of error in the neural network represents minimal (26). DL is based on a multi-layer, feed-forward, and artificial neural network trained using back-propagation with stochastic gradient descent. This network can include several hidden layers consisting of neurons with the hyperbolic tangent, maxout, and rectifier linear activation functions (27, 28).

- ***Training and testing step***

To develop the model, data were first divided into two parts: training and testing. The data of the training sets (70%) produce the models, and the data of the test sets (30%) (29) evaluate the model's performance and determine the label related to the mentioned samples of instances. To train algorithms, a class variable must include an output field and one or more input fields. The disease existence is defined as the class label in the chosen dataset,, and other attributes are applied as input. It should additionally be noted that all the samples or combinations are completed in binary form (zeros and ones).

The parameters used for each of the algorithms in this study are as follows:

1) RF (the criterion was gain ratio, maximal depth=10, and the number of trees=100.)

2) AutoMPL (training cycle=10 and generation= 10)

3) NN (training cycles= 200, training rate=0.01, momentum=0.9 and hidden layer=1)

4) NB

5) SVM (kernel cache=200 and maximum iteration=200000, Gaussian radial basis function (RBF) kernel function, set the parameter C to 1.)

6) Multi-layer, feed-forward neural network: H2O's DL (activation=rectifier, epochs=10, three hidden layers with 50, 50, and 50 neurons)

- Evaluation step

After modeling, the performance should be evaluated. Evaluation improves the model and makes it usable. There are various indicators such as accuracy, sensitivity, precision, and specificity to evaluate classification methods. A confusion matrix can be used to calculate the number of indicators; this matrix is a valuable tool for analyzing how the classification method identifies data or observes different categories. Ideally, most of the data associated with observations should be on the original diameter

of the matrix, and the rest of the matrix values should be zero or near zero.

True-negative (TN) indicates the number of records whose real category is negative, and the classification algorithm has correctly identified the category as negative.

True-positive (TP) indicates the number of records whose actual category is positive, and the classification algorithm has correctly identified the category as positive.

False-positive (FP) indicates the number of records whose actual category is negative and the classification algorithm has erroneously detected the positive category.

False-negative (FN) indicates the number of records whose actual category is positive, and the classification algorithm has erroneously detected the positive category (30-32).

Sensitivity = TP/ TP+FN (1)

Specificity = TN / FP+TN (2)

Accuracy = TP +TN / TP+TN+FP+FN (3)

Precision = TP / TP+FP (4)

Results

Six classification algorithms and classifiers were considered for diagnosing whether a person has breast cancer or not based on several main extracted attributes. First, the authors presented a statistical interpretation of the features of the selected dataset in Figure 2. Data in the chosen dataset were divided into the training set and test set. The training set was utilized to build the classifier, and the test set was applied to validate the model. The validation number was equal to 10-fold-cross-validation. Training and testing steps were conducted with a sample ratio (0.7), or split size (70% training and 30% testing in each cross-validation), and stratified sampling type. The related results of experiments are presented in Table 1.

All the essential indicators were evaluated to investigate the performance of the created models. These indicators are placed in the table with the micro-average details. Table 1 presents the applied six model fitness in cross-validation folds.

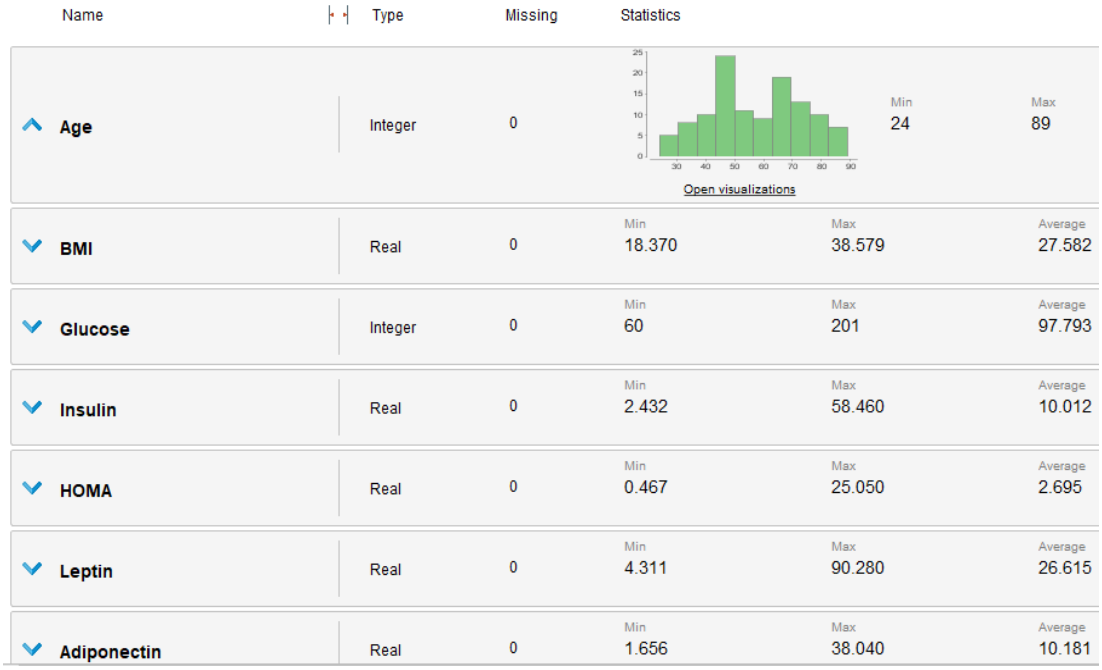

Figure 2: Breast cancer features and their values to classify patients

Table 1: Comparison of model fitness and main specific metrics of six applied classifiers

Classifiers	RF	SVM	NB
Accuracy	79.16% +/- 5.00% (micro-average: 69.29%)	76.78% +/- 5.90% (micro-average: 66.61%)	72.30% +/- 6.81% (micro-average: 62.50%)
Recall	72.42% +/- 11.06% (micro-average: 71.90%)	61.91% +/- 10.27% (micro-average: 61.61%)	41.97% +/- 10.39% (micro-average: 41.64%)
Precision	72.48% +/- 8.82% (micro-average: 71.90%)	75.39% +/- 12.29% (micro-average: 72.35%)	78.70% +/- 7.53% (micro-average: 78.40%)
AUC	0.774 +/- 0.044 (micro-average: 0.774)	0.774 +/- 0.044 (micro-average: 0.774)	0.725 +/- 0.048 (micro-average: 0.725)
Sensitivity	72.42% +/- 11.06% (micro-average: 71.90%)	61.91% +/- 10.27% (micro-average: 61.61%)	41.97% +/- 10.39% (micro-average: 41.64%)
Specificity	66.96% +/- 10.34% (micro-average: 66.14%)	74.58% +/- 17.12% (micro-average: 72.45%)	86.68% +/- 5.95% (micro-average: 86.69%)
	NN	**AutoMLP**	**DL**
Accuracy	79.42% +/- 5.29% (micro-average: 69.51%)	78.07% +/- 6.34% (micro-average: 68.24%)	81.89% +/- 4.10% (micro-average: 71.83%)
Recall	70.04% +/- 8.31% (micro-average: 69.21%)	70.38% +/- 14.55% (micro-average: 70.42%)	69.65% +/- 6.42% (micro-average: 69.84%)
Precision	74.69% +/- 11.14% (micro-average: 73.65%)	71.16% +/- 7.15% (micro-average: 70.87%)	76.48% +/- 7.77% (micro-average: 75.80%)
AUC	0.777 +/- 0.046 (micro-average: 0.777)	0.751 +/- 0.060 (micro-average: 0.751)	0.795 +/- 0.048 (micro-average: 0.795)
Sensitivity	70.04% +/- 8.31% (micro-average: 69.21%)	70.38% +/- 14.55% (micro-average: 70.42%)	69.65% +/- 6.42% (micro-average: 69.84%)
Specificity	70.76% +/- 13.84% (micro-average: 69.88%)	65.81% +/- 11.38% (micro-average: 65.65%)	74.57% +/- 10.01% (micro-average: 74.14%)

Comparative ROC curves based on attributes of breast cancer are shown in Figures 3 and 4. DL, SVM, and RF had the maximum AUC scores (Figure 3). As it is clear, one ROC curve for DL had outperformed with AUC 0.795. Overall, the results of AUC ensure better performance of DL, SVM, and RF classification algorithms. In addition, in Figure 4, the comparative ROC curves for NB, NN, and Auto MLP are presented.

The results of one-way ANOVA regarding performance in six modeling methods showed no statistically significant difference between the methods (P-value <0.05), (Table 2).

Table 2: The results of one-way analysis of variance (ANOVA) for comparing the performance of methods

Performance	ANOVA				
	Sum of squares	df	Mean square	F	Sig.
Between groups	338.533	5	67.707	.815	.548
Within groups	2492.099	30	83.070		
Total	2830.632	35			

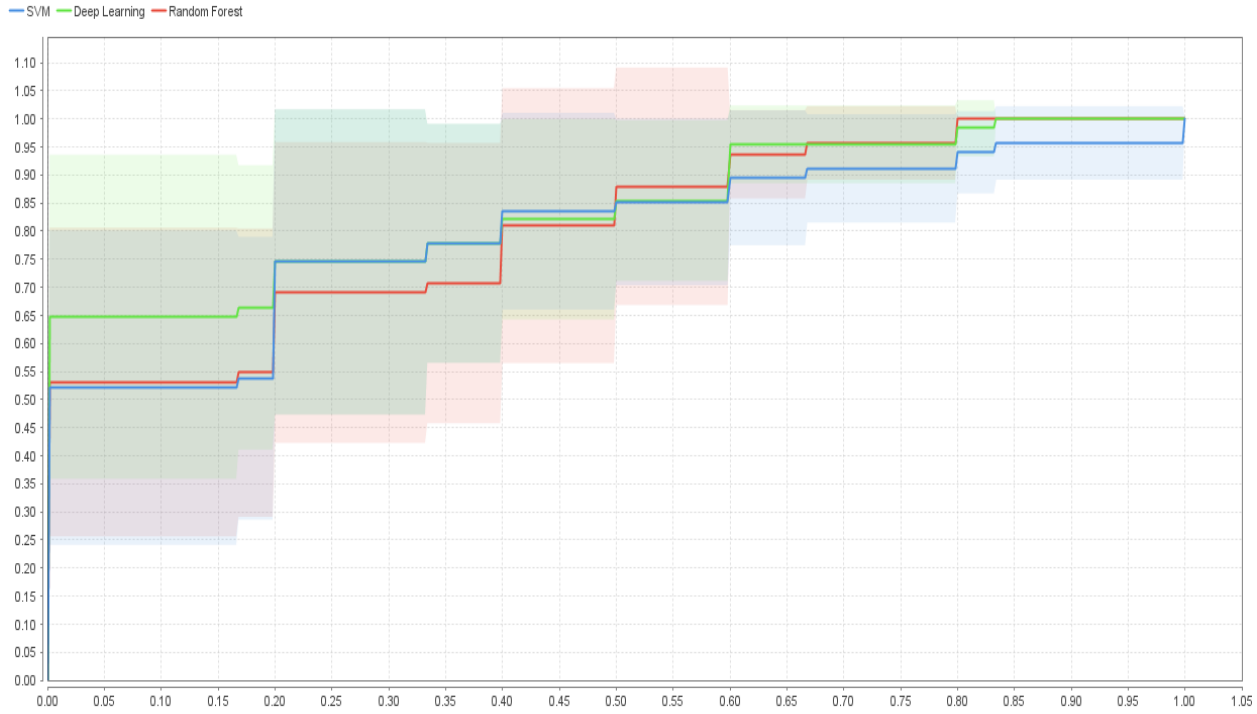

Figure 3: ROC curve for RF, SVM, and DL

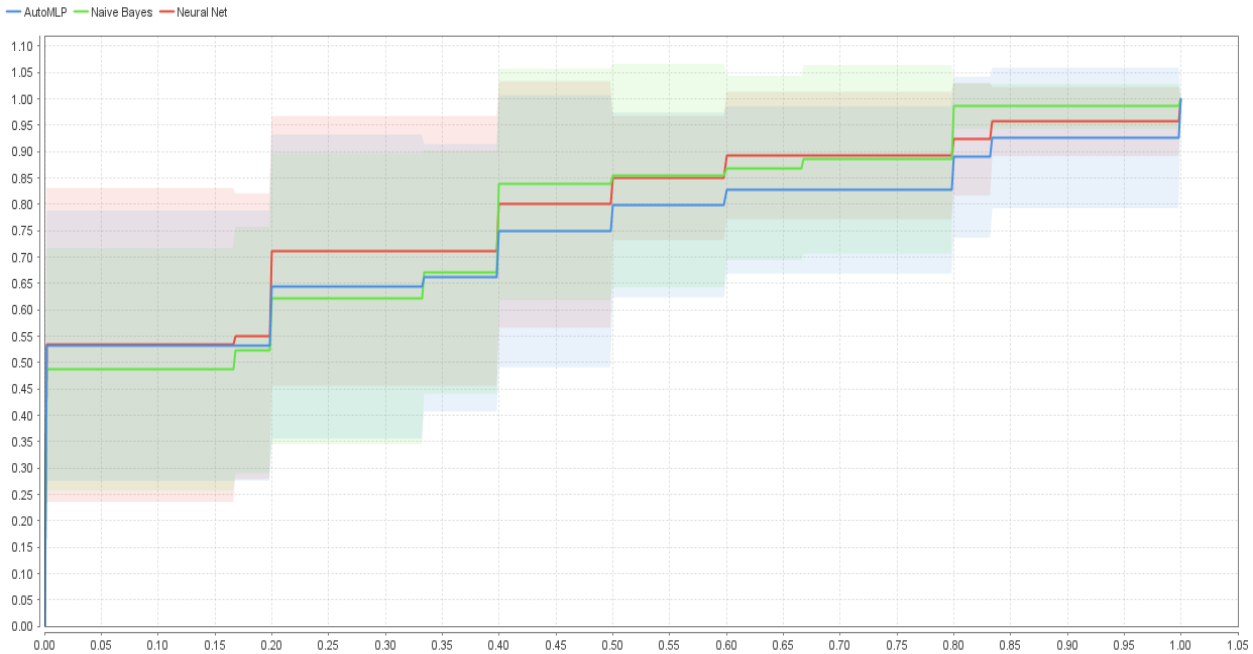

Figure 4: ROC curve for NB, NN, and AutoMLP

Discussion

This study aimed to use data mining approaches to diagnose patients with breast cancer. The output of the clinical decision support system was to classify individuals into two groups: patient and healthy. The input of the system was data related to 116 instances from the UCI repository with nine predictor values. Data mining approaches applied for classification included RF, NN, SVM, AutoMLP, NB, and DL.

The evaluation results demonstrated that DL had the best accuracy and AUC compared to other methods.

In a study by Diz et al. on breast cancer diagnosis, data mining algorithms including K-NN, SVMs, DT: J48 and RF, and NB have been selected for classification (33). In this study, NB had the best performance (83.1%) in detecting the masses texture among the different data mining approaches. The Diz study's results are not consistent with the present study. In this research, the NB classifier had the lowest accuracy among the different approaches. This classifier has several positive features, such as its simplicity and efficiency. On the other hand, one of the weaknesses of this method is that the attribute independence assumption harms its classification performance (34, 35). One way to increase the accuracy of this method is to combine this method with other classification methods such as DT or solve this problem with attribute weighting (34, 35).

The study by Kaladhar et al. had similar results to the present study. NB had the lowest accuracy compared to other methods .In this study, the mentioned method had the worst performance with 41% accuracy compared to Classification And Regression Trees (CART), RF, and Logistic Model Tree (LMT) methods (36).

In the present study, DL performed best in diagnosing breast cancer. DL, a method based on artificial neural networks, has emerged as a powerful tool for machine learning in recent years. Besides, DL can be used to solve problems that are difficult to solve with traditional artificial intelligence methods. This technique is a compelling approach that can be characterized by high computing power, high-speed data storage, parallelism, the semantic interpretation of input data, high power in predicting, and automatic creation of optimized high-level features (37). These features have led to the rapid acceptance of this technique.

The capability of this method to automatically create a set of features without human intervention can have many benefits and applications. For example, in the field of medical imaging, this approach can be used to create these very complex features (38). Another example of the application of this technique is in diagnosing tumors using the detection of irregularities in tissue morphology (39, 40).

A systematic review conducted by Bakator and Radosav on the application of DL in medical diagnosis has observed that the DL method is superior to other techniques that perform very well (41).

The performance of deep learning is compared with other machine learning methods in a study by Tabares-Soto. The results of this study showed that DL performed better than KNN, SVM, NB, DT, and K-means methods (42).

Therefore, based on the results of this study, DL outperformed other data mining methods. A review of existing studies shows that data mining has been used to diagnose breast cancer so far. However, one of the strengths of the present study is using standard data for the data mining process. In this study, a sufficient number of (six techniques) classification algorithms for disease identification were evaluated and compared using specific indicators. Results suggested that DL methods were the best classification between healthy and ill people. Therefore, this algorithm can be used in designing a decision support system by a specialist.

Conclusion

Data mining of medical data is enormously important, and designing decision support systems to assist physicians in diagnosing specific diseases with the help of data mining can help save human lives. In this regard, in the present study, deep learning had the best results for identifying breast cancer patients. The results also revealed that the accuracy of

predicting DL techniques, NN, and RF on the studied data is higher than other widely used and popular methods. The selection of computer diagnostic methods with optimal performance can provide a comprehensive system for the early detection of breast cancer. Thus, by reducing the cost of patient treatment and increasing the quality of services provided, a practical step can be taken to systematically improve medical diagnosis.

Disclaimer Statements

- **Conflict of interest:** Authors declared no conflict of interest.
- **Financial Support:** In this paper, authors did not have any financial support.
- **Authors' contributions:** Authors SR, KM, and SS wrote the first draft of the manuscript. Authors SR, KM, and SS performed dataset selection, analysis, and extracted main characteristics. All authors reviewed, provided critical feedback. All authors read and approved the final manuscript
- **Protection of human and animal subjects:** Not applicable.

References

1. Dafni U, Tsourti Z, Alatsathianos I. Breast cancer statistics in the European Union: incidence and survival across European countries. Breast Care. 2019;14(6):344-53.
2. Ahmad A. Breast cancer statistics: recent trends. Breast Cancer Metastasis and Drug Resistance: Springer; 2019. 1-7.
3. Ghislain I, Zikos E, Coens C, Quinten C, Balta V, Tryfonidis K, et al. Health-related quality of life in locally advanced and metastatic breast cancer: methodological and clinical issues in randomised controlled trials. The Lancet Oncology. 2016;17(7):e294-e304.
4. Amin MS, Chiam YK, Varathan KD. Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics. 2019;36:82-93.
5. Sahani R, Rout C, Badajena JC, Jena AK, Das H. Classification of intrusion detection using data mining techniques. Progress in computing, analytics and networking: Springer; 2018. p. 753-64.
6. Peral J, Maté A, Marco M. Application of data mining techniques to identify relevant key performance indicators. Computer Standards & Interfaces. 2017;54:76-85.
7. Miao F, Fu N, Zhang Y-T, Ding X-R, Hong X, He Q, et al. A novel continuous blood pressure estimation approach based on data mining techniques. IEEE journal of biomedical and health informatics. 2017;21(6):1730-40.
8. Salo F, Injadat M, Nassif AB, Shami A, Essex A. Data mining techniques in intrusion detection systems: A systematic literature review. IEEE Access. 2018;6:56046-58.
9. Nalluri S, Saraswathi RV, Ramasubbareddy S, Govinda K, Swetha E. Chronic Heart Disease Prediction Using Data Mining Techniques. Data Engineering and Communication Technology: Springer; 2020. p. 903-12.
10. Abdelghafar S, Darwish A, Hassanien AE. Intelligent health monitoring systems for space missions based on data mining techniques. Machine learning and data mining in aerospace technology: Springer; 2020. p. 65-78.
11. Chaurasia V, Pal S, Tiwari B. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology. 2018;12(2):119-26.
12. Alshammari M, Mezher M. A Comparative Analysis of Data Mining Techniques on Breast Cancer Diagnosis Data using WEKA Toolbox. (IJACSA) International Journal of Advanced Computer Science and Applications. 2020:224-9.
13. Meera C, Nalini D. Breast cancer prediction system using Data mining methods. International Journal of Pure and Applied Mathematics. 2018;119(12):10901-11.
14. 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.
15. Cios KJ, Pedrycz W, Swiniarski RW. Data mining methods for knowledge discovery. Springer Science & Business Media; 2012 Dec 6.
16. Larose DT, Larose DT. Data mining methods and models: Wiley Online Library; 2006.
17. Lei-da Chen TS, Frolick MN. Data mining methods, applications, and tools. Information systems management. 2000;17(1):67-8.
18. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intelligent Systems and their applications. 1998;13(4):18-28.
19. Mangasarian OL, Musicant DR, editors. Active support vector machine classification. Advances in neural information processing systems; 2001.
20. Saritas MM, Yasar A. Performance analysis of ANN and Naive Bayes classification algorithm for data classification. International Journal of Intelligent Systems and Applications in Engineering. 2019;7(2):88-91.

21. Rish I. An empirical study of the naive Bayes classifier. InIJCAI 2001 workshop on empirical methods in artificial intelligence 2001 Aug 4 (Vol. 3, No. 22, pp. 41-46).
22. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18-22.
23. Oshiro TM, Perez PS, Baranauskas JA. How many trees in a random forest?. InInternational workshop on machine learning and data mining in pattern recognition 2012 Jul 13 (pp. 154-168). Springer, Berlin, Heidelberg.
24. Hansen LK, Salamon P. Neural network ensembles. IEEE transactions on pattern analysis and machine intelligence. 1990;12(10): 993-1001.
25. Hung S-L, Adeli H. A model of perceptron learning with a hidden layer for engineering design. Neurocomputing. 1991;3(1):3-14.
26. Solla SA, Winther O. Optimal perceptron learning: an online Bayesian approach. On-Line Learning in Neural Networks. Cambridge University Press, Cambridge. 1998 Apr.
27. Lee CY, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-supervised nets. InArtificial intelligence and statistics 2015 Feb 21 (pp. 562-570). PMLR.
28. Sun D, Yao A, Zhou A, Zhao H. Deeply-supervised knowledge synergy. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019 (pp. 6997-7006).
29. Gholamy A, Kreinovich V, Kosheleva O. Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. 2018.
30. Hossin M, Sulaiman M. A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process. 2015;5(2):1.
31. Zhang J, Hodge BM, Florita A, Lu S, Hamann HF, Banunarayanan V. Metrics for evaluating the accuracy of solar power forecasting. National Renewable Energy Lab.(NREL), Golden, CO (United States); 2013 Oct 1.
32. Zhu W, Zeng N, Wang N. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. NESUG proceedings: health care and life sciences, Baltimore, Maryland. 2010;19:67.
33. Diz J, Marreiros G, Freitas A. Applying data mining techniques to improve breast cancer diagnosis. Journal of medical systems. 2016;40(9):203.
34. Wu J, Cai Z. Attribute weighting via differential evolution algorithm for attribute weighted naive bayes (wnb). Journal of Computational Information Systems. 2011;7(5):1672-9.
35. Zaidi NA, Cerquides J, Carman MJ, Webb GI. Alleviating naive Bayes attribute independence assumption by attribute weighting. The Journal of Machine Learning Research. 2013;14(1): 1947-88.
36. Kaladhar D, Chandana B, Kumar PB. Predicting cancer survivability using Classification algorithms. International Journal of Research and Reviews in Computer Science. 2011;2(2):340.
37. Ravì D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep learning for health informatics. IEEE journal of biomedical and health informatics. 2016;21(1):4-21.
38. Raj RJS, Shobana SJ, Pustokhina IV, Pustokhin DA, Gupta D, Shankar K. Optimal Feature Selection-Based Medical Image Classification Using Deep Learning Model in Internet of Medical Things. IEEE Access. 2020;8:58006-17.
39. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep learning to improve breast cancer detection on screening mammography. Scientific reports. 2019;9(1): 1-12.
40. Fakoor R, Ladhak F, Nazi A, Huber M. Using deep learning to enhance cancer diagnosis and classification. InProceedings of the international conference on machine learning 2013 Jun (Vol. 28, pp. 3937-3949). ACM, New York, USA.
41. Bakator M, Radosav D. Deep learning and medical diagnosis: A review of literature. Multimodal Technologies and Interaction. 2018; 2(3):47.
42. Tabares-Soto R, Orozco-Arias S, Romero-Cano V, Bucheli VS, Rodríguez-Sotelo JL, Jiménez-Varón CF. A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data. PeerJ Computer Science. 2020;6:e270.