Hindawi
Mobile Information Systems
Volume 2022, Article ID 7340689, 8 pages
https://doi.org/10.1155/2022/7340689

Research Article

Application of Machine Learning Models to the Detection of Breast Cancer

Nasser Binsaif

Department of E-Commerce, Faculty College of Administrative and Financial Sciences, Saudi Electronic University, Riyadh, Saudi Arabia

Correspondence should be addressed to Nasser Binsaif; nbinsaif@seu.edu.sa

Received 20 January 2022; Revised 31 January 2022; Accepted 7 February 2022; Published 2 March 2022

Academic Editor: Hasan Ali Khattak

This work aims to build a binary breast cancer classifier algorithm based on the blood test and anthropometric data (age, body mass index, glucose, insulin, homeostasis model assessment, leptin, adiponectin, resistin, and monocyte chemotactic protein-1) of 116 subjects. For this study, a performance comparison of the following machine learning models was performed: decision tree, random forest, *K*-nearest neighbors, artificial neural networks, vector machines of support, and logistical regression. The methodologies used in the data were as follows: k-fold cross-validation (k = 10); splitting data into 80% training and 20% testing. For the first, the mean of accuracy and sensitivity were evaluated in the second, values of accuracy, sensitivity, specificity, and area under some tests. In addition, most mammograms are performed on benign tumors. With this, it is clear that these exams can use other tools to assist in decision-making, and machine learning can offer great utility and good cost/benefit in the diagnostic process of breast cancer. Many research papers for breast cancer biomarkers have been reported over the years. The present work will analyze the potential quantitative variables: age, receiver operating characteristic curve. Furthermore, the p value, Pearson correlation coefficient, and, depending on the input variable, the test only with variables with a significance threshold of 5% are computed from the normal distribution assessment (calculated from Kolmogorov–Smirnov test (KS test)) which were as follows: glucose, insulin, resistin, and homeostasis assessment model. As the best final classifier, the random forest was used in the training/test method and with nine variables, with 83.3% accuracy, 100% sensitivity, 64% specificity, and 0.881 of area under the curve.

1. Introduction

Breast cancer, common among women, will affect approximately 2.2 million people in 2020, accounting for 11.7% of cancer patients worldwide and 6.9% of cancer-related deaths in the same year [1]. As a result, early diagnosis is critical as the speed with which it is made is directly proportional to the patient's chances of healing [2, 3]. Even though most women are aware of the disease, apprehension about performing specific tests early (mammography, ultrasound, and self-examination) is expected, owing to a combination of factors, including a lack of standard recommendations, the absence of visible symptoms, and feelings of insecurity or fear [4, 5].

According to studies, between 10% and 30% of women diagnosed with breast cancer have benign tumors, indicating that some tests are ineffective or misinterpreted. Furthermore, the majority of mammograms are performed on benign tumors. With this in mind, it is clear that these exams can benefit from using other tools to aid decision-making, and machine learning can provide significant utility and a favorable cost-benefit ratio in the diagnostic process of breast cancer [1]. Over the years, several candidates for breast cancer biomarkers have been reported. The following quantitative variables will be examined in this study: age, BMI, glucose, insulin, homeostasis model assessment (HOMA), leptin, adiponectin, resistin, and monocyte chemotactic protein-1 (MCP-1) [6].

To address this public health issue, the study examines the performance of six machine learning algorithms for data classification: decision tree, random forest, *K*-nearest neighbors, support vector machines, artificial neural networks, and

logistic regression [7]. The purpose of this paper is to discuss the accuracy and efficiency of predicting the occurrence of breast cancer in individuals based on input variables [8, 9], which can be used as a diagnostic aid by the medical community. Thetic of breast cancer represents a quick and efficient solution that organizes patients so that more targeted measures can be taken, easing the doctors' work. It is also worth noting that the results obtained through the models used are not the individuals' final diagnoses.

1.1. Review of the Literature. Machine learning is being used in a wide range of fields in the twenty-first century, thanks to several advances in data analysis and classification techniques [10]. This method can detect breast cancer more accurately and at a lower cost. As a result, so-called biomarkers are frequently used as attributes in machine learning models to classify breast cancer. A log regression algorithm was developed in 2008, and it used two inputs: specific antigen 15-3 and insulin-like growth factor-binding protein-3. The receiver operating characteristic (ROC) metric produced an area under the curve (AUC) of 0.86, with a sensitivity of 85% and specificity of 62%.

In 2013, Dalamaga et al. [11] used serum resistin as a biomarker for postmenopausal breast cancer, finding an AUC of 0.71 with a 95% confidence interval. In 2015, the algorithms logistic regression, random forest, and support vector machine examined the same nine attributes used in this study. The Monte Carlo validation methodology was used [3]. The logistic regression model had 0.81 learning, 76% sensitivity, and 86% specificity on the ROC curve. These values were 0.83, 85%, and 77% for the random forest, respectively. Finally, the support vector machine model yielded 0.85, 81, and 84% results. However, the best results were obtained using only four variables, resistin, glucose, age, and BMI, which were better evaluated by the Gini coefficient [12].

Sensitivity was between 82 and 88%, while specificity was between 85 and 90% [13]. In 2021, Wen et al. [11] also applied machine learning to classify cancer through the attributes glucose, BMI, resistin, age, HOMA, leptin, and adiponectin. The classifier models were random forest and multiple logistic regressions. The second presented the best results, with 75% of accuracy and 0.849 of learning in the ROC curve.

2. Methodology

We used the Breast Cancer Database of Coimbra, made available by the UCI (University of California Irvine) repository, published on March 6, 2016 [5], which has data from 116 individuals, obtained through blood tests and anthropometry, 64 of whom were diagnosed with cancer (using mammography), and 52 were healthy.

The database has 116 rows (individuals) and ten columns, in which the last column corresponds to the output class (1 = healthy and 2 = patient), and the remaining ones are variable. Quantitative levels are as follows: age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistin, and MCP-1. In an attempt to reduce the number of input variables, which could imply the need to perform fewer tests on people, BioEstat 5.0 software was initially used. It was necessary to evaluate the normality of the attributes/variables. To this end, the Kolmogorov–Smirnov test was applied, with a significance level of 5%. The standard variables, whose p value was more significant than this level and which allowed the use of Pearson's correlation, were age, BMI, and MCP-1, while the abnormal ones, with p value lower than the level and which allowed the Spearman correlation, were glucose, insulin, HOMA, leptin, adiponectin, and resistin. Table 1 presents the mean (standard deviation) of the normal variables, as well as the median (interquartile range) for the abnormal ones; it also shows the p value (a 5% significance level was adopted) and the correlation coefficient (Pearson or Spearman, depending on the attribute). When evaluating the results and the test with all nine variables, it was also performed with 4, glucose, insulin, HOMA, and resistin, as they were the ones that presented p value in the correlation, less than 5%, with correlation coefficients above 0.2. The closest attribute was BMI, with a correlation of 0.1326, but with a p value exceeding the allowed by 10.59%. Another critical point is that the analysis of all algorithms was performed with the help of the Google platform, using the Python language [14].

2.1. Cross-Validation and Training/Testing Split. Two methodologies were applied for future performance evaluation of the algorithms. The first consisted of cross-validation using the k-fold technique, which used the entire database and consisted of performing "k" partitions (1 for testing and $k-1$ for training), alternating training and testing data for "k" times. $k = 10$ was used in all 6 classification models [15, 16]. In the second configuration, the database was divided into training and testing in the following proportion: 20% for testing and 80% for training, by the "split" command, using the parameters "stratify" and "random state," the latter being equal to 300 (random choice), to ensure that each time the code was initialized, the division was the same.

2.2. Performance Metrics. The confusion matrix (Table 2), for the addressed problem, consists of a 2×2 matrix assigned to the binary classification. The lines represent the accurate outputs in the database; the columns are the outputs predicted by the algorithm. On the main diagonal are the data correctly classified by the algorithm (true positive = sick person classified as sick and true negative = healthy person classified as healthy), and on the secondary diagonal are the data incorrectly classified (false positive = healthy person classified as unhealthy and false negative = sick person classified as healthy). From this matrix, it is possible to calculate several metrics, which are ways of analyzing the performance of the models.

In the cross-validation, the performance evaluations of the models were obtained by averaging the accuracy and sensitivity in the ten iterations. In the training/test division, the accuracy, sensitivity, specificity, and area under the curve (AUC) were used, with the receiver operating characteristic (ROC) curve.

Table 1: Statistical parameters, for 64 patients and 52 controls, in addition to the correlation test and resulting p value.

Input variables	Patients	Control		
Normal distribution	Mean (standard deviation)	Mean (standard deviation)	Correlation p value	Correlation coefficient
Age (years)	58.182 (18.195)	56.48 (13.92)	0.6624	−0.0369
BMI (Kg/m^2)	28.971 (4.162)	28.21 (5.33)	0.1624	−0.1425
MCP-1 (pg/dL)	563.13 (384.10)	499.92 (292.24)	0.3547	0.0948
Abnormal distribution	Median (interquartile range)	Median (interquartile range)		
Glucose (mg/dL)	98.5 (17.02)	87.64 (10.15)	<0.0001	0.4561
Insulin (pU/mL)	7.59 (11.71)	5.61 (2.69)	0.0261	0.2061
HOMA	2.24 (3.34)	1.22 (0.94)	0.0028	0.2794
Resistin (ng/mL)	14.64 (14.61)	8.33 (6.12)	0.0016	0.2918
Adiponectin (μg/mL)	8.37 (6.64)	8.24 (5.45)	0.7684	0.0234
Leptin (ng/mL)	18.91 (24.79)	21.45 (24.64)	0.9461	0.0063

Table 2: Confusion matrix.

	Predicted value	
Real value	Yes	No
Yes	True positive (TP)	False negative (FN)
No	False positive (FP)	True negative (TN)

The accuracy, represented by equation (1), indicates a general performance of the model, evaluating, among all the classifications, how many the model correctly classified (between sick and healthy). In contrast, equation (2) estimates the number of correct answers of people with cancer concerning all those who have cancer. The specificity, equation (3), refers to the hits of healthy individuals, about all healthy individuals. It is important to note that, for disease detection applications, the algorithm needs to have good sensitivity efficiency as this measure represents the people who have the disease. Therefore, the error must be minimal for this metric.

$$\text{Specificity} = \frac{TN}{TN + FP}, \tag{1}$$

$$\text{sensibility} = \frac{\text{TP}}{FP + FN}, \tag{2}$$

$$\text{accuracy} = \frac{\text{TP+} = TN}{\text{TP} + FN + FP + TN}. \tag{3}$$

As shown in Table 2, another metric included in the algorithm was the AUC-ROC curve, where ROC is a graph of sensitivity (rate of true positives) as a function of the rate of false positives (1 – specificity). The AUC is the area under the ROC curve, having a variation from 0 to 1.

2.3. Machine Learning Models. To define the most suitable parameters for each machine learning model, the "GridSearch" was applied, which swaps between several previously chosen values (under bibliographic research and/or personal choice) and returns the ones that resulted in better performance. For the current situation of detecting the disease, the baseline metric was sensitivity.

2.3.1. K-Nearest Neighbors. The K-nearest neighbor algorithm is one of the simplest supervised learning models [9]. Its classification method does not require training time. It is based on calculated distances between two points, evaluated by the model assuming that issues of the same class would be located close to each other. In this way, the nearest neighbors will dictate the presence of breast cancer in the samples under analysis.

To perform such classification, the choice of parameter K is made, which symbolizes the number of nearest neighbors (not to be confused with the "k-fold" of the validation or crusade). In this sense, the (integer) number chosen must satisfy certain conditions. In turn, the weight determines whether nearest neighbors will have more relevance when choosing the sample class or if they will all have a uniform bearing. Both were evaluated in this model, and the weight for greater relevance was more efficient.

2.3.2. Logistic Regression. Logistic regression aims to generate a model that, through observations of independent variables (input attributes), is capable of predicting the probability of an event to occur, which is usually represented by a binary variable [10, 17].

To create this model, two parameters were used: solver and the maximum number of iterations. The solver is the algorithm used by the model, and the second parameter defines the number of times the solver will be executed. Variations of these parameters, chosen arbitrarily, were tested, only for the solver, "lbfgs," "newton," "liblinear," "sag," and "saga." The values used for the maximum number of iterations were 500, 1000, 1500, and 2000. After using the "GridSearch" tool, the values used for the mentioned parameters were solver = "lbfgs" and maximum number of iterations = 500.

2.3.3. Support Vector Machine. To understand how the support vector machine works, it is necessary to know 4 concepts: the separation hyperplane, the maximum margin hyperplane, the margin smooth, and the kernel function. However, the separation hyperplane is the equivalent of a separation line in a dimension greater than 2. Its role is to define the boundary between samples so that similar are together, and when new data are inserted, they are classified correctly [18]. The maximum margin hyperplane is what differentiates the support vector machine from other hyperplane-based classifiers. There are several ways to separate two groups. However, the maximum margin hyperplane is considered the best. For this to happen, the classifier calculates the shortest distance between two samples from different groups, finds the mean value, and then traces the hyperplane [19, 20].

The soft margin occurs when the classifier uses a margin that accepts wrong classifications for the training data. During the tests, the classifications have a low percentage of error. This situation occurs in databases where at least one sample from a group X is close to the samples from a group Y [21]. The kernel function solves problems where creating a hyperplane or line is impossible without future classification errors by increasing the data dimension. For example, the kernel function increases to two dimensions for a one-dimensional dataset where the hyperplane cannot be traced. The new dimension is the square of the initial values. In this way, it is possible to draw a corresponding hyperplane for that dataset, which was previously impossible [22]. For this model, three parameters were used: kernel, gamma, and C. After using the "GridSearch" tool, the values chosen for the parameters used were kernel = "poly," $C = 10$, and gamma = 1.

2.3.4. Decision Tree. The decision tree algorithm consists of classifying data through the analysis of its attributes to efficiently represent the knowledge obtained through the input set. In the model in question, the nodes represent the tests performed on the attribute values; the arcs indicate the possible output for a given test, and finally, the leaves show the final classification of the tree over the dataset [23]. Three parameters were used to implement the decision tree algorithm: criterion, random state, and maximum depth. The criterion is the function that measures the quality of a division; two parameters are supported: Gini (Gini impurity) and entropy (information gain). In addition, the tree has the maximum tree depth variant, which will define how far the tree will be branched, and the parameters range from 0 to infinity. Finally, the decision tree also has the random parameter state, where the main objective is to control the randomness of the data; that is, if the number 80 is assigned to the random state, the data output will always be the same [24]. The following values were used in the model: maximum depth = 2, criterion = entropy, and random state = 100. It is worth mentioning that these parameters were obtained through the "GridSearch" tool.

2.3.5. Random Forest. The random forest model is characterized by performing classification or regression based on the decision tree model [25]. However, some differences arise when analyzing the criteria for branching the nodes. In its application, the algorithm randomly selects the features that will compose the roots of the trees, thus constituting different models. Then, the branches are performed using the same impurity calculations present in the decision tree model. At the end of this process, the test data will be classified under the criteria of "n" trees (evaluated from 1 to 300), and by statistical analysis, the sample class will be inferred. In this work, 19 trees performed better in the classification. An important parameter to highlight about this model is creating a "bootstrap," which means the generation of a subset of data [26, 27]. In it, the algorithm randomly selects training data samples, possibly even repeated ones, and applies them in the design of the trees. This parameter has the function of reducing the occurrence of "overfitting" and improving the algorithm's stability. Another parameter used was the maximum depth of the trees, responsible for dictating how many subdivisions each will not do, that is, the maximum number of subclassifications made before classifying the sample finally. It was evaluated in the range of 1 to 10, with the number 7 obtaining the highest performance. Finally, to avoid randomization of results, the model declaration still has a parameter called "random state," whose function is to standardize training input selections. In this sense, evaluating the range from 1 to 500, the value 50 was chosen.

2.3.6. Artificial Neural Networks (ANNs). Artificial neural networks are constituted by simple units (neurons) and are based on nonlinear mathematical functions to obtain an organization and generalization of the data. Like the biological nervous system, neurons are organized by one or more layers, interconnected by numerous connections (synapses). In the artificial neural network, the synapses represent the synaptic weight, responsible for the weighting of the input data in each neuron; the learning process of a neural network occurs through numerous iterations and successive corrections of synaptic weights. Such correction is only possible after the network provides an output and performs the comparison with the real output, which represents the error function. Then, the network will propagate the data back to the input and correct the applied weights, a step called "backpropagation." For the implementation of the algorithm, six parameters were used: solver, size of hidden layers, initial learning rate, activation function, the maximum number of iterations, and state random. The best combination was solver = "Adam," a hidden layer, with 3 neurons, initial learning rate = 0.1, activation = "logistics," maximum number of iterations = 200, and random state = 100.

3. Results and Discussion

This section will be split into two. In the first one, the results of the application of the k-fold cross-validation will be presented and the second from the training/testing division. It is noteworthy that two considerations were made. The idea

is to verify if this dimensionality reduction will improve/worsen or maintain the same values of performance metrics as using all 9.

3.1. Cross-Validation. The data are shown in Figure 1, and Figure 2 represents the average of the accuracy and sensitivity metrics, respectively, of the ten iterations since cross-validation was used (to evaluate the generalization capacity of each of the six models), by the *k*-fold method, with a value of "*k*" equal to 10. Experimentally, this value admits few prediction errors ("bias" and variance).

According to Figure 1, the support vector machine model obtained the highest accuracy, for nine variables, in the cross-validation (76%). For 4 variables, the model with the highest accuracy was *K*-nearest neighbors (80.45%). Then, it can be seen that even with fewer quantitative learning attributes (which reduce computer processing), it was possible to obtain better accuracy. In terms of sensitivity, in Figure 2, the logistic regression model reached the best result for 9 variables, about the other models, reaching 86.4%. For 4 variables, the artificial neural network model presented the highest sensitivity, 100%. Again, better performance was obtained for the dimensionality reduction test, representing an increase of 13.6%.

3.2. Training/Tests

3.2.1. Metrics by the Confusion Matrix. Figure 3 represents the performance values for 9 and 4 variables. It is noteworthy that, for the present application, the metric sensitivity is of considerable importance as it is necessary to reduce as much as possible the possibility of misclassifying individuals with breast cancer. As a result, relatively lower values of specificity can be allowed because if a healthy person is classified as sick, the most that will happen is that an additional test will be requested, given that it is necessary. It is worth remembering that the output of the proposed models does not represent the absolute truth. On the contrary, if someone who is sick is defined as healthy, that would be a severe mistake.

In terms of accuracy (total of correct answers for sick and healthy people) and considering all attributes, the artificial neural network model obtained the best performance (83.33%). Furthermore, the artificial neural network algorithm reached 100% specificity (healthy correctness). Concerning sensitivity (patient correctness), the random forest and decision tree models reached the best possible result (100%). However, the random forest reached a higher accuracy (83.3%), and for this reason, it was chosen as the best model for 9 variables.

Regarding the performance of the models using the 4 variables, the support vector machine model obtained the best performance in terms of accuracy (79%). Among the six, three models reached 100% sensitivity: logistical regression, decision tree, and artificial neural network. Finally, support vector machine reached 91%, the highest among the models in terms of specificity.

Again, using the sensitivity and the balance for accuracy, it is understood that the decision tree achieved greater efficiency in the proposed classification task. It is also remarkable to verify that the reduction in the number of variables used did not result in such a significant general improvement, analyzing all models.

3.2.2. AUC-ROC Curve. The ROC curve is the relationship between the true positive rate and the false positive rate. The AUC ranges from 0 to 1, and the closer to 1 it is (far away, towards the top, from the random prediction AUROC = 0.5) , the more generalist in learning is the model; the models that achieved the highest AUC values, for 9 and 4 variables, respectively, were random forest (AUROC = 0.881) and logistic regression/support vector machine (AUROC = 0.860). On the contrary, the decision tree model obtained the worst result (for 9 variables), with AUROC = 0.825, and for 4 variables, the model with the lowest value was the artificial neural network, with AUROC = 0.119, which represents a very small learning value.

As the idea was to compare the performance between 9 and 4 input attributes, the parameters found by the "GridSearch" tool, in the initial situation (with all the variables), were applied in the reduction, 0 to 4. More extreme values, such as the ANN, become possible with this. Thus, in comparison with the article, in which its best model (random forest) obtained sensitivity and specificity of 85% and 77%, respectively, in addition to AUROC = 0.85, the model by artificial neural networks present in work, for example, showed 100% sensitivity for 4 variables in the cross-validation. Furthermore, in the training/test division for 9 attributes, 2 models reached 100% sensitivity, these being the random forest and decision tree models, while in the use of 4 attributes, 3 models also obtained the maximum performance: logistical regression, decision tree, and artificial neural networks. Furthermore, the support vector machine model achieved specificity greater than 77% for the divisions of 9 and 4 attributes. Finally, it is worth mentioning that the logistic regression model for AUROC reached a result of 0.86, surpassing the model of the cited article.

The AUROC = 0.881, obtained by the random forest (with 9 entries), exceeded the values of 0.86, 0.71, and 0.849, found, respectively, by [25–27]. Table 3 shows the best models. For cross-validation, the artificial neural network, with 4, was better for presenting 100% of sensitivity, and the logistical regression, with 9, reached 86.4%. By training/test, with 4, logistical regression achieved 100% sensitivity and AUROC 0.860, and with 9, random forest achieved 100% sensitivity and AUROC 0.881. The final choice of the best topology between the two evaluation methods is the random forest. By validation and 4 attributes, the artificial neural network had just over 50% of accuracy, and the random forest had 83.3%.

4. Conclusions

The research problem is to compare different machine learning models in the job of identifying the existence of breast carcinoma. Two techniques were applied: cross-

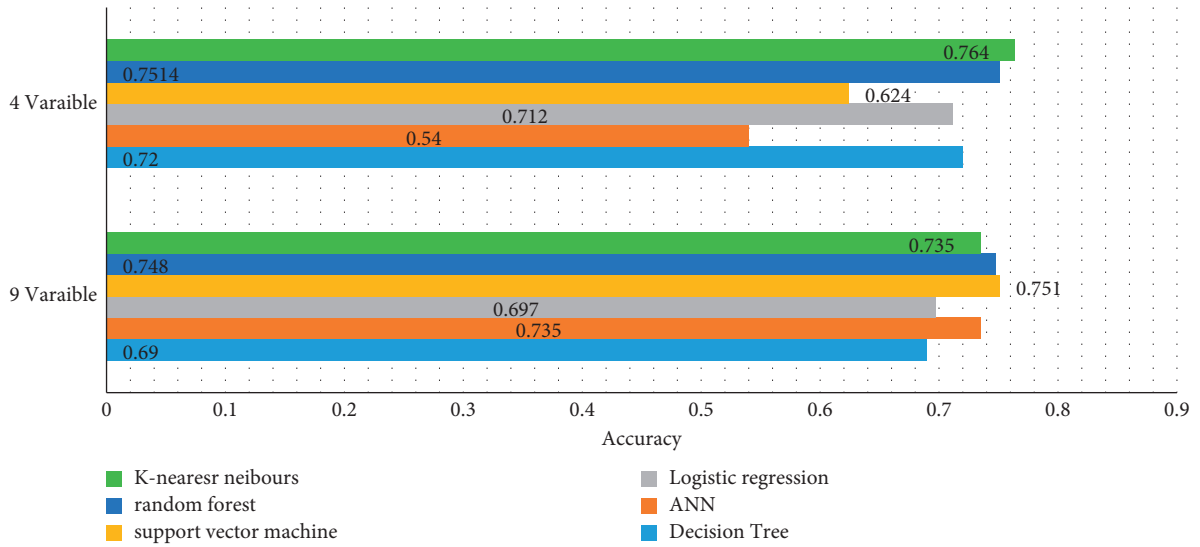

Figure 1: Cross-validation accuracies of the models.

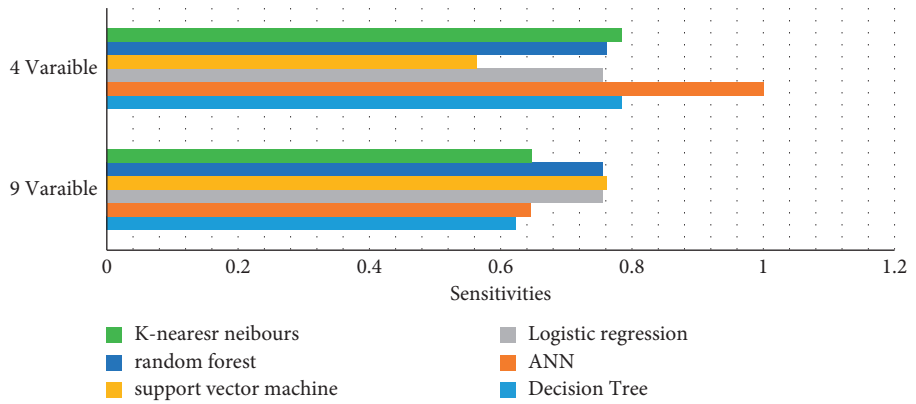

Figure 2: Sensitivities of the cross-validation of models.

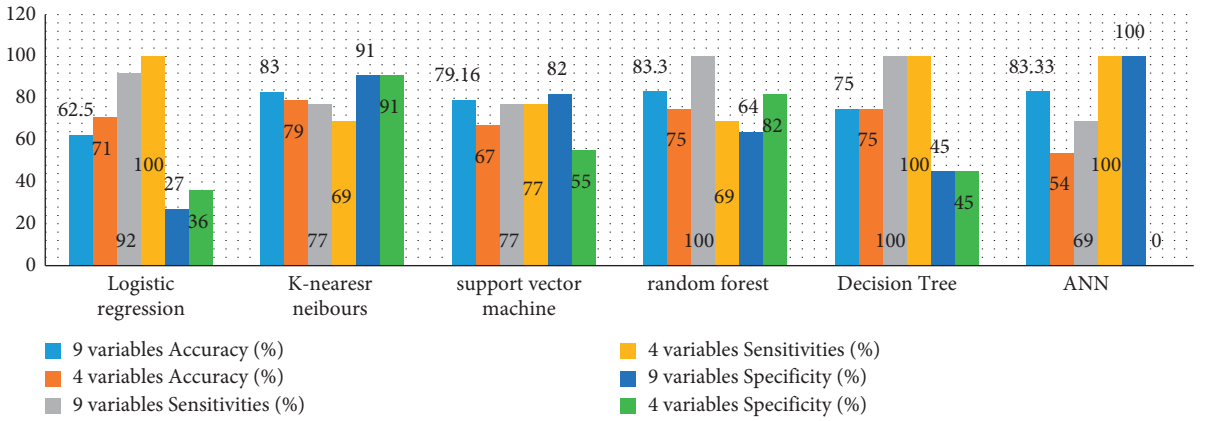

Figure 3: Test results.

Table 3: Best learning models.

Method of validation			
Cross-validation, $k = 10$		80% training, 20% testing	
9	4	9	4
ANN	Logistic regression	Logistic regression	Logistic regression

validation, $k = 10$, and data division into 80% training and 20% testing. Using the Kolmogorov–Smirnov test and consequent correlation (Pearson or Spearman), considering a significance level of 5%, tests were also carried out with the reduction of 9 attributes to 4: glucose, insulin, HOMA, and resistin. The programming was performed on the Google Colab platform. The performance evaluation metrics were accuracy and sensitivity for cross-validation and accuracy, sensitivity, specificity, and AUROC for training/testing. The parameters of the learning models were found by previous bibliographic research and personal choice, in addition to using the "GridSearch" command of the Python language. The results showed, in general, an approximation of the results for 9 and 4 attributes, not representing a great improvement in saving some exams. As a final result, the random forest machine learning model, with 19 trees, 7 subdivisions, and random state 50, obtained the best overall evaluation among all, obtaining 100% sensitivity, 83.3% accuracy, 64% specificity, and 0.881 AUROC. The program's response can help in the decision-making by the professional of the health area.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] S. Gündoğdu, "Improving breast cancer prediction using a pattern recognition network with optimal feature subsets," *Croatian Medical Journal*, vol. 62, no. 5, pp. 480–487, 2021.

[2] S. Lei, R. Zheng, S. Zhang et al., "Global patterns of breast cancer incidence and mortality: a population-based cancer registry data analysis from 2000 to 2020," *Cancer Communications*, vol. 41, no. 11, pp. 1183–1194, 2021.

[3] J. Ferlay, M. Colombet, I. Soerjomataram et al., "Cancer statistics for the year 2020: an overview," *International Journal of Cancer*, vol. 149, 2021.

[4] M Al-Hashimi, "Trends in Breast Cancer Incidence in Iraq During the Period 2000-2019," *Asian Pacific journal of cancer prevention: APJCP*, vol. 22, pp. 3889–3896, 2021.

[5] A. Habeeb, H. Al-Attar, A. Maqtoof, and O. Habib, "Knowledge of women from basrah about breast cancer: its risk and preventive factors," *Iraqi National Journal of Medicine*, vol. 2, pp. 117–123, 2020.

[6] M. Patrício, J. Pereira, J. Crisóstomo et al., "Using Resistin, glucose, age and BMI to predict the presence of breast cancer," *BMC Cancer*, vol. 18, no. 1, p. 29, 2018.

[7] S. B. Faisal, S. A. Mustafa, I. K. Mohammed, and M. A. Maha, "Breast cancer decisive parameters for Iraqi women via data mining techniques," *Journal of Contemporary Medical Sciences*, vol. 5, pp. 260–264, 2019.

[8] S. Jha, S. Ahmad, H. A. Abdeljaber, A. A. Hamad, and M. B. Alazzam, "A post COVID Machine Learning approach in Teaching and Learning methodology to alleviate drawbacks of the e-whiteboards," *Journal of Applied Science and Engineering*, vol. 25, no. 2, pp. 285–294, 2021.

[9] S. F. Khorshid, A. M. Abdulazeez, and A. B. Sallow, "A comparative analysis and predicting for breast cancer detection based on data mining models," *Asian Journal of Research in Computer Science*, vol. 8, no. 4, pp. 45–59, 2021.

[10] M. Bader Alazzam, H. Mansour, M. M. Hammam et al., "Machine Learning of Medical Applications Involving Complicated Proteins and Genetic Measurements," *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 1094054, 6 pages, 2021.

[11] M. Dalamaga, G. Sotiropoulos, K. Karmaniolas, N. Pelekanos, E. Papadavid, and A. Lekka, "Serum resistin: a biomarker of breast cancer in postmenopausal women? Association with clinicopathological characteristics, tumor markers, inflammatory and metabolic parameters," *Clinical Biochemistry*, vol. 46, no. 7-8, pp. 584–590, 2013.

[12] L. Chakravarti and Roy, *Handbook of Methods of Applied Statistics*, U.S. Environmental Protection Agency, Las Vegas, NV, USA, 1964.

[13] K. Hajian-Tilaki, "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation," *Caspian journal of internal medicine*, vol. 4, no. 2, pp. 627–635, 2013.

[14] A. Colaprico, C. Cava, G. Bertoli, G. Bontempi, and I. Castiglioni, "Integrative analysis with Monte Carlo cross-validation reveals miRNAs regulating pathways cross-talk in aggressive breast cancer," *BioMed Research International*, vol. 2015, Article ID 831314, 17 pages, 2015.

[15] R. Wen, K. Zheng, Q. Zhang et al., "Machine learning-based random forest predicts anastomotic leakage after anterior resection for rectal cancer," *Journal of Gastrointestinal Oncology*, vol. 12, no. 3, pp. 921–932, 2021.

[16] J. Crisóstomo, P. Matafome, D. Santos-Silva et al., "Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer," *Endocrine*, vol. 53, pp. 433–442, 2016.

[17] I. H. Sarker, "Machine learning: algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, p. 160, 2021.

[18] Z. Zhang, "Introduction to machine learning: K-nearest neighbors," *Annals of Translational Medicine*, vol. 4, no. 11, p. 218, 2016.

[19] L. Liu, "Research on logistic regression algorithm of breast cancer diagnose data by machine learning," in *Proceedings of the 2018 International Conference on Robots & Intelligent System (ICRIS)*, pp. 157–160, Changsha, China, May, 2018.

[20] D. V. Soundari, N. Nanthini, P. Krishnaanand, R. Padmapriya, C. Thirumariselvi, N. Nanthini, and K. Priyadharsini, Detection of breast cancer using machine learning support vector machine algorithm," *Journal of Computational and Theoretical Nanoscience*, vol. 16, no. 2, pp. 441–444, 2019.

[21] R. S. A. Maabreh, M. B. Alazzam, and A. S. AlGhamdi, "Machine learning algorithms for prediction of survival curves in breast cancer patients," *Applied Bionics and Biomechanics*, vol. 2021, Article ID 9338091, 12 pages, 2021.

[22] A. S. Al-Obeidi, A. A. Hamad, S. F. Al-Azzawi, M. L. Thivagar, Z. Meraf, and S. Ahmad, "A novel of new 7D hyperchaotic system with self-excited attractors and its hybrid synchronization," *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 3081345, 11 pages, 2021.

[23] J. Keerthika, D. Sruthi, D. Swathi, S. Swetha, and R. Vinupriya, "Diagnosis of breast cancer using decision tree data mining technique," in *Proceedings of the 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 1530–1535, Coimbatore, India, March, 2021.

[24] B. Dai, R.-C. Chen, S.-Z. Zhu, and W.-W. Zhang, "Using random forest algorithm for breast cancer diagnosis," in

Proceedings of the 2018 International Symposium on Computer, Consumer and Control (IS3C), pp. 449–452, Taichung, Taiwan, December, 2018.

[25] D R. Kumar, "Review of machine learning algorithm on cancer classification for cancer prediction and detection," *International Journal of Analytical and Experimental Modal Analysis*, vol. 11, pp. 3177–3186, 2020.

[26] M. B. Alazzam, A. A. Hamad, and S. A. Ahmed, "Dynamic mathematical models' system and synchronization," *Mathematical Problems in Engineering*, vol. 2021, Article ID 6842071, 7 pages, 2021.

[27] A. A. Hamad, M. LellisThivagar, M. B. Alazzam, F. Alassery, M. M. Khalil, and V. Ramesh, Z. Meraf and V. Kumar, Dynamic systems enhanced by electronic circuits on 7D," *Advances in Materials Science and Engineering*, vol. 2021, Article ID 8148772, 11 pages, 2021.