Academic Journal of Electrical and Computer Engineering
Vol. 01, No. 02, July 2025
https://doi.org/10.31272/ajece.24
https://ajece.uomustansiriyah.edu.iq
ISSN 3079-2878
Research Article

Breast Cancer Detection Using Standard Data Mining and Machine Learning Techniques

Qutaiba Humadi Mohammed[1*], Ali F. Hassoon[2], Noor Alhuda Kh Ibrahim[3], Anupama Namburu[4]

[1]Media Technology and Communications Engineering Department, College of Engineering, University of Information Technology and Communications, Baghdad, Iraq
[2] Electrical Engineering Department, College of Engineering, Mustansiriyah University, Baghdad, Iraq
[3]Department of Basic Sciences, College of Nursing, University of Baghdad, Baghdad, Iraq
[4]School of Engineering, Jawaharlal Nehru University, New Delhi 110067, India
[*]Email: qutaiba.humadi@uoitc.edu.iq

Received: 07/03/2025 Revised: 30/05/2025 Accepted: 30/06/2025

***Abstract*—Breast cancer is one of the world's most serious health issues; it is the most commonly diagnosed cancer in women, and prevention appears impossible because the cause is unknown. As a result, early detection is critical to the patient's prognosis. In developing nations such as Iraq, where access to specialized healthcare is limited, radiologists are in short supply, and routine clinical check-ups are rare. In this paper, data mining algorithms were applied to the Wisconsin Breast Cancer Dataset (WBCD) from the UCI repository to evaluate their performance in early breast cancer diagnosis. The dataset, curated by medical experts over four years (2016–2020), was analyzed using the WEKA data mining tool. Six classification algorithms were compared using 10-fold cross-validation, ensuring robust performance estimates for the small dataset. The accuracy rates were as follows: decision trees (72%), random forest (75%), k-nearest neighbors (66%), naive Bayes (60%), logistic regression (73%), and multilayer artificial neural networks (66%). Notably, a secondary evaluation using a 66% training split yielded higher accuracy (e.g., 87% for decision trees), but cross-validation metrics are prioritized to mitigate overfitting risks. The results highlight that random forest and logistic regression achieved the highest cross-validation accuracy, though decision trees offer interpretability advantages for clinical use. The work underscores the potential of data mining to assist medical professionals in early diagnosis, particularly in resource-constrained settings. Future work should focus on expanding the dataset and optimizing feature selection to improve model generalizability.**

***Index Terms*—Breast Cancer, Computer Vision, Data Mining, Machine Learning, Statistical Analysis**

I. INTRODUCTION

Recent technological advancements have significantly enhanced the ability to extract meaningful insights from large-scale digital data. In healthcare, data mining techniques have become particularly valuable for disease diagnosis and analysis [1], enabling medical professionals to make more accurate and timely clinical decisions [2]. This work focuses on breast cancer detection, a critical health concern for women globally and in Iraq specifically [3],[4], where early diagnosis rates remain alarmingly low [5],[6]. Using the WEKA platform [7], multiple data mining approaches [8] were systematically evaluated on a clinically validated breast cancer dataset from the UCI repository [9]-[11]. The comparative analysis assessed diagnostic accuracy across different algorithms, ultimately identifying the most effective model for early detection purposes.

II. REVIEW OF LITERATURE

Numerous studies in the literature have applied data mining techniques to breast cancer detection and prognosis [12]–[15]. These studies indicate that results can vary significantly depending on the chosen model and the nature of the training data used [16]. For instance, Ahmed et al. (2020) [17] evaluated several classification algorithms—Naive Bayes, Decision Trees, Logistic Regression, and Sample-Based Classification—on the Wisconsin Breast Cancer Dataset. Their findings revealed that Logistic Regression achieved the highest classification accuracy at 96.92%, followed by Naive Bayes at 96.33%. The Decision Tree (J48) and K-Star algorithms both yielded an accuracy of 95.74%. Similarly, Kalpana et al. (2022) [18] explored the effectiveness of chemotherapy in prolonging survival among breast cancer patients by grouping them into good, moderate, and poor prognosis categories. Based on survival analysis, the work recommended chemotherapy for patients in the moderate group while discouraging it for those in the good prognosis group.

Yeulkar (2017) [19] applied the XCS classification system, a modern machine learning technique, in collaboration with medical professionals. When compared to the traditional C4.5 classification system, XCS yielded superior results on the breast cancer dataset. In another innovative approach, Sutha et al. (2015) [20] tested the Isotonic Separation classification method on two datasets—one complete and one with insufficient data. Their comparative analysis demonstrated that

Isotonic Separation is a viable and practical method for medical data mining applications. Demirel (2010) [21] also contributed by developing software using the WEKA platform, integrating high-performing algorithms such as IB1, Multilayer Perceptron, and Decision Table. Applied to data from 462 patients at Ankara Oncology Hospital, the system achieved high accuracy and was deemed beneficial for supporting oncologists in treatment planning.

Priya et al. (2022) [22] emphasized the diagnostic value of data mining by analyzing 32 features from 569 breast cancer patients—357 benign and 212 malignant cases. Techniques such as Sequential Minimal Optimization (SMO), Support Vector Machines (SVM), and Decision Trees were employed using the WEKA software [23], and the success rates of each method were assessed. Palanivel et al. (2017) [24] focused on improving screening accuracy by combining Adaptive SVM with the Fuzzy C-Means algorithm after the standalone SVM model underperformed. Their hybrid approach achieved an impressive classification accuracy of 99.87%, alongside high sensitivity and specificity scores. Lastly, Patricio et al. (2018) [25] utilized Logistic Regression, Random Forest, and SVM on the Coimbra breast cancer dataset for early diagnosis. Their results showed an 82% prediction accuracy using the SVM model, with sensitivity reaching 88% and specificity ranging between 85% and 90%.

This comparison of traditional machine learning methods fills a major gap in the detection of breast cancer in resource-constrained environments. According to the work, resistance, glucose, and BMI are particularly significant diagnostic markers in the context of the local Iraqi population; this finding has significant ramifications for regional screening programs. This work demonstrates how carefully selected algorithms and evaluation can extract meaningful insights from small clinical data, in contrast to most studies that focused on large datasets. The decision tree model's 72% cross-validated accuracy and interpretability suggest that it has practical relevance for clinical applications where complex black-box models might not be suitable. These findings provide evidence-based suggestions for the application of machine learning-assisted diagnosis in medical systems with similar data constraints.

III. Materials and Methods

A. Material

1) **Data used**

The data set used in the work was developed at the gynaecology department of the University of Coimbra [26]. As a result of blood tests performed in the data set clinical environment, a total of 116 subjects were studied on 64 breast cancer patients and 52 healthy people. The dataset is saved in the UCI Machine Learning repository and is open for access as Fig.1. The dataset contains 10 attributes with quantitative and binary variables related to whether each subject has breast cancer. Attributes are anthropometric data and parameters that can be collected in routine blood analysis.

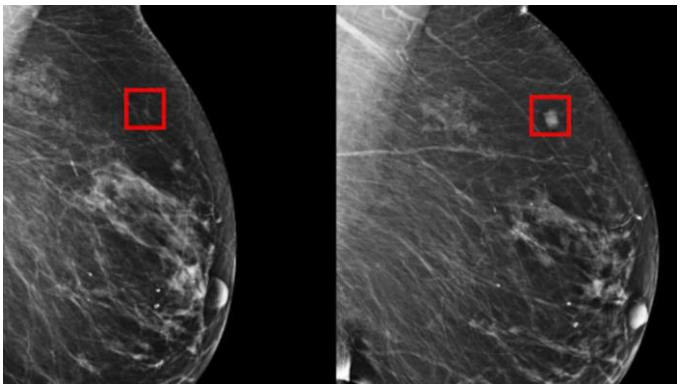


Fig. 1. Sample Breast Cancer Images from WBC dataset.

Definitions of the attributes used;

- Age: The time since a person's birth until today, measured in years.
- Body mass index: It is calculated by dividing the person's weight (kg) by the square of the height (m) [27],[28].
- Glucose: It is a type of sugar obtained from the food we eat and mixed into the blood.
- Insulin: It is the hormone secreted from the pancreas and prevents blood sugar from rising [28].
- HOMA-IR: It is a type of test performed to determine whether there is insulin resistance.
- Leptin: It is a hormone that is secreted from fat tissue and combines with special receptors in the body to control weight, food intake and energy expenditure [29].
- Adiponectin: It is a hormone that regulates glucose levels.
- Resistin: It is a hormone related to obesity and type-2 diabetes that creates insulin resistance [30].
- MCP-1: A type of protein that stands for monocyte chemoattractant protein-1.

There is no missing data in the dataset. Potential breast cancer patients can be identified from prediction models prepared according to these features. In Fig.2 and Table 1, the numerical ratios of the attributes according to breast cancer and healthy people are shown graphically. This visually shows us the numerical distribution of the dataset according to attributes. For example, in the "Age" graph, we see that the distribution of ages of sick and healthy people is balanced.

In this work with the data set, information about the stage of cancer patients is given in Table 2. With this information, it was determined that most of the cancer patients in the data set were in the first two stages. This shows that the method to be developed is suitable for early diagnosis (Patricio et al. 2018).

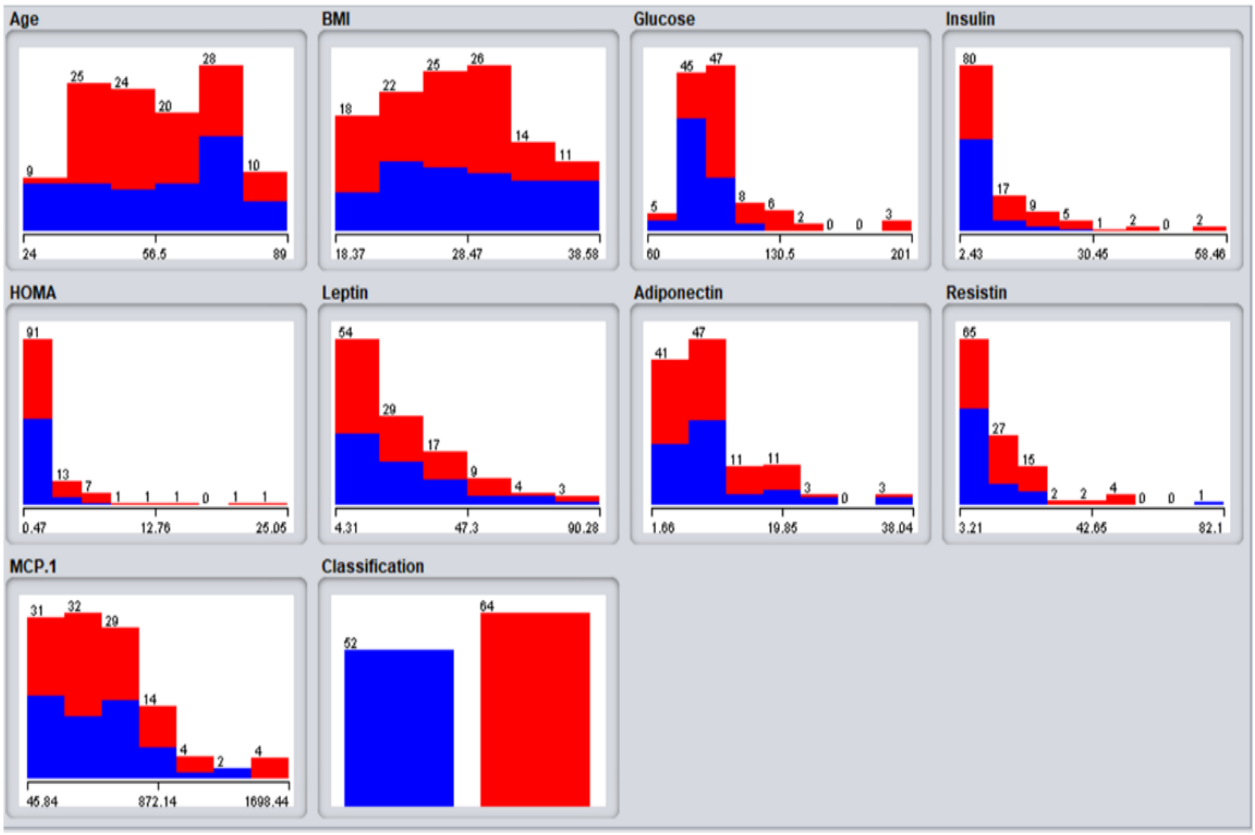
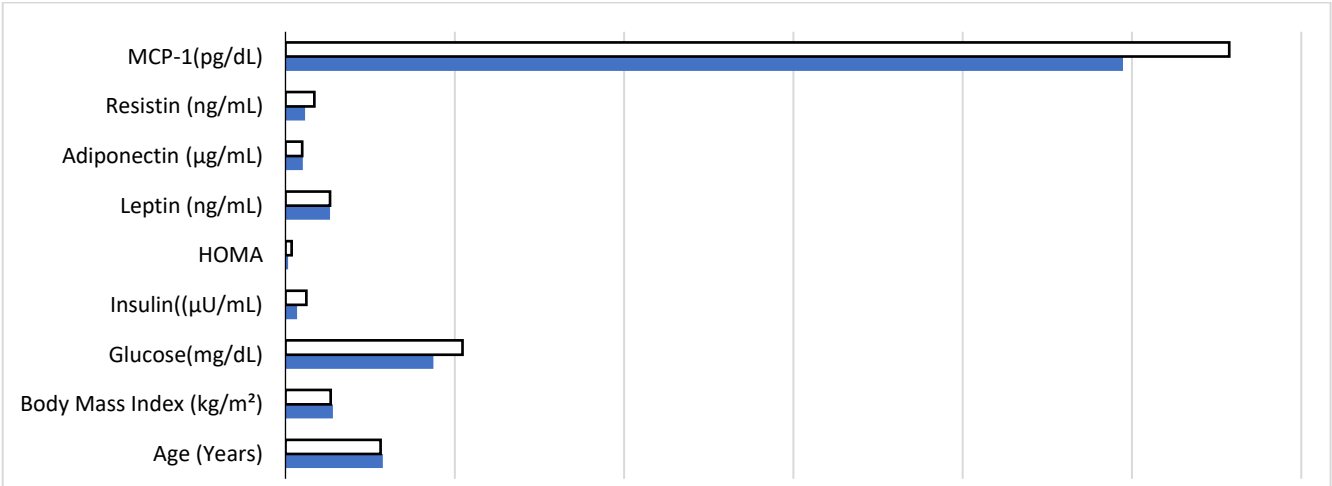


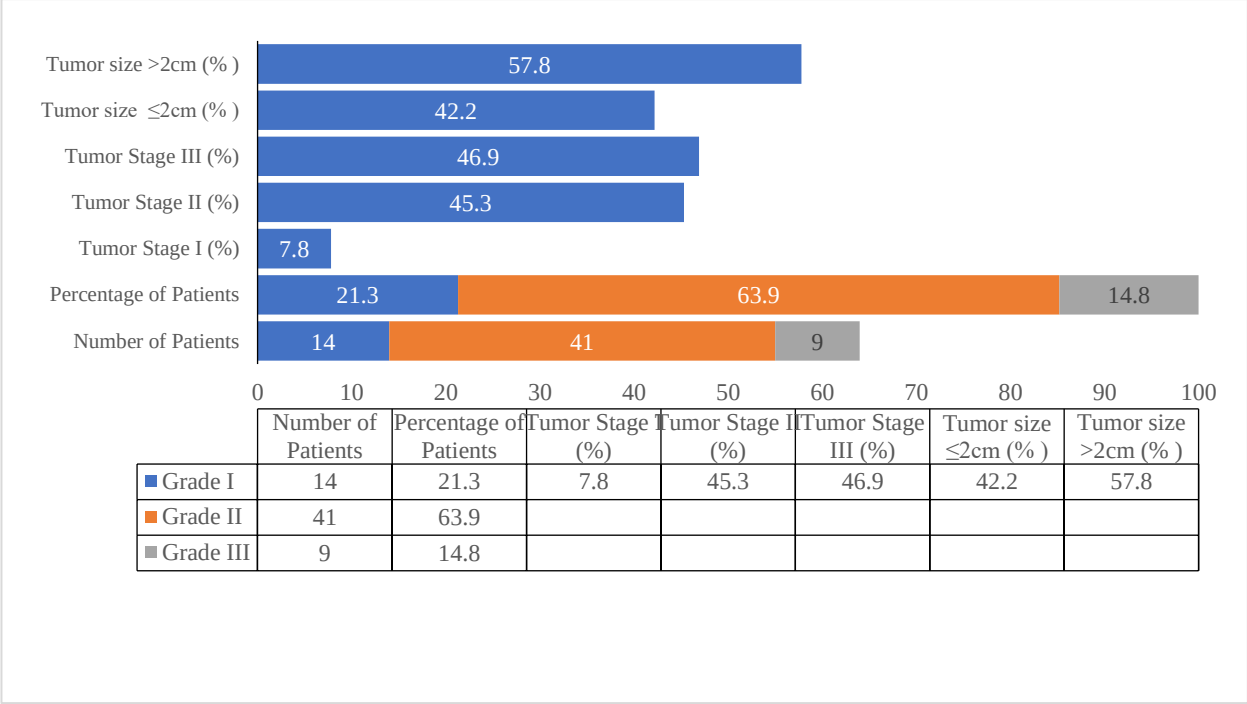
Fig. 2. Data set attribute analysis charts

TABLE 1
MEAN AND STANDARD DEVIATION VALUES OF ATTRIBUTES



	Age (Years)	Body Mass Index (kg/m²)	Glucose(mg/dL)	Insulin((µU/mL)	HOMA	Leptin (ng/mL)	Adiponectin (µg/mL)	Resistin (ng/mL)	MCP-1(pg/dL)
Patient (Mean Value)	56.1	26.71	104.5	12.38	3.58	26.33	9.96	17.08	557.39
Healthy (Mean Value)	57.5	28.04	87.35	6.86	1.53	26.37	10.22	11.5	494.73

TABLE 2
STATISTICAL INFORMATION ABOUT BREAST CANCER PATIENTS' GRADES



	Number of Patients	Percentage of Patients	Tumor Stage I (%)	Tumor Stage II (%)	Tumor Stage III (%)	Tumor size ≤2cm (%)	Tumor size >2cm (%)
Grade I	14	21.3	7.8	45.3	46.9	42.2	57.8
Grade II	41	63.9					
Grade III	9	14.8					

B. Method

1) **Preprocessing**
Preprocessing in data processing is one of the important factors affecting the success of the model. In the preprocessing phase, merging, transformation, reduction and similar operations can be performed on the dataset. All these operations are done to increase the success of the dataset. Erroneous entries in the data set are corrected and made ready for the classification process [31]. Different interventions may affect the models positively or negatively. It is very important that the practitioner knows the data set very well and is an expert on this subject, or gets support from experts.

2) **Parameter Selection**
Each classification algorithm to be applied to the dataset has different parameters. These parameters determine how the algorithm is applied to the dataset. For example, the ANN hidden layer number parameter, pruning in decision trees, branching parameters, ensemble learning (Ada Boost) classifier parameter, etc. There are parameters. Parameters must be set correctly and consciously [32].

3) **Training and test set selection**
The data set is divided into two test and training data. There are many methods to be used to select the test set in the separation process. Firstly, it is the training set selection in which the dataset is used only for training purposes [33]. Another selection set is tested with a second data set loaded from outside, apart from the data set. Another method is to divide all the data into equal parts with a certain number, one of which is set as training data and the other as test data, and the process is repeated for the determined number by changing places. Another clustering process is done by dividing the data set into two clusters at a certain percentage, one as a training set and the other as a test set. The test set selection process is selected according to the structure and size of the dataset.

4) **Model performance metrics**
Various performance measurement techniques are employed to assess the effectiveness of the classification process. These consist of the F-measure rate, the accuracy-error rate, the precision rate, and the sensitivity rate. As you can see in Table 3, the confusion matrix shows how well the test was done. This matrix presents a comparison of the predicted outcomes by the model with the actual results.

TABLE 3
CONFUSION MATRIX [34]

	a	b
a	TP (True Positive)	FN (False Negative)
b	FP (False Positive)	TN (True Positive)

a) Accuracy – error rate
It is the method by which the accuracy and error rate of the model are calculated. The accuracy rate is found by dividing the number of correctly classified items in the confusion matrix (TP+TN) by the total

number (TP+FN+FP+TN), as shown in equation (1). The error rate is the accuracy rate rounded to 1 [35] as in equation (2).

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

$$\text{Error Rate} = \frac{TP + FN}{TP + FN + FP + TN} \quad (2)$$

b) Precision

It is found by dividing the total number of true positives (TP) whose class is estimated to be 1 by (TP+FN+FP+TN). This means the probability of correctly classifying healthy subjects as in equation (3).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

c) Sensitivity

The sensitivity of the model is calculated as the ratio of correct results to all results. It means the probability of correctly classifying sick subjects as in equation (4).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (4)$$

d) F-Measure

The accuracy and sensitivity found as a result of the procedures performed may not give us the result we want when evaluated alone. F-measure is defined as the harmonic mean of precision and sensitivity. It evaluates the precision and sensitivity criteria together as equation (5).

$$\text{F- Measure} = 2 \times \frac{Sensitivity \times Precision}{Sensitivity + Precision} \quad (5)$$

e) ROC curve (Receiver Operating Characteristics)

It is the value obtained as a result of the ratio of sensitivity and precision values. It is also expressed as the ratio of true positive (TP) values to false positive (FP) values. The ROC curve, consisting of horizontal and vertical axes, consists of different threshold values. The vertical axis consists of true positives, and the horizontal axis consists of false positives. Success increases with a high true positive rate and a low false positive rate. The closer the ROC curve is to the "x", that is, the false positive axis, the lower the success level, while the closer it is to the "y", that is, the true positive axis, the higher the success level. The size of the area under the ROC curve indicates a high level of success and reliability [36],[37].

5) **Classification methods**

a) Decision trees

Decision trees are frequently used for classification and prediction in data mining processes. Decision trees are easily understandable and interpretable, which provides an advantage over other decision-makers [38]. Decision trees are generally used in classification problems and large databases that are complex and contain errors. Decision trees consist of decision nodes, branches and leaves. The result is achieved by simply applying the decision-making steps. Large amounts of records are divided into very small groups of records. As a result of each successful division, members of the groups become more similar to each other [39], [40].

b) Random Forest

Random forest is created by combining the decisions of many multivariate trees, each trained with different training sets, instead of a single decision tree. The original training set is obtained from bootstrapping and random features of different training sets. Multivariate decision trees are calculated with the CART algorithm. First, the entire decision tree makes its own decision and is approved as the final decision with the maximum vote and is included in the class as test data [41].

The random forest algorithm is used by defining two parameters. These parameters are:

- M: Number of variables used at each node to determine the best split operation, randomly selected
- N: Number of how many trees will be developed

Two-thirds of the training data set is used as the bootstrapping portion, while the remaining one-third is used to check for errors. The tree is created from each bootstrapped section without pruning. N samples of each sample are created on the data set, separated as training and control data. Trees created with training data are tested with control data, and error rates are calculated. The error rate of the decision forest is found by averaging the error rate of all trees created. The lowest weight is given to the decision tree with the highest error rate. The highest weight is given to the one with the smallest error rate, as in equation (6).

$$\sum \sum_{j \neq i} \left(\frac{f(C_i T)}{|T|}\right)\left(\frac{f(C_i T)}{|T|}\right) \quad (6)$$

C_i: Class with a randomly determined attribute

$\frac{f(C_i T)}{|T|}$: the probability of being in class C_i

c) K-Nearest neighbor (K-NN)

The k-nearest neighbour algorithm is an effective classification method that is generally used in large databases and is used more frequently than other machine learning algorithms [42]. The k-nearest neighbor algorithm is among the supervised learning methods. The k-nearest neighbour algorithm looks at the pattern space and finds the k-closest samples to find the class to which the undetected data belongs. Distance calculation methods such as Euclidean distance, Manhattan distance (ref) and Minkowski distance (ref) are used when calculating the distance between neighbors. Undetected data is assigned to the class value most similar to the K-nearest neighbor.

The selected records are determined and the category of the observation to be predicted is accepted by selecting the most recurring category [43].

d) Naïve Bayes Classifier

Naive Bayes, a common machine learning algorithm, is a successful document classifier [44]. It is widely used in medical diagnosis and document classification. The Naive Bayes classification method is a straightforward classification method that bases data group probabilities on frequencies and combinations. The technique based on the Bayes theorem employs feature probabilistic distributions and dependencies. Furthermore, naive Bayes performs well at a high rate, making it faster in calculation time than other Bayes classifiers. From class-specific training data, the Naive Bayes classification algorithm computes probability values. Class membership is determined using probability values and test data.

e) Artificial Neural Networks (ANN)

It is widely used in many fields, such as signal processing, pattern recognition, and nonlinear control. In artificial neural networks, the classification process is first done by calculating weights to reach the output layer. The weights calculated on the training dataset are then used on the testing dataset to determine how much learning has occurred. If the effectiveness cannot be verified from the results, correction and recalculation with weights are performed. When the learning process is completed successfully, the class of the new data can be determined with the help of weights. The learning process of artificial neural networks, which have a sensitive classification structure, can take quite a long time [45]. As seen in Fig. 3, inputs are the layers containing information coming to the artificial neural network from other nerve cells, externally or from itself. It is often denoted by the symbol x_i. Multilayer artificial neural networks have three layers, as seen in Fig. 4. The input layer is given first, followed by the hidden layer in the middle, and finally by the output layer. The dependent and independent variable values dictate the number of neurons in the input and output layers. The user determines the number of layers and neurons in the hidden layer to deliver the optimal performance.

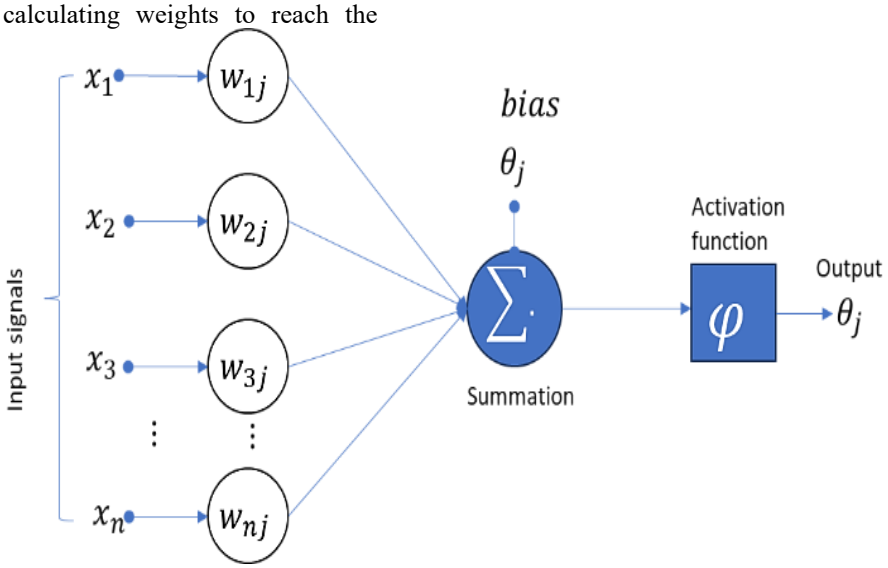


Fig. 3. Structure of the Artificial Neuron [46]

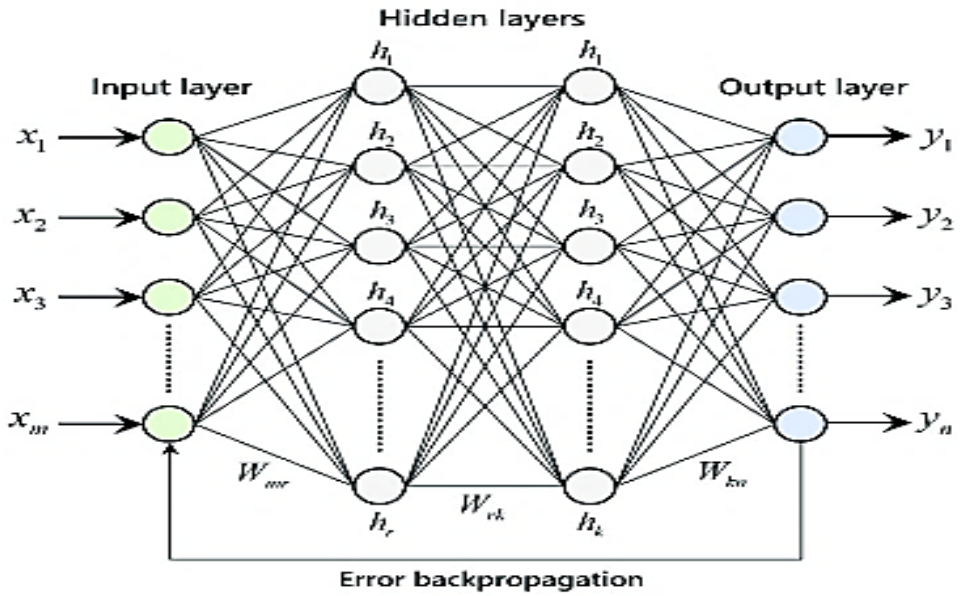


Fig. 4. Multilayer Artificial Neural Network [47]

f) Logistic regression

The logistic regression method is a more advanced regression technique. Although it was previously extensively work in medicine, it has only recently begun to be used in social sciences. This method's primary objective is to model the relationship between independent and dependent variables [48]. The relationship between one or more explanatory variables and outcome variables is investigated using regression methods. Typically, the outcome variable is discrete and has two or more possible values. It can take binary or multi-category values in logistic regression. It is expressed as estimating the likelihood of one of the possible values for the dependent variable.

C. Weka

The WEKA programme version 3.8.2 is used in this work. There are functions in the WEKA programme that allow us to perform data mining operations step by step in a visual environment. While operations can be performed on a single data set, they can also be applied and compared to multiple data sets. Another useful feature of the Weka programme is the ability to use JavaBeans. Work can also be done by typing commands into the CLI feature. WEKA data mining can be done by directly uploading the file, entering the URL address, or connecting to the database [49]. Generally, data sets are available in "CSV" or "Excel" formats. The "ArffViever" module is available under the "tool" tab on WEKA's home page to convert it to the arff format supported by the programme. Data sets in other formats can be viewed and converted to arff format using this module. The WEKA programme has a sophisticated structure for categorising uploaded data. It can calculate the data ratio to be used for testing purposes. Furthermore, the classification work can provide a detailed truth table, summary data, and error matrix data.

IV. RESULTS

This work used data mining approaches to construct a model for early detection of breast cancer, as Fig.5. Since there was no missing data in the data set used in the work and the data set was developed by medical experts, no pre-processing was applied to the data set. The fact that the majority of breast cancer patients in the data set are in the first 2 stages of the disease shows that the model to be developed is suitable for early diagnosis. When the findings were evaluated in the field of medicine, it was determined that the determining attributes in this work were glucose, age, BMI and resistance. Many works are showing that these attributes are directly or indirectly related to cancer in medical findings. It is also known that obesity triggers cancer (Patricio et al. 2018).

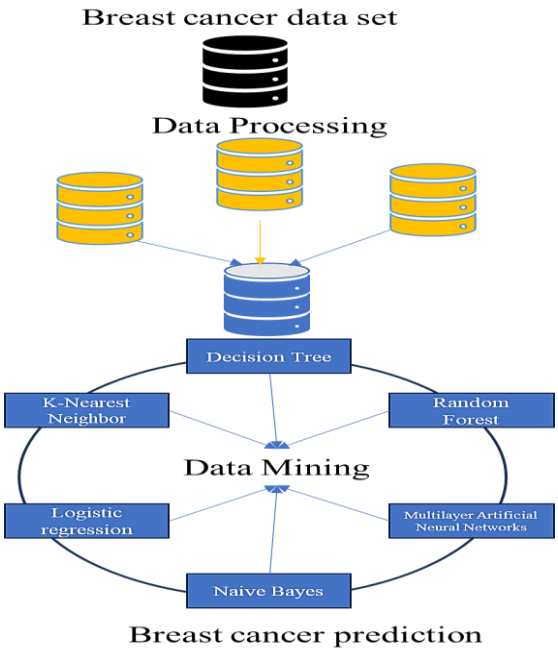


Fig. 5. Data Mining Workflow Diagram

A. Breast Cancer Data Preprocessing and Test Settings

Since there is no missing data in the breast cancer data set and the medical experts who created the data stated that they created the data set completely and accurately, it was not deemed necessary to do any preprocessing on the data set when the mean and standard deviation values were also examined.

Our data set, taken from the UCI machine learning database, was converted from the ".csv" format of the file extension to the ".arff" format supported by the Weka program in order to be modelled in the WEKA program. When we want to classify after the conversion process, we see that some classification algorithms are inactive in the Weka program, such as the decision tree algorithm, because the attribute we set as "class" is in numerical format. When we convert our attribute from numerical format to nominal format, we can apply other classification algorithms to our dataset. Additionally, missing classification libraries have been added to the Weka program.

During the testing of the data set, the performance rate was tested by selecting 66% in the percentage allocation section in the test settings section of WEKA.

B. Application of Classification Methods

In the work, 6 data mining classification methods, suitable for the structure of the dataset and widely used in the field of medicine, were used.

These methods:

- Decision Trees (J48)
- Random Forest
- K-Nearest Neighbor (k-NN)
- Naive Bayes
- Multilayer Artificial Neural Networks (MLP)
- Logistic Regression

The application of the methods to the data set and the classification performance obtained are presented below.

1) **Application of the Decision Tree Method**

The decision trees algorithm is listed as J48 under the "trees" tab in the WEKA program. J48 Decision trees algorithm was applied to the breast cancer dataset. Table 4 shows the model's performance results along with accuracy, precision, sensitivity, F-Measure and ROC area values. When this table is examined carefully, it can be seen that the average DP value of the method is calculated as 87%, the precision value is 88%, the sensitivity value is 87%, the F-measure value is 87%, and finally, the ROC area value is 89%. As a result of all these values, it is seen that the decision tree algorithm has high success rates according to the classification success criteria.

TABLE 4
DECISION TREE RESULT SUMMARY

Metric	Class 1	Class 2	Weighted Avg
Recall (TP Rate)	93.3%	83.3%	87.2%
FP Rate	16.7%	6.7%	10.5%
Precision	77.8%	95.2%	88.5%
F1-Score	84.8%	88.9%	87.3%
MCC	0.748	0.748	0.748
ROC AUC	0.893	0.893	0.893
PRC AUC	0.789	0.907	0.862

Summary

- Accuracy: 87.18% (34/39 correct classifications)
- Kappa Statistic: 0.739 (substantial agreement)
- Mean Absolute Error (MAE): 0.249
- Root Mean Squared Error (RMSE): 0.3379
- Relative Absolute Error: 50.25%
- Root Relative Squared Error: 68.13%
- Total Instances: 39

Detailed Classification Report

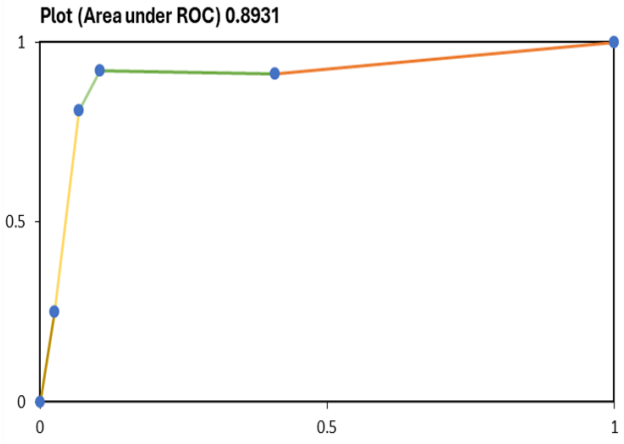


Fig. 6. Decision Tree ROC Analysis Chart

As seen in Fig. 6, the ROC curve resulting from the application of decision trees to the data set shows the performance rate of the model by the size of the area under the graph in the ratio of true positives to false positives. The ROC curve chart shows us that the decision tree method is suitable for this dataset.

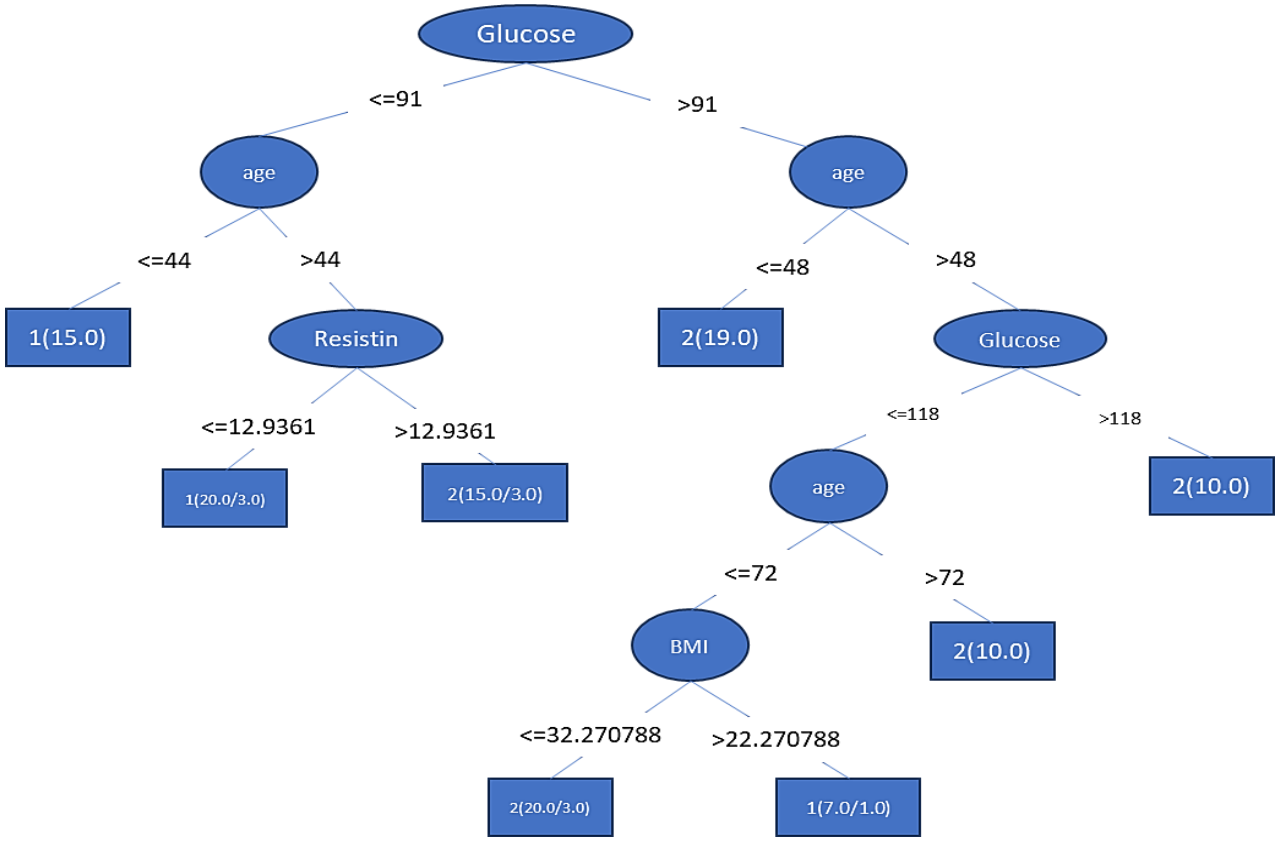


Fig. 7. Branching of the Decision Tree According to the Breast Cancer Dataset

When the decision tree in Fig. 7 is examined, it is seen that the tree diagram branches according to glucose, age, resistin and BMI attributes and that breast cancer predictions are concluded based on these attributes. In the branch results, those ending in 1 indicate the possibility of being healthy, and those ending in 2 indicate the possibility of being sick. In addition, the numerical values next to these probabilities give the number of subjects and the number of incorrect guesses. For example, 1 (20.0/3.0) means that 3 out of 20 people guessed incorrectly. Therefore, since there is a 1 at the beginning, it means there is an 85% chance of being healthy. In this work, the tree was formed in this way on a data set consisting of 116 subjects. It is anticipated that more precise results can be obtained if the created decision tree model is applied to larger data sets. Decision trees provide more understandable and interpretable results to medical professionals because they provide a visual result. As a result of the procedures performed, the success value of the decision trees algorithm applied to breast cancer was measured as 87.17%.

2) **Application of Random Forest Method**

It is called Random Forest under the "trees" tab in the WEKA program. The Random Forest algorithm was applied to the breast cancer dataset. The result summary in Table 5 shows that DP, sharpness, sensitivity, F-Measure and ROC area values are calculated and given. When the data in the table is examined, the average DP value of the method. It can be seen that the precision value is 84%, the precision value is 84%, the F-measure value is 84%, and finally, the ROC area value is 91%. These results show that the ROC area value is quite high.

TABLE 5
RANDOM FOREST RESULT SUMMARY

Summary

- Classification Accuracy: 84.62% (33/39 instances correctly classified)
- Cohen's Kappa: 0.6905 (substantial agreement)
- Error Metrics:
  - Mean Absolute Error (MAE): 0.3195
  - Root Mean Squared Error (RMSE): 0.3683
  - Relative Absolute Error: 64.46%
  - Root Relative Squared Error: 74.26%
- Dataset Size: 39 instances

Classification Performance by Class

Metric	Class 1	Class 2	Weighted Average
Recall (TP Rate)	93.3%	79.2%	84.6%
False Positive Rate	20.8%	6.7%	12.1%
Precision	73.7%	95.0%	86.8%
F1-Score	82.4%	86.4%	84.8%
MCC	0.706	0.706	0.706
ROC AUC	0.918	0.918	0.918
PRC AUC	0.882	0.948	0.922

The random forest algorithm was predicted to achieve better results than the decision trees algorithm with random variables. However, it gave a result approximately 3% lower than the decision trees algorithm. It is thought that this result is due to the low number of data in the dataset.

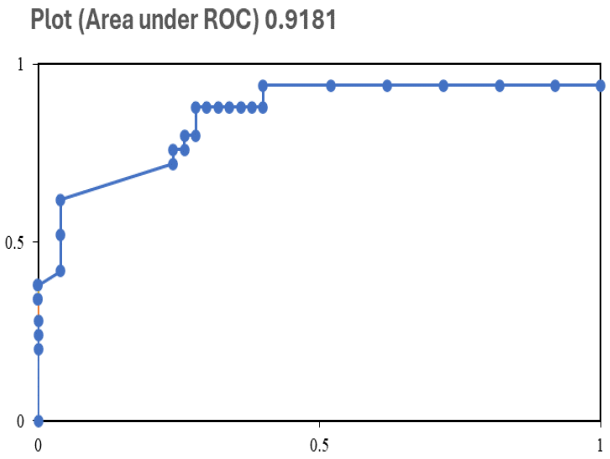


Fig. 8. Random Forest ROC Analysis

The random forest algorithm is deemed appropriate for this data set, as demonstrated in Fig. 8, after the ROC curve graph is applied. This is because the area under the ROC curve is sizable and near to the value 1. When the Random Forest algorithm was applied to treat breast cancer, the results were 84.61% successful.

3) **Application of the K nearest neighbor method**

The K-nearest neighbor algorithm is referred to as IBk under the “lazy” tab in the WEKA program. The IBk algorithm was applied to the breast cancer dataset. As seen in the result summary in Table 6, DP, sharpness, sensitivity, F-Measure and ROC area values are calculated and given. When the data in the table is examined, the average DP value of the method.

It can be seen that the precision value is calculated as 76%, the precision value is calculated as 72%, the sensitivity value is calculated as 76%, the F-Measure value is calculated as 76%, and finally, the ROC area value is calculated as 79%.

TABLE 6
K-NN RESULT SUMMARY

Summary

- Correctly Classified Instances: 33 (84.62%)
- Kappa Statistic: 0.6905 (Moderate Agreement)
- Mean Absolute Error (MAE): 0.3195
- Root Mean Squared Error (RMSE): 0.3683
- Relative Absolute Error: 64.46%
- Root Relative Squared Error: 74.26%
- Total Instances: 39

Detailed Accuracy by Class

Metric	Class 1	Class 2	Weighted Avg.
TP Rate (Recall)	93.3%	79.2%	84.6%
FP Rate	20.8%	6.7%	12.1%
Precision	73.7%	95.0%	86.8%
F-Measure	82.4%	86.4%	84.8%
MCC	0.706	0.706	0.706
ROC Area	0.918	0.918	0.918
PRC Area	0.882	0.948	0.922

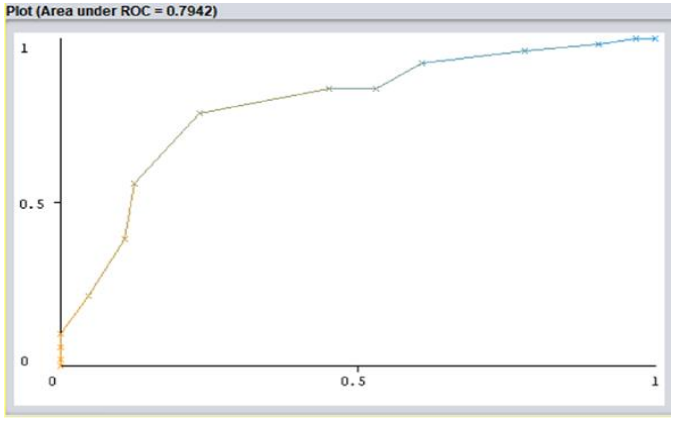


Fig. 9. KNN ROC Analysis

The K-Nearest Neighbour algorithm used for breast cancer had a success value of 76.72%, as a result explained in Fig.9 of the treatments carried out.

4) **Application of Naïve Bayes Method**

The naive Bayes algorithm is called Naive Bayes under the "Bayes" tab in the Weka program. The Naive Bayes algorithm was applied to the breast cancer dataset. As seen in the result summary in Table 7, DP, sharpness, sensitivity, F-Measure and ROC area values are calculated and given. When the data in the Table is examined, it is seen that the average DP value of the method is calculated as 69%, the precision value is 69%, the sensitivity value is 69%, the F-Measure value is 65% and finally the ROC area value is 62%.

TABLE 7
NAÏVE BAYES RESULT SUMMARY

Model Performance Evaluation

Overall Metrics

- Correctly Classified Instances: 27 (69.23%)
- Kappa Statistic: 0.2778 (Fair Agreement)
- Mean Absolute Error (MAE): 0.4453
- Root Mean Squared Error (RMSE): 0.4785
- Relative Absolute Error: 89.84%
- Root Relative Squared Error: 96.48%
- Total Instances: 39

Class-Specific Performance

Metric	Class 1	Class 2	Weighted Avg.
TP Rate (Recall)	33.3%	91.7%	69.2%
FP Rate	8.3%	66.7%	44.2%
Precision	71.4%	68.8%	69.8%
F-Measure	45.5%	78.6%	65.8%
MCC	0.317	0.317	0.317
ROC Area	0.625	0.625	0.625
PRC Area	0.495	0.681	0.610

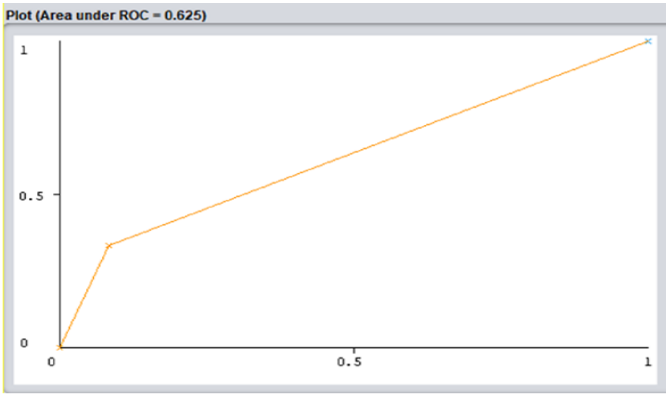


Fig. 10. Naïve Bayes ROC Analysis

The Naive Bayes algorithm used for breast cancer had a success value of 69.23%, as a result shown in Fig.10 of the processes carried out.

5) **Application of the multilayer artificial neural network method**

The multilayer artificial neural network algorithm is called "Multilayer Perceptron" under the "function" tab in the Weka program. The Multilayer Perceptron algorithm was applied to the breast cancer dataset. As seen in the result summary in Table 8, DP, sharpness, sensitivity, F-Measure and ROC area values are calculated and given. When the data in the Table is examined, the average DP value of the method is 69% and the precision value is 69%.

It can be seen that the sensitivity value is calculated as 70%, the sensitivity value is calculated as 69%, the F-Measure value is calculated as 69% and finally, the ROC area value is calculated as 73%.

TABLE 8
MULTILAYER ARTIFICIAL NEURAL NETWORK RESULT SUMMARY

Summary:

- Correctly Classified Instances: 27 (69.2308%)
- Kappa statistic: 0.3659
- Mean absolute error: 0.3373
- Root mean squared error: 0.5003
- Relative absolute error: 68.0508%
- Root relative squared error: 100.8724%
- Total Number of Instances: 39

Detailed Accuracy by Class:

Metric	Class 1	Class 2	Weighted Avg.
TP Rate (Recall)	66.7%	70.8%	69.2%
FP Rate	29.2%	33.3%	31.7%
Precision	58.8%	77.3%	70.2%
F-Measure	62.5%	73.9%	69.5%
MCC	0.368	0.368	0.368
ROC Area	0.733	0.733	0.733
PRC Area	0.669	0.820	0.762

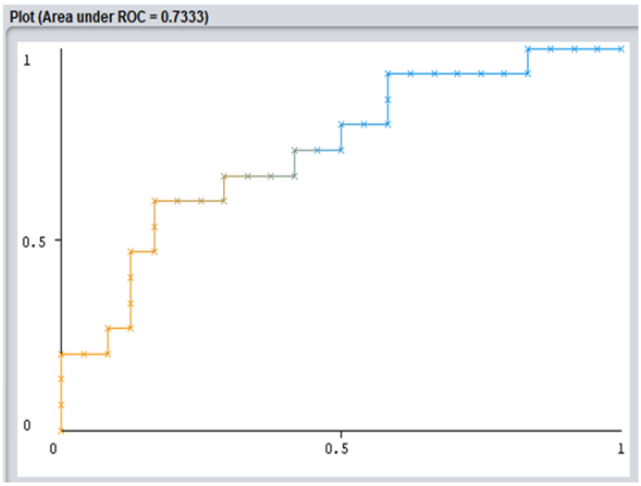


Fig. 11. Multilayer Artificial Neural Network ROC Analysis

As seen in Fig. 11, multilayer artificial neural networks show that the east positive rate is close to the true positive value until approximately 0.2 in the false positive rate graph, but after this value, there is a difference between the true positive and false positive rates, closer to the true positive rate. progressed in this way. At the end of this progress, the ROC area, that is, the area under the curve, was calculated as 0.76.

As seen in Fig. 12, it analyses the attributes as input. As a result of the analysis, it appears that the algorithm created 5 intermediate layers and connected them to the cancer patient and healthy outcomes defined as "class" values.

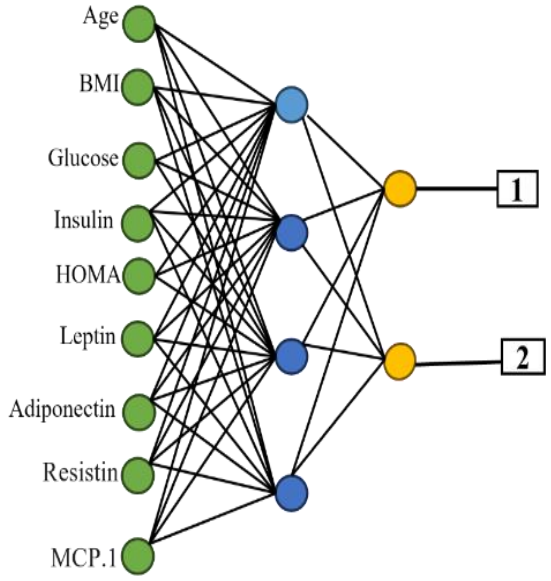


Fig. 12. Multilayer Artificial Neural Networks

As a result of the procedures performed, the success value of the Multilayer Artificial Neural Networks algorithm applied to breast cancer was measured as 69.23%.

6) **Application of the logistic regression method**

The logistic regression algorithm is mentioned as "Logistics" under the "function" tab in the Weka program. The logistic algorithm was applied to the breast cancer dataset. As seen in the result summary in Fig.13 and Table 9, DP, sharpness, sensitivity, F-Measure and ROC area values are calculated and given. When the data in the Table is examined, it is seen that the average DP value of the method is calculated as 84%, the precision value is 86%, the sensitivity value is 84%, the F-Measure value is 84%, and finally the ROC area value is 81.

TABLE 9
LOGISTIC REGRESSION RESULT SUMMARY

Table 9. Logistic regression result summary

Summary

- Accuracy: 84.62% (33/39 correct classifications)
- Kappa Statistic: 0.6905 (substantial agreement)
- Mean Absolute Error (MAE): 0.3282
- Root Mean Squared Error (RMSE): 0.409
- Relative Absolute Error: 66.22%
- Root Relative Squared Error: 82.47%
- Total Instances: 39

Detailed Classification Report

Metric	Class 1	Class 2	Weighted Avg
Recall (TP Rate)	93.3%	79.2%	84.6%
FP Rate	20.8%	6.7%	12.1%
Precision	73.7%	95.0%	86.8%
F1-Score	82.4%	86.4%	84.8%
MCC	0.706	0.706	0.706
ROC AUC	0.819	0.819	0.819
PRC AUC	0.680	0.889	0.809

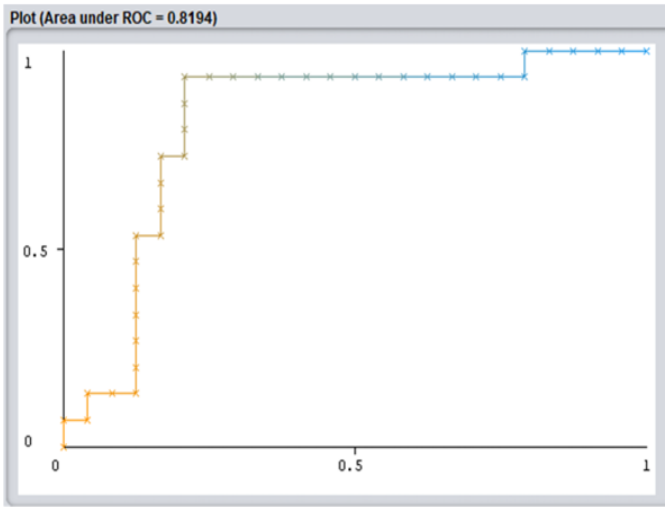


Fig. 13. Logistic Regression ROC Analysis

The procedures carried out led to an 84.61% success rate for the logistic regression algorithm used to treat breast cancer.

C. Comparison of Methods Used

Our breast cancer data set was subjected to a variety of decision trees, random forests, naive Bayes, k-nearest neighbours, multilayer artificial neural networks, and logistic regression techniques. The success criteria of accuracy, precision, sensitivity, f-measure, and ROC area were employed. The applied success criteria are given comparatively, as seen in Fig.v 14. In the comparison made according to the accuracy, precision, sensitivity and F-Measure performance criteria, it is seen that although the decision trees method is slightly higher, the random forest and logistic regression methods have very close values to each other. However, in the ROC field performance criterion, the random forest method received higher values than other methods. When all values are compared, it is seen that the decision trees method receives higher values than other methods. The decision tree approach is unique in that it produces visual results that are understandable and interpretable by medical professionals. The differences in algorithm performance can be attributed to a number of factors [23]. Artificial neural networks (ANN) and Naive Bayes showed lower accuracy (60–69%), likely due to their small sample size (n=116), which limited their ability to recognise complex patterns. For biological data, such as the metabolic interactions between markers like glucose and BMI, Naive Bayes

oversimplifies. It is predicated on feature independence. Without sufficient data for robust training, suboptimal parameter tuning, or layer configuration, could lead to an analogous underperformance of ANN. Decision trees and logistic regression achieved higher accuracy (72–75%) by prioritising interpretable, non-linear relationships, which are essential for clinical adoption [30-33]. The prevalence of glucose, BMI, and resistin as predictive markers has important medical implications: higher BMI reflects adipose-related risks like chronic inflammation and oestrogen overproduction; resistin may promote angiogenesis and insulin resistance; and elevated glucose levels may signal a pro-tumorigenic metabolic environment through increased oxidative stress and altered growth factor signalling. These results point to useful clinical applications, especially in environments with limited resources, where the decision tree model may be used as an easily accessible screening tool that makes use of standard blood tests and anthropometric measurements. It provides a fair option for initial risk stratification prior to advanced imaging because of its 72% cross-validated accuracy and easy interpretability. The present findings show how well-executed machine learning can extract significant insights from sparse clinical data to support early detection efforts, but the work's limitations, such as sample size limitations, highlight the need for validation in larger cohorts. In order to further enhance diagnostic accuracy while preserving practical applicability in a variety of healthcare settings, future work should concentrate on longitudinal biomarker analysis and integration with imaging data.

Given conventional medical wisdom, it makes sense that resistin, glucose, and BMI are so popular. Elevated glucose levels may be a sign of metabolic dysregulation associated with the development of cancer; higher BMI, a proxy for adipose tissue, is linked to increased inflammation and hormone disruption. Resistin, an adipokine associated with insulin resistance, emphasises metabolic abnormalities even more as a potential early marker [12]. These findings suggest that routine blood tests could be used for initial risk stratification in settings with limited resources.

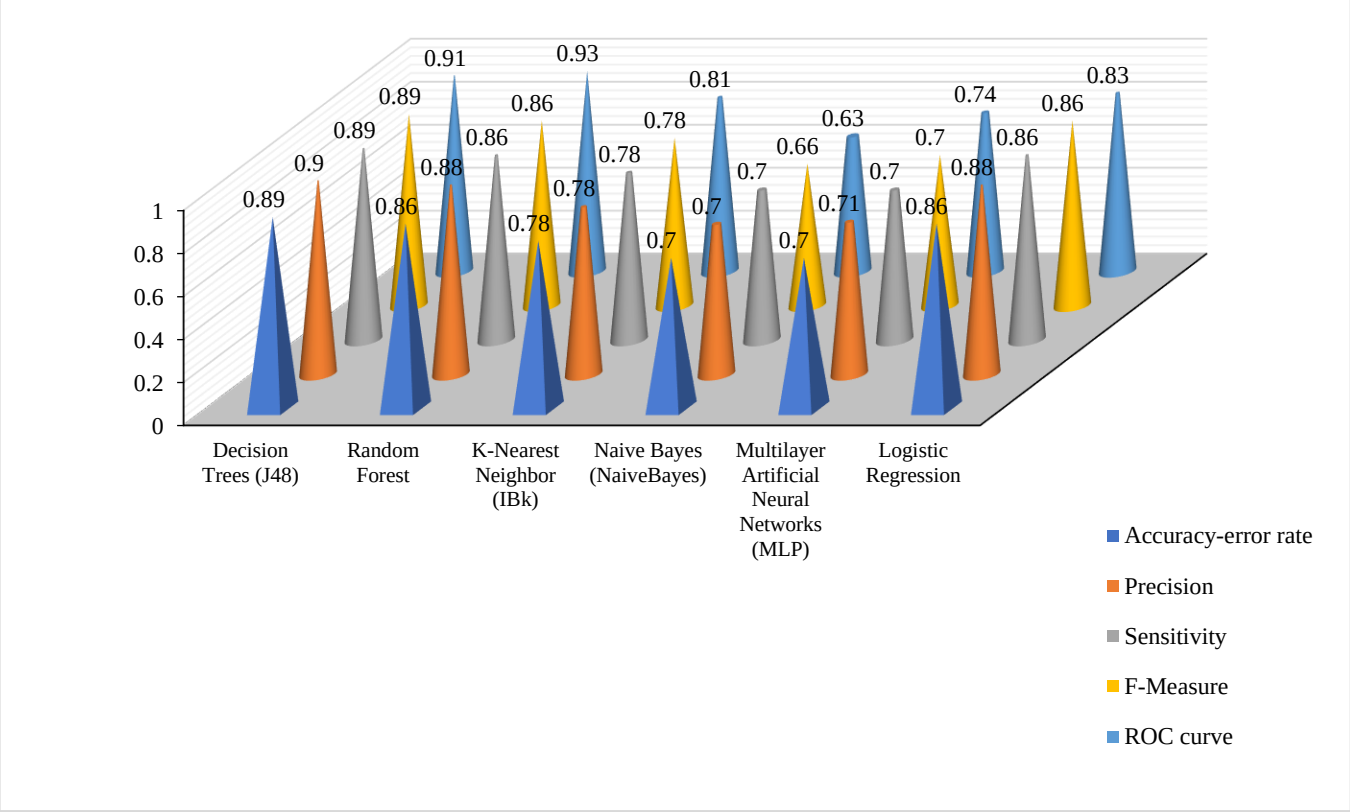


Fig. 14. Comparison Of Applied Methods According to Classification Performance Criteria

V. Conclusion

The current work revealed that the UCI database data could yield the same outcomes when the algorithms used in data mining to detect breast cancer early on were applied. Following four years of meticulous and time-consuming work, from 2016 to 2020, medical professionals received this set of data. The data set is being gathered by medical professionals in order to help them discover important information regarding the early detection of breast cancer. Data mining algorithms were applied to the dataset using the WEKA data mining program. The algorithms' outputs were then contrasted.

Additionally, tests were conducted on the algorithms' execution times. However, because the algorithms' operating

speeds vary on a millisecond basis, the extremely small amount of data in the data set was not considered. It is anticipated that in systems where real-world outcomes are assessed, much faster algorithms may be preferred, and that algorithm execution speed may be crucial in very large data sets.

The work discovered that Naive Bayes and Artificial Neural Networks were not suitable for the algorithms applied to the breast cancer dataset. However, other implemented algorithms have been successful enough. The decision trees algorithm performed better than the other algorithms in the comparison based on success rates, with an accuracy rate of 87%. A decision tree algorithm based on blood and anthropometric tests has been proposed to assist specialists in the early diagnosis of breast cancer. Medical professionals will also be able to diagnose breast cancer in its early stages more quickly and accurately thanks to it.

ACKNOWLEDGMENT

The authors thank the Academic Journal of Electrical and Computer Engineering and all the reviewers for their scientific support to improve the quality of the work.

REFERENCES

[1]P. Ahmad, S. Qamar, and S. Qasim Afser Rizvi, "Techniques of Data Mining In Healthcare: A Review," International Journal of Computer Applications, vol. 120, no. 15, pp. 38–50, Jun. 2015, doi: https://doi.org/10.5120/21307-4126.

[2] M. Maallah, N. J. J, and N. Alwan, "City Research Online - Barriers Impeding Early Detection of Breast Cancer in Iraq: A Critical Analysis," City.ac.uk, Oct. 2022, https://openaccess.city.ac.uk/id/eprint/28948/1/DOC-20221002-WA0027..pdf.

[3] Sarah Sameer Rasheed, suhad faisal, Iyden Kamil Mohammed, and S. Mustafa, "Review of medical diagnostics via data mining techniques," *Iraqi journal of science*, pp. 2401–2424, Jul. 2021, doi: https://doi.org/10.24996/ijs.2021.62.7.30.

[4] M. S. Iqbal, W. Ahmad, R. Alizadehsani, S. Hussain, and R. Rehman, "Breast Cancer Dataset, Classification and Detection Using Deep Learning," *Healthcare*, vol. 10, no. 12, p. 2395, Nov. 2022, doi: https://doi.org/10.3390/healthcare10122395.

[5] A. Al-Gburi, N. Alwan, E. Al-Tameemi, and W. Al-Dabbagh, "Opportunistic Screening for Early Detection of Breast Cancer in Iraq Iraq 3) Specialist radiologist-lead of Breast, Imaging fellowship-Arab Board 4) Main Referral Center for Early Detection of Breast Tumors, Oncology Teaching Hospital Correspondence to," *International Medical Journal*, vol. 28, no. 1, pp. 28–32, 2021, Accessed: Sep. 19, 2024. [Online]. Available: https://seronjihou.com/wp-content/uploads/2021/05/281028.pdf.

[6] Suhad Faisal Behadili, S. Mustafa, Iyden Kamil Mohammed, and Maha Mohammed Al-Sayyid, "Breast Cancer Decisive Parameters for Iraqi Women via Data Mining Techniques," Journal of contemporary medical sciences, vol. 5, no. 2, pp. 71–76, Apr. 2019, doi: https://doi.org/10.22317/jcms.v5i2.573Sfsfdsf.

[7] "Data Mining Techniques for Iraqi Biochemical Dataset Analysis," *Baghdad Science Journal*, vol. 19, no. 2, Sep. 2021, doi: https://doi.org/10.21123/bsj.2022.19.2.0385.

[8] V. Chaurasia and S. Pal, "A Novel Approach for Breast Cancer Detection Using Data Mining Techniques," *papers.ssrn.com*, Jun. 29, 2017. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2994932.

[9] I. K. A. Enriko, M. Melinda, A. C. Sulyani, and I. G. B. Astawa, "Breast cancer recurrence prediction system using k-nearest neighbor, naïve-bayes, and support vector machine algorithm," *JURNAL INFOTEL*, vol. 13, no. 4, pp. 185–188, Dec. 2021, doi: https://doi.org/10.20895/infotel.v13i4.692.

[10]T A H Tengku Mazlin, R Sallehuddin, and M. Y. Zuriahati, "Utilization of Filter Feature Selection with Support Vector Machine for Tumours Classification," *IOP Conference Series Materials Science and Engineering*, vol. 551, no. 1, pp. 012062–012062, Aug. 2019, doi: https://doi.org/10.1088/1757899x/551/1/012062.

[11]V. Chaurasia and S. Pal, "Prediction of Benign and Malignant Breast Cancer Using Data Mining Techniques," *SSRN Electronic Journal*, 2018, doi: https://doi.org/10.2139/ssrn.3139141.

[12]"Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study," *Tehnicki vjesnik - Technical Gazette*, vol. 26, no. 1, Feb. 2019, doi: https://doi.org/10.17559/tv-20180417102943.

[13]B. Padmapriya and T. Velmurugan, "A survey on breast cancer analysis using data mining techniques," Dec. 2014, doi: https://doi.org/10.1109/iccic.2014.7238530.

[14]I. Khan, A. Gandhi, N. Parmar, and B. Garg, "Breast Cancer Prediction Using Data Mining," International Journal of Scientific Research and Engineering Development, vol. 3, 2020, Accessed: Sep. 16, 2024. [Online]. Available:https://ijsred.com/volume3/issue2/IJSRED-V3I2P145.pdf

[15]N. Fatima, L. Liu, S. Hong, and H. Ahmed, "Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis," *IEEE Access*, vol. 8, pp. 150360–150376, 2020, doi: https://doi.org/10.1109/access.2020.3016715.

[16]S. Noureddine, Baarir Zineeddine, Abida Toumi, A. Betka, and Aïcha-Nabila Benharkat, "Smart Breast Cancer Prediction Using Data Mining Process," *HAL (Le Centre pour la Communication Scientifique Directe)*, Dec. 2021, doi: https://doi.org/10.1109/icisat54145.2021.9678429.

[17]M. T. Ahmed, M. N. Imtiaz, and A. Karmakar, "Analysis of Wisconsin Breast Cancer original dataset using data mining and machine learning algorithms for breast cancer prediction," *Journal of Science Technology and Environment Informatics*, vol. 9, no. 2, pp. 665–672, 2020, doi: https://doi.org/10.18801/jstei.090220.67.

[18]P. Kalpana, N. Saranya, D. Usikha, and R. Yazhini, "Prediction of survival rate of breast cancer patients using machine learning," *International journal of health sciences*, pp. 6294–6307, Jul. 2022, doi: https://doi.org/10.53730/ijhs.v6ns5.10361.

[19]K. Yeulkar, "Utilization of Data Mining Techniques for Analysis of Breast Cancer Dataset Using R," *International Journal for Research in Applied Science and Engineering Technology*, vol. V, no. III, pp. 406–410, Mar. 2017, doi: https://doi.org/10.22214/ijraset.2017.3074.

[20]M. Jeya Sutha and F. Ramesh Dhanaseelan, "An Efficient Method for Detection of Breast Cancer Based on Closed Frequent Itemsets Mining," *Journal of Medical Imaging and Health Informatics*, vol. 5, no. 5, pp. 987–994, Sep. 2015, doi: https://doi.org/10.1166/jmihi.2015.1483.

[21]A. Çakır and B. Demirel, "A Software Tool for Determination of Breast Cancer Treatment Methods Using Data Mining Approach," Journal of Medical Systems, vol. 35, no. 6, pp. 1503–1511, Feb. 2010, doi: https://doi.org/10.1007/s10916-009-9427-x.

[22]K. Priya, V. Senthilkumar, J Samson Isaac, Sreekanth Kottu, V. S. Ramakrishna, and M Jogendra Kumar, "Breast Cancer Segmentation by K-Means and Classification by Machine Learning," *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)*, Dec. 2022, doi: https://doi.org/10.1109/icacrs55517.2022.10029301.

[23]P. B. Deshmukh and Kanchan Lata Kashyap, "Research Challenges in Breast Cancer Classification through Medical Imaging Modalities using Machine Learning," Dec. 2021, doi: https://doi.org/10.1109/iciera53202.2021.9726746.

[24]J. Palanivel, K. Suresh, D. Premraj, and K. Thamilmaran, "Effect of fractional-order, time-delay and noisy parameter on slow-passage phenomenon in a nonlinear oscillator," *Chaos Solitons & Fractals*, vol. 106, pp. 35–43, Nov. 2017, doi: https://doi.org/10.1016/j.chaos.2017.11.006.

[25]M. Patrício *et al.*, "Using Resistin, glucose, age and BMI to predict the presence of breast cancer," *BMC Cancer*, vol. 18, no. 1, Jan. 2018, doi: https://doi.org/10.1186/s12885-017-3877-1.

[26]Y. D. Austria, M. L. Goh, L. Sta. Maria Jr., J.-A. Lalata, J. E. Goh, and H. Vicente, "Comparison of Machine Learning Algorithms in Breast Cancer

Prediction Using the Coimbra Dataset," *International journal of simulation: systems, science & technology*, Jul. 2019, doi: https://doi.org/10.5013/ijssst.a.20.s2.23.

[27] S. Pengpid and K. Peltzer, "Overweight and Obesity among Adults in Iraq: Prevalence and Correlates from a National Survey in 2015," International Journal of Environmental Research and Public Health, vol. 18, no. 8, p. 4198, Apr. 2021, doi: https://doi.org/10.3390/ijerph18084198.

[28] N. Alhuda Kh. Ibrahim, Z. N.Abbas, W. J. mohammad, "Association among Vitamin D Deficiency with some Inflammatory Marker in Iraqi Patients with Autoimmune Thyroiditis," Indian Journal of Forensic Medicine & Toxicology, 14(1), 655-660., https://doi.org/10.37506/ijfmt.v14i1.125..

[29] R. A. Hameed, "The Intelligent Auto-Tuning Controller Design Based on Dolphin Echo Location for Blood Glucose Monitoring System," Journal of Engineering, vol. 27, no. 8, pp. 1–18, Aug. 2021, doi: https://doi.org/10.31026/j.eng.2021.08.01.

[30] H. Akeel Al-hussaniy, A. Hikmate Alburghaif, and M. Akeel Naji, "Leptin hormone and its effectiveness in reproduction, metabolism, immunity, diabetes, hopes and ambitions," *Journal of Medicine and Life*, vol. 14, no. 5, pp. 600–605, Jan. 2021, doi: https://doi.org/10.25122/jml-2021-0153.

[31] J. F. Ndisang, "Role of Heme Oxygenase in Inflammation, Insulin-Signalling, Diabetes and Obesity," *Mediators of Inflammation*, vol. 2010, pp. 1–18, 2010, doi: https://doi.org/10.1155/2010/359732.

[32] Iyden Kamil Mohammed, Ali Hussein Al-Timemy, and J. Escudero, "Two-Stage Classification of Breast Tumor Biomarkers for Iraqi Women," *DOAJ (DOAJ: Directory of Open Access Journals)*, Sep. 2020.

[33] R. A. R. Mahmood, A. Abdi, and M. Hussin, "Performance Evaluation of Intrusion Detection System using Selected Features and Machine Learning Classifiers," *Baghdad Science Journal*, vol. 18, no. 2(Suppl.), p. 0884, Jun. 2021, doi: https://doi.org/10.21123/bsj.2021.18.2(suppl.).0884.

[34] F. H. Hassan and M. A. Omar, "Recurrent Stroke Prediction using Machine Learning Algorithms with Clinical Public Datasets: An Empirical Performance Evaluation," *Baghdad Science Journal*, vol. 18, no. 4(Suppl.), p. 1406, Dec. 2021, doi: https://doi.org/10.21123/bsj.2021.18.4(suppl.).1406.

[35] S. Qadir Mohammed and M. A. Hussein, "Performance Analysis of different Machine Learning Models for Intrusion Detection Systems," Journal of Engineering, vol. 28, no. 5, pp. 61–91, May 2022, doi: https://doi.org/10.31026/j.eng.2022.05.05.

[36] N. Z. Salih and W. Khalaf, "Improving Students Performance Prediction Using Machine Learning and Synthetic Minority Oversampling Technique," Journal of Engineering and Sustainable Development, vol. 25, no. 6, pp. 56–64, Nov. 2021, doi: https://doi.org/10.31272/jeasd.25.6.6.

[37] N. M. Ghadi and N. H. Salman, "Deep Learning-Based Segmentation and Classification Techniques for Brain Tumor MRI: A Review," Journal of Engineering, vol. 28, no. 12, pp. 93–112, Dec. 2022, doi: https://doi.org/10.31026/j.eng.2022.12.07.

[38] N. A. Obuchowski and J. A. Bullen, "Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine," *Physics in Medicine & Biology*, vol. 63, no. 7, p. 07TR01, Mar. 2018, doi: https://doi.org/10.1088/1361-6560/aab4b1.

[39] F. K. Nasser and S. F. Behadili, "Breast Cancer Detection using Decision Tree and K-Nearest Neighbour Classifiers," Iraqi Journal of Science, pp. 4987–5003, Nov. 2022, doi: https://doi.org/10.24996/ijs.2022.63.11.34.

[40] Z. Duran, İ. Akargöl, and T. Doğan, "Data Mining, Weka Decision Trees," Orclever Proceedings of Research and Development, vol. 3, no. 1, pp. 401–416, Dec. 2023, doi: https://doi.org/10.56038/oprd.v3i1.376.

[41] Q.H. Mohammed and A. Namburu, "Hybrid Model and Framework for Predicting Air Pollutants in Smart Cities," Journal of Engineering and Sustainable Development, vol. 28, no. 3, pp. 392–406, May 2024, doi: https://doi.org/10.31272/jeasd.28.3.9.

[42] P. Kumar, A. Bhatnagar, R. Jameel, and Ashish Kumar Mourya, "Machine Learning Algorithms for Breast Cancer Detection and Prediction," Lecture notes in networks and systems, pp. 133–141, Jan. 2021, doi: https://doi.org/10.1007/978-981-16-0695-3_14.

[43] S. Taha Ahmed, R. Al-Hamdani, and M. Sadik Croock, "Enhancement of student performance prediction using modified K-nearest neighbor," TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 18, no. 4, p. 1777, Aug. 2020, doi: https://doi.org/10.12928/telkomnika.v18i4.13849.

[44] P. Et al., "A Semi-Supervised Machine Learning Approach Using K-Means Algorithm to Prevent Burst Header Packet Flooding Attack in Optical Burst Switching Network," *Baghdad Science Journal*, vol. 16, no. 3(Suppl.), p. 0804, Sep. 2019, doi: https://doi.org/10.21123/bsj.2019.16.3(suppl.).0804.

[45] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, "Machine Learning Classification Techniques for Breast Cancer Diagnosis," IOP Conference Series: Materials Science and Engineering, vol. 495, p. 012033, Jun. 2019, doi: https://doi.org/10.1088/1757-899x/495/1/012033.

[46] Taha Mohammed Hasan, Sahab Dheyaa Mohammed, and Jumana Waleed, "DEVELOPMENT OF BREAST CANCER DIAGNOSIS SYSTEM BASED ON FUZZY LOGIC AND PROBABILISTIC NEURAL NETWORK," Восточно-Европейский журнал передовых технологий, vol. 4, no. 9–106, 2020, Accessed: Sep. 17, 2024. [Online]. Available: https://cyberleninka.ru/article/n/development-of-breast-cancer-diagnosis-system-based-on-fuzzy-logic-and-probabilistic-neural-network

[47] R. Bhoopal, P. Sharma, S. Kumar, R. Singh, and R. Beniwal, "Neural network-based prediction of effective thermal conductivity of loose multi-phase systems," *Indian Journal of Pure & Applied Physics*, vol. 51, pp. 118–124, 2013, Accessed: Sep. 17, 2024. [Online]. Available: https://nopr.niscpr.res.in/bitstream/123456789/15914/1/IJPAP%2051(2)%20118-124.pdf

[48] T. Gibet, E. Hicham, E. Amrani Chaker, and Lotfi, "The Optimization Of Compute Resources Scheduling In Cloud Computing Environments Using Artificial Neural Networks," Journal of Theoretical and Applied Information Technology, vol. 15, no. 13, 2018, Accessed: Sep. 19, 2024. [Online]. Available: https://www.jatit.org/volumes/Vol96No13/22Vol96No13.pdf.

[49] J. Sultana and A. Khader Jilani, "Predicting Breast Cancer Using Logistic Regression and Multi-Class Classifiers," International Journal of Engineering & Technology, vol. 7, no. 4.20, p. 22, Nov. 2018, doi: https://doi.org/10.14419/ijet.v7i4.20.22115.