OPEN ACCESS
Computer Science & IT Research Journal
P-ISSN: 2709-0043, E-ISSN: 2709-0051
Volume 5, Issue 7, P.1605-1620, July 2024
DOI: 10.51594/csitrj.v5i7.1306
Fair East Publishers
Journal Homepage: www.fepbl.com/index.php/csitrj

Prediction of breast cancer based on machine learning

Hind I. Mohammed[1], Sabah A. Abdulkareem[2], & Shaimaa Khamees Ahmed[3]

[1]Department of Mathematics, Al-Muqdad, College of education,
University of Diyala, Diyala, Iraq.
[2, 3]Department of Computer Engineering, College of Engineering,
University of Diyala, Diyala, Iraq

*Corresponding Author: Hind I. Mohammed
Corresponding Author Email: hindim@uodiyala.edu.iq / sbh_anwar@uodiyala.edu.iq
Article Received: 07-02-24 **Accepted:** 05-05-24 **Published:** 17-07-24

ABSTRACT

Breast cancer is a frequent cancer that develops when normal cells in the breast transform into malignant cells. Breast cancer can arise from glandular tissue, muscular tissue, or fatty tissue in the breast. Many variables contribute to the risk of breast cancer, including genetics, environmental exposure, food, and lifestyle. Breast cancer should be detected early through breast self-examination, regular clinical evaluation, and mammography to identify any abnormal changes, In recent years, early detection of breast cancer in women has emerged as a beacon of hope and a pivotal point in the treatment of this dangerous disease, and its timely identification has become paramount. Modern advancements in technology, especially artificial intelligence algorithms, have played a vital role in developing systems that facilitate automated disease detection, diagnosis, rapid response, and a reduced risk of fatalities. This paper delves into a comparative study of various machine learning (ML) techniques, namely logistic regression (LR), support vector machines (SVM), linear SVM, Gaussian Naive Bayes (GNB), and artificial neural networks

(ANNs). The evaluation metrics used in this study are accuracy and elapsed time. The results show that Gaussian Naive Bayes achieved the highest accuracy of 94.07% in just 0.005495 seconds, outperforming SVM (91.85%), linear SVM (90.19%), logistic regression (87.04%), and ANN (37.04%). These findings highlight the potential of Gaussian Naive Bayes in aiding the early detection of breast cancer, leading to more effective and timely interventions, ultimately improving patient outcomes.

Keywords: Breast Cancer, Machine learning (ML), Logistic Regression (LR), Support Vector Machine (SVM), Linear SVM, Gaussian Naive Bayes (GNB) and Artificial Neural Networks (ANNs).

INTRODUCTION

Breast cancer is a type of cancer that develops in the breast tissue cells and ranks second as the most prevalent cancer among women in the United States, just behind skin cancer. Although breast cancer can affect both men and women, it is more frequently diagnosed in women (Fatima, et. al., 2020). Significant backing for breast cancer awareness and research funding has been instrumental in driving progress in the Diagnosis and treatment of breast cancer (Sun, et. al., 2017). Over time, breast cancer survival rates have shown significant improvement, and the mortality rates linked to the disease have steadily declined. Several key factors have contributed to these positive trends, including early detection practices, advancements in personalized treatment approaches, and a deeper comprehension of the disease (Lu, et. al., 2018). In 2020, breast cancer emerged as the most prevalent form of cancer, with over 2.2 million reported cases. Alarmingly, about one in 12 women are projected to develop breast cancer at some point in their lives. Tragically, breast cancer also ranked as the leading cause of cancer-related deaths among women, with approximately 685,000 women succumbing to the disease in the same year. The burden of breast cancer is disproportionately higher in low- and middle-income countries. There exist significant disparities between high-income and low- and middle-income countries regarding breast cancer outcomes. In high-income countries, the 5-year survival rate after breast cancer exceeds 90%, while in India, it remains at around 66 %, and in South Africa, it is as low as 40%.Africa and Polynesia exhibit the highest age-standardized death rates from breast cancer. Notably, in sub-Saharan Africa, a concerning fact is that half of all breast cancer deaths occur among women under the age of 50. Encouragingly, substantial strides have been made in breast cancer treatment since 1980. High-income countries have witnessed a remarkable 40% decline in age-standardized death rates from breast cancer between the 1980s and 2020 (https://www.who.int/ar/news-room/fact-sheets/detail/breast-cancer). Breast cancer has different types, as in Figure 1 (Fatima, et. al., 2020). Breast cancer encompasses different types depending on how the affected cells and tissues spread within the body. One such type is Ductal Carcinoma in Situ (DCIS), where abnormal cells spread outside the breast, but the cancer remains non-invasive. Another type is Invasive Ductal Carcinoma (IDC), also known as Infiltrative Ductal Carcinoma, which occurs when abnormal breast cells spread throughout the breast tissue. Interestingly, IDC is typically found in men as well (Zhang, et. al., 2019). The third type of breast cancer is Mixed Neoplasia Breast Cancer (MTBC), also referred to as Invasive Breast Cancer (Hou, et. al., 2019).

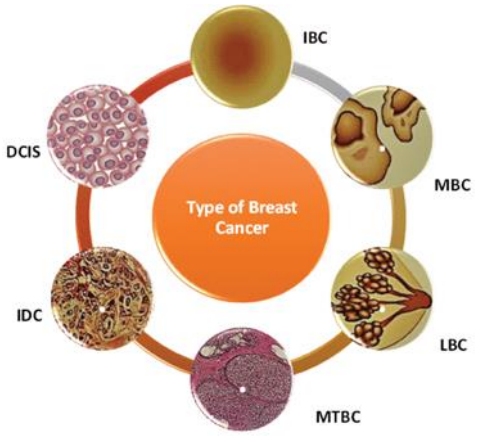

Figure 1: Demonstration of Major Types of Breast Cancer (Fatima, et. al., 2020)

ML falls under the umbrella of AI. Is dedicated, to developing systems of learning and evolving based on data and information (Mohammed, et. al., 2023). ML along with Deep Learning (DL) methods has been widely applied across sectors in the field of healthcare. The medical industry heavily depends on analyzing data for diagnostics; automating these procedures can greatly enhance result accuracy. ML classification models have played a role in minimizing errors that inexperienced medical professionals may make and providing accurate outcomes. Nonetheless challenges can arise when employing ML techniques, including issues, like classification models, ineffective validation processes and unnecessary unweighted features. Additionally, breast cancer classification and prediction in medical imaging encounter complexities, making it a challenging and crucial task. Nonetheless, integrating ML and DL methods has shown promising potential in revolutionizing medical diagnostics and treatment (Hirra, et. al., 2021, Elsadig, et. al., 2023)

The present study is divided into different sections. Section (2) is about the Related Works that are being used for breast cancer prediction, section (3) is about Methods and Materials, section (4) is about the data pre-processing, section (5) is about Performance Measure Parameters, section (6) is the implementation and Result Analysis and section (7) provides the conclusion of this paper.

Related Works

AI technologies are commonly used to find and diagnose diseases automatically. According to the authors in (Zhang, et. al.,2019), breast cancer is the most widespread cancer in women worldwide and is considered a many-factor disease. These may be diverse social, economic, clinical and lifestyle factors (Rabiei, et. al., 2022).

In a search introduced by the authors (Liu, et. al., 2018), they used a dataset consisting of 569 inputs to verify the application of an LR classifier model. The primary goal of this study was the Wisconsin diagnosis of breast cancer (WDBC). Feature selection techniques were also adopted by the authors to improve the system performance, including the accuracy. Based on two selected saliant features, texture and perimeter, LR model achieved accuracy of 96.5%. Thus, LR model harnessing feature selection scheme showed improved performance, comparing to LR alone. In (Khourdifi, et. al., 2018), several machine learning models were adopted to identify the cancer as benign or malignant. The authors showed that support vector machine (SVM) outperformed all other models tested in the study, with reported accuracy of 97.9% on WDBC dataset.

Using Weka data mining tool, another comparative study was conducted by the authors of (Keleş, et. al., 2019) who compared 22 machine learning models on antenna dataset. The findings revealed that Random Forest, Random Committee, Bagging, SimpleCART and IBk algorithms achieved superior performance in terms of accuracy metric evaluated on 10-folds. In (Alghunaim, et. al.,2019), Gene Expression (GE) and DNA methylation data were used to train ML models and predict the breast cancer. Three ML models were trained on the data, including RF, SVM, and DT, and evaluated using accuracy classification metric. The results showed that SVM demonstrated the best performance, achieving 98,03% and 99.68%, on both datasets, respectively.

K-Nearest Neighbour (KNN), NB, and J48 models were assessed, and their performance were compared in (Maliha, et. al.,2019) to predict different types of breast cancer. The KNN outperformed the two other models, achieving accuracy of 98.8%. Using modified recursive feature selection technique in IOT environment, the authors of (Memon, et. al.,2019) trained and deployed SVM model adopting various kernel types, including polynomial, sigmoid, RBF kernels. SVM with the three kernel types achieved accuracy of 84%, 97%, and 99%, respectively.

Moreover, the authors of (Bharat, et. al., 2018) compared the performance of SVM, NB, DT, and KNN on WDBC dataset. Their findings showed that SVM outperformed all the rest models with execution time of 0.07. However, KNN reported shorter execution time comparing to SVM. In another comparative study (Bayrak, et. al., 2019).

, the authors compared the performance of SVM and ANN algorithms on WDBC dataset. They found that SVM indicates better performance, reporting accuracy of 96.9%, while ANN achieved accuracy of 95.4%.

Computer aided detection system was implemented in (Omondiagbe, et. al., 2019) where the author used ANN, NB, SVM and WDBC dataset for breast cancer diagnosis. The reported accuracy results revealed performance of 97.0%, 91%, and 96.4% on the three classifiers, respectively. Using Weka data mining tool, the researchers in (Bharati, et. al., 2018) trained several ML models and reported highest performance from KNN algorithm based on Kappa statistics, true positive and false positive rates, and precision evaluation metrics. The authors in (Aruna, et. al., 2011) utilized the CSSFFS algorithm to choose 15 features for enhancing the accuracy rate for other machine learning algorithms. The results indicated that the accuracy rate obtained 92.9%, 92.9%, 92.6% and 93.6% for Simple CART algorithm, J48, Naive Bayes and RBF network, respectively.

The authors (Zheng, et. al.,2014) proposed a hybrid technique combining the K-Mean and SVM algorithms. The K-SVM method achieved an accuracy rate of 97.38% after undergoing K Fold cross-validation, showcasing its superior performance with a minimal error rate.

The authors concluded (Kumar, et. al., 2013) that SVM proved to be the more suitable algorithm for breast cancer prediction. SVM achieved an accuracy rate of 97.59%, surpassing the accuracy of other techniques in their study.

A group of researchers in (Bataineh, et. al., 2019) employed nonlinear machine learning algorithms, and the MLP (Multi-Layer Perceptron) algorithm achieved an accuracy rate of 96.70%, surpassing KNN, NB, and CART algorithms.

The authors in (Padmapriya, et. al., 2016) found that the CART algorithm was the most suitable for breast cancer analysis among the algorithms compared. CART achieved an accuracy rate of 98.50%, while J48 and A DT algorithms achieved 98.10% and 97.70%, respectively.

The authors employed (Williams, et. al., 2015) the J48 DT algorithm and NB, which handles class numbers based on probabilistic theory. The accuracy rate of J48 DT was .3

measured at 94.2%, while Naive Bayes achieved an accuracy rate of 82.6%.

In (Bharati, et. al.,2018), a comparative analysis of ML algorithms was introduced. The accuracy rates for MLP and LR were 64.6% and 68.8%, respectively, while RF, NB, and KNN achieved accuracy rates of 69.5%, 71.6%, and 72.3%, respectively.

In another study (Asri, et. al.,2016), the authors provided a comparative analysis of various algorithms for breast cancer prediction, considering accuracy rates, efficiency, and effectiveness. SVM stood out with the highest accuracy rate of 97.13% and the lowest error rate.

In (Saleh, et. al.,2022) , ML and DL techniques were utilized to predict breast cancer. Regular ML models and optimized deep RNN were applied to selected features, with the univariate method showing the best performance for cross-validation and testing results when applied to the optimized deep RNN with selected features.

The authors, through (El Massari, et. al., 2022), employed ML algorithms and utilized two test modes: 10-fold cross-validation and percentage split. They also used several performance measures. The results showed that the ontological model exhibited superior accuracy, even without feature selection.

(Islam, et. al., 2020) Also tackled the problem using ML techniques. The findings revealed that SVM achieved an accuracy of 97.14%, precision of 95.65%, and F1 score of 0.9777, while ANN obtained the highest accuracy of 98.57%, precision of 97.82%, and F1 score of 0.9890.

Another paper in (Hou, et. al.,2020), the performance of three different ML algorithms was evaluated and compared. The results indicated that all three ML algorithms outperformed LR in accuracy. Notably, XGBoost emerged as the best choice for predicting breast cancer.

In (Naji, et. al.,2021), ML algorithms were used, and the findings showed that SVM achieved the highest accuracy of 97.2% among all other classifiers. Similarly,

The reference (Elsadig, et. al.,2023) ML algorithms were employed, and the results demonstrated that SVM was the best classifier, surpassing even the stack classifier with an accuracy of 97.7% and classification errors of 0.019 false positive (FP) and 0.029 false negative (FN).

Despite of great effort already made by the researchers in the literature for automatic breast cancer detection, we believe that there are still open rooms for further improvements.

MATERIAL & METHODS

Various ML techniques are available to aid in predicting whether an individual is affected by benign or malignant cancer, ensuring an efficient and error-free process.

Linear Support Vector Machine (LSVM)

Linear SVM is employed to handle linearly separable data, which means data that can be divided into two classes using a single straight line. When a dataset can be distinctly separated in this manner, it is referred to as linearly separable data, and the Linear SVM classifier is the appropriate choice for such scenarios (Zareef, et. al., 2020, Kurani, et. al., 2023).

Support Vector Machine (SVM)

One of the most powerful algorithms in ML and artificial intelligence for prediction, particularly in supervised learning, is SVMs (Tharwat, et. al., 2015). To predict point labels, SVMs create a sample from the dataset during training. To distinguish between two classes depending on a set of binaries, they learn a boundary of a linear decision known as training vectors. The model is then evaluated using the derived linear classification rule to classify additional test cases. Linear SVMs are proficient in solving optimization problems, the simplest form of SVM is the solid margin classifier, which identifies the most significant geometric margin with the linear classifier rule (Mohammd, et. al., 2021, Hussein, et. al.,2022).

In real-world datasets, it's common for data to be non-linearly separable, necessitating adjustments to SVM. Using the soft margin principle, this modification balances increasing the geometric margin and minimizing the classification error on the training data points. The soft margin allows for a superplane that accommodates incorrect classifications for complex states while increasing the distance between them and the nearest fully separated data samples (Abdulkareem, et. al.,2022).

Artificial Neural Networks (ANNs)

The multi-layer perceptron architecture known as ANNs (Abdulkareem, et. al.,2022) is employed for training and classifying input samples to produce the desired output. In our ANN model, we used a three-layer network consisting of 32 neurons in the input layer, 34 in the hidden layer, and two in the output layer. The number of neurons in the hidden layer was selected experimentally to achieve optimal performance. The ANN model was trained on the data, and backpropagation was utilized as an improvement technique to adjust the weights of network through gradient descent, reducing error at each stage to enhance the model's performance (Thomas, et. al., 2020). The function of soft-max activation was applied in the output layer to obtain the eventual prediction. The network's input consisted of thirty-two criteria, while the output layer distinguished between the breast cancer presences (0) and absence (1). Equation (1) is defined the cross-entropy loss function as follows:

$$L = -\sum_{i=1}^{2} t_i \log(p_i) \qquad (1)$$

In the context of a specific sample (i) of the ground truth value (t) and the probability (p) obtained through the soft-max activation function (Abdulkareem, et. al.,2022) for that sample, the following relationship holds.

Naive Bayes (NB) Classifier

The Naive Bayes (NB) method, a statistical pattern recognition technique, was applied to detect breast cancer cases by utilizing Bayesian probability. It relies on an acceptable assumption about the data generation process, assuming that all sample attributes are independent (Rahma, et. al.,2022). However, if this assumption is incorrect, the classification hypothesis becomes a mere approximation. Despite this limitation, the NB classifier has demonstrated high accuracy in its predictions, even though the evaluation of the function may be achieved with lower accuracy (Wu, et. al., 2015, Ibrahim, et. al., 2017, Manikandan, et. al., 2023).

Logistic Regression

Logistic Regression, a significant analytical modelling technique, is prominent among various ML algorithms. It relates the probability level to a set of explanatory variables, making it suitable for analysing datasets with one or more independent variables to determine a binary outcome with two possible results. This method is commonly used for binary predictions (1/0, Yes/No, True/False). The LR model can be described using the following equations (2) and (3):

$$x = c_0 + \sum_{i=1}^{n} c_i x_i \quad (2)$$

$$p_{(x)} = \frac{e^x}{1+e^x} \quad (3)$$

In the context of Logistic Regression, the participation quantity of variables x_i (where i = 1, n) and their corresponding regression coefficients c_i are used to determine the highest probability of an event along with its regular errors. In LR, a specific event's probability (P) is calculated using the Bernoulli test and is associated with the sampling event (Dong, et. al., 2016).

DATA PRE-PROCESSING

Data Set Description

The dataset of breast cancer was uploaded from the UCI ML repository (Breast Cancer Wisconsin (Original) Data Set. Available from (https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data). There are 569 persons with thirty-two attributes representing symptoms that may cause breast cancer, where the Diagnosis is either benign (B) or malignant (M). For such situations, 357 (62.741%) are benign, and 212 (37.258%) are malignant. Table (1) shows examples of data applied in the research.

Exploratory Data Analysis (EDA)

In statistics, exploratory data analysis (EDA) is used to thoroughly examine data sets and succinctly represent their essential characteristics, frequently utilizing visual methods. Whether or not a statistical model is used, the main objective of EDA is to reveal and identify significant insights within the data that go beyond formal modelling or hypothesis testing (Camizuli, et. al., 2018); below table 1that shows the results after applied EDA.

Table 1

*The Data set after EDA (**https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data**)*

	count	mean	std	min	25%	50%	75%	max
id	569	3.04E+07	1.25E+08	8670	869218	906024	8.81E+06	9.11E+08
radius mean	569	1.41E+01	3.52E+00	6.981	11.7	13.37	1.58E+01	2.81E+01
texture mean	569	1.93E+01	4.30E+00	9.71	16.17	18.84	2.18E+01	3.93E+01
perimeter mean	569	9.20E+01	2.43E+01	43.79	75.17	86.24	1.04E+02	1.89E+02
area mean	569	6.55E+02	3.52E+02	143.5	420.3	551.1	7.83E+02	2.50E+03
smoothness means	569	9.64E-02	1.41E-02	0.05263	0.08637	0.09587	1.05E-01	1.63E-01
compactness means	569	1.04E-01	5.28E-02	0.01938	0.06492	0.09263	1.30E-01	3.45E-01

concavity means	569	8.88E-02	7.97E-02	0	0.02956	0.06154	1.31E-01	4.27E-01
concave points mean	569	4.89E-02	3.88E-02	0	0.02031	0.0335	7.40E-02	2.01E-01
symmetry mean	569	1.81E-01	2.74E-02	0.106	0.1619	0.1792	1.96E-01	3.04E-01
fractal_ dimension_ mean	569	6.28E-02	7.06E-03	0.04996	0.0577	0.06154	6.61E-02	9.74E-02
radius_ se	569	4.05E-01	2.77E-01	0.1115	0.2324	0.3242	4.79E-01	2.87E+00
texture_ se	569	1.22E+00	5.52E-01	0.3602	0.8339	1.108	1.47E+00	4.89E+00
perimeter_ se	569	2.87E+00	2.02E+00	0.757	1.606	2.287	3.36E+00	2.20E+01
area_ se	569	4.03E+01	4.55E+01	6.802	17.85	24.53	4.52E+01	5.42E+02
smoothness_ se	569	7.04E-03	3.00E-03	0.001713	0.005169	0.00638	8.15E-03	3.11E-02
compactness_ se	569	2.55E-02	1.79E-02	0.002252	0.01308	0.02045	3.25E-02	1.35E-01
concavity_ se	569	3.19E-02	3.02E-02	0	0.01509	0.02589	4.21E-02	3.96E-01
concave points_ se	569	1.18E-02	6.17E-03	0	0.007638	0.01093	1.47E-02	5.28E-02
symmetry_ se	569	2.05E-02	8.27E-03	0.007882	0.01516	0.01873	2.35E-02	7.90E-02
fractal_ dimension_ se	569	3.79E-03	2.65E-03	0.000895	0.002248	0.003187	4.56E-03	2.98E-02
radius_ worst	569	1.63E+01	4.83E+00	7.93	13.01	14.97	1.88E+01	3.60E+01
texture_ worst	569	2.57E+01	6.15E+00	12.02	21.08	25.41	2.97E+01	4.95E+01
perimeter_ worst	569	1.07E+02	3.36E+01	50.41	84.11	97.66	1.25E+02	2.51E+02
area_ worst	569	8.81E+02	5.69E+02	185.2	515.3	686.5	1.08E+03	4.25E+03
smoothness_ worst	569	1.32E-01	2.28E-02	0.07117	0.1166	0.1313	1.46E-01	2.23E-01
compactness_ worst	569	2.54E-01	1.57E-01	0.02729	0.1472	0.2119	3.39E-01	1.06E+00
concavity_ worst	569	2.72E-01	2.09E-01	0	0.1145	0.2267	3.83E-01	1.25E+00

concave points_ worst	569	1.15E-01	6.57E-02	0	0.06493	0.09993	1.61E-01	2.91E-01
symmetry_ worst	569	2.90E-01	6.19E-02	0.1565	0.2504	0.2822	3.18E-01	6.64E-01
fractal_ dimension_ worst	569	8.39E-02	1.81E-02	0.05504	0.07146	0.08004	9.21E-02	2.08E-01

Training and Testing Phase

Divide the data set into the training set and the test set and the percentage (95%, Number of Instances (540)) for training and (0.05%, Number of Instances (29) for testing.

Normalize Data

Data normalization is an essential pre-processing before modelling. The scaling is as follows:

y = ((x - min) / (max-min)) (4) (Nagarajan, et. al.,2013).

Normalized Data with Min_ Max_ Scaler

Several techniques exist for data normalization, such as min-max normalization, z-score normalization, and normalization by decimal scaling. Min-max normalization involves applying a linear transformation to the original data, using the minimum value (min) and the maximum value (max) of the attribute (Al Shalabi, et. al.,2006).

Performance Measure Parameters

ML algorithm's baseline performance was evaluated using a confusion matrix (CM), to measure the performance of the classification model (Deng, et. al., 2016), which is an ML structure that aims to predict information about a classification model's actual and expectant classifications. The CM matrix has two dimensions to index the class: the object's actual class index and the classifier's predicted class index (Castaneda, et. al.,2019). The sample number categorized as class A_j but actually belonging to class Ai is acted by N_{ij} in the confusion matrix (Nuzzi, et. al.,2019).

The number of classification performance measures is evaluated using different evaluation factors. Accuracy or classification Rate, Recall or Sensitivity, specificity, Precision, and F1-score or F-measure are among the evaluation metrics, which are defined as: (Horn, et. al., 2017).

Accuracy or Classification Rate

The relationship between the actual accurate classification numbers and the total number of test samples applied during training and testing refers to accuracy and is calculated by using equation (5).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (5) \text{ (Salih, et. al., 2019)}$$

Where truth positive (TP), false positive (FP), true negative (TN), and false negative (FN), each one has a particular meaning in the confusion matrix.

IMPLEMENTATION AND RESULT ANALYSIS

Experimental Setup

To classify cells as benign or malignant, we employed five ML techniques: LR, SVM, linear SVM, GNB, and ANN. Each method was individually implemented under specific system requirements, including hardware processor with Core i7- CPU 8550U running at 2.00 GHz, 8 GB of RAM and the Windows-10 operating system. The programming language used for the implementation was Python (version 3.7.10, 64-bit).

RESULTS AND DISCUSSION

The data set was divided into 95% for training and 0.05 for testing for all five methods, where we used 29 cases to test all models. The CM of the ML strategies used is shown in Tables 2, 3, 4, 5, and 6, which provide predictive results for LR, SVM, linear SVM, GNB, and ANN, respectively.

Table 2
CM for LR

		Malignant	**Benign**
Actual	Malignant	14	2
	Benign	2	11
	Predicted		

Table 3
CM for SVM

		Malignant	**Benign**
Actual	**Malignant**	12	4
	Benign	0	13
	Predicted		

Table 4
CM for Linear SVM

		Malignant	Benign
Actual	**Malignant**	13	2
	Benign	1	13
	Predicted		

Table 5
CM for GB

		Malignant	Benign
Actual	Malignant	14	2
	Benign	0	13
	Predicted		

Table 6
CM for ANN

		Malignant	**Benign**
Actual	**Malignant**	7	7
	Benign	12	3
	Predicted		

In table 7. The combined CM depicts that the GNB and LR model predicts the highest number of true positives (14 out of 29 test samples) among the five techniques. In addition, the SVM and

GNB models predict the highest number of true negatives and the lowest number of false negatives (13 of 29 test samples) and (0 of 29 test samples), respectively.

Table7
The Combined Confusion Matrix

	LR	SVM	Linear SVM	GNB	ANN
TP	14	12	13	14	7
TN	11	13	13	13	3
FP	2	4	2	2	7
FN	2	0	1	0	12

The measures of calculated performance are shown in Figure 2., & Table 8, where they show that GNB has outperformed all other ML techniques in terms of accuracy and elapsed time for execution then the other four algorithms are, respectively SVM, Linear SVM, LR, and ANN, which are considered to have failed to build this model with the least accuracy and the highest execution time, according to the values in Table 8.

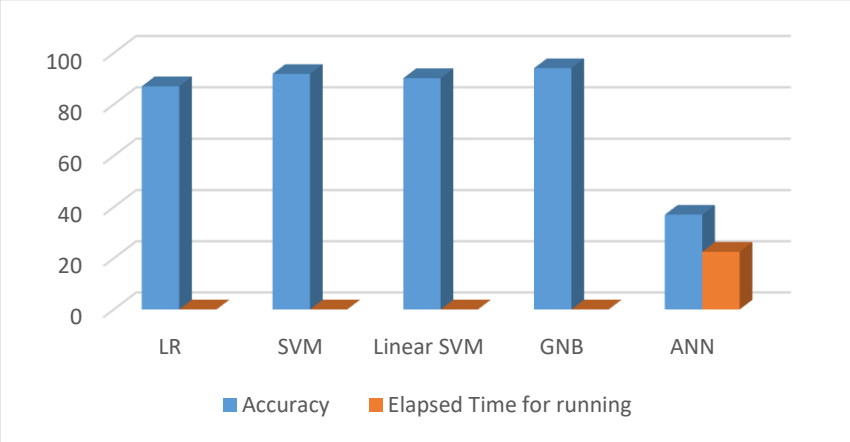

Figure 2: Comparison of Algorithms

Table 8
ML Techniques in Accuracy and Elapsed Time

NO.	Algorithm	Accuracy	Elapsed Time
1.	LR	87.04	0.020136
2.	SVM	91.85	0.016133
3.	Linear SVM	90.19	0.007103
4.	GNB	94.07	0.005495
5.	ANN	37.04	22.523813

CONCLUSION

The current research compared five machine learning (ML) techniques for breast cancer prediction: LR, SVM, linear SVM, GNB, and ANN. The study provided an overview of each ML

approach's fundamental features and working principles. Among the five techniques, GNB achieved the highest accuracy of 94.57%, while ANN showed the lowest accuracy of 37.04%. It's worth mentioning that ANN excels in accuracy when dealing with image-related tasks. For the remaining three algorithms, SVM achieved an accuracy of 91.85%, linear SVM achieved 90.19%, and LR achieved 87.04%. Diagnostics in the medical field is costly and time-consuming. The system suggested that the ML technology could act as a clinical assistant for breast cancer diagnosis and would be very useful to Medical diagnostics are known for being costly and time-consuming. However, the introduction of ML technology as a clinical assistant for breast cancer diagnosis has shown promising potential to benefit physicians and new doctors, particularly in misdiagnosis cases. Among the ML models studied, the one developed by GNB displayed remarkable consistency, surpassing other technologies. This advancement has the potential to revolutionize breast cancer prediction practices. The research findings lead to the conclusion that ML techniques can automatically detect diseases with high accuracy.

References

Abdulkareem, S. A., Radhi, H. Y., Fadil, Y. A., & Mahdi, H. (2022). Soft computing techniques for early diabetes prediction. *Indonesian Journal of Electrical Engineering and Computer Science, 25*(2), 1167-1176.

Al Shalabi, L., & Shaaban, Z. (2006). Normalization as a preprocessing engine for data mining and the approach of preference matrix. In International Conference on Dependability of Computer Systems (207-214). IEEE.

Alghunaim, S., & Al-Baity, H.H. (2019). On the scalability of machine-learning algorithms for breast cancer prediction in big data context. *IEEE Access*, *7*, 91535-91546.

Aruna, S., & Rajagopalan, S. P. (2011). A novel SVM based CSSFFS feature selection algorithm for detecting breast cancer. *International Journal of Computer Applications, 31*(8), 14-20.

Asri, H., Mousannif, H., & Al Moatassime, H. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science, 83*, 1064-1069.

Bataineh, A. A. (2019). A comparative analysis of nonlinear machine learning algorithms for breast cancer detection. *International Journal of Machine Learning and Computing, 9*(3), 248-254.

Bayrak, E. A., Kırcı, P., & Ensari, T. (2019, April). Comparison of machine learning methods for breast cancer diagnosis. In *2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT)* (1-3). IEEE.

Bharat, A., Pooja, N., & Reddy, R. A. (2018, October). Using machine learning algorithms for breast cancer risk prediction and diagnosis. In *2018 3rd International conference on circuits, control, communication and computing (I4C)* (1-4). IEEE.

Bharati, S., Rahman, M. A., & Podder, P. (2018). Breast cancer prediction applying different classification algorithm with comparative analysis using WEKA. In 2018 4th International Conference on Electrical Engineering and Information and Communication Technology (iCEEiCT) (581-584). IEEE.

Bharati, S., Rahman, M. A., & Podder, P. (2018, September). Breast cancer prediction applying different classification algorithm with comparative analysis using WEKA. In *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)* (581-584). IEEE.

Breast Cancer Wisconsin (Original) Data Set. Available from https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

Camizuli, E., & Carranza, E. J. (2018). Exploratory data analysis (EDA). In The Encyclopedia of Archaeological Sciences (1-7).

Castaneda, G., Morris, P., & Khoshgoftaar, T. M. (2019). Evaluation of maxout activations in deep learning across several big data domains. *Journal of Big Data*, *6*, 1-35.

Deng, X., Liu, Q., Deng, Y., & Mahadevan, S. (2016). An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences*, *340*, 250-261.

Dong, L., Wesseloo, J., Potvin, Y., & Li, X. (2016). Discrimination of mine seismic events and blasts using the Fisher classifier, naive Bayesian classifier and logistic regression. *Rock Mechanics and Rock Engineering*, *49*, 183-211.

El Massari, H., Gherabi, N., Mhammedi, S., Ghandi, H., Qanouni, F., & Bahaj, M. (2022). An ontological model based on machine learning for predicting breast cancer. *International Journal of Advanced Computer Science and Applications*, *13*(7).

Elsadig, M. A., Altigani, A., & Elshoush, H. T. (2023). Breast cancer detection using machine learning approaches: a comparative study. *International Journal of Electrical & Computer Engineering (2088-8708)*, *13*(1).

Elsadig, M. A., Altigani, A., & Elshoush, H. T. (2023). Breast cancer detection using machine learning approaches: a comparative study. *International Journal of Electrical & Computer Engineering*, *13*(1), 2088-8708.

Fatima, N., Liu, L., Hong, S., & Ahmed, H., (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, *8*, 150360-150376.

Hirra, I., Ahmad, M., Hussain, A., Ashraf, M.U., Saeed, I.A., Qadri, S.F., Alghamdi, A.M., & Alfakeeh, A.S., (2021). Breast cancer classification from histopathological images using patch-based deep learning modeling. *IEEE Access*, *9*, 24273-24287.

Horn, Z. C., Auret, L., McCoy, J. T., Aldrich, C., & Herbst, B. M. (2017). Performance of convolutional neural networks for feature extraction in froth flotation sensing. *IFAC-Papers Online*, *50*(2), 13-18.

Hou, C., Zhong, X., He, P., Xu, B., Diao, S., & Yi, F. (2020). Predicting breast cancer in Chinese women using machine learning techniques: algorithm development. *JMIR Medical Informatics*, *8*(6).

Hou, R., Mazurowski, M.A., Grimm, L.J., Marks, J.R., King, L.M., Maley, C.C., Hwang, E.S.S., & Lo, J.Y., (2019). Prediction of upstaged ductal carcinoma in situ using forced labeling and domain adaptation. *IEEE Transactions on Biomedical Engineering*, *67*(6), 1565-1572.

Hussein, R. R. A., & Najeeb, H. D. (2022). Improving measurement of effectiveness of blended learning in Iraqi education using SVM. *Iraqi Journal of Science, 63*(9), 4057-4066.

Ibrahim, A. A. E., Hashad, A. I., & Shawky, N. E. M. (2017). A comparison of open source data mining tools for breast cancer classification. In Handbook of Research on Machine Learning Innovations and Trends (636-651). IGI Global.

Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science, 1*, 1-14.

Keleş, M. K. (2019). Breast cancer prediction and detection using data mining classification algorithms: a comparative study. *Tehnički Vjesnik, 26*(1), 149-155.

Khourdifi, Y., & Bahaj, M. (2018, December). Applying best machine learning algorithms for breast cancer prediction and classification. In *2018 International conference on electronics, control, optimization and computer science (ICECOCS)* (1-5). IEEE.

Kumar, G. R., Ramachandra, G. A., & Nagamani, K. (2013). An efficient prediction of breast cancer data using data mining techniques. *International Journal of Innovations in Engineering and Technology (IJIET), 2*(4), 139.

Kurani, A., Doshi, P., Vakharia, A., & Shah, M. (2023). A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. *Annals of Data Science, 10*(1), 183-208.

Liu, L. (2018, May). Research on logistic regression algorithm of breast cancer diagnose data by machine learning. In *2018 International Conference on Robots & Intelligent System (ICRIS)* (157-160). IEEE.

Lu, Y., Li, J.Y., Su, Y.T., & Liu, A.A. (2018). A review of breast cancer detection in medical images. *2018 IEEE Visual Communications and Image Processing (VCIP)*, 1-4.

Maliha, S.K., Ema, R.R., Ghosh, S.K., Ahmed, H., Mollick, M.R.J., & Islam, T., (2019, July). Cancer disease prediction using naive bayes, K-nearest neighbor and J48 algorithm. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (1-7). IEEE.

Manikandan, P., Durga, U., & Ponnuraja, C. (2023). An integrative machine learning framework for classifying SEER breast cancer. *Scientific Reports, 13*(1), 5362.

Memon, M. H., Li, J. P., Haq, A. U., Memon, M. H., & Zhou, W. (2019). Breast cancer detection in the IOT health environment using modified recursive feature selection. *Wireless Communications and Mobile Computing, 2019*(1), 5176705.

Mohammed, H. I., Waleed, J., & Albawi, S. (2021). An inclusive survey of machine learning-based hand gestures recognition systems in recent applications. In IOP Conference Series: Materials Science and Engineering (1076, 1, 070012).

Mohammed, H.I., & Waleed, J., (2023, March). Hand gesture recognition using a convolutional neural network for arabic sign language. In *AIP Conference Proceedings* 2475 (1). AIP Publishing.

Nagarajan, S., & Subashini, T. S. (2013). Static hand gesture recognition for sign language alphabets using edge-oriented histogram and multi-class SVM. *International Journal of Computer Applications, 82*(4).

Naji, M. A., El Filali, S., Aarika, K., & Benlahmar, E. L. H. (2021). Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Computer Science, 191*, 487-492.

Nuzzi, C., Pasinetti, S., Lancini, M., Docchio, F., & Sansoni, G. (2019). Deep learning-based hand gesture recognition for collaborative robots. *IEEE Instrumentation & Measurement Magazine, 22*(2), 44-51.

Omondiagbe, D.A., Veeramani, S., & Sidhu, A.S., (2019, June). Machine learning classification techniques for breast cancer diagnosis. In *IOP conference series: materials science and engineering* (495, 012033). IOP Publishing.

Padmapriya, B., & Velmurugan, T. (2016). Classification algorithm based analysis of breast cancer data. *International Journal of Data Mining Techniques and Applications, 5*(1), 43-49.

Rabiei, R., Ayyoubzadeh, S. M., Sohrabei, S., Esmaeili, M., & Atashi, A. (2022). Prediction of breast cancer using machine learning approaches. *Journal of Biomedical Physics & Engineering*, *12*(3), 297.

Rahma, M. M., & Salman, A. D. (2022). Heart disease classification–based on the best machine learning model. *Iraqi Journal of Science, 63*(9), 3966-3976.

Saleh, H., Abd-El Ghany, S.F., Alyami, H., & Alosaimi, W., (2022). Predicting breast cancer based on optimized deep learning approach. *Computational Intelligence and Neuroscience*, *2022*(1),1820777.

Salih, M. M., Ahmed, M. A., Al-Bander, B., Hasan, K. F., Shuwandy, M. L., and Al-Qaysi, Z. T. (2023). Benchmarking framework for COVID-19 classification machine learning method based on fuzzy decision by opinion score method. *Iraqi Journal of Science, 64*(2), 922-943.

Sun, Y.S., Zhao, Z., Yang, Z.N., Xu, F., Lu, H.J., Zhu, Z.Y., Shi, W., Jiang, J., Yao, P.P., & Zhu, H.P., (2017). Risk factors and preventions of breast cancer. *International Journal of Biological Sciences*, *13*(11), 1387.

Tharwat, A., Gaber, T., Hassanien, A. E., Shahin, M. K., & Refaat, B. (2015). Sift-based Arabic sign language recognition system. In Afro-European Conference for Industrial Advancement: Proceedings of the First International Afro-European Conference for Industrial Advancement AECIA 2014 (334, 359-370). Springer International Publishing.

Thomas, T., Pradhan, N., & Dhaka, V. S. (2020). Comparative analysis to predict breast cancer using machine learning algorithms: a survey. In International Conference on Inventive Computation Technologies (ICICT) (192-196). IEEE.

Williams, K., Idowu, P. A., Balogun, J. A., & Oluwaranti, A. I. (2015). Breast cancer risk prediction using data mining classification techniques. *Transactions on Networks and Communications, 3*(2), 01-11.

Wu, W., Nagarajan, S., & Chen, Z. (2015). Bayesian machine learning: EEG/MEG signal processing measurements. *IEEE Signal Processing Magazine, 33*(1), 14-36.

Zareef, M., Chen, Q., Hassan, M. M., Arslan, M., Hashim, M. M., Ahmad, W., & Kutsanedzie, F. Y. H. (2020). An overview on the applications of typical non-linear algorithms coupled with NIR spectroscopy in food analysis. *Food Engineering Reviews, 12*, 173-190.

Zhang, X., Shengli, S. U., & Hongchao, W. A. (2019). Intelligent diagnosis model and method of palpation imaging breast cancer based on data mining. *Big Data Research*, *5*(1), 2019005.

Zhang, X., Shengli, S.U., & Hongchao, W.A., (2019). Intelligent diagnosis model and method of palpation imaging breast cancer based on data mining. *Big Data Research*, *5*(1), 2019005.

Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications, 41*(4), 1476-1482.

Acknowledgements

I am grateful to the College of Education Al-Miqdad and the College of Engineering / University of Diyala / Iraq for providing support and encouragement to publish research in reputable scientific journals.

Conflict of Interest Statement

No conflict of interest between the three researchers was declared as they have the same scientific specialization and will continue research cooperation in the future