Leveraging Machine Learning for Effective Breast Cancer Diagnosis

RAHMA ABU SALMA[1], HAYEL KAFAJEH[2], RAED ALAZAIDAH[2], MAHMOUD ASSASFEH[3],
ALA'A SAEB AL SHERIDEH[3], NAWAF ALSHDAIFAT[4]

[1]Department of Computer Science, Faculty of IT,
Zarqa University,
Zarqa,
JORDAN

[2]Department of AI, Faculty of IT,
Zarqa University,
Zarqa,
JORDAN

[3]Department of Cyber Security, Faculty of IT,
Zarqa University,
Zarqa,
JORDAN

[4]Faculty of IT,
Applied Science Private University,
Amman,
JORDAN

Abstract: - Breast cancer is a prevalent global health concern, constituting 25% of female cancer cases. Early diagnosis through mammogram screening is effective, but limitations exist, particularly in dense breast cases. Machine Learning (ML) emerges as a promising tool for precise diagnosis. This study aims to identify optimal ML strategies, classifiers, and feature selection techniques for breast cancer diagnosis. This study analyses three breast cancer datasets, categorizing them by location, type (benign or malignant), and recurrence. We evaluate twenty-two classifiers across six ML strategies, taking into account accuracy, precision, recall, and ROC area metrics. We employ five feature selection techniques on 50% of the features. The results are promising for the adoption of ML in breast cancer diagnostics, with accuracy reaching higher than 93% for some applications. It is found that the HT and J48 classifiers from the Trees strategy and the NB classifier from the Bayesian strategy revealed promising results in the diagnostics and detection of breast cancer compared to other analyzed classifiers. Using feature reduction techniques in detecting the type of breast cancer (Benign/Malignant), the correlationAtriEval technique was found to have the highest performance, while the RellieffAttriEval technique has the highest performance when employed for feature reduction in detecting types of breast cancer (recursive/non-recursive).

Key-Words: - Algorithms, breast cancer, classifiers, feature selection, machine learning, prediction.

Received: March 15, 2024. Revised: August 17, 2024. Accepted: September 17, 2024. Available online: October 21, 2024.

1 Introduction

Breast cancer is an abnormal growth of breast cells. Breast cancer is one of the most common cancers among women worldwide, accounting for approximately 25% of all female cancer cases, [1]. One in three new females will receive a cancer diagnosis at some point in their lives in 2023, [2]. The prevalence of breast cancer increased and mortality decreased due to improved diagnostic criteria, [3].

Early and regular checkups for breast cancer are crucial. A diagnostic mammogram can effectively assess the suspected area of tumor development, and mammogram screening reduces mortality to 0%, [4]. However, a mammogram is not effective in the diagnosis of dense breast cancer, [5].

Reducing death rates and improving patient outcomes depend heavily on early detection and precise diagnosis. However, modeling can effectively assist radiologists in breast cancer diagnosis and classification, as well as in identifying high-risk patients, [6]. ML has become a potent tool with great promise for improving diagnostic precision and enabling individualized treatment plans, [7]. ML, a modeling approach within Artificial Intelligence (AI), allows machines to learn through experience, eliminating the need for explicit programming, by exposing them to various datasets, [8], [9]. In recent decades, ML techniques have become widely used in the development of prediction models to aid in efficient decision-making. Cancer research could use these methods to distinguish between benign and malignant cancers by identifying various patterns in data collection. ML represents the process of extracting knowledge from data and discovering hidden relationships, [10], [11].

Classification techniques play an important role in breast cancer. Much research has demonstrated the importance of breast cancer prediction with different techniques. Accuracy, Recall, Precision, and Area We can evaluate the efficacy of these methods using metrics under the ROC, [12]. Therefore, the current study's objectives are as follows:

1. The goal is to pinpoint a successful educational approach for diagnosing breast cancer.
2. We aim to identify the most efficient classifiers for diagnosing breast cancer.
3. The goal is to determine the most effective feature selection technique for the different machine learning models used in breast cancer diagnosis.

The second section of the paper discusses related work, while the third and fourth sections cover the research methodology, analysis, and results. The fifth section provides a summary of the results. The final section discusses the Future Work and Conclusions.

2 Related work

Breast cancer is a highly perilous condition that poses a significant threat to women's well-being. Breast cancer poses significant challenges to the well-being and physical condition of women. Researchers and institutions are making significant efforts in the diagnosis and treatment of breast cancer. The conventional breast cancer diagnosis procedure necessitates medical professionals to repeatedly analyze patient data. The diagnosis of breast cancer involves the examination of cell morphology through pathological analysis, as well as the use of imaging techniques such as mammography, magnetic resonance imaging (MRI), CT scans, and ultrasound. These unique instruments give clinicians the opportunity to examine patients' afflicted organs, which would otherwise be imperceptible without the aid of technology. In this scenario, doctors utilize algorithmic technology to receive prompt feedback and a highly probable reference outcome, a crucial step in improving diagnostic efficiency and reducing their workload, [13]. Advancements in computer performance and machine learning have made it evident that machines are increasingly replacing humans. Intelligent algorithms have demonstrated the ability to substitute human behavior and decision-making in certain domains. Machine learning technology has the capability to automate data processing and utilize precise mathematical models to provide a comprehensive description of the data. The medical industry currently uses machine learning technology to aid in the identification of various diseases, [14].

The use of algorithmic technology is particularly important in enhancing the precision of a physician's diagnosis of preexisting breast cancer. In [15], the researcher conducts performance verification experiments using Wisconsin breast cancer data from the UCI database. The objective is to assess the effectiveness of the Whale Optimization Algorithm-Support Vector Machine (WOA-SVM) algorithm in enhancing the accuracy of breast cancer recognition. Experimental results show that the WOA-SVM model outperforms the conventional breast cancer recognition model in terms of recognition accuracy.

Metastatic breast cancer (MBC) is a prominent contributor to cancer-related fatalities in women. We use the Anaconda-Jupiter notebook to develop Python programming modules for text mining, data processing, and machine learning, aiming to establish a non-invasive breast cancer classification system that can detect cancer metastases. We conduct the evaluation of the ML models' prediction performance using classification model cross-validation criteria, such as accuracy, AUC, and ROC. The EMR created a text-mining framework that facilitated the segregation of blood profile data and the identification of patients with metastatic breast cancer (MBC). The removal of outliers from the blood profile data significantly enhanced the accuracy of ML models. The Decision Tree (DT) classifier achieved an Area Under the Curve (AUC) of 0.87 with an accuracy of 83%. We then employed

Flask to install decision tree classifiers and construct a web application for the reliable diagnosis of metastatic breast cancer (MBC) patients. In summary, the researchers determined that machine learning models utilizing blood profile data could assist doctors in identifying intensive-care MBC patients, leading to enhanced overall survival rates, [16].

The goal is to create a precise model that can accurately detect and classify cancers. Support Vector Machine (SVM) was compared to five other machine learning algorithms in [17]. These were K Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree Classifier (CART), Naive Bayes Classifier (NB), and Linear Discriminant Analysis (LDA). The evaluation was based on estimations of accuracy, ROC area, precision, and recall. ML algorithms have proven to be highly effective in categorizing data, to the extent that the medical industry extensively utilizes them for diagnostic purposes. The SVM technique demonstrated the highest accuracy and the most optimal performance, [17].

The study in [2] uses the original Wisconsin breast cancer datasets to compare how well support vector machines (SVM), decision trees (C4.5), naive bayes (NB), and K-nearest neighbors (KNN) work as machine learning methods. By employing the WEKA data mining tool in a simulated environment, we assessed the effectiveness and efficiency of each algorithm based on metrics such as accuracy, precision, sensitivity, and specificity. The objective was to ascertain the accuracy of the data categorization process. Based on empirical evidence, support vector machines (SVM) demonstrate the lowest error rate and the highest level of accuracy, reaching 97.13% [2]. The research objective in [18] is to use the significant capabilities of machine-learning algorithms for early CC prediction. We have employed three renowned feature selection and ranking strategies to identify the key qualities that significantly contribute to the diagnosis process. In addition, we trained and thoroughly evaluated eighteen different classifiers from six learning methodologies using a dataset of five hundred images. We are also conducting a study on disparate class distributions, a common occurrence in medical datasets. The results, based on four distinct evaluation metrics, indicated that the Random Forest and LWNB classifiers demonstrated superior performance overall. The medical industry found that logistic classifiers and LWNB are the most effective options for addressing the problem of uneven class distribution, [18].

[19], carried out a comprehensive examination of the outcomes and evaluations of various machine learning models employed in the detection of breast cancer. We developed the approach using the Wisconsin Breast Cancer Diagnostic (WBCD) dataset. We examined the data and used it in several machine-learning models. The prediction task utilized random forest, logistic regression, decision tree, and K-nearest neighbor algorithms. The researchers discovered that the logistic regression model yields the most optimal outcomes, achieving an accuracy rate of 98%. produces the best results with 98% Accuracy.

3 Research Methodology

The primary objective of this analytic study is to determine the effective classifiers for breast cancer diagnosis. The methodology of the research approach is shown in Figure 1.

3.1 Data

The first step is the data collection step, where the breast cancer datasets are collected and downloaded from the UCI Machine Learning Repository, [19].

The characteristics of the datasets are shown in Table 1, and categorized into three diagnostic criteria**:**
Dataset1: location of breast cancer (right, left, or both),
Dataset2: type (benign, malignant), and
Dataset3: type (recursive, non-recursive).

Secondly, dataset collection step is followed by the pre-processing stage which contains cleaning, handling missing data, and data reduction. Notably, the datasets used in this paper are properly prepared and don't require any pre-processing.

The third step is the main focus of this research, which is to find the best classifier to use with the three datasets of breast cancer. Therefore, twenty-two ML classifiers belonging to six learning strategies were considered and compared based on their predictive performance.

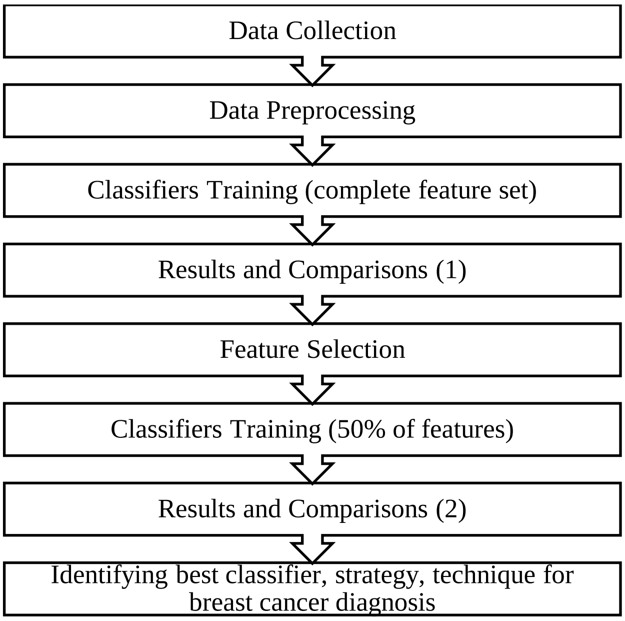
Data Collection
Data Preprocessing
Classifiers Training (complete feature set)
Results and Comparisons (1)
Feature Selection
Classifiers Training (50% of features)
Results and Comparisons (2)
Identifying best classifier, strategy, technique for breast cancer diagnosis

Fig. 1: Basic phases of the Research Methodology

Table 1. Datasets characteristic

Name	Instances	Features	No. of Classes	Reference
Dataset1	625	5	3	[19]
Dataset2	286	10	2	[19]
Dataset3	699	10	2	[19]

3.2 Machine Learning Strategies and Classifiers

Classification is a fundamental supervised learning job in machine learning. The goal of this job is to precisely forecast the category assigned to an instance that has not been previously observed, [20], [21], [22], [23]. Typically, classification falls into two main categories: single-label classification (SLC) and multi-label classification (MLC) [24]. Each instance or example in the collection must be associated with a single class label, according to the first requirement. Therefore, the class labels in SLC (Statistical Learning Classifier) are always mutually exclusive, preventing their overlap or assignment to multiple classes, [24]. Moreover, we can further categorize SLC into two subtypes: multi-class classification (MCC) and binary classification (BC). The former only considers datasets with two class labels, whereas the latter considers datasets with more than two class labels, [25], [26].

This study utilizes six machine learning classifier techniques.

Tree-based learning systems offer a very efficient approach to decision-making because they present the problem and its potential outcomes in a structured manner. It can be used by developers to examine the potential ramifications of a choice, and as a classifier gains access to additional data, it can forecast results for future data.

The Bayes learning strategy computes conditional probabilities, or the likelihood of an event occurring given another event.

The rules-learning technique entails the creation of rules based on data, as well as pre-existing rules or models. Rule learning encompasses various styles of reasoning, such as inductive, deductive, and analogical reasoning, with inductive rule learning being the most widely used.

Logistic regression is a highly effective modeling technique that extends the principles of linear regression. We employ logistic regression to evaluate the probability of a disease or health condition based on a risk factor and variables. Both simple and multiple logistic regression analyzes the relationship between an independent variable (Xi), also known as exposure or predictor variables, and a dichotomous dependent variable (Y), also known as the outcome or response variable. Its primary purpose is to forecast binary or multiclass-dependent variables.

In machine learning, a lazy learning strategy delays processing training data until a prediction is required. Lazy learning classifiers defer model construction until they receive a new query, as opposed to building models during training. This approach entails storing and comparing training instances during the prediction process.

6. Meta-learning strategy: helps models learn new, unseen tasks on their own. Meta-learning aims at discovering ways to dynamically search for the best learning strategy as the number of tasks increases, [27].

Creating an ideal model in machine learning entails identifying a set of parameters, a training dataset, and an effective learning classifier to get the greatest performance.

In addition, this study utilizes twenty-two classifiers from those methodologies. The tree learning strategies include Decisions ump (DS), Hoeffding Tree (HT), J48, LMT, and Random Forest (RF). BayesNet (BN), Naive Bayes (NB), and Naive Bayes Updateable (NBU) are algorithms derived from the Bayes learning technique. The Decision Table (DT), JRip, and OneR are all examples of rules-learning strategies. The Functions method implements the Logistic (L), Simple Logistic (SL), and SMO algorithms. The lazy learning technique includes IBK, KStar, and LWL. The meta-learning strategy includes bagging (Ba), classification via regression (CVR), logit boost

(LB), multiclass classifier (MCC), and random committee (RC).

Weka (Waikato Environment for Knowledge Analysis) received the data for training to determine the most suitable classifier for handling the datasets. The evaluation was based on four metrics: accuracy, precision, recall, and ROC Area, [28], [29].

3.3 Feature SelectionTechniques

The objective of the next stage is to identify the most effective feature selection strategy for analyzing datasets related to breast cancer. For this stage, we utilized five distinct techniques trained in WEKA to prioritize the features.

The strategies mentioned are InfoGainAttributeEval (IG), ClassifierAttributeEval (CA), Principal Component Analysis (PCA), CorrelationAttributeEval (CoA), and RellieffAttriEval (RA) [25]. Furthermore, the IG algorithm assesses an attribute's value by quantifying the amount of information obtained in relation to the class, [30], [31], [32]. The CA algorithm assesses the value of a certain attribute by quantifying the amount of information it provides about the class, [33], [34]. RA assesses the value of an attribute by iteratively choosing an instance and examining the value of the specified attribute for the closest instance of the same or different class, [31]. The PCA algorithm assesses an attribute's value by calculating the gain ratio with respect to the class. CoA evaluates an attribute's value by quantifying the correlation (using Pearson's method) between the attribute and the class, [29].

3.4 Evaluation Metrics

Four metrics—Accuracy, Precision, Recall, and Receiver Operating Characteristics (ROC) Area—are used to judge the 22 classifiers during the evaluation step. These measures are described and figured out in the following ways:

Accuracy The percentage of all predicted samples that were right, including both positive and negative samples, in the whole sample [30]. This is the formula:

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

Precision: number that shows how many of the expected positive samples are actually positive samples. The formula is:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

Recall: It's also called the true positive rate. For the first samples, the recall shows how many of the positive samples were right predicted. Here's the formula:

$$\text{Recall } = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

The **ROC Area** was assessed by graphing the FP rate at the x-axis and the TP rate at the y-axis to determine the best cut-off value for the diagnosis of breast cancer.

Where,
TP: True positive.
TN: True negative.
FP: False positive.
FN: False negative.

4 Analysis and Results

The results obtained by training datasets are demonstrated in this part grouped by each evaluation metric.

Additionally, the results based on each metric are found for training 50% of the features using the five feature selection techniques mentioned earlier.

4.1 Accuracy Results

The Accuracy results for training the three datasets are shown in Table 2. In detecting the location of the breast cancer (Dataset1), the Accuracy is highest (90.56%) when the HT classifier is employed with a slight preference over the NB and NBU classifiers (90.4%).

However, the classifier that resulted in the best Accuracy in detecting the type of cancer either benign or malignant (Dataset2) is the J48 (75.53%) followed by both LMT and SL classifiers with an Accuracy of 75.18% both.

Finally, the BN classifier is found to have the highest Accuracy in detecting the type of cancer whether it is recursive or non-recursive (Dataset 3) with an Accuracy of 97.14%.

Table 2 shows the evaluation results for the considered classifiers with respect to the Accuracy metric.

In demonstrating the performance of the five feature selection techniques as applied to 50% of the features of the datasets, it is revealed that the IG and PCA techniques have the highest Accuracy in terms of detecting the location of breast cancer as shown in Table 3. The classifiers associated with the highest accuracies are HT, NB, and NBU. They showed comparable performance in terms of Accuracy in detecting the location of the breast cancer, however, the NB and NBU preceded the HT with 0.16%.

In assessing the Accuracy of detecting the type of cancer either benign or malignant, Table 4 shows that the CoA technique has the highest Accuracy and it is achieved by applying either the LMT or SL classifiers.

Table 2. The Accuracy results of the classifiers for the three datasets

Strategy	Classifier	Dastaset1	Dataset2	Dataset3
Trees	DS	55.040	73.700	92.418
	HT	**90.560**	69.930	95.994
	J48	76.640	**75.525**	94.564
	LMT	89.760	75.175	95.994
	RF	81.440	69.580	96.567
Average		78.688	72.782	95.107
Rules	DT	73.120	73.427	95.279
	JRip	79.040	70.979	95.422
	OneR	56.320	65.734	92.704
Average		69.493	70.047	94.468
Meta	Ba	83.360	69.231	96.424
	CVR	87.680	71.329	95.708
	LB	88.000	72.378	95.708
	MCC	85.280	68.881	96.567
	MCU	88.000	69.930	96.710
	RC	78.080	67.483	95.851
Average		85.067	69.872	96.161
Bayes	BN	72.320	72.028	**97.138**
	NB	90.400	71.678	95.994
	NBU	90.400	71.678	95.994
Average		84.373	71.795	96.376
Function	L	89.600	68.881	96.567
	SL	87.840	75.175	95.994
	SMO	87.680	69.580	96.996
Average		88.373	71.212	96.519
Lazy	IBK	86.560	72.378	95.136
	KStar	88.480	73.427	95.422
	LWL	55.200	72.378	90.272
Average		76.747	72.727	93.610

Table 3. The highest Accuracy results for 50% of the features in the first dataset

Attribute Evaluation Technique	Classifier	Strategy	Accuracy
CA	NB	Bayes	70.40
	NBU	Bayes	
	HT	Trees	70.24
CoA	NB	Bayes	70.40
	NBU	Bayes	
	HT	Trees	70.24
RA	SMO	Functions	70.24
	NBU	Bayes	70.08
IG	NB	Bayes	**70.56**
	NBU	Bayes	
	KStar	Lazy	70.40
PCA	NB	Bayes	**70.56**
	NBU	Bayes	
	KStar	Lazy	70.40

Finally, the accuracy of detecting the type of cancer to be either recursive or non-recursive is highest when the RA technique is applied. However, the CoA and IG show comparable results with accuracies not less than 0.5% of that of the RA as shown in Table 3, Table 4 and Table 5.

Table 4. The highest Accuracy results for 50% of the features in the second dataset

Attribute Evaluation Technique	Classifier	Strategy	Accuracy
CA	IBK	Lazy	73.776
CoA	LMT	Trees	**76.224**
	SL	Functions	
RA	J48	Trees	72.028
IG	DT	Rules	75.874
	SL	Functions	75.525
PCA	J48	Trees	74.825

Table 5. The highest Accuracy for 50% of the features in the third dataset

Attribute Evaluation Technique	Classifier	Strategy	Accuracy
CA	MCU	Meta	95.279
	SMO	Functions	95.136
CoA	BN	Bayes	96.853
RA	SMO	Functions	**96.996**
	MCU	Meta	96.853
IG	BN	Bayes	96.567
PCA	JRip	Rules	96.137

4.2 Precision Results

As shown in Table 6, the highest Precision in detecting the cancer location is achieved using the MCU classifier. While the type of cancer, either benign or malignant is best detected using the J48 classifier. However, the BN classifier showed the highest Precision in detecting the type of cancer whether it is recursive or non-recursive. However, the Precision of SMO is comparable to that of the BN and lower with 0.2%.

The performance of the feature selection techniques as applied to 50% of the features in the datasets is represented in Table 7, Table 8, and Table 9. The results show that the CA and CoA techniques have the highest Precision in terms of detecting the location of breast cancer as shown in Table 7. The classifier associated with the highest Precision obtained is BN.

In assessing the Precision of detecting the type of cancer either benign or malignant, Table 8 shows that the CoA technique has the highest Precision and it is achieved by applying either the LMT or SL classifiers.

Table 6. The Precision results of the classifiers for the three datasets

Strategy	Classifier	Dastaset1	Dataset2	Dataset3
Trees	DS	0.540	0.68	0.929
	HT	0.904	0.676	0.962
	J48	0.732	**0.752**	0.946
	LMT	0.859	0.737	0.960
	RF	0.826	0.664	0.966
Average		0.772	0.701	0.953
Rules	DT	0.736	0.712	0.953
	JRip	0.807	0.688	0.955
	OneR	0.555	0.624	0.927
Average		0.699	0.675	0.945
Meta	Ba	0.778	0.641	0.965
	CVR	0.878	0.688	0.957
	LB	0.873	0.702	0.957
	MCC	0.872	0.668	0.966
	MCU	**0.931**	0.676	0.967
	RC	0.818	0.644	0.959
Average		0.858	0.670	0.962
Bayes	BN	0.724	0.707	**0.972**
	NB	0.907	0.704	0.962
	NBU	0.907	0.704	0.962
Average		0.846	0.705	0.965
Functions	L	0.908	0.668	0.966
	SL	0.868	0.737	0.960
	SMO	0.886	0.671	**0.970**
Average		0.887	0.692	0.965
Lazy	IBK	0.825	0.699	0.951
	KStar	0.818	0.714	0.954
	LWL	0.556	0.703	0.910
Average		0.733	0.705	0.938

Table 7. The highest Precision results for 50% of the features in the first dataset

Attribute Evaluation Technique	Classifier	Strategy	Precision
CA	BN	Bayes	**0.766**
CoA	BN	Bayes	**0.766**
RA	MCU	Meta	0.739
IG	DT	Rules	0.723
	MCU	Meta	0.718
PCA	DT	Rules	0.723
	MCU	Meta	0.718

Table 8. The highest Precision results for 50% of the features in the second dataset

Attribute Evaluation Technique	Classifier	Strategy	Precision
Ca	IBK	Lazy	0.718
	LWL	Lazy	0.704
CoA	LMT	Trees	**0.756**
	SL	Functions	
	LWL	Lazy	0.738
RA	SMO	Functions	0.702
	J48	Trees	0.698
IG	DT	Rules	0.747
	SL	Functions	0.745
PCA	J48	Trees	0.736
	KStar	Lazy	0.713

Table 9. The highest Precision results for 50% of the features in the third dataset

Attribute Evaluation Technique	Classifier	Strategy	Precision
CA	MCU	Meta	0.953
	JRip	Rules	0.952
	SMO	Functions	0.951
	J48	Trees	0.950
	BN	Bayes	
	NB		
	SL	Functions	
	KStar	Lazy	
CoA	BN	Bayes	**0.969**
	NB		0.964
	NBU		
	JRip	Rules	0.963
	MCC	Meta	
	MCU		
	L	Functions	
	SMO		
RA	SMO	Functions	**0.970**
	MCU	Meta	**0.969**
	BN	Bayes	
IG	BN	Bayes	0.966
PCA	JRip	Rules	0.963

Finally, the Precision of detecting the type of cancer to be either recursive or non-recursive is highest when the RA technique is applied. However, the CoA shows relatively similar results with a Precision of less than 0.1% of that of the RellieffAttriEval as shown in Table 9. The classifiers associated with the highest Precision values are the SMO, MCU, and BN.

4.3 Recall Results

By setting the Recall as an evaluation metric, the results are shown in Table 10. The HT classifier has the highest value in detecting the location of the breast cancer outperforming the NB and NBU classifiers with 0.2%. The J48 classifier has the best Recall values in detecting the type of cancer to be either benign or malignant, although it is close to the performance of the LMT and SL classifiers where both have a Recall lower by 0.3% compared to the J48. In detecting the type of cancer to be either recursive or non-recursive, the highest Recall belongs to the BN classifier followed by SMO which is lower than the BN Recall by 0.1%.

Based on the Recall metric, the performance of the feature selection techniques as applied to 50% of the features in the datasets is represented in Table 11, Table 12 and Table 13. Notably, the values for all strategies have no significant differences. However, the InfoGainAttributeEval and principal Components techniques have the highest Recall values in terms of detecting the location of breast cancer. The classifiers associated with the highest accuracies are NB and NBU from the Bayes strategy.

Table 10 depicts the evaluation results for the considered classifiers on the three datasets with respect to the Recall metric.

Table 10. The highest Recall results of the classifiers for the three datasets

Strategy	Classifier	Dastaset1	Dataset2	Dataset3
Trees	DS	0.550	0.69	0.924
	HT	**0.906**	0.699	0.960
	J48	0.766	**0.755**	0.946
	LMT	0.898	0.752	0.960
	RF	0.814	0.696	0.966
Average		0.787	0.717	0.951
Rules	DT	0.731	0.734	0.953
	JRip	0.790	0.710	0.954
	OneR	0.563	0.657	0.927
Average		0.695	0.700	0.945
Meta	Ba	0.834	0.692	0.964
	CVR	0.877	0.713	0.957
	LB	0.880	0.724	0.957
	MCC	0.853	0.689	0.966
	MCU	0.880	0.699	0.967
	RC	0.781	0.675	0.959
Average		0.851	0.699	0.962
Bayes	BN	0.723	0.720	**0.971**
	NB	0.904	0.717	0.960
	NBU	0.904	0.717	0.960
Average		0.844	0.718	0.964
Functions	L	0.896	0.689	0.966
	SL	0.878	0.752	0.960
	SMO	0.877	0.696	**0.970**
Average		0.884	0.712	0.965
Lazy	IBK	0.866	0.724	0.951
	KStar	0.885	0.734	0.954
	LWL	0.552	0.724	0.903
Average		0.768	0.727	0.936

Table 11. The highest Recall results for 50% of the features in the first dataset

Attribute Evaluation Technique	Classifier	Strategy	Recall
CA	SMO	Functions	0.705
CoA	SMO	Functions	0.705
	NB	Bayes	0.704
	BN		
RA	SMO	Functions	0.702
IG	NB	Bayes	**0.706**
	NBU		
PCA	NB	Bayes	**0.706**
	NBU		
	KStar	Lazy	0.704
	LB	Meta	0.702

In assessing the Recall of detecting the type of cancer either benign or malignant, Table 12 shows that the correlationAtriEval technique has the highest Recall and it is achieved by applying either the LMT or SL classifiers.

Moreover, the Recall value for detecting the type of cancer to be either recursive or non-recursive is highest when the RellieffAttriEval technique is applied.

Table 12. The highest Recall results for 50% of the features in the second dataset

Attribute Evaluation Technique	Classifier	Strategy	Recall
CA	IBK	Lazy	0.738
CoA	LMT	Trees	**0.762**
	SL	Functions	
	LWL	Lazy	0.752
RA	J48	Trees	0.720
	SMO	Functions	0.706
IG	DT	Rules	0.759
	LMT	Trees	0.755
	LWL	Lazy	0.752
PCA	J48	Trees	0.748
	KStar	Lazy	0.734
	DT OneR	Rules	0.724

Table 13. The highest Recall results for 50% of the features in the third dataset

Attribute Evaluation Technique	Classifier	Learning Strategy	Recall
CA	MCU	Meta	0.953
	SMO	Functions	0.951
	BN	Bayes	0.950
	NB		
	KStar	Lazy	
CoA	BN	Bayes	**0.969**
	NB		0.963
	NBU		
	SMO	Functions	0.963
	L		
RA	SMO	Functions	**0.970**
	BN	Bayes	**0.969**
IG	BN	Bayes	0.966
	SMO	Functions	0.961
PCA	JRip	Rules	0.963

However, the correlationAtriEval shows a relatively similar result with a Precision of less than 0.1% of that of the RellieffAttriEval as shown in Table 13. The classifiers associated with the highest Precision values are the SMO and BN.

4.4 ROC Area Results

The results based on the ROC area metric are shown in Table 14.

The LMT classifier showed the best performance in detecting the location of breast cancer in terms of ROC area. However, in detecting the type of cancer either benign or malignant, both NB and NBU classifiers showed the highest performance. Five classifiers are found to have the highest ROC area in detecting the type of cancer to be either recursive or non-recursive, which are LMT, MCCL, and SLL.

Table 14. The ROC area results of the classifiers for the three datasets

Strategy	Classifier	Dastaset1	Dataset2	Dataset3
Trees	DS	0.584	0.590	0.905
	HT	0.972	0.650	0.986
	J48	0.811	0.584	0.955
	LMT	**0.981**	0.675	**0.993**
	RF	0.944	0.634	0.989
Average		0.858	0.626	0.966
Rules	DT	0.837	0.658	0.987
	JRip	0.836	0.598	0.973
	OneR	0.595	0.542	0.908
Average		0.756	0.599	0.956
Meta	Ba	0.934	0.649	0.990
	CVR	0.954	0.659	0.991
	LB	0.962	0.676	0.992
	MCC	0.973	0.646	**0.993**
	MCU	0.972	0.596	0.965
	RC	0.883	0.631	0.987
Average		0.946	0.643	0.986
Bayes	BN	0.886	0.698	0.992
	NB	0.971	**0.701**	0.986
	NBU	0.971	**0.701**	0.986
Average		0.943	0.700	0.988
Functions	L	0.976	0.646	**0.993**
	SL	0.974	0.675	**0.993**
	SMO	0.879	0.590	0.968
Average		0.943	0.637	0.985
Lazy	IBK	0.928	0.628	0.973
	KStar	0.951	0.645	0.988
	LWL	0.767	0.638	0.977
Average		0.882	0.637	0.979

Table 15. The highest ROC Area results for 50% of the features in the first dataset

Attribute Evaluation Technique	Classifier	Strategy	ROC area
CA	NB	Bayes	0.793
	NBU		
	HT	Trees	
CoA	NB	Bayes	0.793
	NBU		
	HT	Trees	
RA	HT	Trees	**0.805**
	NB	Bayes	
IG	NB	Bayes	0.797
	NBU		
	HT	Trees	
PCA	NB	Bayes	0.797
	NBU		

Table 16. The highest ROC Area results for 50% of the features in the second dataset

Attribute Evaluation Technique	Classifier	Learning	ROC area
CA	NB	Bayes	0.683
	NBU		
CoA	DT	Rules	0.682
	NB	Bayes	0.679
	NBU		
RA	SL	Functions	0.667
IG	NB	Bayes	**0.704**
	NBU		
PCA	KStar	Lazy	0.680

Table 17. The highest ROC Area results for 50% of the features in the third dataset

Attribute Evaluation Technique	Classifier	Strategy	ROC area
AE	BN	Bayes	0.983
	NB		
	L	Functions	
	SL		
	MCC	Meta	
CoA	MCC	Meta	**0.993**
	L	Functions	
	SL		
	LMT	Trees	
	BN	Bayes	**0.992**
	NB		
	NBU		
RA	MCC	Meta	**0.993**
	SL	Functions	
	L		
	LMT	Trees	
	BN	Bayes	**0.992**
	LB	Meta	
	NB	Bayes	**0.991**
IG	MCC	Meta	**0.991**
	SL	Functions	
	L		
PCA	MCC	Meta	**0.990**
	L	Functions	
	SL	Functions	0.989
	BN	Bayes	0.987

Based on the ROC area metric, the performance of the feature selection techniques as applied to 50% of the features in the datasets is represented in Table 15, Table 16 and Table 17.

The RellieffAttriEval technique has the highest ROC area value in terms of detecting the location of breast cancer. The classifiers associated with the highest accuracies are NB and HT.

In assessing the ROC area of detecting the type of cancer either benign or malignant, Table 16 shows that the InfoGainAttributeEval technique has the highest ROC area and it is achieved by applying NB and NBU classifiers.

Moreover, the ROC area values for detecting the type of cancer to be either recursive or non-recursive have no significant difference among the feature selection techniques. However, correlationAtriEval and RellieffAttriEval techniques have the relatively highest ROC area results as shown in Table 17. The classifiers associated with the highest Precision values are the L, SL, LMT, and MCC.

5 Summary and Discussion

In this study, the performance of different ML strategies was evaluated for breast cancer diagnosis. Three distinct datasets were analyzed using 22 classifiers across six ML strategies, and evaluated by Accuracy, Precision, Recall, and ROC area metrics. Additionally, five feature reduction techniques have been applied and analyzed.

In detecting the location of the Breast cancer using the complete features set it was found that HT classifier showed the highest Accuracy and Recall values (90.56% and 0.906 respectively). While MCU classifier showed the highest Precision (0.931), LMT classifier had the highest ROC area (0.981). In the meanwhile, to detect the location of Breast cancer using feature selection techniques (50% of the features), the Accuracy, Precision, Recall, and ROC area are significantly lower than that of the complete feature set. The results show that InfoGainAttributeEval and principal components analysis techniques have the highest Accuracy and Recall (70.65% and 0.706 respectively) while CA and CoA have the highest Precision (0.702). However, the RellieffAttriEval technique has the highest ROC area value (0.805).

In detecting the type of Breast cancer either Benign or malignant using the complete features set, the **J48** classifier showed the highest Accuracy, Precision, and Recall values (75.53%, 0.752, and 0.755 respectively). While NB and NBU classifiers have the highest ROC area (0.793). On the other hand, to detect the type of Breast cancer either Benign or malignant using feature selection techniques (50% of the features), the four metrics have similar values to that of the complete features results. CoA technique showed the highest Accuracy, Precision, and Recall values (76.224%, 0.756, 0.762) while the highest ROC Area was reached by the InfoGainAttributeEval technique (0.702).

In detecting the type of Breast cancer either Recursive or Non-recursive using the complete features set, BN classifier shows the highest Accuracy, Precision, and Recall values (97.14%, 0.972, and 0.723 respectively). For the ROC area metric, the highest value is obtained by applying LMT, MCCL, and SLL classifiers (0.992).

Similarly, to detect the type of Breast cancer either Recursive or Non-recursive using feature selection techniques (50% of the features), the results are almost equal to those of the complete features set.

RellieffAttriEval technique shows the highest Accuracy, Recall, Precision, and ROC area values (96.996%, 0.970, 0.970, and 0.993 respectively).

Table 18 summarizes the best classifiers as found in this study for each dataset, while Table 19 summarizes the best feature reduction technique implemented for each dataset.

Table 18. Summary of Classifiers having the highest performance for each dataset for complete features analysis

Name	Accuracy	Precision	Recall	Roc-Area
Dataset1	HT	MCU	HT	LMT
Dataset2	J48	J48	J48	NB, NBU
Dataset3	BN	BN	BN	LMT, MCCL, SLL

Table 19. Summary of feature selection techniques having the highest performance for each dataset for 50% of feature analysis

Name	Accuracy	Precision	Recall	Roc-Area
Dataset1	IG	CA	IG	RA
Dataset2	CoA	CoA	CoA	IG
Dataset3	RA	RA	RA	RA

Notably, the analysis using feature selection techniques showed nearly identical results to the analysis based on the complete features datasets. That was valid for detecting the type of breast cancer (Benign/Malignant) and (Recursive/Non-recursive), but not applicable in the case of detecting the location of the breast cancer. That could be interpreted by the small number of features in the original dataset corresponding to the location which contained five features. Accordingly, any further reduction in the features will cause a significant loss in the total information included in the dataset.

6 Conclusion and Future Work

According to the results, there is no advantage or superiority of a certain ML strategy for breast cancer diagnosis in general. Different classifiers within the same strategy show different performances for the four metrics adopted in this study. However, it has been shown in this study that HT and J48 classifiers from the Trees strategy and NB classifier from the Bayesian strategy revealed promising results in the diagnostics and detection of breast cancer compared to other analyzed classifiers. CoA has the highest performance when employed for feature reduction in detecting the type of breast cancer (Benign/Malignant). The classifiers associated with the high performance in this technique are the LMT, SL, and NB classifiers.

RA has the highest performance when employed for feature reduction in detecting types of breast cancer (Recursive/Non-recursive). The classifiers associated with the high performance in this technique are the SMO, MCU, and BN.

Future research should focus on fine-tuning and optimizing the classifiers found in this study by experimenting with different hyperparameter configurations and ensemble techniques. Additional examination is required to assess the resilience and applicability of these classifiers on various datasets, taking into account practical aspects such as patient demographics and data imbalance. Additionally, investigating multi-modal data integration with proteomics, imaging, or genomics may improve the overall Accuracy of a breast cancer diagnosis. Together, these research areas enhance ML applications for breast cancer diagnosis.

References:

[1] P. Jagadeesh, P. ShyamalaBharathi, Amudha, G. Ramkumar, T. J. Nagalakshmi, and N. Nalini, “A Multi-Embedded Learning Algorithm for Breast Cancer Diagnosis,” in *2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, IEEE, Jan. 2022, pp. 1–6. doi: 10.1109/ACCAI53970.2022.9752594.

[2] V. Chaurasia, S. Pal, and B. Tiwari, “Prediction of benign and malignant breast cancer using data mining techniques,” *J Algorithm Comput Technol*, vol. 12, no. 2, pp. 119–126, Jun. 2018, doi: 10.1177/1748301818756225.

[3] S. A. Mokhtar and Alaa. M. Elsayad, “Predicting the Severity of Breast Masses with Data Mining Methods,” May 2013.

[4] J. Fan, Y. Wu, M. Yuan, D. Page, J. Liu, I. M. Ong, D. Burnside, "Structure-Leveraged Methods in Breast Cancer Risk Prediction," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2956-2970, 2016.

[5] A. Staff, “Cancer Facts & Figures 2018,” Atlanta, 2018.

[6] M. Alzyoud, R. Alazaidah, M. Aljaidi, G. Samara, M. Qasem, M. Khalid, and N. Al-Shanableh, "Diagnosing Diabetes Mellitus Using Machine Learning Techniques," *International Journal of Data and Network Science*, vol. 8, no. 1, pp. 179-188, 2024.

[7] D. Bazazeh and R. Shubair, “Comparative study of machine learning algorithms for breast cancer detection and diagnosis,” in *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, IEEE, Dec. 2016, pp. 1–4. doi: 10.1109/ICEDSA.2016.7818560.

[8] A. Mughaid, I. Obeidat, S. AlZu’bi, E. A. Elsoud, A. Alnajjar, A. R. Alsoud, & L. Abualigah, (2023). A novel machine learning and face recognition technique for fake accounts detection system on cyber social networks. *Multimedia Tools and Applications*, 82(17), 26353-26378.

[9] N. Jothi, N. A. Rashid, and W. Husain, “Data Mining in Healthcare – A Review,” *Procedia Comput Sci*, vol. 72, pp. 306–313, 2015, doi: 10.1016/j.procs.2015.12.145.

[10] Z. Salah, K. Salah & E. Elsoud, (2024). Spatial domain noise removal filtering for low-resolution digital images. *Indonesian Journal of Electrical Engineering and Computer Science, 34*(3), 1627-1642.

[11] M. Maabreh, I. Obeidat, E. A. Elsoud, A. Alnajjar, R. Alzyoud, & O. Darwish, (2022). Towards Data-Driven Network Intrusion Detection Systems: Features Dimensionality Reduction and Machine Learning. *International Journal of Interactive Mobile Technologies*, 17(14).

[12] M. Morrow, J. Waters, and E. Morris, “MRI for breast cancer screening, diagnosis, and treatment,” *The Lancet*, vol. 378, no. 9805, pp. 1804–1811, Nov. 2011, doi: 10.1016/S0140-6736(11)61350-0.

[13] Z. Jia, J. Li, S. Wang, L. Wang, S. Zhao, X. Li, and D. Zhang, "Multi-View Spatial-Temporal Graph Convolutional Networks with Domain Generalization for Sleep Stage Classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1977–1986, 2021, doi: 10.1109/TNSRE.2021.3110665.

[14] X. Jia, X. Sun, and X. Zhang, “Breast Cancer Identification Using Machine Learning,”

Math. Probl. Eng, vol. 2022, pp. 1–8, Oct. 2022.

[15] M. Botlagunta, M. D. Botlagunta, M. B. Myneni, D. Lakshmi, A. Nayyar, J. S. Gullapalli, and M. A. Shah, "Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms," *Scientific Reports*, vol. 13, no. 1, p. 485, 2023.

[16] M. Gupta and B. Gupta, "A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques," in *2018 Second International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, Feb. 2018, pp. 997–1002. doi: 10.1109/ICCMC.2018.8487537.

[17] G. Samara, (2020, November). Wireless sensor network MAC energy-efficiency protocols: a survey. *In 2020 21st International Arab Conference on Information Technology (ACIT)* (pp. 1-5). IEEE.

[18] M. Monirujjaman Khan, S. Islam, S. Sarkar, F. I. Ayaz, M. M. Kabir, T. Tazin, ... & F. A. Almalki, Machine Learning Based Comparative Analysis for Breast Cancer Prediction," *Journal of Healthcare Engineering*, vol. 2023, pp. 1–1, Aug. 2023, doi: 10.1155/2023/9870523.

[19] M. Lichman, "UCI Machine Learning Repositry", [Online]. https://archive.ics.uci.edu/ (Accessed Date: November 25, 2023).

[20] R. Alazaidah, M. A. Almaiah, and M. Al-Luwaici, "Associative classification in multi-label classification: An investigative study," *Jordanian Journal of Computers and Information Technology*, vol. 2, no. 7, 2021.

[21] H. A. Owida, H. S. Migdadi,,O. S. Hemied, N. F. F. Alshdaifat, S. F. A. Abuowaida, & R. S. Alkhawaldeh, (2022). Deep learning algorithms to improve COVID-19 classification based on CT images. *Bulletin of Electrical Engineering and Informatics,* 11(5), 2876-2885.

[22] G. Samara, (2020). "Intelligent reputation system for safety messages in VANET." *Int J Artif Intell* 9, no. 3 (2020): 439-447..

[23] I. Hussain, G. Samara, I. Ullah, & N. Khan, (2021, December). Encryption for end-user privacy: a cyber-secure smart energy management system. *In 2021 22nd International Arab Conference on Information Technology (ACIT)* (pp. 1-6). IEEE.

[24] Z. Salah & E. A. Elsoud, (2023). Toward Effective Framework for Wireless Intrusion Detection System in Detecting Krack and kr00k attacks in IEEE 802.11, doi: 10.20944/preprints202307.1619.v1.

[25] V. Gancheva, I. Georgiev, & V. Todorova, (2023). X-Ray Images Analytics Algorithm based on Machine Learning. *WSEAS Transactions on Information Science and Applications*, vol. 20, pp.136-145, https://doi.org/10.37394/23209.2023.20.16.

[26] R. Alazaidah, F. Ahmad, and M. Mohsin, "Multi Label Ranking Based on Positive Pairwise Correlations Among Labels," *The International Arab Journal of Information Technology*, vol. 17, no. 4, pp. 440–449, Jul. 2020.

[27] R. Vilalta and Y. Drissi, "A Perspective View and Survey of Meta-Learning," *ArtifIntell Rev*, vol. 18, no. 2, pp. 77–95, 2002.

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[29] A. Y. Alhusenat,H. A. Owida, H. A. Rababah, J. I. Al-Nabulsi, & S. Abuowaida, (2023). A Secured Multi-Stages Authentication Protocol for IoT Devices. *Mathematical Modelling of Engineering Problems*, 10(4).

[30] A. Ghaben, M. Anbar, I. H. Hasbullah, and S. Karuppayah, (2021). Mathematical Approach as Qualitative Metrics of Distributed Denial of Service Attack Detection Mechanisms. In September 2021, date of current version September 13, 2021. *IEEE Access.* DOI: 10.1109/ACCESS.2021.3110586.

[31] E. Elbasi, & A. I. Zreikat, (2023). Heart Disease Classification for Early Diagnosis based on Adaptive Hoeffding Tree Algorithm in IoMT Data. *International Arab Journal of Information Technology, 20*(1),38-48.

[32] A. Sheta,W. El-Ashmawi, A. Baareh, "Heart Disease Diagnosis Using Decision Trees with Feature Selection Method", *The International Arab Journal of Information Technology (IAJIT)*,Vol. 21, Number 03, pp. 427 - 438, May 2024, doi: 10.34028/iajit/21/3/7.

[33] M. Maree, M. Eleyat, E. Mesqali, "Optimizing Machine Learning-based Sentiment Analysis Accuracy in Bilingual Sentences via Preprocessing Techniques", *The International Arab Journal of Information Technology (IAJIT)*,Vol. 21, Number 02, pp. 257 - 270, March 2024, doi: 10.34028/iajit/21/2/8.

[34] J. Li, R. Wang, "An Anomaly Detection Method for Weighted Data Based on Feature Association Analysis", *The International Arab Journal of Information Technology (IAJIT)*, Vol. 21, Number 01, pp. 117 - 127, January 2024, doi: 10.34028/iajit/21/1/11.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This research has been funded by Zarqa university.

Conflict of Interest

The authors have no conflicts of interest to declare.