International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys.13, No.1 (Apr-23)

http://dx.doi.org/10.12785/ijcds/130156

An Improved Model for Breast Cancer Diagnosis by Combining PCA and Logistic Regression Techniques

Djihane Houfani [1], Sihem Slatnia [1], Okba Kazar [1,2], Ikram Remadna [1], Hamza Saouli [3], Guadalupe Ortiz [4] and Abdelhak Merizig [1]

[1]*Computer Science Department, Smart Computer Science Laboratory (LINFI), University of Biskra, Algeria*
[2]*Department of Information Systems and Security, College of Information Technology, United Arab Emirate University, UAE*
[3]*Customs Bridge Startup, Lille, France*
[4]*University of Càdiz, School of Engineering, UCASE Software Engineering Group, Avda. de la Universidad de Càdiz 10, Puerto Real, 11519, Spain*

Received 21 Feb. 2022, Revised 19 Dec. 2022, Accepted 6 Feb. 2023, Published 16 Apr. 2023

Abstract: Breast cancer is weighed one of the most life-threatening illnesses confronting women. It happens when the multiplication of cells in breast tissue is uncontrollable. Several studies have been performed in the healthcare field for early breast cancer diagnosis. However, traditional methods can generate incomplete or misleading outcomes. To overcome these limitations, computer-aided diagnosis (CAD) systems are extensively exploited in the healthcare domain. It is designed to improve accuracy, decrease complexity, and reduce misclassification costs. The goal of this study is to present a breast cancer CAD system based on combining the Principal Component Analysis (PCA) method for feature reduction and Logistic Regression (LR) for BC tumors classification. The experiments have been conducted on Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Original Breast Cancer (WOBC) datasets from UCI repository using different training and testing subsets. Moreover, we carried out extensive comparisons of our approach with other existing approaches. Multiple metrics like precision, F1 score, recall, accuracy, and Area Under Curve (AUC) were used in this study. Experimental results indicate that the proposed approach records a remarkable performance rate with an accuracy of 1.00 and 0.98 for WDBC and WOBC respectively and outperforms the previous works by decreasing the number of features, improving the data quality, and reducing the response time.

Keywords: Computer-Aided Diagnosis, Breast Cancer, Machine Learning, Logistic Regression, PCA, Feature Selection

1. INTRODUCTION

Breast cancer is one of the most prevalent illnesses in women, impacting 2.1 million in 2018 [1]. Its occurrence poses a major threat to women's lives as it is the second major cause of death among women [2]. Early diagnosis has a vital role in improving BC survival and promoting timely clinical treatment to patients to manage this disease effectively. In this scope, researchers and physicians in the medical field look for solutions to detect this disease through early diagnosis to intervene at the right time. Due to diagnosis limits and problems, data scientists and engineering try to contribute to this matter. Artificial Intelligence is extensively applied in the medical field [3]. Recently, many computer-based solutions to predict or diagnose have been proposed, these machine learning ML-based solutions provide reports or images to help physicians in their decision-making process. However, in traditional ideas, some technical issues related to human errors and imaging quality augment BC's misdiagnosis by physicians. Computer-aided diagnosis systems (CADs) refer to pattern recognition software that assists physicians in medical images interpretation [4]. The primary goal of CADs is to decrease observational oversights and the error rates of physicians interpreting medical images. In the literature, several researches on the early diagnosis of BC have been conducted and many CADs have been developed. These studies aim to improve the achievement of physicians in distinguishing between malignant and benign tissue. However, it stills a challenging task and more studies should be proposed to improve BC diagnosis, detection, and prediction. The main goal this study is not only to improve accuracy but also to reduce complexity and time response and required storage space and decrease misclassification costs. Through the present paper, we will describe a new approach for BC tumors classification; the principal contributions of this paper are summarized as follows:

- We proposed BC early diagnosis by classifying malignant and benign tumors using Logistic regression machine learning method.

E-mail address: djihane.houfani@univ-biskra.dz , sihem.slatnia@univ-biskra.dz, o.kazar@univ-biskra.dz , http://journals.uob.edu.bh
ikram.remadna@univ-biskra.dz , hamzasaouli87@gmail.com , guadalupe.ortiz@uca.es, a.merizig@univ-biskra.dz

- We compared the performance of PCA with relief algorithm and ISOMAP for feature dimensionality reduction.
- We tested our approach on different subsets of reduced features.
- We used the PCA method for data dimensionality reduction. This step allows us to enhance data, and increase the performance of our classifier in terms of precision, accuracy, recall, F1-score, and AUC.
- We evaluate the proposed approach on WDBC and WOBC datasets from UCI repository.
- We test our model on different training and testing subsets (80–20%, 75%-25%, 70%-30%, 50–50%, and 75%-25% training-test partition.
- We conduct several comparisons of our approach with other ML classifiers and DL approaches, and some previous studies .

The experimental findings demonstrate the effectiveness of our approach which records promising performance measures on WDBC and WOBC datasets and shows a good generalization capability.

In this paper, section 2 provides a background where we discuss CAD systems and ML methods for BC diagnosis. Section 3 illustrates the related work. A detailed description of the proposed approach is supplied in section 4. Section 5 presents The obtained results. In section 6, we close the study and discuss perspectives.

2. BACKGROUND

In this section we will define CAD systems and present the most used ML techniques in medical prediction and diagnosis.

A. CAD systems

A CAD system is a tool designed to assist radiologists in detecting suspicious features on the images [4], its fundamental goal is to highlight regions of images that present abnormalities and alert the clinician to these regions during image interpretation in order to decrease observational oversights and the error rates of physicians interpreting medical images and decision making [5].

B. Machine learning approaches for BC diagnosis

In the last decades, ML algorithms have been widely applied in medical diagnosis to improve its performance. ML approaches are able to extract key features and potential rules; this allows reducing time and memory consumption. According to the amount and type of supervision they get during training, ML systems are categorized into two major classes (supervised and unsupervised learning as shown in Figure 1) [6].

1) Description of supervised classification techniques

In supervised learning, the desired outputs are included in the training data; this serves as a guide to the algorithms. In the literature, several algorithms that used ML methods were proposed. The k nearest neighbor (K-NN) is a non-parametric algorithm that gives the correct predictions by giving the separation between the test data and inputs [7].

Linear and logistic regression are statistical methods. The aim of linear regression is to establish a linear relation between tow variables by finding the best-fitting line through points. Logistic regression is applied for datasets with independent variables analysis. It finds the best fitting model between inputs and outputs.

The concept of support vector machine (SVM) can be defined as a nonlinear, nonparametric method for classification. Its principal goal is to find a hyperplane for data separation. It is advantageous for high dimensional spaces [7].

Decision trees (DTs) are flowchart-like structures useful for classification and prediction [8]. Random forest (RF) is a set of DTs where the output is obtained by calculating the mode of classes found separately by each tree [7].

An Artificial Neural Network (ANN) is reasoning model inspired from the human brain. It is a hierarchy of layers: input, hidden, and output layers. The input layer acquires the data then transfers them to a hidden layer for processing and supplying the training results to the output layer. The output layer displays the results.

2) Description of unsupervised classification techniques

Unsupervised learning consists of finding new transformations of the input data without labeled responses for data visualization, data compression, or for better understanding of the correlations present in the data [8].

K means is a clustering algorithm for partitioning dataset observations into a set of k clusters, where k is predefined. It allows classifying observations into mutually exclusive clusters (an observation can only be found in one cluster at a time) [9]. This algorithm is scalable, and offers the possibility to handle a large amount of data. However, it can converge to a local minimum.

The main purpose of PCA is to decrease the size of a dataset, improve interpretability, and decrease information loss. These goals are assured by generating new uncorrelated variables that successively maximize the variance [10].

Hierarchical Cluster Analysis (HCA) is an alternative approach to k-means used for clusters identification in a dataset. It creates a hierarchy of clusters without specifying their number [9].

3. RELATED WORK

Wang et al. [11] developed an improved RF-based rule extraction (IRFRE) for classification rules derivation from a

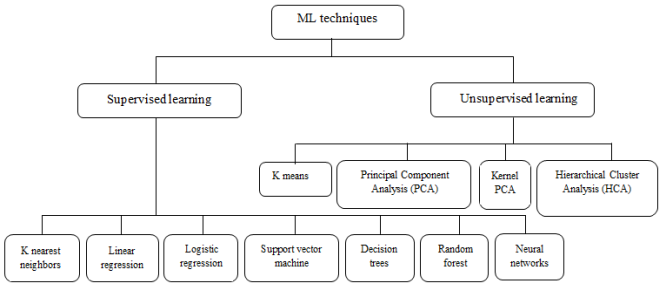


Figure 1. Overview of ML techniques[6]

decision tree for BC diagnosis. This approach was evaluated using three benchmark datasets: WOBC, WDBC, and the Surveillance, Epidemiology and End Results (SEER) BC dataset.

Mesut et al. [12] combined convolutional neural network (CNN) models and the auto encoder network model to classify invasive ductal carcinoma BC.

Mesut et al. [13] proposed BreastNet, a CNN- based model for tumors classification. This model comprises the attention module, hypercolumn technique, and residual block. It was implemented using histopathological images of breast tumor and achieves an accuracy of 98.80% on BreakHis data.

Moloud et al. [14] proposed a datamining technique for BC prediction by applying SVM and ANN to analyze Wisconsin Breast Cancer Dataset (WBCD).

Abir et al. [15] designed an automated CAD by combining genetic-fuzzy algorithm on the Saudi breast cancer diagnosis database. The system is employed to assist physicians for BC early detection.

Reza et al. [16] developed a CAD system for the breast dynamic contrast enhanced magnetic resonance imaging DCE-MRI on a real dataset of 112 patients. The system is based on a mixed ensemble of CNN (ME-CNN) for breast tumors classification. The diagnosis process is divided into two important stages: i) tumor segmentation and ii) tumor classification using CNNs.

Umit et al. [17] proposed a diagnosis system for breast tumors based on a fully convolutional network (FCN) for high-level feature extraction, and bidirectional long short term memory (Bi-LSTM) for tumors detection. The BreaKHis public database was used to implement the system.

Sami and Hushang [18] developed software for early BC detection. The process is based on algorithms and techniques for thermal breast images analysis. The goal of this study is to detect the signs of BC using CNNs optimized by the Bayes algorithm. They obtained an accuracy of 98.95% using images of 140 individuals.

Ruholla et al. [19] presented a Life-Sensitive Self-Organizing Error Drive (LSSOED) ANN for BC diagnosi on the WBCD and WOBC datasets. This approach improved the decision-making quality by reducing misclassification costs.

Liu et al. [20] applied the K2 algorithm and statistical computation methods to perform a BN modeling approach for breast tumor classification (benign or malignant). The used data were clinical dataset from a Chinese hospital and fine-needle aspiration cytology (FNAC) dataset.

Liu et al. [21] proposed an intelligent approach for BC diagnosis. They used information gain directed simulated annealing genetic algorithm wrapper (IGSAGAW) for feature selection. The proposed method minimized the SAGASW algorithm's complexity and misclassification cost, and improved accuracy by extracting the optimal feature subset. The proposed system was tested on WOBC and WDBC datasets.

Sahu et al. [22] combined PCA and ANN to classify BC tumors. The hybrid method was applied on WBCD and compared to other classification algorithms. It outperformed other proposed works and gave good accuracy, sensitivity, and F measure.

Gopal et al. [23] combined ML techniques with IoT (Internet of Things) technology for BC early diagnosis. They applied PCA for feature extraction and MLP, Logistic regression, and Random forest for breast cancer tumors classification. Their experiments have been carried out using the WBCD dataset.

The objective of Meerja [24] is to build a decision support system by applying an ensemble model based on BN and Radial Basis Function (RBF) for BC data classification. This model outperformed existing methods with an accuracy of 97.42% on the WBCD dataset.

ML techniques are widely used for early BC diagnosis, and many research works are proposed and have shown their ability to improve classification and prediction accuracy. Despite the advantages of the proposed works, we can observe certain limitations: the use of small data sets, data quality problems, and computational cost [25]. In the Table I we discussed the performance and the limitations of each proposed work.

4. PROPOSED APPROACH

We proposed a breast cancer CAD system based on PCA for feature reduction and LR (PCA-LR) for breast tumors binary classification. In this section, we will detail the architecture and the process of our approach.

A. Data preprocessing

This step has an important role in achieving more accurate results. In the following sub-sections, we will

TABLE I. Comparative table

reference	Performance	Limitations
Wang et al. [11]	Good accuracy and interpretability	Training performance and application to the other diseases are limited by the dataset.
Mesut et al. [12]	98.59% of classification accuracy	High computation cost
Mesut et al. [13]	98.80% of classification accuracy	Evaluated on one dataset. Computational cost.
Mouloud et al. [14]	Good accuracy Overfitting issue is overcome Flexible model	Computational cost
Abir et al.[15]	Accuracy 97% Good degree of confidence 91%	Evaluated on one dataset Not applied to complex real-world diagnosis problems
Reza et al. [16]	classification accuracy 96.39% Fast execution time	Small dataset
Umit et al.[17]	Good performance in terms of accuracy	Computational cost
Sami and Hushang [18]	98.95% of accuracy rate	Small dataset
R. Jafari et al. [19]	Good performance Good results in decision making	Computational cost Model complexity
Liu et al. [20]	Applicable to other diseases diagnosis	Computational cost
Liu et al. [21]	Running time and accuracy are improved	Model complexity
Sahu et al.[22]	Good performance	Tested on small dataset
Gopal et al. [23]	Good performance	Tested on one dataset
Meerja [24]	Good accuracy	Tested on small dataset

report the applied techniques to improve the quality of the used dataset.

1) Missing Data

The missing data is a trivial issue in almost all studies. It may occur due to many reasons: human errors, equipment damages, false measurements... etc. To handle missing values many strategies have been developed [26]:

- Imputation: replacing the missing values with other values (mean, median, constant...)
- Multiple imputation: consists on replacing missing values by probable values that contain the natural variability and uncertainty of the right values.
- Predictive modeling: uses several prediction algorithms to replace missing values.
- Missing data deletion: the simplest method, it consists of deleting all the cases with missing values. This is the method chosen for our proposal, since the missing data are not relevant and don't affect the obtained results.

For this study, we opted for Knn Imputer to deal with missing data in WOBC dataset.

2) Data normalization

In this stage, the independent variables of the dataset are standardized within specific range. For this study, we applied z-score standardization method. its equation is given in (1). μ_m is the mean of data, δ_m is the standard deviation, $x_m$is the raw data, and $x_m^{'}$ is the result [13]. This step aims

to reduce computing complexity.

$$x_m^{'} = \frac{x_m - \mu_m}{\delta_m} \quad (1)$$

B. Diagnosis

Diagnosis includes two main steps which are:

1) Feature extraction

This process aims to determine the most relevant features of the dataset to reduce its volume [27]. Feature extraction has a crucial impact on the system performance, memory size, and computational cost. In this study, we opted for PCA for the feature dimensionality reduction.

2) Classification

This step is primordial to predict BC tumor class (benign or malignant) by analyzing combinations of different values in selected features. In this study, we propose to use logistic regression to classify data. It consists of two significant steps: model building and model training. A logistic regression model aims to find a function that gives the relationship between independent inputs and outputs; it uses sigmoid function $\sigma(x)$ to estimate probabilities[6].

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

Where x represents the linear combination of weights and sample features and can be calculated as $x = b_0 + b_1x_1 + ... + b_nx_n$. The activation function is illustrated in figure 2.

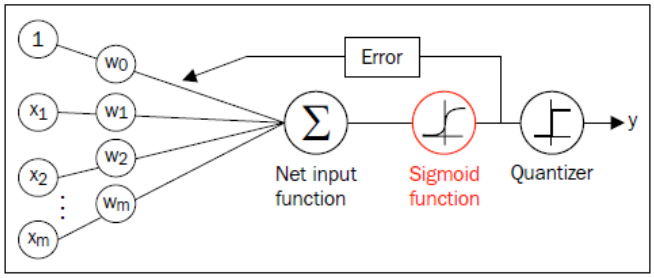


Figure 2. Logistic regression model [28]

As we can see in figure 3 and the sequence diagram in figure 4, before using data, we first proceeded to its preprocessing by eliminating the missing values. Then, we normalized it by applying the z-score standardization method. After that, we extract the most relevant features for the diagnosis using PCA. Then, we utilized the logistic regression method for breast tumor classification. Finally, we evaluate the system performance by plotting the ROC curve and confusion matrix.

Pseudo-codes for the proposed approach are given in algorithms 1 and 2.

Algorithm 1 Proposed algorithm

Input: WBCD dataset
Output : Y = tumor class (benign or malignant)

1. Data acquisition;
2. Missing values imputation;
3. Normalization ▹ Equation 1
4. Feature extraction ▹ Algorithm 2
5. Data split
X˙train, Y˙train, X˙test, Y˙test = split(x,y) ▹ 70% training, 30% testing
6. Classifier training ▹ Logistic regression

To reduce automatically the dimensionality of the dataset, we opted for PCA unsupervised algorithm. Its main idea consists of transforming the correlated variables into new variables called Principal components (see algorithm 1).

Algorithm 2 PCA algorithm

Input: Normalized dataset X with size N × M ▹ $X_i = (x_{1i}, x_{2i}, ..., x_{Mi}),\ i = 1, 2, 3, ..., N$
Output : Reduced data with size N × K

1: Compute the mean of each column, putting it into matrix **B**. $\mu \leftarrow \frac{1}{N}\sum_{i=1}^{N} X_i$
2: Compute covariance matrix of the dataset $\mathbf{C} \leftarrow \frac{1}{N}B^TB$.
3: Compute the Eigen values (λ_j) and Eigen vectors (v_j) of **C**, $Cv_j = \lambda_j v_j,\ j = 1, 2, 3, ..., M$
4: Estimate high-valued Eigen vectors
(i) Choose a threshold θ
(ii) Select K Eigen vectors corresponding to selected high-valued λ_j ▹ Reject those with Eigen value less than θ
5: Reduce the high dimensionality of feature matrix from M to K

5. Experiment

In this section, we will describe the used dataset and environment to perform the proposed PCA and LR based approach for BC diagnosis. Then, we will evaluate its performance.

To run and compile this experiment, we used Wisconsin diagnosis breast cancer dataset and "Google Colab" cloud service, with python language and scikit-learn bibliography.

A. Datasets description

In this study, both Wisconsin Breast Cancer datasets (original and diagnosis) from the UCI Machine Learning Repository are used.

1) Wisconsin diagnosis breast cancer dataset (WDBC)

WDBC dataset from the UCI repository was utilized to implement this system. It consists of 569 instances counting 357 (62.7%) benign and 212 (37.3%) malignant (see figure 9 of Appendix)

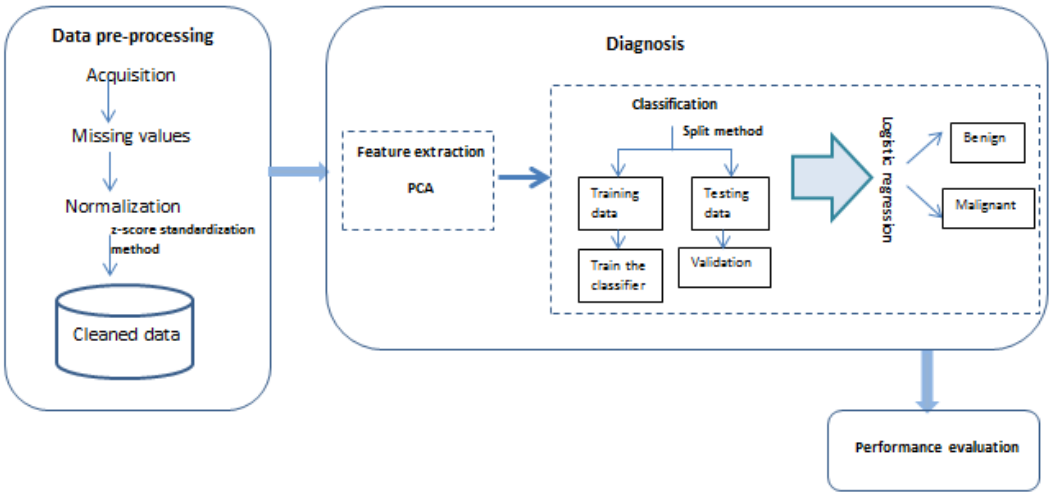


Figure 3. PCA–LR based CAD for BC tumors classification

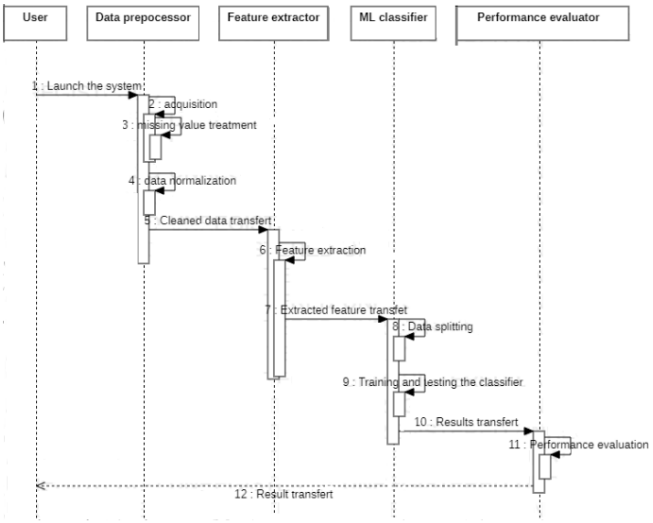


Figure 4. Sequence diagram of the proposed approach.

TABLE II. WDBC dataset description [29]

No	Attribute
1	Radius : mean of distances from center to points on the perimeter
2	Texture: standard deviation of gray-scale values
3	Perimeter
4	Area
5	Smoothness: local variation in radius lengths
6	Compactness: perimeter^2 / area - 1.0
7	Concavity: severity of concave portions of the contour
8	Concave points: number of concave portions of the contour
9	Symmetry
10	Fractal dimension: coastline approximation" - 1

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features (table II) are computed , resulting in 32 features.

2) Wisconsin original breast cancer (WOBC) dataset

The WOBC dataset contains 699 records and nine features obtained from fine needle aspirates (FNA). Its description is given in table III

B. Evaluation metrics definition

To validate the efficiency of a machine learning system, performance evaluation is a primordial step. For this purpose, different metrics are used: Recall, precision, f1 score, accuracy, confusion matrix, and Area Under Curve (AUC), their definitions and equations are depicted below [31].

- Recall: represents the fraction of positive examples that are correctly classified.
  $Recall = \frac{TP}{TP+FN}$
- Precision is the proportion of positive correctly classified samples to the total number of samples classified as positive.
  $Precision = \frac{TP}{TP+FP}$
- F1 score refers to the function of precision and recall.
  $F1score = 2 \times \frac{precision \times recall}{precision+recall}$
- Accuracy measures the fraction of the total number of predictions that were correct
  $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

TABLE III. WOBC dataset description[30]

No	Attribute	value
1	Clump Thickness	1 - 10
2	Uniformity of Cell Size standard deviation of gray-scale values	1 - 10
3	Uniformity of Cell Shape	1 - 10
4	Marginal Adhesion	
5	Single Epithelial Cell Size	1 - 10
6	Bare Nuclei	1 - 10
7	Bland Chromatin	1 - 10
8	Normal Nucleoli	1 - 10
9	Mitoses	1 - 10
10	Class	2 for benign, 4 for malignant

Where:

- **TP:** correctly classified malignant(M)tumors (M identified as M) ;
- **FP:** incorrectly classified (B) benign tumors (B identified as B);
- **FN:** incorrectly classified (M) tumors;
- **TN:** correctly classified (B) tumor.

TABLE IV. Confusion matrix

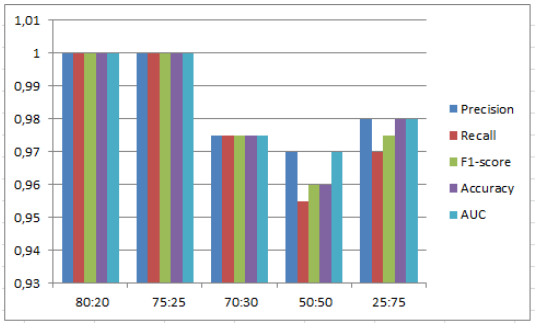
Actual	Predicted	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

A receiver operative characteristic (ROC) is also a widespread tool for visualization, organization, and selection of the classifiers based on their performance, it is plotted with TPR against the FPR and gives the value of AUC [6].

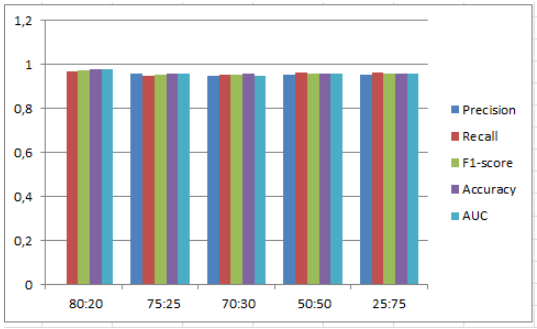
C. Experiments

To evaluate the effectiveness of our system, we performed several experiments on WDBC and WOBC datasets. We can sum up them briefly:

- We applied our model on different training and testing subsets: 80–20%, 75%-25%, 70%-30%, 50–50%, and 25–75% training-test partition.
- We tested it on various sets of reduced features.



(a) On WDBC dataset



(b) On WOBC dataset

Figure 5. Obtained results using different training-testing partitions

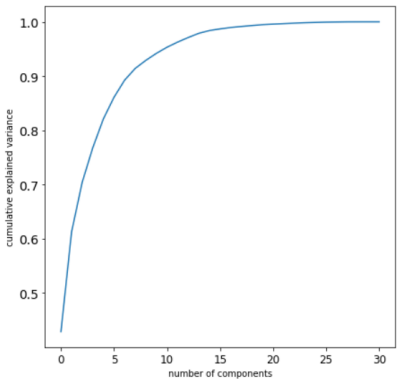
- We used different approaches for feature dimensionality reduction such as PCA, Relief algorithm, and ISOMAP method.

1) *Obtained results using different training-test subsets on WDBC and WOBC datasets*

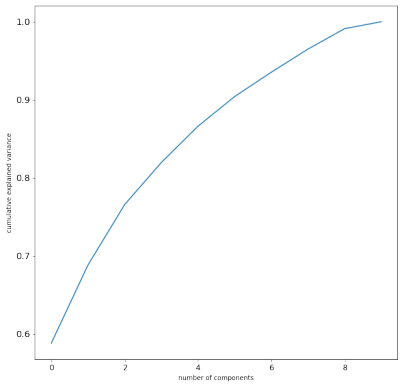
In our experiment, different training/testing ratio have been considered including (80:20, 75:25, 70:30, 50:50, 25:75). The traning-testing subsets 75:25 and 80:20 on WDBC and WOBC respectively offers the most promising results in terms of precision, recall, F1-score, accuracy, and AUC with values of (0.98, 0.97, 0.975, 0.98, and 0.98) respectively on the WOBC dataset and 1 for all metrics on the WDBC dataset (figure 5). By observing these results we notice that using large training partition allows the system to learn more and to generalize better. The obtained classification results are detailed in table VIII (see appendix)

2) *Obtained results using different sets of reduced features on WDBC and WOBC datasets*

To check the effect of feature reduction on the efficiency of our proposed classifier, we have tested it on various subsets of reduced features (85% 90% 95% 97%, and 99%) on WDBC and WOBC datasets (see Table IX). For this study, 95% of cumulative explained variance represents the optimal results for WDBC and the original features were reduced to 11. Regarding WOBC, 85% of explained variance is selected by PCA and features are reduced to 5 (see Figure 6 and IX of Appendix)

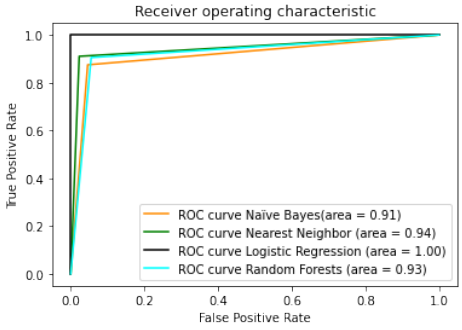


(a) WDBC dataset

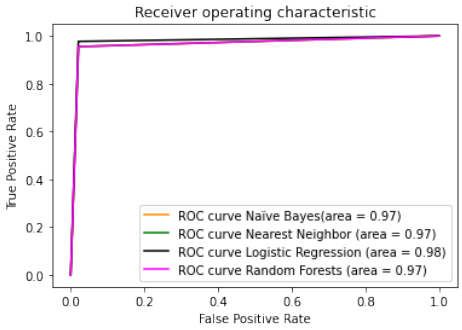


(b) WOBC dataset

Figure 6. The curve of cumulative explained variance with number of components of the WDBC dataset



(a) ROC curve of different classifiers using WDBC dataset



(b) ROC curve of different classifiers using WOBC dataset

Figure 7. ROC curves

These selected features are most relevant and allow to improve performance measures and to reduce time consumption and computational cost. For WDBC, by selecting 11 features our system achieves its highest performance with precision, recall, F1 score, accuracy, and AUC of 1.00. For WOBC, we can notice that with different % of variance the classifier performance is constant for all metrics with a precision of 0.98,a recall of 0.97, F1 score of 0.98, accuracy of 0.98, and AUC of 0.98.

3) Obtained results using different feature reduction methods on WDBC and WOBC datasets

Table V shows experimental results obtained by combining ISOMAP, relief algorithm, and PCA for feature dimentionality reduction task with LR classifier. We can easly notice that PCA outperforms other reduction techniques in terms of precision, recall, F1 score, accuracy, and AUC with respectively (1.00, 1.00, 1.0, 1.00, and 1.00) on WDBC and (0.98, 0.97, 0.98, 0.98, and 0.98) on WOBC dataset.

By combining ISOMAP and Relief algorithm with LR, the best selection is 6 and 9 features on WDBC. Thus, a reduced number of features allows to minimize the time consumption. However, we notice that classification performance is lower than our proposed approach that selects 11 features. This indicates that the proposed model accomplished the highest classification accuracy by improving the data quality, decreasing the number of attributes without losing the main objective information and increasing the performance rate.
As shown in the table, For WOBC dataset Relief algorithm and PCA gave the same and the highest results for all metrics. However, PCA outperforms relief algorithm in both dataset and ensures generalization capability. Our approach is promising for BC diagnosis by using different datasets.

4) Obtained results by various ML classifiers on WDBC and WOBC datasets

By analysing the results given in Table VI, confusion matrices in figure 10 and 11 (see appendix), and the ROC curves in figures 8, we can notice that our proposed approach outperforms, by getting no missclassified instance against other KNN, NB, and RF classifiers which predict incorrectly 7, 11, and 10 instances respectively for WDBC dataset. It also shows a slight improvement on WOBC dataset compared to other classifiers by predicting incorrectly three instances instead of four.

5) Comparison of our approach and other ML based approaches from the literature

In the Table VII, we have made a comparison of our approach with existing studies in the literature in different operating conditions (same test rate for each proposed work). The reviewed works are based on different ML

TABLE V. Comparison of PCA with other reduction methods

Dataset	Reduction method	Number of selected features	Class	Precision	Recall	F1 score	Accuracy	AUC
WDBC	ISOMAP + LR	6	M	0.96	0.98	0.97	0.98	0.98
			B	0.99	0.98	0.98		
			Mean	0.98	0.98	0.975		
	Relief Algo + LR	9	M	0.98	0.96	0.97	0.98	0.98
			B	0.98	0.99	0.98		
			Mean	0.98	0.98	0.98		
	Proposed approach (PCA + LR)	11	M	1.00	1.00	1.00	1.00	1.00
			B	1.00	1.00	1.00		
			Mean	1.00	1.00	1.00		
WOBC	ISOMAP + LR	5	M	0.98	0.93	0.95	0.97	0.97
			B	0.97	0.99	0.98		
			Mean	0.97	0.96	0.97		
	Relief Algo + LR	5	M	0.98	0.96	0.97	0.98	0.98
			B	0.98	0.99	0.98		
			Mean	0.98	0.97	0.98		
	Proposed approach (PCA + LR)	5	M	0.98	0.96	0.97	0.98	0.98
			B	0.98	0.99	0.98		
			Mean	0.98	0.97	0.98		

TABLE VI. Comparison of the proposed approach with other ML classifiers

Dataset	classifier	class	Precision	Recall	F1 score	Accuracy	AUC
WDBC	PCA +NB	M	0.88	0.92	0.90	0.92	0.91
		B	0.95	0.92	0.94		
		Mean	0.915	0.92	0.92		
	PCA+ RF	M	0.91	0.91	0.91	0.93	0.93
		B	0.94	0.94	0.94		
		Mean	0.925	0.925	0.925		
	PCA +KNN	M	0.91	0.96	0.94	0.95	0.94
		B	0.98	0.94	0.96		
		Mean	0.945	0.95	0.95		
	Proposed approach (PCA + LR)	M	1.00	1.00	1.00	1.00	1.00
		B	1.00	1.00	1.00		
		Mean	1.00	1.00	1.00		
WOBC	PCA +NB	M	0.96	0.96	0.96	0.97	0.97
		B	0.98	0.98	0.98		
		Mean	0.97	0.97	0.97		
	PCA+ RF	M	0.96	0.96	0.96	0.97	0.97
		B	0.98	0.98	0.98		
		Mean	0.97	0.97	0.97		
	PCA +KNN	M	0.98	0.98	0.98	0.97	0.97
		B	0.96	0.96	0.96		
		Mean	0.97	0.97	0.97		
	Proposed approach (PCA + LR)	M	0.98	0.96	0.97	0.98	0.98
		B	0.98	0.99	0.98		
		Mean	0.98	0.97	0.98		

techniques such as MPL, RF, LR, least square support vector machine (LSSVM) classifier, ANN and PCA, BN and RBF for BC diagnosis. We can notice that combining PCA with LR classifier results very promising performance for WDBC dataset in terms of precision, recall, F1 score, accuracy, and AUC with a value of 1.00 each one. Moreover,

TABLE VII. Comparison of the proposed approach with previous studies

Dataset	Classification method	Train : Test ratio	Class	Precision	Recall	F1 score	Accuracy	AUC
WDBC	MLP [25]	75:25	M	0.99	0.97	0.98	0.98	0.98
			B	0.95	0.98	0.97		
			Average	0.97	0.975	0.975		
	RF [25]		M	0.96	0.99	0.97	0.96	0.97
			B	0.98	0.92	0.95		
			Average	0.97	0.955	0.96		
	LR [25]		M	0.99	0.98	0.99	0.98	0.98
			B	0.97	0.98	0.98		
			Average	0.98	0.98	0.985		
	Proposed approach (PCA + LR)		M	1.00	1.00	1.00	1.00	1.00
			B	1.00	1.00	1.00		
			Average	1.00	1.00	1.00		
WOBC	PCA + ANN [23]	80:20	M		-		0.97	-
			B		-			
			Average	0.95	0.95	0.95		
	LSSVM classifier [32]		M		-		0.97	-
			B		-			
			Average		0.97			
	Proposed approach (PCA+LR)		M	0.98	0.96	0.97	0.98	0.98
			B	0.98	0.99	0.98		
			Average	0.98	0.97	0.975		
	BN+RBF [24]	75:25	M		-		0.97	-
			B		-			
			Average	0.993	0.97	0.98		
	Proposed approach (PCA+LR)		M	0.95	0.95	0.95	0.97	0.96
			B	0.97	0.97	0.97		
			Average	0.95	0.96	0.96		
	LSSVM classifier [32]	50:50	M		-		0.958	-
			B		-			
			Average	-	0.948	-		
	Proposed approach (PCA+LR)		M	0.93	0.96	0.94	0.96	0.95
			B	0.98	0.96	0.97		
			Average	0.95	0.96	0.96		

on WOBC dataset,it achieves a slight improvement in term of precision, accuracy, F1 score and, AUC and gives the best recall values compared to other studies. For BC early diagnosis, The physicians are more interested in predicting positive cases (malignant BC tumors). Therefore, recall is considered as an important metric since it indicates the rate of correctly identified malignant samples.

D. Comparison of our approach and DL-based models

Figure 10 shows the comparison of accuracies of some DL-based models proposed in the literature and our proposed model. The obtained accuracy by combining PCA and logistic regression algorithm (100%) is better than the accuracies obtained in other DL-based models. It can be explained by the fact that:

- PCA improves the performance of Logistic regression algorithm and overcomes overfitting issue.

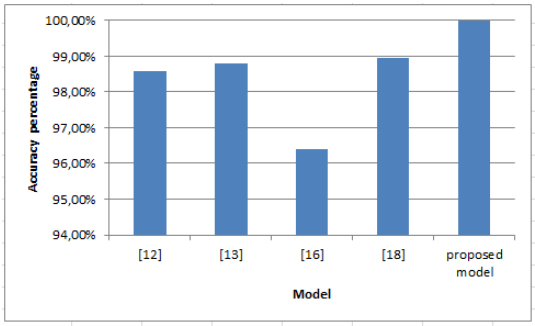


Figure 8. Proposed model and DL based models accuracy comparison.

- DL methods can be inefficient when databases are small.

6. Conclusions and Future Work

Breast cancer continues to affect women around the world; it represents a large number of new cancer cases and deaths. Early detection and diagnosis is primordial to decrease death rates. In the literature, many studies based on ML techniques for breast cancer early diagnosis are proposed. However, it stills challenging and more researches should be conducted in order to improve performance, time response, computational cost, and data quality to help specialists in diagnosis and early detection.

In this paper, we have reported the most used ML approaches and their applications in BC diagnosis. We explained that ML approaches show a remarkable power to improve classification in terms of accuracy.

The experimental results demonstrate that feature extraction by applying the PCA method is advantageous because it improved logistic regression classification performance by improving the data quality, decreasing the number of features without losing the main objective information from the original dataset. As a result, it has the advantage of reducing computational cost, memory usage and processing time. Our proposed system outperformed other research works proposed in the literature by achieving a high performance on both WDBC and WOBC datasets. For future work, we intend to switch to deep learning technique for tabular data (TabNet) [33] in order to handle efficiently large amounts of data and ensure patient's data privacy. To tackle the problem of data size and feature dimension, we will use data augmentation techniques and feature engineering. We also aim to integrate our work in clinical BC diagnostic to assist radiologists in decision making and apply it to other diseases diagnoses.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries ca," *A Cancer Journal for Clinicians*, pp. 394–424, 2018.

[2] https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/, [Online; accessed March 15, 2020].

[3] D. Houfani, S. Slatnia, O. Kazar, H. Saouli, and A. Merizig, "Artificial intelligence in healthcare: a review on predicting clinical needs," *International Journal of Healthcare Management*, vol. 19, pp. 1–9, 2021.

[4] R. A. Castellino, "Computer aided detection (cad): an overview," *International Cancer Imaging Society*, vol. 5, pp. 17–19, 2005.

[5] N. Petrick, B. Sahiner, S. G. Armato, A. Bert, L. Correale, S. Delsanto, M. T. Freedman, D. Fryd, D. Gur, L. Hadjiiski, Z. Huo, Y. Jiang, L. Morra, S. Paquerault, V. Raykar, F. Samuelson, R. M. Summers, G. Tourassi, Y. Hiroyuki, B. Zheng, C. Zhou, and H. Chan, "Evaluation of computer-aided detection and diagnosis systems," *Medical Physics*, vol. 8, pp. 1–16, 2013.

[6] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2017.

[7] K. Amit and KS.Bikash, "A case study on machine learning and classification," *International Journal Information and Decision Sciences*, vol. 9, pp. 97–208, 2017.

[8] F. Chollet, *Deep Learning with Python*. Manning Publications, 2018.

[9] B. Boehmke and B. Greenwell, *Hands-On Machine Learning with R*. the R series, 2020.

[10] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *the Royal Society publishing*, 2016.

[11] S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, and Y. Jin, "An improved random forest-based rule extraction method for breast cancer diagnosis," *Applied Soft Computing Journal*, vol. 86, p. 105941, 2019.

[12] M. Toğaçar, B. Ergen, and Z. Cömert, "Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders," *Medical Hypotheses*, vol. 135, p. 109503, 2019.

[13] M. Toğaçar, K. B. Özkurt, and Z. C. B. Ergen, "Breastnet: A novel convolutional neural network model through histopathological images for the diagnosis of breast cancer," *Physica A*, vol. 545, p. 123592, 2019.

[14] M. Abdar and V. Makarenkov, "Cwv-bann-svm ensemble learning classifier for an accurate diagnosis of breast cancer," *Measurement*, vol. 146, pp. 557–570, 2019.

[15] A. Alharbi and F. Tchier, "Using a genetic-fuzzy algorithm as a computer aided diagnosis tool on saudi arabian breast cancer database," *Mathematical Biosciences*, vol. 286, pp. 39–48, 2017.

[16] R. Rasti, M. Teshnehlab, and S. L. Phung, "Breast cancer diagnosis in dce-mri using mixture ensemble of convolutional neural networks," *Pattern Recognition*, vol. 72, pp. 381–390, 2017.

[17] Budak, Z. Cömert, Z. N. Rashid, A. Şengür, and M. Çıbuk, "Computer-aided diagnosis system combining fcn and bi-lstm model for efficient breast cancer detection from histopathological images," *Applied Soft Computing Journal*, vol. 85, p. 105765, 2019.

[18] S.Ekici and H. Jawzal, "Breast cancer diagnosis using thermography and convolutional neural networks," *Medical Hypotheses*, vol. 137, p. 109542, 2020.

[19] R. Jafari-Marandi, S. Davarzani, M. S. Gharibdousti, and B. K. Smith, "An optimum ann-based breast cancer diagnosis: Bridging gaps between ann learning and decision-making goals," *Applied Soft Computing*, vol. 72, pp. 108–120, 2018.

[20] S. Liu, J. Zeng, H. Gong, H. Yang, J. Zhai, Y. Cao, J. Liu, Y. Luo, Y. Li, L. Maguire, and X. Ding, "Quantitative analysis of breast cancer diagnosis using a probabilistic modelling approach," *Computers in Biology and Medicine*, vol. 92, pp. 168–175, 2018.

[21] N. Liu, E. Qi, M. Xu, B. Gao, and G. Liu, "A novel intelligent classification model for breast cancer diagnosis," *Information Processing and Management*, vol. 56, pp. 609–623, 2019.

[22] B. Sahu, S. Mohanty, and S. Rout, "A hybrid approach for breast cancer classification and diagnosis," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 20, pp. 1–8, 2019.

[23] V. N. Gopal, F.Al-Turjman, R.Kumar, L. Anand, and M. Rajesh, "Feature selection and classification in breast cancer prediction using iot and machine learning," *Measurement*, vol. 178, p. 109442, 2021.

[24] M. A. Jabbar, "Breast cancer data classification using ensemble machine learning," *Engineering and Applied Science Research*, vol. 48, pp. 65–72, 2021.

[25] D. Houfani, S. Slatnia, O. Kazar, N. Zerhouni, A. Merizig, , H. Saouli, and I. Remadna, "Breast cancer classification using machine learning techniques: a comparative study," *Medical Technologies Journal*, vol. 4, pp. 535–544, 2020.

[26] D. Houfani, S. Slatnia, O. Kazar, N. Zerhouni, A. Merizig, , and H. Saouli, *Machine Learning Techniques for Breast Cancer Diagnosis: Literature Review*. Marrakesh, Morocco: Springer, 2019, pp. 247–254.

[27] H. Kang, "The prevention and handling of the missing data," *Korean J Anesthesiol*, vol. 5, pp. 402–406, 2013.

[28] S. Raschka, *Python Machine Learning*. Packt Publishing.

[29] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+\Wisconsin+ (Diagnostic), [Online; accessed June 1,2021].

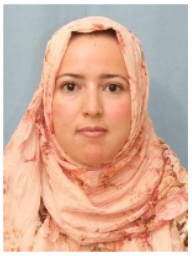
[30] https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+ (original), [Online; accessed June 13,2022].

[31] J.Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd International Conference on Machine Learning - ICML*, 2016, pp. 233–240.

[32] K. Polat and S. Güneş, "Breast cancer diagnosis using least square support vector machine," *Digital signal processing*, vol. 17, pp. 694–701, 2007.

BIOGRAPHIES

Djihane Houfani received her Master degree in Computer Science from University of Biskra, Algeria in 2017. She is now a PhD student in Artificial Intelligence at the same University. Her current research interest includes medical prediction, Deep Learning, Machine Learning, multi-agent systems, and optimization.



Sihem Slatnia followed her high studies at the University of Biskra, Algeria at the Computer Science Department and obtained the engineering diploma in 2004. After that, she obtained Master diploma in 2007 (option: Artificial intelligence and advanced system's information). She obtained PhD degree from the same university in 2011. Presently she is an associate professor at computer science department of Biskra University. She is interested to the artificial intelligence, emergent complex systems and optimization.

Okba Kazar professor in the Computer Science Department of Biskra, he helped to create the laboratory LINFI at the University of Biskra. He is a member of international conference program committees and the "editorial board" for various magazines. His research interests are artificial intelligence, multi-agent systems, web applications and information systems.

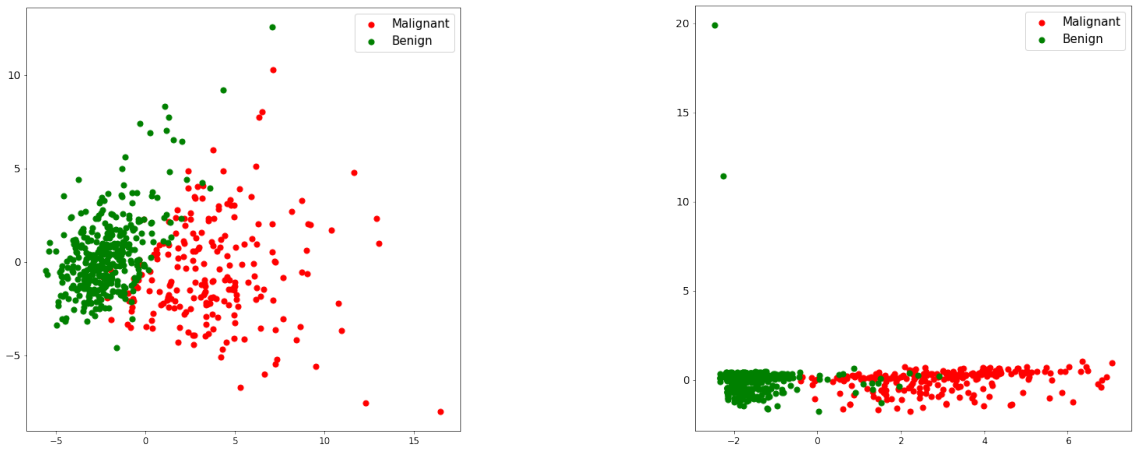
Ikram Remadna received her Master degree in Computer Science from University of Biskra, Algeria in 2016. She is now a PhD student in artificial intelligence at the University of Biskra and her current research interest includes Prognostics and Health Management and Deep learning.

Hamza Saouli received the Master and Doctorate degrees in Computer Science from University of Mohamed Khider Biskra (UMKB), the Republic of Algeria in 2010 and 2015, respectively. He was a university lecturer (2015-2019) and his research interest includes artificial intelligence, web services and Cloud Computing. Since 2019, he is Data scientist AI expert in Customs Bridge startup (Lille, France).

Guadalupe Ortiz is Associate Professor in Computer Science and Engineering at the University of Cádiz. She has participated in various programs and organization committees of scientific workshops and conferences and acts as a reviewer for several journals. Her research interests embrace software architectures for context-aware services and their adaptation to edge devices, as well as the integration of CEP in service-oriented architectures in the scope of the IoT and sustainable smart cities.

Abdelhak Merizig obtained his Master degree by 2013 from Mohamed Khider University, Biskra, Algeria; He is working on an artificial intelligence field. He obtained his PhD degree from the same university in 2018. He is now a university lecturer at the computer science department of Biskra University. Also, he is a member of LINFI Laboratory at the same University. His research interest includes multi-agent systems, service composition, Cloud Computing and Internet of Things.

Appendix



(a) WDBC dataset (b) WOBC

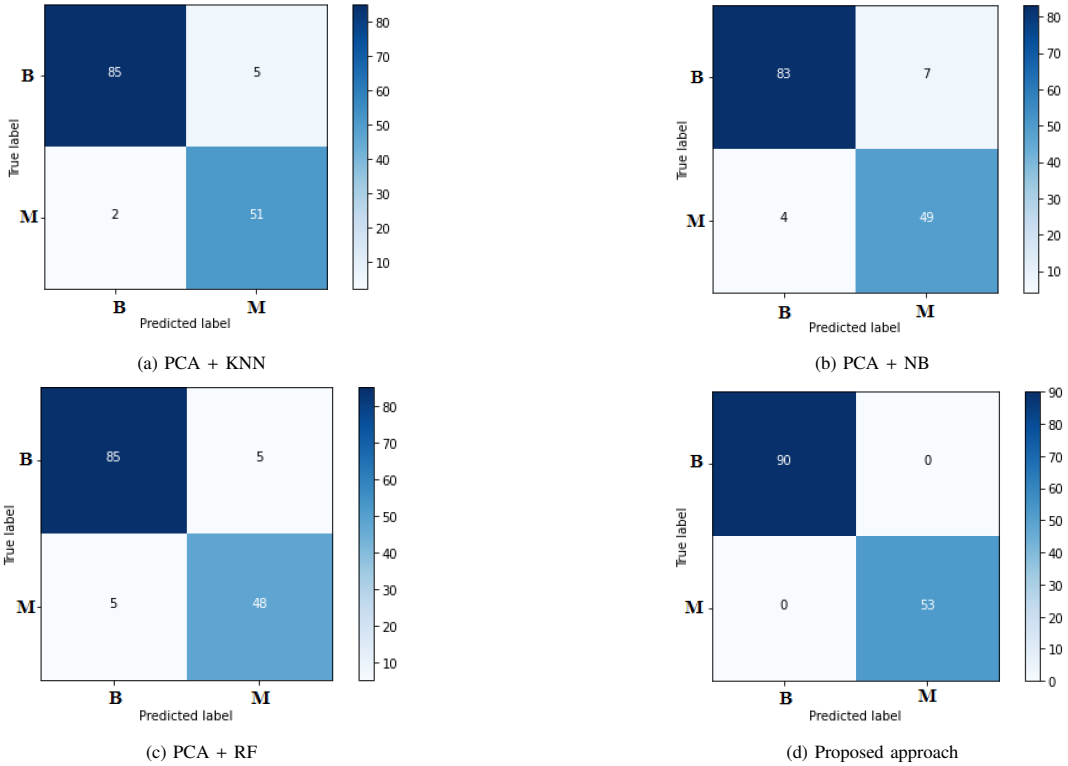
Figure 9. Benign and malignant classes distribution for WDBC and WOBC datasets

TABLE VIII. Obtained results using 80–20%, 75%-25%, 70%-30%, 50–50%, and 25–75% training-test

Dataset	Train : Test ratio	Class	Precision	Recall	F1 score	Accuracy	AUC
WDBC	80 : 20	M	1.00	1.00	1.00	1.00	1.00
		B	1.00	1.00	1.00		
		Mean	1.00	1.00	1.00		
	75 : 25	M	1.00	1.00	1.00	1.00	1.00
		B	1.00	1.00	1.00		
		Mean	1.00	1.00	1.00		
	70 : 30	M	0.97	0.97	0.97	0.98	0.97
		B	0.98	0.98	0.98		
		Mean	0.975	0.975	0.975		
	50 : 50	M	0.98	0.92	0.95	0.96	0.97
		B	0.96	0.99	0.97		
		Mean	0.97	0.955	0.96		
	25 : 75	M	1.00	0.94	0.97	0.98	0.98
		B	0.96	1.00	0.98		
		Mean	0.98	0.97	0.975		
WOBC	80 : 20	M	0.98	0.96	0.97	0.98	0.98
		B	0.98	0.99	0.98		
		Mean	0.98	0.97	0.975		
	75 : 25	M	0.95	0.93	0.94	0.96	0.96
		B	0.97	0.97	0.97		
		Mean	0.96	0.95	0.955		
	70 : 30	M	0.93	0.95	0.94	0.96	0.95
		B	0.97	0.96	0.97		
		Mean	0.95	0.955	0.955		
	50 : 50	M	0.93	0.97	0.95	0.96	0.96
		B	0.98	0.96	0.97		
		Mean	0.955	0.965	0.96		
	25 : 75	M	0.93	0.96	0.95	0.96	0.96
		B	0.98	0.97	0.97		
		Mean	0.955	0.965	0.96		

TABLE IX. Obtained results results with 90%, 95%, 97%, and 99% of variance

Dataset	% variance	Number of selected features	Precision	Recall	F1 score	Accuracy	AUC
WDBC	85%	6	0.99	0.99	0.99	0.99	0.99
	90%	8	0.97	0.97	0.97	0.97	0.97
	95%	11	1.00	1.00	1.00	1.00	1.00
	97%	13	1.00	1.00	1.00	1.00	1.00
	99%	18	1.00	1.00	1.00	1.00	1.00
WOBC	85%	5	0.98	0.97	0.98	0.98	0.98
	90%	6	0.98	0.97	0.98	0.98	0.98
	95%	8	0.98	0.97	0.98	0.98	0.98
	97%	9	0.98	0.97	0.98	0.98	0.98
	99%	9	0.98	0.97	0.98	0.98	0.98



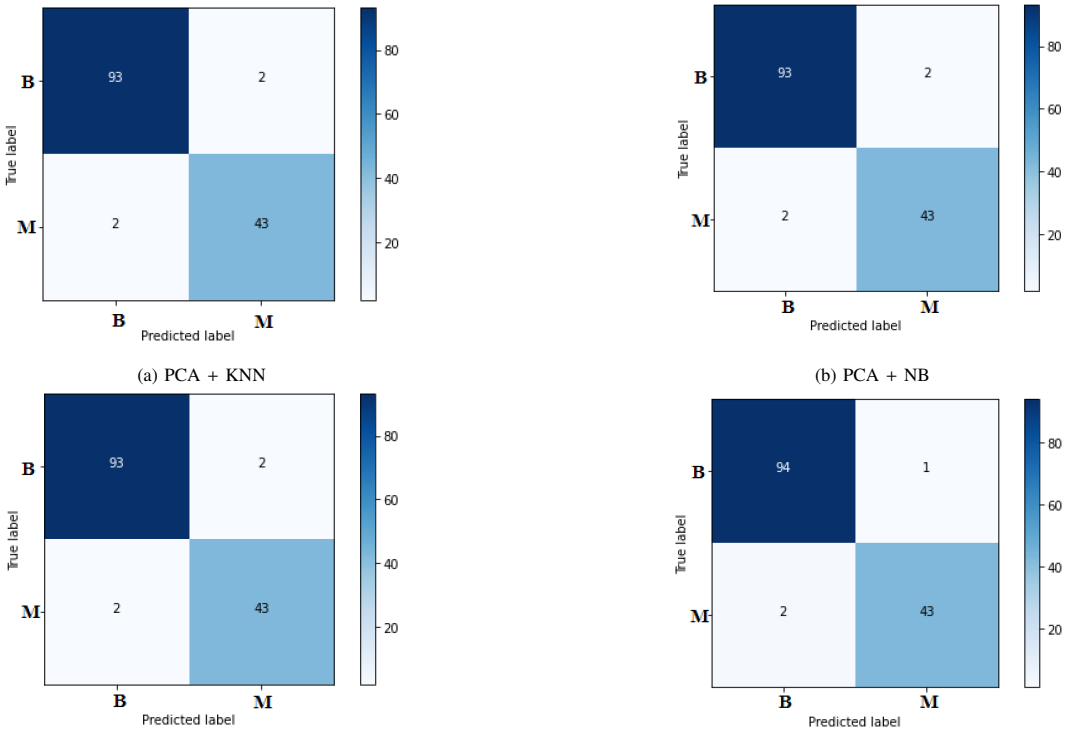
(a) PCA + KNN

(b) PCA + NB

(c) PCA + RF

(d) Proposed approach

Figure 10. Confusion matrices of different classifiers using WDBC dataset



(a) PCA + KNN

(b) PCA + NB

(c) PCA + RF

(d) Proposed approach

Figure 11. Confusion matrices of different classifiers using WOBC dataset