EAI Endorsed Transactions
on Pervasive Health and Technologies
Research Article EAI.EU

PCA-DNN: A Novel Deep Neural Network Oriented System for Breast Cancer Classification

P. Rani[1], R. Kumar[2 id], A. Jain[3,*], R. Lamba[4], R. K. Sachdeva[5], and T. Choudhury[6,*]

[1]MMICTBM, Maharishi Markandeshwar (Deemed to be University), Mullana, Ambala, Haryana, India
[2]Department of Computer Engineering, Maharishi Markandeshwar Engineering College, Maharishi Markandeshwar (Deemed to be University), Mullana, Ambala, Haryana, India
[3]School of Computer Sciences, University of Petroleum and Energy Studies, Dehradun, India
[4]Electronics & Communication Engineering Department, Maharishi Markandeshwar Engineering College, Maharishi Markandeshwar (Deemed to be University), Mullana, Ambala, Haryana, India
[5]Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
[6]Professor, Department of Computer Science & Engineering, Symbiosis Institute of Technology, Symbiosis International University, Lavale Campus, Pune, India

Abstract

INTRODUCTION: The number of women diagnosed with breast cancer has risen rapidly in recent years worldwide, which is anticipated to continue. After lung cancer, it is the second most common cause of mortality worldwide, and most women are diagnosed with it in their lives. Accurate breast cancer classification has become a challenging task in the healthcare sector. Breast cancer is a malignant tumor found in the breast tissue due to abnormal cell proliferation inside the breast. Early detection of breast cancer can reduce the death rate.
OBJECTIVES: This article proposes a principal component analysis deep neural network (PCA-DNN) for breast cancer classification.
METHODS: PCA-DNN is developed using features extracted through Principal component analysis (PCA) with deep neural network (DNN).In addition to PCA-DNN, conventional DNN and machine learning classifiers, including support vector machine (SVM), naive bayes (NB), random forest (RF), and adaptive boosting (AdaBoost), are used to perform classification. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset available at the University of California, Irvine (UCI) is used to conduct experiments.
RESULTS: PCA-DNN provided 98.83% of accuracy and 10.36% of loss. The area under the receiver operating characteristic curve (AUROC) equals 99.3%.
CONCLUSION: Results provided by PCA-DNN are better than conventional DNN and traditional machine learning classifiers. Compared to conventional DNN, it offered accuracy improvements of 3.68% and loss reductions of 29.37%.

Keywords: Breast Cancer, Principal Component Analysis, Naive Bayes, Support Vector Machine, Random Forest, Adaptive Boosting

Received on 05 July 2023, accepted on 15 October 2023, published on 18 October 2023

doi: 10.4108/eetpht.9.3533

* Corresponding authors (Anurag Jain dr.anuragjain14@gmail.com and Tanupriya Choudhury tanupriya1986@gmail.com)

1. Introduction

Thousands of women die each year from breast cancer (BC), one of the most common causes of mortality in women. BC is the second-most frequent type of cancer worldwide. More than 23% of all female cancers worldwide are breast cancer cases. In Western countries, one out of eight to nine women will eventually develop breast cancer [1]. According to a United States Cancer Society estimate, 1.3 million American women have BC diagnoses, and 0.5 million died from cancer yearly [2]. In Asia, there are about 90,000 cases of this disease each year, and forty thousand people pass away

as a result [3]. According to a US report, there are 3.8 million women who are alive but have breast cancer. [4].
Breast cancer manifests in various ways when cells affected by cancer spread across the body. Ductal Carcinoma in Situ (DCIS) occurs due to the spreading of abnormal cells outside the breast [5]. Invasive Ductal Carcinoma (IDC) [6] cancer occurs due to the spreading of abnormal breast cells across the breast tissues, and it mostly affects men [7, 8]. Mixed Tumors Breast Cancer (MTBC) occurs due to abnormal duct and lobular cells. Lobular Breast Cancer (LBC) occurs within the lobule.
Mucinous breast cancer (MBC) is caused by invasive ductal cells. It is caused due to abnormal tissues across the duct [9]. The last type is inflammatory breast cancer (IBC), which causes reddening and swelling of the breast. It is a rapidly growing cancer that is caused by the blockage of lymph vessels in the broken cell [10].
Breast cancer has no known origin, and the best treatment depends on when the cancer is diagnosed. The possibility of a patient's survival improves when the disease is detected early. As a result, tumor diagnosis has become a critical and urgent issue in the medical field [11-13].
Breast cancer is caused by uncontrolled cell proliferation. A typical cell develops in size, divides into new cells, and dies at the appropriate period during its life cycle. Cancerous cells, on the other hand, act differently from normal cells. Normal cells can become cancerous because of any mutations in DNA. Some genes regulate normal cell function, such as cell development, division of cells, and repair or death at the appropriate time. A proto-oncogene is a type of gene that regulates cell proliferation. It becomes a "bad" gene called Oncogenes when too many copies exist. Furthermore, tumor suppressor genes reduce the pace of cell division. Uncontrolled cell development occurs when certain genes do not operate properly, which has the potential to cause cancer [14].
Certain DNA mutations also increase breast cancer risk. The reason behind breast cancer-causing mutations is unknown. Cancerous cells can clump together to create a tumor. Benign tumors are those that are not cancerous. Malignant tumors are those that are cancerous. Malignant tumors can migrate to other body regions, causing them to spread [15]. Abnormal cell growth inside the breast leads to benign tumor. However, they do not expand.

Beyond the breast, and do not pose a threat to human life. The types of tumors are shown in Figure 1.
The relevance and urgency of the topic addressed in this study lie in the profound impact of breast cancer on public health globally. Developing accurate classification methods is vital to improve early detection and effective treatment. This research aims to address issues related to late-stage diagnoses, which frequently result in difficult and inefficient treatment. Improved survival rates and better treatment outcomes depend on early detection of BC. Numerous techniques have been developed to diagnose BC and reduce the number of fatalities from the disease, and many computer-aided approaches have been employed to improve diagnostic accuracy. However, accurately classifying benign and malignant tumors is challenging [16].

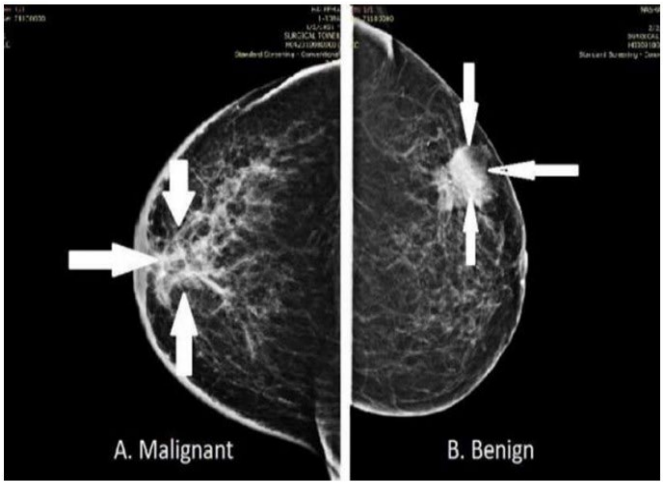

Figure 1. Malignant and Benign Tumors [15]

The research objectives of this paper are as follows:

1. A literature review of the various classification schemes for breast cancer is conducted to identify the research gap.
2. PCA-DNN is proposed in this research as a strategy for diagnosing breast cancer by merging DNN and PCA.
3. The performance of PCA-DNN is compared to conventional DNN and machine learning classifiers in terms of accuracy, specificity, sensitivity, precision, and F-measure.
4. The performance of PCA-DNN is compared to existing systems, using fuzzy logic and other supervised and semi-supervised techniques in recent literature.

The remaining sections of the paper are as follows: Section 2 describes the work done by researchers. Materials and methods follow it in section 3. The results and discussions are in the next section. The conclusion and future scope are given in Section 5.

2. Literature Survey

Abdel-Zaher and Eldeib [17] used a deep belief network (DBN) to classify breast cancer. Backpropagation neural networks with the levenberg-marquardt learning function were used to develop this system. The supervised path of backpropagation followed the unsupervised path of DBN. The DBN path was used to initialize the weights.
Asri et al. [18] classified breast cancer using four algorithms: SVM, NB, decision tree (DT), and k nearest neighbors (KNN) on Wisconsin datasets. The experiments were done on the Weka machine-learning tool. SVM outperformed with 97.13 % of accuracy.

Peng et al. [19] proposed a method based on artificial immunity and achieved 98% accuracy on the WDBC dataset. Computer immunology is based on the concept of the immune system of biology. One of the main challenges in diagnosis systems based on supervised learning is obtaining labelled data. The proposed system reduced the requirement for labelled data.

Nilashi et al. [20] developed a system with fuzzy logic. The problem of multicollinearity was solved using PCA. Fuzzy rules were produced using the classification and regression tree (CART) algorithm. The system achieved 94.1% accuracy on the Mammographic mass dataset and 93.2% on the WDBC dataset.

Huang et al. [21] developed SVM ensembles for BC classification. The best features from the dataset were chosen using a genetic algorithm (GA). SVM ensembles were developed using bagging and boosting methods. SVM classifiers with different kernels were used while constructing SVM ensembles. An ensemble based on the bagging method performed best on a small dataset, whereas an ensemble based on the method performed best on a large dataset.

Dora et al. [22] proposed the Gauss-Newton representation-based algorithm (GNRBA). Sparse representation was used with training samples. Optimal weights of training samples were found using the gauss newton-based method. Implementation was done in Matlab software. The system achieved 98.48% of accuracy.

Alikovi and Subasi [23] proposed a system with two stages. In the first stage, unnecessary features were removed using GA. GA has selected 14 features. In the second stage, various classifiers were used, and the best classifier was selected. Rotation forest was identified as the best classifier.

Wang et al. [24] developed an ensemble model based on SVM. Twelve SVM having different kernel functions and structures were combined to develop an ensemble. The proposed model achieved 97.68% accuracy.

Sivakumar et al. [25] developed a mixed-mode database miner (MMDBM) classifier for identifying BC. MMDBM was proposed by combining decision trees and supervised learning in quest (SLIQ) algorithms. MMDBM was compared to eighteen classifiers on four datasets. MMDBM provided better results than other classifiers, and an accuracy of 85.45 % was achieved.

Wang et al. [26] classified breast cancer using a context-based probability neural network (CPNN). CPNN was developed by constructing second layer of PNN using the concept of contexts. Parameters of CPNN were optimized using GA. CPNN results outperformed PNN and achieved 97.40% of accuracy.

Zhang et al. [27] proposed a hybrid approach combining the k-means and C5.0 algorithms. Clustering was done using k-means, and informative samples around the cluster's edge were chosen. It resulted in a balanced dataset classified with the boosted C5.0 algorithm. The system obtained 98.2% accuracy.

Dhahri et al. [28] proposed an automated system based on genetic programming. By genetic programming, the best features and optimal values of parameters were identified for the classifiers. The performance of SVM, KNN, DT, NB, RF, AdaBoost, logistic regression (LR), gradient boosting (GB), and linear discriminant analysis (LDA) was evaluated. The AdaBoost achieved the maximum accuracy of 98.24%.

Zhang and Chen [29] developed a hybrid model combining k-means, random oversampling example (ROSE), and SVM methods. The dataset was balanced using ROSE. K-means was used to select samples near the cluster boundary. Using ROSE and k-means along with SVM improved the performance of SVM.

Salod [30] used the breast cancer Coimbra dataset (BCCD), which contains 116 rows having ten features based on breast cancer patients' routine tests. The performance of different algorithms, including SVM, LR, DT, KNN, AdaBoost, RF, and GB, were checked on full features and selected features. Correlation-based feature selection (CFS) was used for selecting features.

Kadam et al. [31] proposed a method based on softmax regression and sparse auto-encoders for classifying breast cancer. An auto-encoder comprises a decoder, an artificial neural network, and an encoder. In sparse auto-encoder, sparseness constraints are applied on all hidden nodes.

Khan et al. [32] classified BC images using transfer and deep learning. Features were extracted using VGGNet, GoogleNet, and ResNet. A total of 8000 images were used for training and testing. A maximum accuracy of 97.25% was obtained.

Al Ghunaim et al. [33] compared machine learning algorithms using two types of big data. Algorithms were applied to individual datasets and combined datasets. SVM, DT, and RF were used to develop a model using three datasets. Results show that SVM in the spark platform provided the best performance.

Memon et al. [34] used the SVM with a recursive feature elimination technique to detect breast cancer. The performance of SVM on various kernels was evaluated. SVM achieved 98% on the RBF kernel, 84% accuracy on the sigmoid kernel, and 97% accuracy on the polynomial kernel.

Abdar et al. [35] performed various experiments using SVM and ANN. The performance of SVM was evaluated with various values of hyperparameters. It was identified that these hyperparameters helped in improving the performance of SVM. CWV-BANNSVM model was proposed by combining boosting ANNs (BANN) with SVM using the confidence-weighted voting method (CWV). Hyperparameters selected during the first experiment were used to develop CWV-BANSVM. The model was evaluated on a dataset having 669 records.

Zheng et al. [36] divided images into MRI, ultrasound, and digital images. The authors performed classification using CNN, autoencoders, and long short-term memory (LSTM). Adaboost high-level learning model (DLA-EABS) was proposed, which provided 97.2% accuracy.

Abdar et al. [37] proposed a nested ensemble mechanism based on voting and stacking. There were meta classifiers and classifiers in nested ensemble classifiers. Each meta-classifier contained two or more classification algorithms.

The proposed classifier outperformed other classifiers, achieving 98.07% accuracy.

Supriya and Deepa [38] proposed an optimized artificial neural network (OANN) model. Firstly, data was preprocessed using replacing missing attributes (RMA) and normalization methods. Important features were selected using the modified dragonfly algorithm (MDF). The classification was done using OANN, which was optimized using the grey wolf optimization (GWO) algorithm. The OANN model achieved an accuracy of more than 96%.

Kumar et al. [39] predicted BC using twelve classifiers: decision table, AdaBoost, J48, J-Rip, lazy K-star, lazy IBK, LR, RF, NB, multilayer perceptron, multiclass classifier, and random tree. Experiments were performed on the Wisconsin dataset. All of the classifiers performed well and most of them provided accuracy of more than 94%. NB has provided the worst performance, and lazy IBK has the best.

Naji et al. [40] diagnosed breast cancer using SVM, KNN, NB, DT, and LR classifiers. Three best-performing classifiers, SVM, KNN and LR, were used to develop an ensemble model using a majority voting mechanism. The ensemble model provided an accuracy of 98.1%.

Al-Azzam and Shatnawi [41] applied various semi-supervised and supervised learning methods on the WDBC dataset. LR, NB, SVM, DT, RF, gradient boosting, and extreme gradient boosting (XGBoost) classifiers were evaluated, and their performance was compared. Performance was assessed using k-fold validation. The highest accuracy of 98% was obtained using the KNN classifier.

According to a literature review, multiple systems have been given to detect breast cancer. These systems have been developed using various techniques, including machine learning, deep learning, and fuzzy logic. Some systems have used images for diagnosis, while others have used clinical data from medical test results.

Most systems combine machine learning algorithms with feature selection and feature extraction techniques. In existing systems, feature extraction methods are not utilized with DNN. This study incorporates the PCA feature extraction approach with DNN to address this research gap.

3. Materials and Methods

3.1 Dataset

This research used the WDBC dataset available at the UCI repository. This dataset has 569 incidences, 357 of which are benign and 212 of which are malignant [42]. One class attribute, an ID number, and 30 real-value attributes comprise the 32 included features. These features are derived from an image of a fine needle aspiration technique performed on a breast mass and are used to define the properties of the cell nuclei. The class attribute has two possible values: malignant and benign.

3.2 Methods

The underlying techniques for PCA-DNN are described in this section.

3.2.1 Principal Component Analysis

The PCA feature extraction method and traditional DNN are combined in the PCA-DNN approach. A matrix with n features is transformed using PCA into a new dataset with fewer features. In other words, it reduces the number of features by introducing new, fewer variables that effectively capture the significant quantity of information in the original features. PCA identifies the eigenvectors of a covariance matrix having the highest eigenvalues to transform the data to fewer dimensions.

3.2.2 Deep Neural Network

The structure of the human brain drives the basic architecture of a DNN. DNN architecture includes multiple computation units. These computational units are connected. The perceptron receives input and provides output. The fundamental concept behind a neural network is that input, i, is combined with a bias, b, and then weighted by, w, before being summed, as shown in equation 1.

$$O = f(\sum(w * i) + b) \qquad (1)$$

$$O = Output$$

$$f = activation\ function$$

$$w = Weight$$

$$i = Input$$

$$b = Bias$$

Weight lies between -1 to 1. If all the weights are made very small, it will take longer to get to a point where anything significant occurs. Conversely, using large initial weights increases the risk of becoming locked in a local optimum too early.

The activation function performs transformation non-linearly and activates and deactivates nodes in DNN. Rectified linear (Relu), sigmoid, and softmax are frequently used activation functions.

The programmer does not need to provide all the computational parameters to the DNNs, which is a significant feature of the DNNs. A DNN is trained by exposing several examples and modifying the internal parameters.

The performance of PCA-DNN was compared to conventional DNN and four machine learning classifiers NB, RF, SVM, and AdaBoost. NB is based upon Bayes theorem [43]. According to the Bayes theorem, P(H|I), or probability that the hypothesis H is true for a sample I, can be computed as in equation 2:

$$P(H|I) = P(I|H)P(H) \div P(I) \quad (2)$$

$P(H|I) =$ Posterior Probability

$P(I|H) =$ Likelihood

$P(H) =$ Class Prior Probability

$P(I) =$ Predictor Prior Probability

If and only if the likelihood of having class n conditioned on I is greater than the likelihood of other classes, an Input I is classified to a class Cn as:

$$P(C_n|I) > P(C_m|I) \text{for } i \leq m \leq k \; n \neq m \quad (3)$$

$P(C_n|I) =$ Probability of input I belong to Class C_n

$P(C_m|I) =$ Probability of input I belong to Class C_m

SVM classifies by creating a hyperplane with all samples from one class on one side and samples from another class on the other;

$$H: x + b = 0 \quad (4)$$

$H =$ Hypeplane

$x =$ vector representing a point in the vector space

$b =$ vector representing a displacement vector

SVM can classify both linear and non-linear data. Finding a straight line separating two classes is impossible with non-linear data. To get data in linear form, low-dimensional data is transformed into high-dimensional data. In SVM, the kernel function is utilized to carry out this task. The extreme points chosen by SVM that assist in creating the hyperplane are known as support vectors. Although SVM has a good overall performance, some nontrivial parameters impact the performance of the SVM model, such as kernel and regularization parameters [44].
RF makes predictions by combining the results of more than one decision tree. Gini index shown is used to decide how branching will be done in different nodes of the decision tree:

$$GIndex = 1 - \sum_{j=1}^{C} (f_j)^2 \quad (5)$$

$GIndex = Gini\ Index$

$c = Number\ of\ Classes$

$f_j =$ Frequency of class j in the dataset

The final output is produced by combining the prediction of each tree using majority voting [45].
AdaBoost employs the boosting concept, which improves weak classifiers' performance. In this approach, the classifier is initially trained using the original dataset. The classifier is then trained many times, with each iteration aiming to correct the mistakes caused by the iteration before it [46].

3.3 Proposed Methodology

This section discusses the system model, architecture, and working of the PCA-DNN approach. Automatic feature extraction and selection are features of a traditional DNN. PCA-DNN surpasses DNN by incorporating explicit feature extraction into DNN. Explicit feature extraction is done via PCA. It is a feature extraction technique that condenses the original dataset into fewer principal components, which are uncorrelated derived variables [50].

3.3.1 System Model

In PCA-DNN, principal components are extracted from the original dataset using the following steps:
- Consider the d+1 dimensional dataset and ignore the labels achieving d dimensional dataset.
- Calculate the mean m of each dimension i with i=1,2,.....d.
- Calculate the covariance matrix of the complete dataset as:

$$COV(A,B) = -\frac{1}{d-1}\sum_{j=1}^{d}(A_j - \overline{A})) + (B_j - \overline{B}) \quad (6)$$

$COV(A,B) = covariance\ between\ A\ and\ B$

$d = Number\ of\ Samples$

$A_j = Value\ of\ Feature\ A\ in\ Sample\ j$

$\overline{A} = Mean\ of\ \ Feature\ A$

$B_j = Value\ of\ Feature\ B\ in\ Sample\ j$

$\overline{B} = Mean\ of\ \ Feature\ B$

-Eigenvectors and eigenvalues are calculated as equation 7. If M is a square matrix, A is a vector, and € is a scalar associated with this eigenvector:

$$MA = €A \quad (7)$$

$A = \ Matrix\ A$

$€ = Scalar\ Value$

The eigenvalues of M are the roots of:

$$\det(M - €I) = 0 \quad (8)$$

$$M = Matrix\ M$$

$$I = Identity\ Matrix$$

$$\det(\ ) = \text{Determinant of the matrix}$$

where I is the identity matrix
– The eigenvectors are sorted based on eigenvalues, and K eigenvectors having the highest eigenvalues are selected. By using these selected eigenvectors, dxK dimensional matrix M1 is formulated.
– The transpose, M2, of matrix M1 is calculated:

$$B = M2 \text{ x } A \quad (9)$$

By following the above steps, eight components were extracted from the dataset, which were applied to DNN input. PCA-DNN has three layers: one input layer, one output layer, and one hidden layer. There are eight nodes in the input layer, fifty in the hidden layer, and one in the output layer. The training of the PCA-DNN was done in a hundred epochs having batch size thirty. An optimum number of layers and nodes were found by performing different experiments.
A relu activation function is used to introduce the non linearity:

$$R(a) = \max(0, a) \quad (10)$$

$$a = \text{Input a}$$

$$R(a) = \text{Relu activation function on input a}$$

In the output layer, the sigmoid activation function is used, which is defined as:

$$f(a) = \frac{1}{1 + e^{-a}} \quad (11)$$

$$a = \text{Real number or matrix of the real number}$$

$$f(a) = \text{Sigmoid activation function on input a}$$

The binary cross-entropy function is used to calculate loss because of the binary classification nature of the problem:

$$BCEF(x) = \sum_{i=1}^{n} AO(a) log PO(a) \quad (12)$$

$$BCEF(x) = Binary\ cross - entropy\ function$$

$$a = Input$$

$$AO(a) = Actual\ output\ for\ input\ a$$

$$PO(a) = Predicted\ output\ for\ input\ a$$

Loss is computed as:

$$Loss = -\frac{1}{M}\sum_{i=1}^{M} C_i * \log(p(C_i)) + (1 - C_i) * \log(1 - p(C_i)) \quad (13)$$

$$M = Total\ Samples$$

$$C_i = \text{Class label of input i}$$

$$\sum_{i=1}^{M} C_i * \log(p(C_i)$$

$$C_i * \log(p(C_i)) = \text{Log probability of class label of input i}$$

3.3.2 Architecture and Working

The flowchart of PCA-DNN is given in Figure 2. The methodology of PCA-DNN is shown in Figure 3. In addition to PCA-DNN, conventional DNN is also studied with the same number of layers as PCA-DNN. The conventional DNN was also trained in a hundred epochs with batch size thirty. Traditional machine learning classifiers, mainly NB, SVM, RF, and AdaBoost, were also used for classification. The performance of PCA-DNN was compared to the conventional DNN and the traditional machine learning classifiers. Breast cancer was classified as malignant or benign 2.

The Pseudocode of PCA-DNN is given in Algorithm 1.

Algorithm 1

Pseudocode of PCA-DNN

```
Algorithm PCA-DNN( )
{
   //Take the dataset with the original set of features
F= Original set of Features
 // Apply PCA on the features F
//obtaining principal components
pc=PCA(F)
no of epochs=100
Construct a neural network with one input layer, one hidden
layer, and one output layer.
Divide the dataset into training and testing data
i=1
 //Train the neural network with 100 epochs
While ( i <= no of epochs)
   {
Train the neural network with training data using pc.
          i=i + 1
   }
Test the neural network with testing data.
```

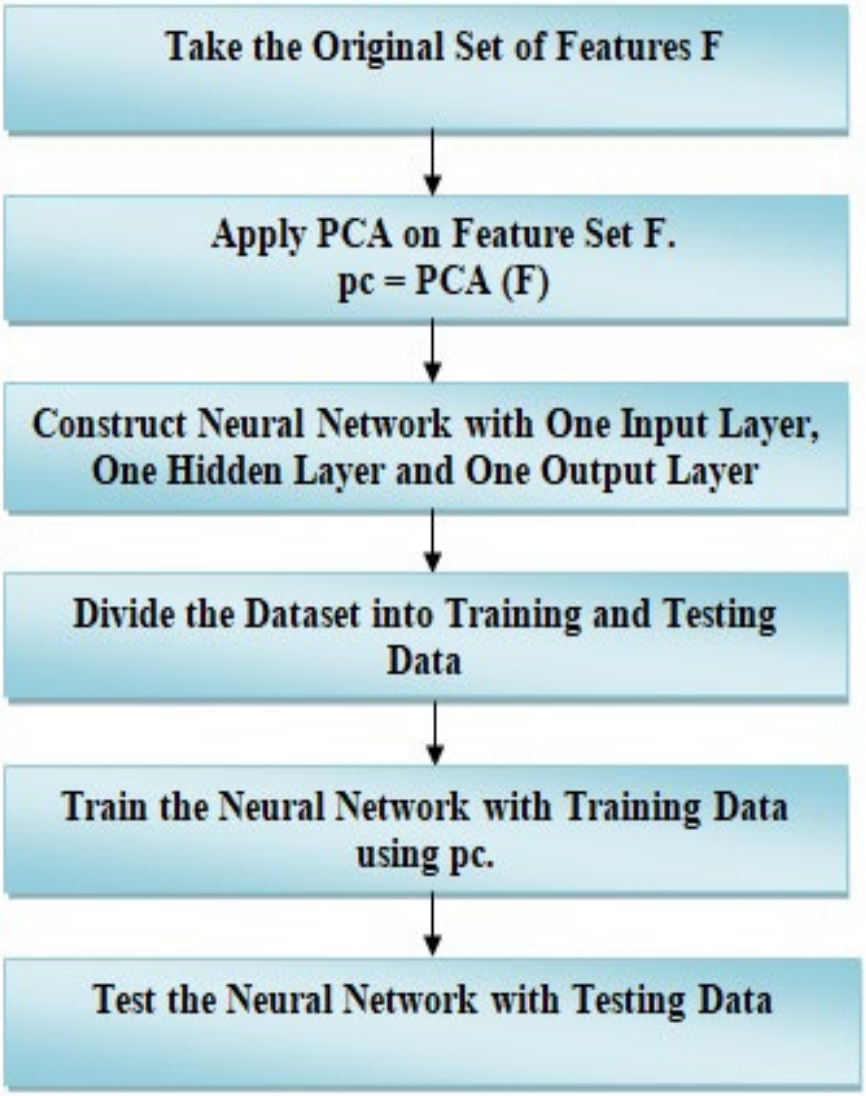

Figure 2. Flowchart of PCA-DNN

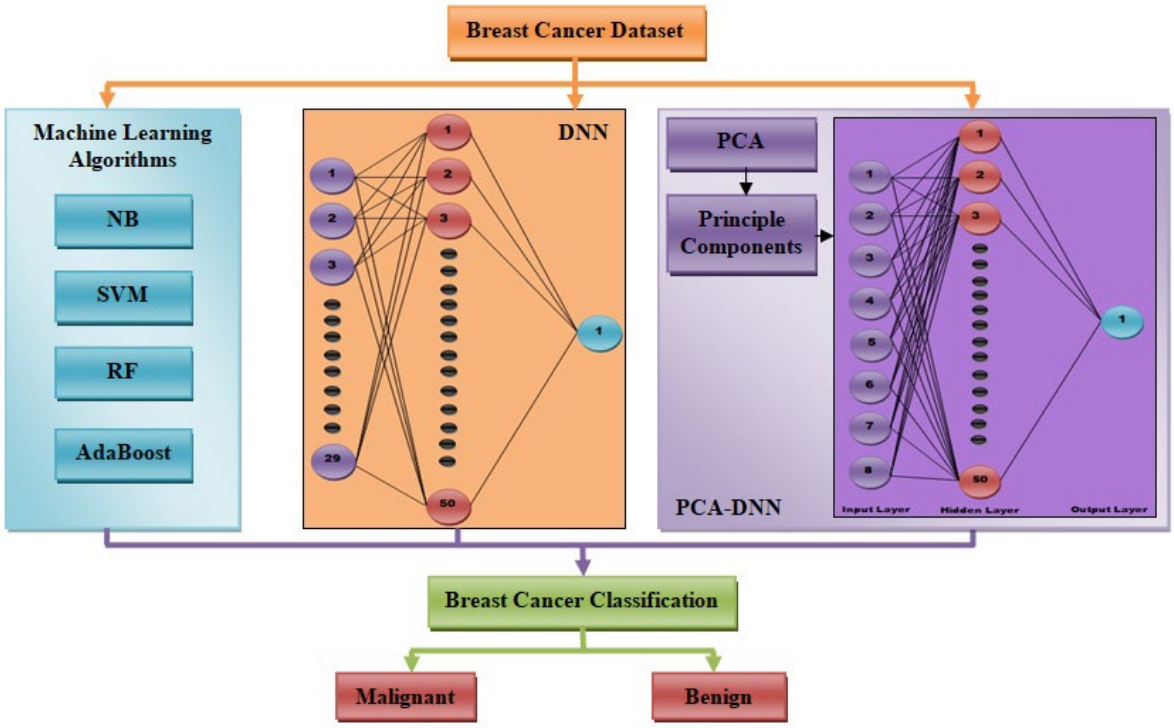

Figure 3. Methodology of the proposed PCA-DNN method for Breast Cancer Classification

4. Experimental Results and Analysis

4.1 Experimental Setup

The authors performed experiments on the system having an i3 processor seventh generation and 8 GB RAM. The programming language used was Python 3.1, and the coding environment used was Jupyter Notebook.

4.2 Experimental Parameters

The following parameters were used to assess the performance:

Accuracy: It indicates the percentage of correct predictions.

$$\text{Accuracy} = \left(\frac{\text{TP} + \text{FP}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}\right) * 100 \quad (14)$$

$$TP = True\ Positives$$

$$FP = False\ Positives$$

$$TN = True\ Negatives$$

$$FN = Fasle\ Negatives$$

Sensitivity: It indicates the percentage of correct positive predictions.

$$\text{Sensitivity} = \left(\frac{\text{TP}}{\text{FN} + \text{TP}}\right) * 100 \quad (15)$$

Specificity: It indicates the percentage of correct negative predictions.

$$\text{Specificity} = \left(\frac{\text{TN}}{\text{FP} + \text{TN}}\right) * 100 \quad (16)$$

Precision: It indicates the percentage of relevant predictions.

$$\text{Precision} = \left(\frac{\text{TP}}{\text{TP} + \text{FP}}\right) * 100 \quad (17)$$

F-Measure: It calculates the harmonic mean of sensitivity and precision.

$$\text{F} - \text{Measure} = 2 * \frac{\text{Sensitivity} * \text{Precision}}{\text{Sensitivity} + \text{Precision}} \quad (18)$$

True positives (TP) denote the number of times the system has correctly identified cancer. On the other hand, true negatives (TN) denote scenarios where a person without cancer is accurately classified. False positives (FP) are the

number of normal people who are mistakenly classified as cancer patients, and false negatives (FN) are the number of cases where a cancer patient is mistaken for a normal person [47].

4.3 Results

In this study, three experiments were performed. Firstly, machine learning algorithms were used to classify breast cancer. Secondly, a DNN was used for classification. Thirdly, the classification was done by using the proposed PCA-DNN. Standard machine learning algorithms, including NB, RF, SVM, and AdaBoost, were used for classification. Ten-fold cross-validation is used to perform validation and evaluate the performance of classifiers. The performance achieved by the algorithms under comparison to classify breast cancer is shown in Table 1. Adaboost obtained the best accuracy of 96.30%, and NB the worst accuracy of 93.49%. After classifying with machine learning classifiers, conventional DNN was used to perform the classification.

The hold-out validation procedure assesses the performance of DNN and PCA-DNN. The dataset is split into 70% training and 30% testing data. The model was trained in a hundred epochs with batch size thirty. Accuracy and loss of the DNN on train and test data were measured. The obtained change in accuracy and loss of the DNN with the increasing number of epochs are shown in Figures 4 and 5.

The performance of the DNN on training and testing data is given in Table 2. The DNN achieved 94.97% accuracy on training data and 95.32% on testing data. It achieved a 17.63% loss in the training data and a 14.67% loss in the testing data.

Table 1. Performance achieved by the machine learning algorithms under study in classifying breast cancer.

Classifier	Accuracy	Sensitivity	Specificity	Precision	F1-Score
NB	93.49	89.15	96.07	93.10	91.08
SVM	88.93	74.05	97.75	95.15	83.28
RF	95.78	92.45	97.75	96.07	94.23
AdaBoost	96.30	94.33	97.47	95.69	95.01

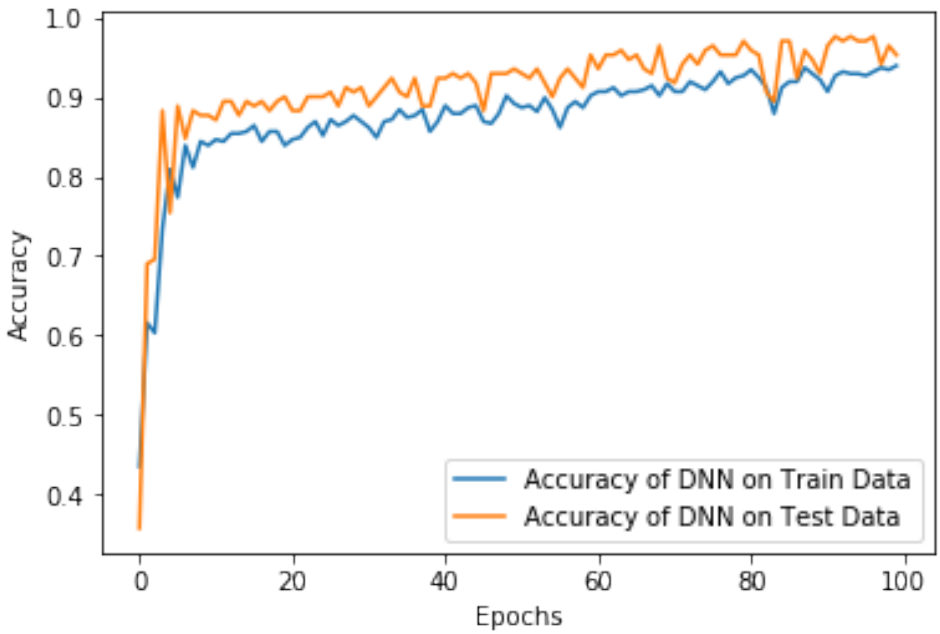

Figure 4. Accuracy versus Epochs in DNN

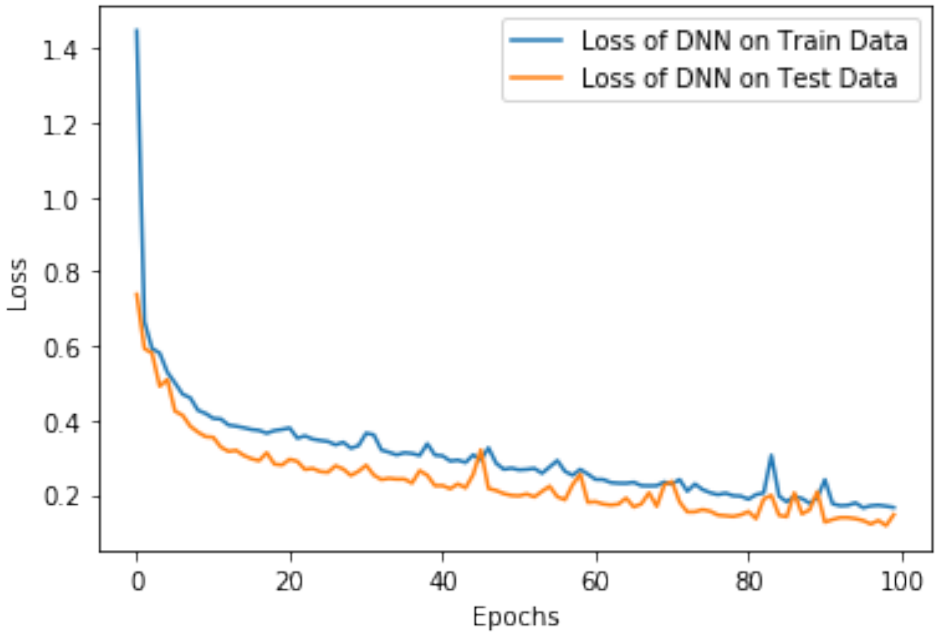

Figure 5. Loss Versus Epochs in DNN

Table 2. Performance of the DNN on the Training and Testing Data

Performance Metric	Training Data	Testing Data
Accuracy	94.97%	95.32%
Sensitivity	96.02%	95.08%
Specificity	94.33%	95.45%
Precision	91.19%	92.06%
F-Measure	93.54%	93.54%
Loss	17.63%	14.67%

The proposed PCA-DNN was also trained with 70% of the data and tested with 30% of the data. The training was done in a hundred epochs with batch size thirty. The obtained change in accuracy and loss of the proposed PCA-DNN with the increasing number of epochs is shown in Figures 6 and 7.

The performance of the PCA-DNN on the training and testing data is given in Table 3. The PCA-DNN obtained 98.24% accuracy on training data and 98.83% on testing data. It obtained a 4.87% loss on the training data and a 10.36% loss on the testing data.

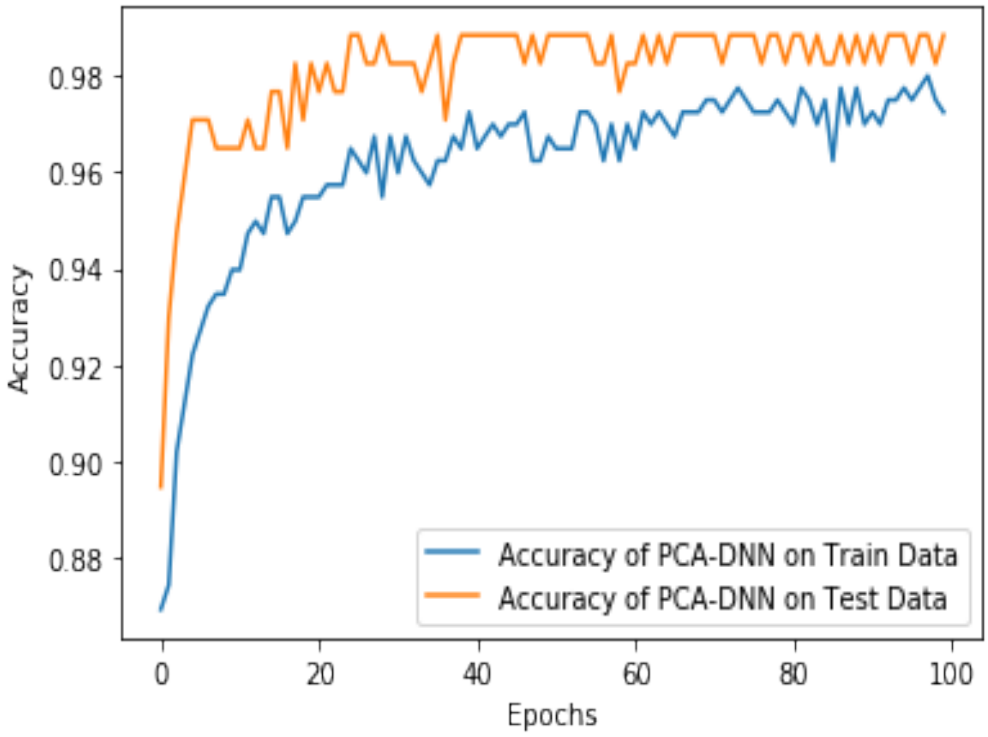

Figure 6. Accuracy versus Epochs in the proposed PCA-DNN

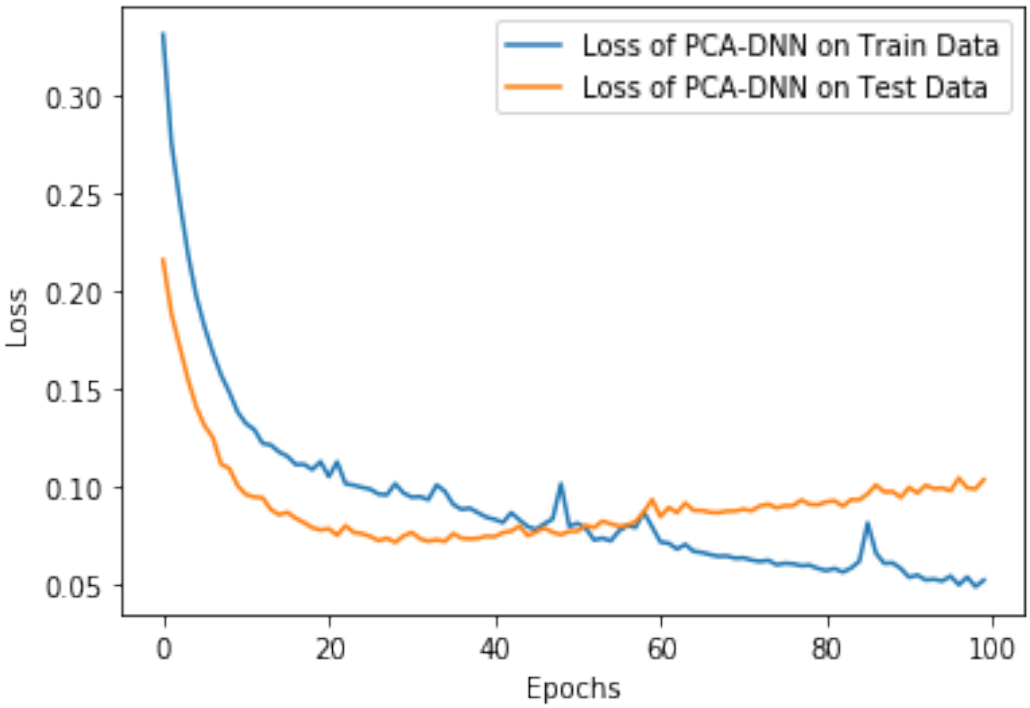

Figure 7. Loss versus Epochs in the proposed PCA-DNN

Table 3. Performance of the proposed PCA-DNN on the Training and Testing Data

Performance Metric	Training Data	Testing Data
Accuracy	98.24%	98.83%
Sensitivity	98.67%	98.36%
Specificity	97.97%	99.09%
Precision	96.75%	98.36%
F-Measure	97.70%	98.36%
Loss	4.87%	10.36%

4.4 Performance Comparison

The proposed PCA-DNN performed better than those obtained by machine learning algorithms and conventional DNN. The performance comparison among the Proposed PCA-DNN, machine learning classifiers, and DNN are shown in Figure 8. The confusion matrix of the machine learning algorithms under comparison is shown in Figure 9. The confusion matrix of the DNN and PCA-DNN is shown in Figure 10.

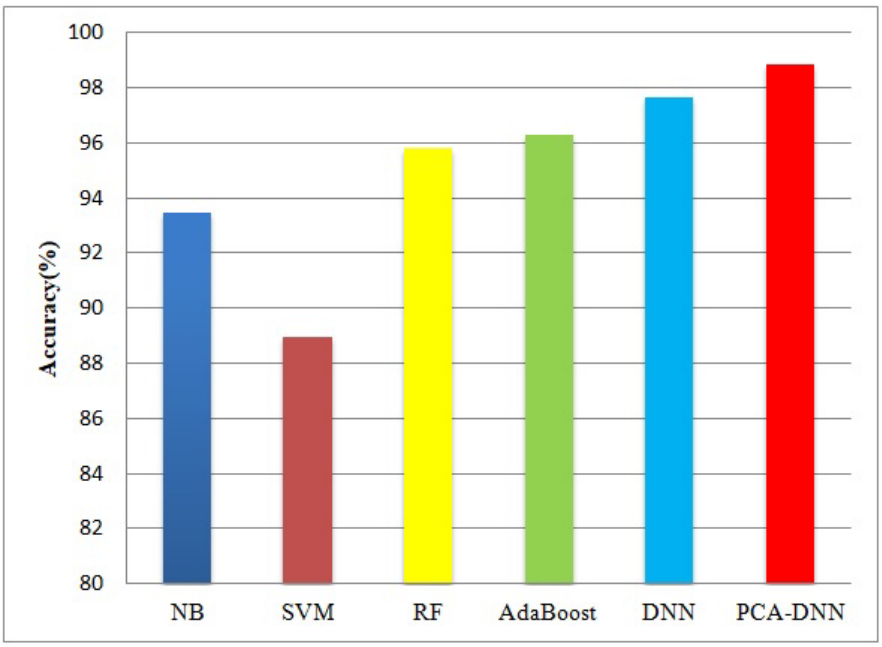

Figure 8. Performance comparison among the proposed PCA-DNN, machine learning classifiers, and DNN

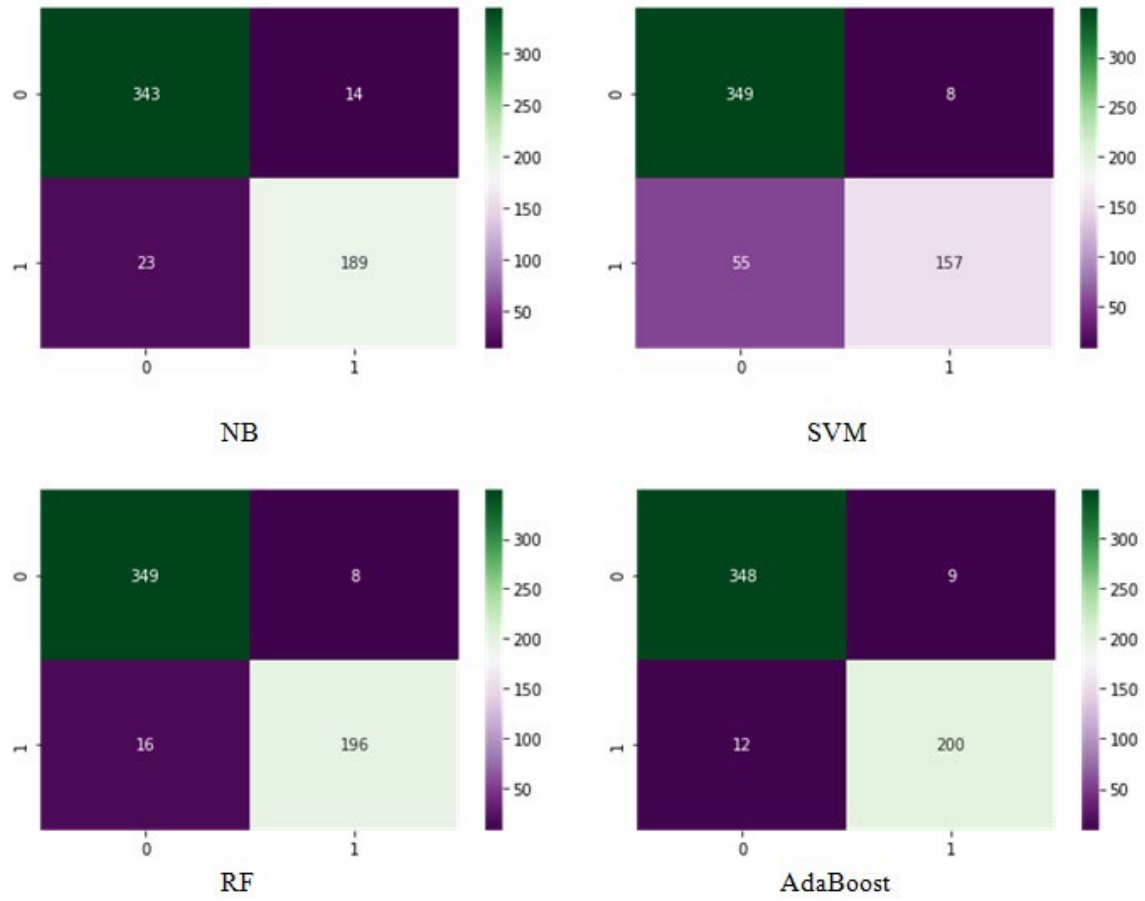

Figure 9. Confusion Matrix of the machine learning algorithms under study

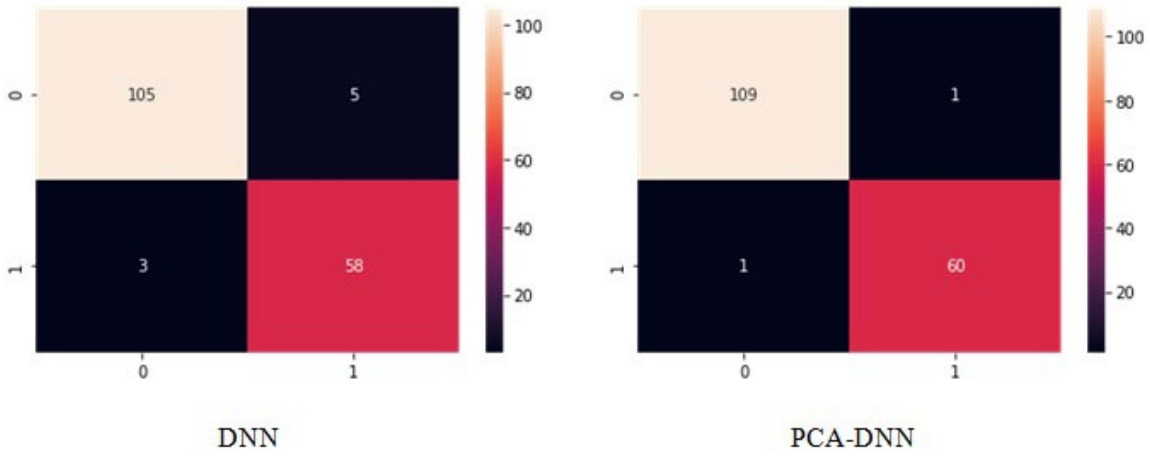

Figure 10. Confusion Matrix of the DNN and proposed PCA-DNN

The receiver operating characteristic (ROC) curve is also important to compare classification performance. It is commonly used to visualize the performance of models by displaying the tradeoffs between the classifier's cost and benefit. It plots the true positive rate against the false positive rate. The model's performance is represented by AUROC, with values near 1 denoting good performance. The ROC curve of different methods is shown in Figure 11. The proposed PCA-DNN obtained the best AUROC value, equal to 99.3%. The proposed PCA-DNN was also compared to existing works to classify in Table 4.

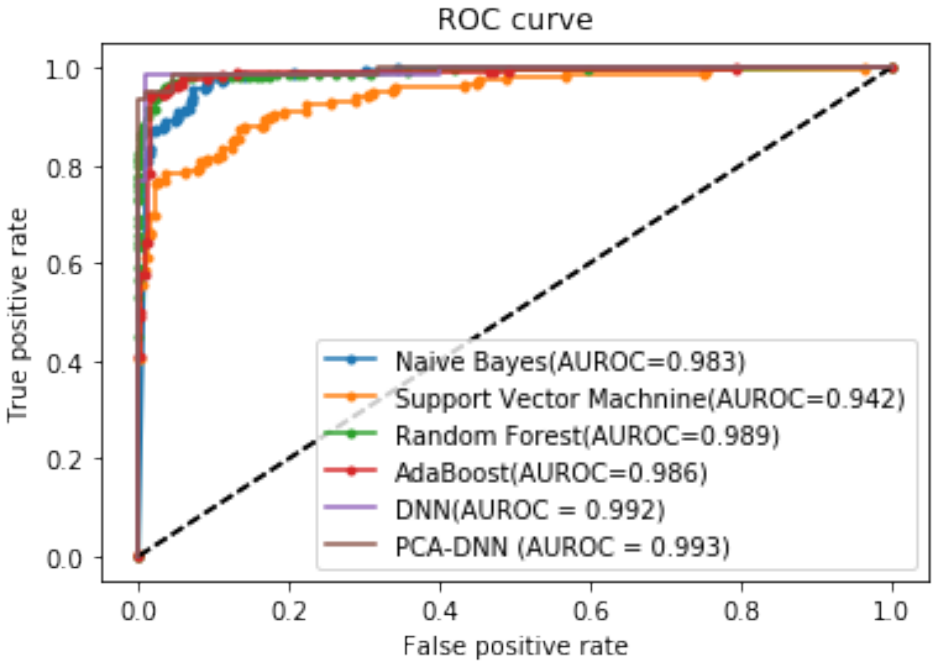

Figure 11. ROC Curve of the machine learning and deep learning methods under study

Table 4. Comparison among the proposed PCA-DNN and existing related works

Authors	Year	Dataset	Method	Accuracy(%)
Peng at al. [19]	2016	WDBC	Semi-supervised learning method based upon artificial immune	98.00
Nilashi et al. [20]	2017	WDBC	Fuzzy logic	93.20
Alikovi and Subasi [23]	2017	WDBC	Rotation forest	97.41
Dora et al. [22]	2017	WDBC	GNRBA	98.48
Wang et al. [24]	2018	WDBC	CPNN	97.40
Dhahri et al. [28]	2019	WDBC	Adaboost with the feature set reduced using genetic algorithm.	98.24
Abdar et al. [37]	2020	WDBC	Ensemble mechanism based on voting and stacking	98.07
Naji et al. [40]	2021	WDBC	Majority Voting	98.10
Al-Azzam and Shatnawi [41]	2021	WDBC	KNN	98.00
Proposed PCA-DNN	2022	WDBC	PCA-DNN	98.83

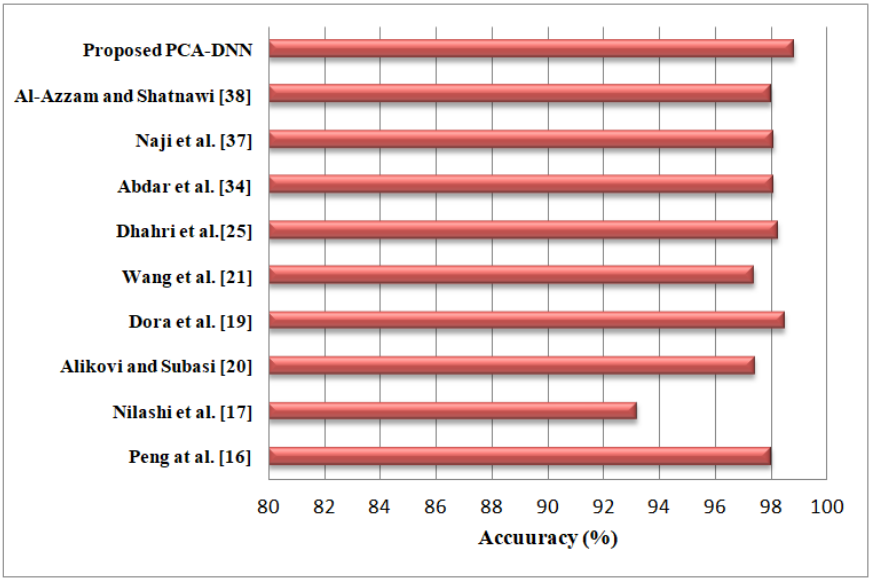

Figure 12. Comparison of proposed PCA-DNN with existing systems

A comparison is shown in Figure 12. Results analysis shows that PCA-DNN performs better than traditional machine learning and deep learning techniques.

The strengths of the PCA-DNN approach compared to conventional DNN techniques and the other classifiers are as follows:

1. **Dimensionality Reduction**: PCA helps reduce the dataset's dimensionality, which can lower computational complexity and improve the DNN's performance.
2. **Feature Extraction**: By giving a more pertinent feature set to the DNN, PCA can boost the performance of the DNN.
3. **Non-linearity Handling**: PCA-DNN effectively captures non-linear correlations in the data, making them appropriate for complicated datasets where conventional classifiers struggle.

Weaknesses of PCA-DNN strengths of the PCA-DNN approach compared to conventional DNN techniques and the other classifiers are as follows:

1. **Loss of Interpretability**: When PCA and DNN are combined, there may be a loss of interpretability, making it difficult to comprehend why the model made certain predictions.
2. **Data preprocessing Overhead**: The PCA procedure necessitates rigorous preprocessing, and DNN requires significant tuning, increasing computational overhead and complexity.

PCA-DNN has also improved performance over systems proposed by existing researchers. It can be concluded that doctors can utilize the PCA-DNN to detect breast cancer efficiently.

5. Conclusion & Future Work

In this research, PCA-DNN is proposed for the classification of breast cancer. The idea of PCA-DNN was put forth by integrating the PCA concept with traditional DNN. It allowed the utilization of explicit feature extraction with DNN. In addition, breast cancer was also classified using conventional DNN and machine learning classifiers. NB obtained an accuracy of 93.49%, SVM of 88.93%, RF of 95.78%, and Adaboost of 96.30%. The DNN obtained an accuracy of 95.32%. The proposed PCA-DNN obtained the highest accuracy, which was 98.83%. The proposed solution achieved reliable results on both training and testing data. Results from PCA-DNN outperformed those from traditional DNN and standard machine learning classifiers. It generated results that were 3.68% more accurate and 29.37% less loss than conventional DNN.

The PCA-DNN model can be used as a reliable tool for breast cancer diagnosis. This has important practical implications, such as increased diagnostic accuracy resulting in prompt interventions. The robust performance of the model is due to the explicit feature extraction capabilities made possible by combining PCA with DNN.

PCA-DNN might be further improved by applying the concept of regularization to DNN. The model's generalizability can be ensured by validating it across multiple datasets. Incorporating imaging data, such as mammograms, can further enrich the feature extraction process and possibly improve the model's performance in early cancer detection. Future research might be focused on turning the proposed methodology into a potentially useful tool for clinicians seeking a second opinion on a breast cancer diagnosis. More optimization techniques can be used to improve the system's performance.

Declarations

Funding: Not Applicable

Conflicts of interest: The authors declare no conflict of interest.

Availability of data and material: Not Applicable

Code availability: Not Applicable

References

1. Nasir MU, Ghazal TM, Khan MA, Zubair M, Rahman AU, Ahmed R, Hamadi HA, Yeun CY. Breast cancer prediction empowered with fine-tuning. Computational Intelligence and Neuroscience. 2022 Jun 9; 2022: 1-9. doi: 10.1155/2022/5918686.
2. Babiker S, Nasir O, Alotaibi SH, Marzogi A, Bogari M, Alghamdi T. Prospective breast cancer risk factors prediction in Saudi women. Saudi Journal of Biological Sciences. 2020 Jun 1; 27(6):1624-1631. doi: 10.1016/j.sjbs.2020.02.012.
3. Nomani A, Ansari Y, Nasirpour MH, Masoumian A, Pour ES, Valizadeh A. PSOWNNs-CNN: a computational radiology for breast cancer diagnosis improvement based on image processing using machine learning methods. Computational Intelligence and Neuroscience. 2022 May 11, 2022: 1-17. doi: 10.1155/2022/5667264.
4. Fatima N, Liu L, Hong S, Ahmed H. Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. IEEE Access. 2020 Aug 14;8:150360-76. doi: 10.1109/ACCESS.2020.3016715.
5. Hou R, Mazurowski MA, Grimm LJ, Marks JR, King LM, Maley CC, Hwang ES, Lo JY. Prediction of upstaged ductal carcinoma in situ using forced labeling and domain adaptation. IEEE Transactions on Biomedical Engineering. 2019 Sep 9;67(6):1565-72. doi: 10.1109/TBME.2019.2940195.
6. Chaudhury AR, Iyer R, Iychettira KK, Sreedevi A. Diagnosis of invasive ductal carcinoma using image processing techniques. In2011 International Conference on Image Information Processing 2011 Nov 3 (pp. 1-6). IEEE.doi: 10.1109/ICIIP.2011.6108877.
7. Lee B, Kim K, Choi JY, Suh DH, No JH, Lee HY, Eom KY, Kim H, Hwang SI, Lee HJ, Kim YB. Efficacy of the multidisciplinary tumor board conference in gynecologic oncology: a prospective study. Medicine. 2017 Dec; 96(48). doi: 10.1097/MD.0000000000008089.
8. Aggarwal K, Bhamrah MS, Ryait HS. Detection of cirrhosis through ultrasound imaging by intensity difference technique. EURASIP Journal on Image and Video Processing. 2019 Dec;2019:1-0.doi: 10.1186/s13640-019-0482-z.
9. Robertson FM, Bondy M, Yang W, Yamauchi H, Wiggins S, Kamrudin S, Krishnamurthy S, Le-Petross H, Bidaut L, Player AN, Barsky SH. Inflammatory breast cancer: the disease, the biology, the treatment. CA: a cancer journal for clinicians. 2010 Nov 1;60(6):351-375.doi: 10.3322/caac.20082.
10. Murtaza G, Shuib L, Abdul Wahab AW, Mujtaba G, Mujtaba G, Nweke HF, Al-garadi MA, Zulfiqar F, Raza G, Azmi NA. Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. Artificial Intelligence Review. 2020 Mar;53:1655-720.doi: 10.1007/s10462-019-09716-5.
11. Aggarwal K, Bhamrah MS, Ryait HS. Texture Analysis of Ultrasound Images of Liver Cirrhosis Through New Indexes. In Innovations in Computational Intelligence 2018 (pp. 93-101). Springer Singapore. doi: 10.1007/978-981-10-4555-4_7.
12. Mert A, Kilic N, Bilgili E, Akan A. Breast cancer detection with reduced feature set. Computational and mathematical methods in medicine. 2015 May 19;2015.doi: 10.1155/2015/265138.
13. Aggarwal K, Bhamrah MS, Ryait HS. Detection of Cirrhosis Through Ultrasound Imaging. InEmerging Trends in Intelligent Computing and Informatics: Data Science, Intelligent Information Systems and Smart Computing 4 2020 (pp. 245-258). Springer International Publishing.
14. Karthik S, Srinivasa Perumal R, Chandra Mouli PV. Breast cancer classification using deep neural networks. Knowledge Computing and Its Applications: Knowledge Manipulation and Processing Techniques: Volume 1. 2018:227-241. doi: 10.1007/978-981-10-6680-1_12.
15. Abdulla SH, Sagheer AM, Veisi H. Breast cancer classification using machine learning techniques: A review. Turkish Journal of Computer and Mathematics Education (TURCOMAT). 2021 Aug 20;12(14):1970-1979.
16. Sachdeva RK, Bathla P, Rani P, Kukreja V, Ahuja R. A systematic method for breast cancer classification using RFE feature selection. In2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) 2022 Apr 28 (pp. 1673-1676). IEEE. doi: 10.1109/ICACITE53722.2022.9823464.
17. Abdel-Zaher AM, Eldeib AM. Breast cancer classification using deep belief networks. Expert Systems with Applications. 2016 Mar 15;46:139-144.doi: 10.1016/j.eswa.2015.10.015.
18. Asri H, Mousannif H, Al Moatassime H, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia

Computer Science. 2016 Jan 1;83:1064-1069.doi: 10.1016/j.procs.2016.04.224.
19. Peng L, Chen W, Zhou W, Li F, Yang J, Zhang J. An immune-inspired semi-supervised algorithm for breast cancer diagnosis. Computer methods and programs in biomedicine. 2016 Oct 1;134:259-265. doi: 10.1016/j.cmpb.2016.07.020.
20. Nilashi M, Ibrahim O, Ahmadi H, Shahmoradi L. A knowledge-based system for breast cancer classification using fuzzy logic method. Telematics and Informatics. 2017 Jul 1;34(4):133-144. doi: 10.1016/j.tele.2017.01.007.
21. Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF. SVM and SVM ensembles in breast cancer prediction. PloS one. 2017 Jan 6;12(1):e0161501.doi:10.1371/journal.pone.0161501.
22. Dora L, Agrawal S, Panda R, Abraham A. Optimal breast cancer classification using Gauss–Newton representation-based algorithm. Expert Systems with Applications. 2017 Nov 1;85:134-145. doi: 10.1016/j.eswa.2017.05.035.
23. Aličković E, Subasi A. Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Computing and applications. 2017 Apr;28:753-763.doi: 10.1007/s00521-015-2103-9.
24. Wang H, Zheng B, Yoon SW, Ko HS. A support vector machine-based ensemble algorithm for breast cancer diagnosis. European Journal of Operational Research. 2018 Jun 1;267(2):687-99. doi: 10.1016/j.ejor.2017.12.001.
25. Sivakumar S, Nayak SR, Vidyanandini S, Kumar JA, Palai G. An empirical study of supervised learning methods for breast cancer diseases. Optik. 2018 Dec 1;175:105-114. doi: 10.1016/j.ijleo.2018.08.112.
26. Wang D, Wan S, Guizani N. Context-based probability neural network classifiers realized by genetic optimization for medical decision making. Multimedia Tools and Applications. 2018 Sep;77:21995-22006. doi: 10.1007/s11042-018-5631-3.
27. Zhang J, Chen L, Abid F. Prediction of breast cancer from imbalance respect using cluster-based undersampling method. Journal of Healthcare Engineering. 2019 Oct 16;2019: 7294582. doi: 10.1155/2019/7294582.
28. Dhahri H, Al Maghayreh E, Mahmood A, Elkilani W, Faisal Nagi M. Automated breast cancer diagnosis based on machine learning algorithms. Journal of Healthcare Engineering. 2019 Nov 3;2019. doi: 10.1155/2019/4253641.
29. Zhang J, Chen L. Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis. Computer Assisted Surgery. 2019 Oct 7;24(sup2):62-72. doi: 10.1080/24699322.2019.1649074.
30. Salod Z, Singh Y. Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol. Journal of Public Health Research. 2019 Dec 4;8(3): 1677. doi: 10.4081/jphr.2019.1677.
31. Kadam VJ, Jadhav SM, Vijayakumar K. Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression. Journal of medical systems. 2019 Aug;43(8):263. doi: 10.1007/s10916-019-1397-z.
32. Khan S, Islam N, Jan Z, Din IU, Rodrigues JJ. A novel deep learning-based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognition Letters. 2019 Jul 1;125:1-6.doi: 10.1016/j.patrec.2019.03.022.
33. Alghunaim S, Al-Baity HH. On the scalability of machine-learning algorithms for breast cancer prediction in big data context. IEEE Access. 2019 Jul 5;7: 91535-91546. doi: 10.1109/ACCESS.2019.2927080.
34. Memon MH, Li JP, Haq AU, Memon MH, Zhou W. Breast cancer detection in the IOT health environment using modified recursive feature selection. wireless communications and mobile computing. 2019 Nov 11;2019:1-9. doi: 10.1155/2019/5176705.
35. Abdar M, Makarenkov V. CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer. Measurement. 2019 Nov 1;146:557-570.doi: 10.1016/j.measurement.2019.05.022.
36. Zheng J, Lin D, Gao Z, Wang S, He M, Fan J. Deep learning assisted efficient AdaBoost algorithm for breast cancer detection and early diagnosis. IEEE Access. 2020 May 8;8:96946-54. doi: 10.1109/ACCESS.2020.2993536.
37. Abdar M, Zomorodi-Moghadam M, Zhou X, Gururajan R, Tao X, Barua PD, Gururajan R. A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognition Letters. 2020 Apr 1;132:123-131. doi: 10.1016/j.patrec.2018.11.004.
38. Supriya M, Deepa AJ. A novel approach for breast cancer prediction using optimized ANN classifier based on big data environment. Health care management science. 2020 Sep;23:414-26. doi: 10.1007/s10729-019-09498-w.
39. Kumar V, Mishra BK, Mazzara M, Thanh DN, Verma A. Prediction of malignant and benign breast cancer: A data mining approach in healthcare applications. InAdvances in Data Science and Management: Proceedings of ICDSM 2019 2020 (pp. 435-442). Springer Singapore. doi: 10.1007/978-981-15-0978-0_43.
40. Naji MA, El Filali S, Bouhlal M, Benlahmar EH, Abdelouhahid RA, Debauche O. Breast cancer prediction and diagnosis through a new approach based on majority voting ensemble classifier.

Procedia Computer Science. 2021 Jan 1;191:481-486.doi: 10.1016/j.procs.2021.07.061.
41. Al-Azzam N, Shatnawi I. Comparing supervised and semi-supervised machine learning models on diagnosing breast cancer. Annals of Medicine and Surgery. 2021 Feb 1;62:53-64. doi: 10.1016/j.amsu.2020.12.043.
42. https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic).
43. Verma K, Bhardwaj S, Arya R, Islam UL, Bhushan M, Kumar A, Samant P. Latest Tools for Data Mining and Machine Learning. International Journal of Innovative Technology and Exploring Engineering. 2019; 8 (9): 1-6.
44. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data. 2021 Dec;8:1-74. doi: 10.1186/s40537-021-00444-8.
45. Lamba R, Gulati T, Jain A, Rani P. A Speech-Based Hybrid Decision Support System for Early Detection of Parkinson's Disease. Arabian Journal for Science and Engineering. 2023 Feb; 48: 2247-2260.doi: 10.1007/s13369-022-07249-8
46. Kumar R, Rani P. Comparative analysis of decision support system for heart disease. Advances in Mathematics: Scientific Journal. 2020; 9(6):3349-3356. doi: 10.37418/amsj.9.6.15.
47. Ramesh TR, Lilhore UK, Poongodi M, Simaiya S, Kaur A, Hamdi M. Predictive analysis of heart diseases with machine learning approaches. Malaysian Journal of Computer Science. 2022 Mar 31:132-48.