IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 29, No. 1, February, 2024, pp. 347-353
Journal homepage: http://iieta.org/journals/isi

# Integration of ML Techniques for Early Detection of Breast Cancer: Dimensionality Reduction Approach

Wial Hanon[1*], Mahdi Abed Salman[2]

[1] Information Technology, Software Department, University of Babylon, Hilla 51001, Iraq
[2] College of Science for Women, Department of Computer Science, University of Babylon, Hilla 51001, Iraq

Corresponding Author Email: wailh@uobabylon.edu.iq



https://doi.org/10.18280/isi.290134

**Received:** 19 October 2023
**Revised:** 16 January 2024
**Accepted:** 23 January 2024
**Available online:** 27 February 2024

***Keywords:***
*Principal Component Analysis PCA, K-nearest neighbor KNN, integrate PCA and KNN DPBC, dimensionality reduction, diagnosis breast cancer DBC,* Breast Cancer Wisconsin medical *dataset BCW*

**ABSTRACT**

Nowadays, the diagnosis of breast cancer (DBC) helps doctors make early detection of breast cancer into non-cancerous (benign B) and cancerous (malignant M). Therefore, using machine learning (ML) algorithms is a solution to diagnosing and predicting symptoms related to DBC. The increased computational complexity, data size, overfitting, and longer training times harm early diagnosis accuracy. In this paper, propose a dimensionality reduction model integrating PCA and KNN for early breast cancer detection. which is used to diagnose and predict breast cancer (DPBC) based on reduced data size by selecting the best features that capture most of the variance in the data. The performance of the proposed model is evaluated with indices such as accuracy, precision, and f-score. Results for the DPBC model were obtained by using the Breast Cancer Wisconsin medical datasets (BCW).

## 1. INTRODUCTION

Globally, breast cancer, sometimes referred to as carcinoma, is the primary cause of death for women. Before affecting any nearby organs, it initially attacks the tissue in the breasts. If not detected in its early stages, it may turn deadly [1, 2]. Breast cancer is classified as either benign or malignant, depending on whether it is cancerous or non-cancerous. Differentiating between the tissue of a benign and malignant breast tumor is difficult [3].

ML-based methods benefit oncologists in making better medical decisions by making treating the condition simple and affordable. The network of neurons became a substitute for choosing the best qualities [4]. In deep learning, features are directly learned from data using several non-linear processing layers [5], but it still needs some of the requirements, such as a large amount of labeled data, expensive, uninterpretable, data bias, and long training time [6].

ML has been widely employed for computation processing because of its proven ability to improve and raise accuracy for both performance and prediction. The most well-known algorithms are neural networks, decision trees, random forests, and support vector machines (SVM). It is possible to use predictions or facts derived from experience. To determine the most precise link between variables, a variety of application methods can be applied, such as early breast cancer prediction, forecasting jobs, and time-series techniques [7, 8].

By developing prediction models, it may be possible to identify diseases earlier and provide patients with more effective treatment. ML models have demonstrated significant performance when used to diagnose breast cancer in earlier research [9, 10].

ML has been widely used in remote sensing because it can provide accurate predicted input-output data with strong correlations. Numerous options for biophysical parameter retrievals and applications are presented by this [11, 12].

The use of Internet of Things (IoTs) devices has become a necessity in our lives today, especially in the fields of health care [13]. In the same context, the increasing volume of data generated needs to be reduced to facilitate the transfer process to applications and cloud centers for the purpose of processing and analysis [14].

The PCA technique, which does not need data labeling and is a common dimensionality reduction method because to its simplicity and ease of implementation, is an example of an unsupervised learning technique. Its primary premise is the separation of feature groups, with the goal of reducing reciprocal correlation and sorting in accordance with a dropping eigenvalue and subsequently a declining variance. Principal components are another name for eigenvectors. They are initially subject to standard normalization because of the different feature domains [15].

In this paper, an integrated PCA and KNN algorithm for the breast cancer prediction model is proposed. The dimensionality reduction is used to enhance accuracy by selecting the best features. The mode is used to increase accuracy and speed in detecting breast cancer using the smallest possible number of features extracted from a CT scan or MRI scan image. Feature selection is embedded by computing the explained variance ratios and cumulative variance to understand how much variance each component explains, and then selecting the best number of components

based on a threshold for cumulative variance. Finally, it uses the new data components in the KNN to predict breast cancer. The proposed model is compared with several models used for the prediction of disease among patients. The main contributions of this paper include:

(1) To enhance breast cancer detection The integration of PCA and KNN algorithms is used.

(2) Using PCA enabled efficient and effective dimensionality reduction that captured most of the variance in the data and a several of components that will be used in the KNN classifier.

(3) Creation of a powerful ML model with the potential for clinical use in enhancing breast cancer detection. The dimensionality reduction approach led to reduced data size, hence easy to transmit to the cloud for storage analysis and processing for the long-term.

## 2. RELATED WORKS

Utilizing a variety of coping mechanisms, physicians are now able to diagnose breast cancer in women. Numerous data science (DS) approaches, in addition to new technologies, make it easier to gather and analyze cancer-related data in order to forecast this potentially fatal condition. The application of machine learning techniques of the treatment of cancer computationally has proved fruitful. Automatic learning systems, for instance, have been demonstrated in research [16] to boost diagnosis accuracy by 79.97%. On the other hand, machine learning achieved 91.1% accurate predictions.

In the study by Saleh et al. [5], the authors proposed a breast cancer prediction model using an improved deep learning methodology. The authors have provided this improved deep recurrent neural network (RNN) model based on RNN and Keras-Tuner Optimization approach for the early detection of breast cancer. An input layer, five hidden layers, five dropout layers, and an output layer make up the optimized deep RNN model.

Nicula et al. [17] using an SVM algorithm and a few chosen abilities that demonstrated how to find out the breast cancer. The performance of the model was verified using the DBC. When compared to other ML models, the experiment's results demonstrate that the suggested SVM has the highest classification accuracy, reaching up to 98.51 percent. ML approaches such as LR, SVM, and RF have been used to develop models for breast cancer prediction. SVM has scores between 82 and 88 percent, making it more sensitive than other models.

Baby et al. [18] presented a model of decisional trees for breast cancer detection. The Gini index was employed by the decision tree to establish the qualities' priority levels. With an accuracy rate of 90.52 percent, the suggested diagnostic method surpassed other models such as artificial neural network (ANN), SVM, KNN, NB, adaptive boosting (AdaBoost), and others.

Desai et al. [19] proposed a model using a Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN) for breast cancer classification and detection. The model's effectiveness is measured by how well it can spot cancer in breast cells. MLP is less accurate than CNN by 98.37 percent.

In the study by Rajaguru and SR [20], to pick features for the BCD dataset, PCA was used. The selected features were used to train and test a Decision Tree and KNN. With a 90.44 percent Mathews Correlation Coefficient, 95.61 percent accuracy, and 95.95 percent sensitivity, the KNN classifier surpassed the Decision Tree in every statistic. Cross-validation was not performed on the KNN and Decision Tree models. Overfitting and sampling bias may therefore have an impact on performance.

A similar study by Saoud et al. [21] utilized the most effective first search approach for feature selection and wrapper models, such as artificial neural networks (ANN), Bayesian networks, SVM, K-NN, Decision Trees, and Logistic Regression (LR). Different numbers of features were chosen by each model from the BCW dataset. The comparative examination of model performance was done by comparing accuracy measures for both models—one with feature selection and the other without. With an accuracy of 97.36 percent, the SVM model without feature selection fared better than the other models.

Omondiagbe et al. [22] were able to identify breast cancer with 98.82 percent accuracy 98.41 percent sensitivity, and 99.07 percent specificity, using the BCW data set with SVM, radial basis kernel, ANN, and Naive Bayes.

Kumar et al. [23] discovered that, when utilizing the WCB dataset, PCA and K-NN had a 96.4 percent accuracy rate in identifying breast cancer. Recently, two distinct datasets related to breast cancer were used to investigate the efficacy of K-NN utilizing different distance functions and k values. Studies are using K-NN, linear SVM, and Chi-squared features without feature selection.

Most of the reviews mentioned above focus on using several types of ML algorithms without considering the size of the data or the number of features that can be obtained. While keeping the rest of the features to use when needed again. This work used minimum features to find the best data that could be used in the DPBC model and integrate multiple MLs to overcome the lack of a general methodology that finds good solutions in all domains.

## 3. THE PROPOSED INTEGRATION APPROACH

The methods used in this paper are presented in this section. It also offers a comprehensive explanation of the breast cancer diagnosis and prediction methodology. The DPBC model is evaluated by using performance measures through Precision, Accuracy, and F1 Score.

### 3.1 Proposed model

The DPBC model recognizes the disease by integrating two algorithms PCA and KNN. Although this model yields promising results, the final best model for classification for decisions will be made based on a comparison of the output of the SVM, RF, and CNN algorithms with the proposed model in this study. The block diagram of the suggested model is shown in Figure 1. The dimensionality reduction PCA approach is used to increase accuracy, and feature selection is then integrated into the model. Whether or not the incoming data is diseased, all four strategies provide output according to case B or M.

The DPBC method begins with the normalization of X features, applying normalization before using ML techniques helps hasten the convergence of ML during training. Features with various scales may cause convergence to occur more slowly or to less ideal solutions. With data of different scales, problems like overflow or underflow may arise; normalization

can help prevent them.

A high-dimensional dataset can be converted into a lower-dimensional space using the approach of PCA, which keeps the most crucial data intact. The components that represent the most volatility in the data are linear combinations of the original attributes. Computation of the explained variance ratios and cumulative variance, followed by selecting the best number of components based on a threshold for cumulative variance (e.g., 95%) to determine the optimal number of features.

The dataset, the use case unique to the problem, and the quantity of information that should be kept may all influence the threshold and the ideal number of components chosen.

### 3.2 Description of the methodology

Overfitting occurs when we feed our model with datasets that are too big (having a lot of features and columns), which causes the model to start being affected by noise and outlier values. We refer to this as the Dimensionality Curse.

Dimensionality reduction is a statistical or machine learning technique that aims to produce a dataset with the ideal number of dimensions by reducing the number of features in the original dataset.

Feature extraction is a popular technique for achieving dimension reduction. It involves mapping a higher-dimensional feature space to a lower-dimensional feature space to minimize the number of dimensions. Principal component analysis (PCA) is the method for feature extraction that is most frequently used.

The number of potential feature subsets in feature selection issues grows exponentially with the number of features. Furthermore, there are a lot of issues with feature selection. Therefore, even with low-dimensional data, it is not viable to conduct an exhaustive search to discover the best solution.

The dimensionality reduction captures most of the variance in the data based on the number of components from the PCA algorithm and creates new data components that will be used in the KNN to predict. The next step is data splitting to values representing the new data components and targets (X_reduce, Y), followed by applying the KNN algorithm to select the best k-neabouers using an implementation loop to obtain high accuracy. The last step is to apply KNN based on the DPBC model to diagnose and predict the BC.

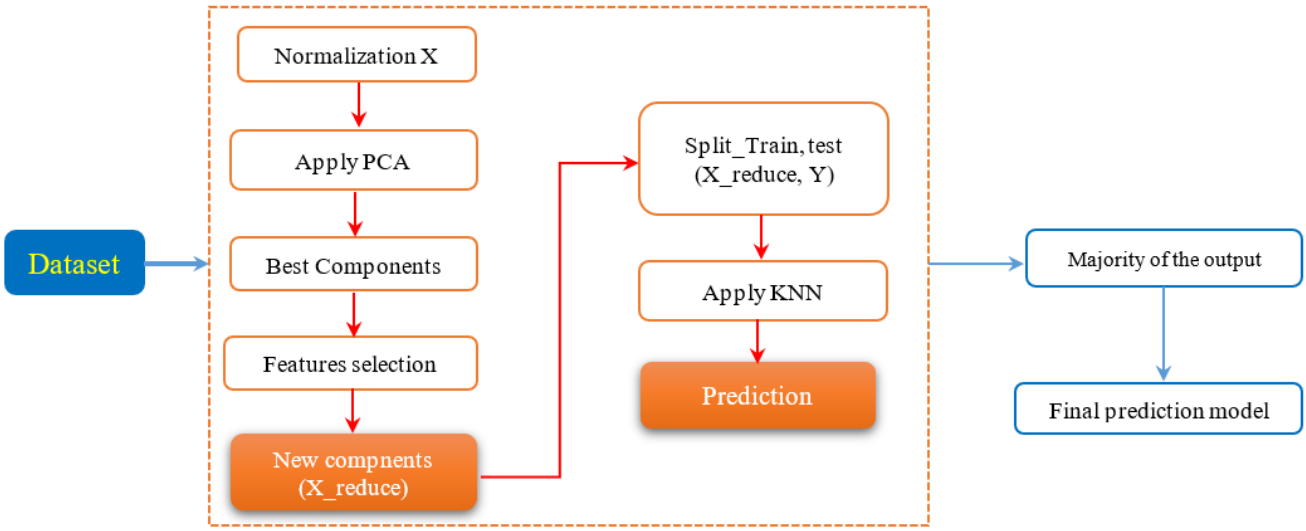


**Figure 1.** The proposed model DPBC

## 4. WORKING OF THE PROPOSED METHOD

The proposed methodology includes six levels of operations as shown in Figure 2. The proposed method starts with data collecting and then moves on to pre-treatment, which includes normalization and data cleansing from null and missing values before utilizing machine learning algorithms. Normalization helps hasten the convergence of ML during training. Features with various scales may cause convergence to occur more slowly or result in less ideal solutions.

The data model is constructed using coordinated data and AI calculations. The suggested methodology compares an improved DPBC model to other approaches to automatically predicting data reduction operations that can also be applied near the data production sources, and then prediction and diagnosis can be applied by the applications in the cloud, which can serve as a central data center for breast cancer data.

Uses a test set, which is a 20% subset of the entire data set, to assess the effectiveness of the model. Following the testing of the templates, we compare the outcomes to ascertain which algorithms yield the best approximation and maximum accuracy for detecting the advancement of cancerous cells.

The integrate KNN and PCA for dimensionality reduction, you can follow these steps after loading the dataset:

(1) Starting by normalizing the data until unifying the data within a specific range.

(2) Constructing the covariance matrix involves determining the relationship between pairs of variables and the amount of variance.

(3) Computing the eigenvectors and eigenvalues.

(4) Sorting the eigenvalues in descending order and selecting the top k eigenvectors corresponding to the highest eigenvalues.

(5) Transforming the feature matrix using the selected principal components by multiplying the original feature matrix by the selected eigenvectors to obtain a reduced-dimensional feature matrix.

(6) Splitting the dataset into training and testing sets by dividing the transformed feature matrix and the target variable into a training set and a testing set of 80% of the data and the

remaining for testing.

(7) Train the KNN model. Fit the KNN model using the training data. Specifying the number of nearest neighbors (k) and any other relevant parameters, such as distance metrics.

(8) Evaluate the model: Use the trained model to predict the testing data. Calculate the appropriate regression evaluation measure, such as mean squared error (MSE).

(9) Iterate through different values of k (the number of principal components) to find the optimal number of features that provides the best regression performance.

(10) Selecting the best-performing k and finalizing the model: Choose the value of k that yields the best model's performance based on the evaluation metrics. Train the final KNN model using this optimal number of principal components.

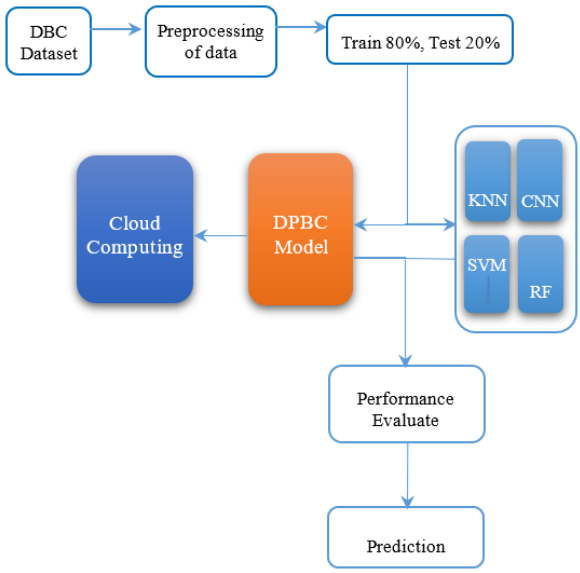


**Figure 2.** Flow diagram of ML models

## 5. EXPERIMENTATION

The ML model is trained on the dataset in advance of being used to predict the target and generate new prediction values when addressing a supervised ML challenge. Using fresh input data, cross-validation assesses the model's capacity to produce reliable and useful results. The primary issue is that the model's error rates don't indicate how well the model would function on fresh data or whether it is substantially biased (underfitted) or greatly variance-overfitted.

To conduct experiments, a computer is used with the specifications mentioned in Table 1 using the Python language and multiple libraries for programming.

**Table 1.** Experimental setup

| Item | Description |
|---|---|
| CPU | 2.70 GHz intel Core i3 |
| RAM | 4 GB |
| VRAM | 1 G |
| OS | Windows 10 pro 64bit |
| IDE | Python 3.10 |

In order to find the optimal hyper-parameters, build a model that is perfect for use with future data, and produce the most accurate predictions, machine learning models are subjected to cross-validation, which is the process of comparing their performance against a different collection of data called the holdout set or validation set.

The data size is necessary to be reduced due to the difficulty of transmission, processing, and storage, so we use a dimensionality reduction by the PCA algorithm for this purpose near data sources. The final diagnosis and prediction can be made in the cloud or in any other application. Feature selection is embedded in DPBC model.

Integrating PCA and KNN can be beneficial for several reasons:

- Dimensionality reduction.
- Feature selection.
- Handling multicollinearity.
- Improved generalization.
- Computational efficiency.

### 5.1 Dataset

The medical data sets from Breast Cancer Wisconsin (BCW) were used in the present study. The Kaggle datasets can be found at (https://www.kaggle.com/datasets) [24, 25]. There are 569 records in the collection. 357 (62.7%) cases of benign breast cancer and 212 cases of malignant breast cancer are reported. Every record contains 30 real-valued input attributes, an ID number, and a diagnostic (dataset label: "B" indicates benign, "M" indicates malignant). An aspirate digital image of the breast mass is used to real-value and evaluate these attributes. A database has 30 real-valued input attributes for 569 cases. There are no missing values in the dataset.

### 5.2 Evaluation measures

Precision, Accuracy, and F1 Score are used as performance measures to assess the model based on the test data. Eq. (1) and (2) reflect the precision value of the classification model, respectively, whereas Eqs. (3)-(4) provide the recall and F1 score, respectively.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

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

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

$$\text{F1 Score} = \frac{2*(\text{precision}*\text{recall})}{\text{precision}+\text{recall}} \quad (4)$$

Where, TP represents the true positive, FP for the false positive, TN for the true negative, and FN for the false negative, and these values are derived from the confusion matrix, see Table 2.

Performance evaluation of the model and accuracy by using four main parameters that are used in computing the accuracy of the models of ML.

- If the right response is positive, TP (true positive) predicts positive;
- If the correct answer is negative, TN (true negative) predicts negative;
- If the correct answer is negative, FP (false positive) predicts positive;
- If the correct answer is positive, FN (false negative) predicts negative.

Accuracy: is a performance measure used in classification problems. The classification aims to determine the class of new instances within the predefined classes. The percentage of accurately anticipated cases relative to all instances is known as accuracy. The enhancement of the data size by calculating the original data then applying PCA and calculating the data size. It achieved a reduction of 68%.

From the analysis of the confusion matrix, as shown in Figure 3 we observe that the infected and undetected cases are zero, meaning there is no benign case that is actually malignant. This is considered an encouraging percentage in diagnosing breast cancer.

The mode achieves a reduction of features from 30 to 8, as shown in Figure 4 where (a) represents the original features and (b) represents the features of applying PCA. By combining various models, these techniques have the potential to increase the accuracy and robustness of prediction and classification models for the problem of breast cancer risk.

**Table 2.** Confusion matrix table

| | Expected M | Expected B |
|---|---|---|
| Real M | TP | FN |
| Real B | FP | TN |

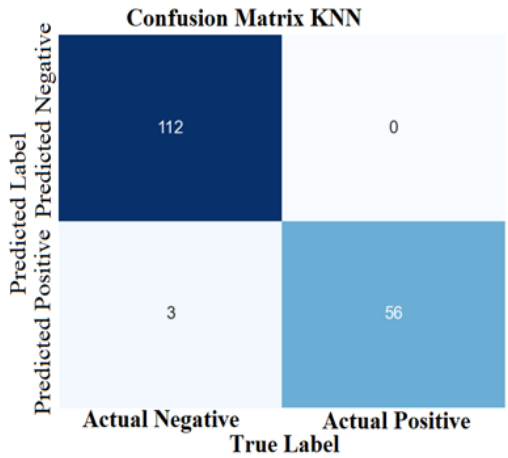


**Figure 3.** Confusion matrix of the DPBC

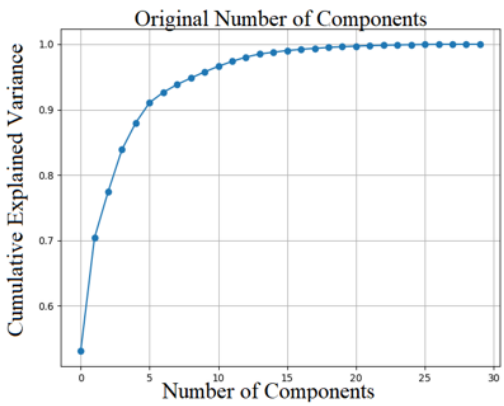


**Figure 4.** Data size reduction

## 6. RESULTS AND DISCUSSION

Implemented the DPBC model and additional ML algorithms for experimenting. by conducting and assessing experiments utilizing Python as a programming language and tools like PyCaret and Sklearn. Each experiment is conducted using a different methodology on the BCW dataset. Table 3 shows results from using the features selection by PCA then KNN, RF, SVM, and CNN in terms of accuracy and without using PCA.

The WCB datasets consist of a large number of features; therefore, PCA dimensionality reduction in integration with KNN helps improve the results. Due to its reliance on distance measures, KNN can suffer from the curse of dimensionality. High-dimensional data can result in more complex computations, more time spent training, and a higher chance of overfitting. KNN becomes more effective and efficient because of PCA, which lowers the data's dimensionality while maintaining crucial information.

On the BCW dataset, the DPBC model's accuracy, recall, precision, and F1 score are compared to those of other ML models. In comparison with other methods, Table 4 describes the comparative accuracy results of the same dataset. The results for the DPBC model, compared to other approaches, appear to have better performance, as shown in the table for DPBC results.

**Table 3.** Results from DPBC model

| Method | Accuracy with Features Selection | Accuracy Without Features Selection |
|---|---|---|
| **KNN** | **99.12** | **94.15** |
| RF | 96.73 | 94.49 |
| SVM | 97.86 | 92.98 |
| CNN | 97.37 | 93.86 |

**Table 4.** Results of the DPBC model with the other models BCW dataset

| Reference | Model | Accuracy% | Precision% | F1% |
|---|---|---|---|---|
| [23] | PCA+KNN | 96.4 | | |
| [20] | PCA+KNN | 95.57 | | |
| [21] | SVM | 97.36 | | |
| [22] | SVM | 98.82 | 99.07 | 98.41 |
| [26] | LR with Area | 98.06 | | |
| [17] | SVM | 98.51 | | |
| [19] | CNN | 98.37 | | |
| [27] | Voting Classifier | 97.61 | | |
| [27] | Polynomial SVM | 99.03 | | |
| [27] | KNN with hyperparameter | 97.35 | | |
| [27] | PCA +Logistic Regression | 94.87 | 94.81 | 92.9 |
| [28] | SVM | | 98 | 96 |
| [28] | KNN | | 94 | 96 |
| [28] | RF | | 96 | 97 |
| [29] | LR+SVM | 98.77 | 98.83 | 98.68 |
| **In this study** | **DPBC** | **99.12** | **99.13** | **99.12** |

In terms of early breast cancer prediction, the DPBC model's results are 4.28% better than those of the other research. These differences result from the DPBC model's use of embedded feature selection approaches, which were applied via PCA and KNN. This demonstrates how the feature selection process affected the model's performance.

The evaluation metrics that were computed by using Eqs. (1) through (4) showed that the DPBC model had remarkably high precision, accuracy, recall, and F1 score. With an accuracy of 0.9912, the model correctly categorized 99.12% of the events. With a recall value of 0.9912, the model successfully classified

99.12% of the actual malignant cases as such. There has been a 68% decrease in the size of the data.

The proposed model was evaluated using different algorithms such as K-Nearest Neighbor, Support Vector Machine, Random Forest, and CNN. As can be seen in the results section, the DPBC model has an accuracy score of 99.12%, while the SVM, RF, and CNN have scores of 96.73%, 96.73%, and 97.86%, respectively. The DPBC model yielded the greatest precision and F1 scores, at 99.13% and 99.12%, respectively. The study's conclusions show that the proposed DPBC model outperforms several classifiers. A comparison of the proposed model and other models from the literature is also presented in this study.

## 7. CONCLUSION

Many lives can be saved if breast cancer is detected early. The integration of PCA and KNN in the DPBC model is presented in this study. ML uses several significant dimensionality reduction techniques. A subset of principle components that captures the majority of the variance in the data can be chosen using the dimensionality reduction approach PCA, which can be used for feature selection. This model makes use of the power of integrating various algorithms to improve prediction accuracy and offer more thorough insights. Furthermore, creating ML classifier models, and dimensionality reduction is an essential stage because it has a big impact on the model's performance, training duration, and interoperability. Enhanced capability for prediction models using numerous ML algorithms. Integrate models can capture a range of features and lower the danger of overfitting by merging the predictions of various models, resulting in more accurate and dependable forecasts. This may increase the precision of breast cancer risk evaluations. This demonstrates that, when compared to other ML models, the DPBC model classifier has the most accurate results, with the highest accuracy, precision, and F score values. Utilizing the DPBC model results in a reduction of the original data size by up to 68.75% when compared to the size after applying PCA.

In the future, improve performance by using several new methods of feature selection. The model designed for other forms of cancer is also utilized. Furthermore, the performance of the proposed model can be tested in real-time.

## REFERENCES


[1] Schmid, P., Cortes, J., Dent, R., et al. (2022). Event-free survival with pembrolizumab in early triple-negative breast cancer. New England Journal of Medicine, 386(6): 556-567. https://doi.org/10.1056/nejmoa2112651

[2] Hao, Y., Zhang, L., Qiao, S., Bai, Y., Cheng, R., Xue, H., Hou, Y., Zhang, W., Zhang, G. (2022). Breast cancer histopathological images classification based on deep semantic features and gray level co-occurrence matrix. Plos One, 17(5): e0267955. https://doi.org/10.1371/journal.pone.0267955

[3] Naji, M.A., El Filali, S., Aarika, K., Benlahmar, E.H., Abdelouhahid, R.A., Debauche, O. (2021). Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Computer Science, 191: 487-492. https://doi.org/10.1016/j.procs.2021.07.062

[4] Islam, M.M., Haque, M.R., Iqbal, H., Hasan, M.M., Hasan, M., Kabir, M.N. (2020). Breast cancer prediction: A comparative study using machine learning techniques. SN Computer Science, 1: 1-14. https://doi.org/10.1007/s42979-020-00305-w

[5] Saleh, H., Alyami, H., Alosaimi, W. (2022). Predicting breast cancer based on optimized deep learning approach. Computational Intelligence and Neuroscience, 2022: 1820777. https://doi.org/10.1155/2022/1820777

[6] Pandian, A.P. (2021). Performance evaluation and comparison using deep learning techniques in sentiment analysis. Journal of Soft Computing Paradigm, 3(2): 123-134. https://doi.org/10.36548/jscp.2021.2.006

[7] Gzar, D.A., Mahmood, A.M., Abbas, M.K. (2022). A comparative study of regression machine learning algorithms: Tradeoff between accuracy and computational complexity. Mathematical Modelling of Engineering Problems, 9(5): 1217-1224. https://doi.org/10.18280/mmep.090508

[8] Benamor, Z., Seghir, Z.A., Djezzar, M., Hemam, M. (2023). A comparative study of machine learning algorithms for intrusion detection in IoT networks. Revue d'Intelligence Artificielle, 37(3): 567-576. https://doi.org/10.18280/ria.370305

[9] Dalwinder, S., Birmohan, S., Manpreet, K. (2020). Simultaneous feature weighting and parameter determination of neural networks using ant lion optimization for the classification of breast cancer. Biocybernetics and Biomedical Engineering, 40(1): 337-351. https://doi.org/10.1016/j.bbe.2019.12.004

[10] Dadheech, P., Kalmani, V., Dogiwal, S.R., Sharma, V.K. (2021). Breast cancer prediction using supervised machine learning techniques. Journal of Information and Optimization Sciences, 44(3): 383-392. https://doi.org/10.47974/jios-1348

[11] Nanglia, S., Ahmad, M., Khan, F.A., Jhanjhi, N.Z. (2022). An enhanced Predictive heterogeneous ensemble model for breast cancer prediction. Biomedical Signal Processing and Control, 72: 103279. https://doi.org/10.1016/j.bspc.2021.103279

[12] Lilhore, U.K., Simaiya, S., Pandey, H., Gautam, V., Garg, A., Ghosh, P. (2022). Breast cancer detection in the IoT cloud-based healthcare environment using fuzzy cluster segmentation and SVM classifier. In: Hu, YC., Tiwari, S., Trivedi, M.C., Mishra, K.K. (eds) Ambient Communications and Computer Systems. Lecture Notes in Networks and Systems, 356. https://doi.org/10.1007/978-981-16-7952-0_16

[13] Janani, S.P., Jebadurai, I.J., Paulraj, G.J.L., Jebadurai, J., Durga, S. (2022). Preparedness for managing pandemic using distributed mobile brokers–Using COVID 19 use case. Materials Today: Proceedings, 51: 2384-2388. https://doi.org/10.1016/j.matpr.2021.11.586

[14] Muneeb, M., Ko, K.M., Park, Y.H. (2021). A fog computing architecture with multi-layer for computing-intensive IoT applications. Applied Sciences, 11(24): 11585. https://doi.org/10.3390/app112411585

[15] Bai, L., Song, C., Zhou, X., Tian, Y., Wei, L. (2023). Assessing project portfolio risk via an enhanced GA-BPNN combined with PCA. Engineering Applications of Artificial Intelligence, 126: 106779. https://doi.org/10.1016/j.engappai.2023.106779

[16] Rahman, M.M., Ghasemi, Y., Suley, E., Zhou, Y., Wang, S., Rogers, J. (2021). Machine learning based computer aided diagnosis of breast cancer utilizing anthropometric

and clinical features. IRBM, 42(4): 215-226. https://doi.org/10.1016/j.irbm.2020.05.005

[17] Nicula, B., Dascalu, M., Newton, N.N., Orcutt, E., McNamara, D.S. (2021). Automated paraphrase quality assessment using language models and transfer learning. Computers, 10(12): 166. https://doi.org/10.3390/computers10120166

[18] Baby, D., Devaraj, S.J., Hemanth, J. (2021). Leukocyte classification based on feature selection using extra trees classifier: Atransfer learning approach. Turkish Journal of Electrical Engineering and Computer Sciences, 29(8): 2742-2757. https://doi.org/10.3906/elk-2104-183

[19] Desai, M., Shah, M. (2021). An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). Clinical eHealth, 4: 1-11. https://doi.org/10.1016/j.ceh.2020.11.002

[20] Rajaguru, H., SR, S.C. (2019). Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer. Asian Pacific Journal of Cancer Prevention: APJCP, 20(12): 3777-3781. https://doi.org/10.31557/APJCP.2019.20.12.3777

[21] Saoud, H., Ghadi, A., Ghailani, M., Abdelhakim, B.A. (2019). Using feature selection techniques to improve the accuracy of breast cancer classification. In: Ben Ahmed, M., Boudhir, A., Younes, A. (eds) Innovations in Smart Cities Applications Edition 2. SCA 2018. Lecture Notes in Intelligent Transportation and Infrastructure. Springer, Cham, 307-315. https://doi.org/10.1007/978-3-030-11196-0_28

[22] Omondiagbe, D.A., Veeramani, S., Sidhu, A.S. (2019). Machine learning classification techniques for breast cancer diagnosis. IOP Conference Series: Materials Science and Engineering, 495: 012033. https://doi.org/10.1088/1757-899X/495/1/012033

[23] Kumar, E.S., Bindu, C.S., Madhu, S. (2023). Deep convolutional neural network-based analysis for breast cancer histology images. In Machine Learning and Deep Learning in Real-Time Applications. IGI Global. https://doi.org/10.4018/978-1-7998-3095-5.ch008

[24] Sharma, D., Kumar, R., Jain, A. (2022). Hybrid Missing Value Imputation Algorithm-KLR. Mathematical Statistician and Engineering Applications, 71(2): 60-74. https://doi.org/10.17762/msea.v71i2.67

[25] Khan, M.M.R., Arif, R.B., Siddique, M.A.B., Oishe, M.R. (2018). Study and observation of the variation of accuracies of KNN, SVM, LMNN, ENN algorithms on eleven different datasets from UCI machine learning repository. In 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), Dhaka, Bangladesh, pp. 124-129. https://doi.org/10.1109/CEEICT.2018.8628041

[26] Khan, F., Khan, M.A., Abbas, S., Athar, A., Siddiqui, S.Y., Khan, A.H., Saeed, M.A., Hussain, M. (2020). Cloud-based breast cancer prediction empowered with soft computing approaches. Journal of Healthcare Engineering, 2020: 8017496. https://doi.org/10.1155/2020/8017496

[27] Sharma, D., Kumar, R., Jain, A. (2022). Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning. Measurement: Sensors, 24: 100560. https://doi.org/10.1016/j.measen.2022.100560

[28] Jaiswal, V., Saurabh, P., Lilhore, U.K., Pathak, M., Simaiya, S., Dalal, S. (2023). A breast cancer risk predication and classification model with ensemble learning and big data fusion. Decision Analytics Journal, 8: 100298. https://doi.org/10.1016/j.dajour.2023.100298

[29] Omotehinwa, T.O., Oyewola, D.O., Dada, E.G. (2023). A Light Gradient-Boosting Machine algorithm with Tree-Structured Parzen Estimator for breast cancer diagnosis. Healthcare Analytics, 4: 100218. https://doi.org/10.1016/j.health.2023.100218