JIRBDAI Journal of Institutional Research, Big Data Analytics and Innovation

Breast Cancer Prediction using Support Vector Machine and Random Forest Algorithms

Mohammed Abdulkadir[1], Aro Taye Oladele[2], Seriki Aliu Adebayo[3], Saka Kayode Kamil[4], Sanusi Sodiq[5], Hammed Abdullahi[6]

[1,2,4,5,6]Department of Computer Science, Al-Hikmah University, Ilorin, Kwara State, Nigeria
[3]Department of Cyber Security, Abdulrasaq Abubakar Toyin University, Ganmo, Kwara State, Nigeria
Corresponding Author: aliuseriki@gmail.com

ABSTRACT

Breast cancer continues to be one of the primary causes of cancer-related death globally, highlighting the necessity of precise and prompt diagnostic techniques. Despite being widely utilized, conventional methods like mammography, ultrasonography, and biopsy are frequently linked to subjectivity, high false-positive rates, and restricted accessibility in places with limited resources. This study aims to evaluate the effectiveness of Support Vector Machine (SVM) and Random Forest Classifiers (RFC) for breast cancer prediction, with a focus on enhancing diagnostic accuracy through advanced feature extraction and dimensionality reduction. The Kaggle repository provided a publicly accessible dataset with 569 occurrences and 33 characteristics. Principal Component Analysis (PCA) was used to minimize dimensionality while maintaining 90–95% of the variance in the data. Standard performance indicators, such as accuracy, precision, recall, F1-score, and ROC AUC, were used to train and assess the model. Confusion matrix analysis, ROC curves, and classification reports provided additional support. According to the results, SVM performed the best, with an accuracy of 97.7% and a ROC AUC of 0.99. RFC came in second with an accuracy of 94.2% and a ROC AUC of 0.98. Both SVM and RFC demonstrated a good balance between recall and precision, reducing diagnostic error.

KEYWORDS: Support Vector Machine (SVM), Principal Component Analysis (PCA), Random Forest Classifier (RFC)

I. INTRODUCTION

Breast cancer has been the most prevalent cancer worldwide. AI has recently demonstrated a great deal of promise for increasing the precision and effectiveness of cancer diagnosis through predictive modeling, image analysis, and genomic data interpretation (Sufyan et al, 2023). Researchers have found patterns and signals in genetic data and medical pictures using machine learning algorithms and deep neural networks that can help with early cancer identification and individualized treatment planning (Jiang et al, 2023). According to projections, the incidence will increase by 12% by 2025 due to differences in healthcare availability, aging populations, and lifestyle factors. Since the 5-year survival rate for localized tumors is 99% and that of metastatic instances is only 30%, early identification is crucial (American Cancer Society, 2023). High false-positive rates (50% during repeated screenings), which result in needless biopsies, continue to be a problem despite advances in imaging technology like 3D mammography and contrast-enhanced MRI (Jiang et al, 2023). Diagnostic consistency is impacted by inter-observer variability among radiologists (Obaid et al., 2018). Conventional techniques like biopsy, ultrasonography, and mammography are limited by their invasiveness, expense, and subjectivity. These difficulties highlight the need for automated, scalable, and reasonably priced diagnostic technologies to support clinical operations. Breast cancer detection using machine learning (2020–2025). The use of machine learning (ML) techniques for breast cancer diagnosis has advanced significantly between 2020 and 2025. Support Vector Machines (SVM), Random Forests, and Deep

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.

Learning (DL) architectures are examples of machine learning techniques that have shown great promise for increasing diagnostic efficiency and accuracy. A supervised learning technique called Support Vector Machines (SVMs) uses the identification of hyperplanes in high-dimensional spaces to classify data. SVMs have shown successful in classifying histopathological images and selecting features from mammograms for the identification of breast cancer. Jabbar et al. (2021), for instance obtained 92% accuracy when using SVM with histogram of oriented gradients (HOG) characteristics to distinguish between benign and malignant cancers. However, SVMs require kernel methods for optimization because they have trouble with huge datasets and non-linear data distributions (Anmol et al, 2024). An ensemble of decision trees called Random Forests (RF) is particularly good at managing high-dimensional data and minimizing overfitting. By utilizing characteristics including cell nucleus texture and radius. This study aims to close these gaps by undertaking a methodical assessment of particular machine learning algorithms for breast cancer prediction and enhancing greater picture resolution with additional dataset preparation.

II. RELATED WORKS

Worldwide, breast cancer is the most prevalent type of cancer that affects women. It occurs when malignant cells in your breasts grow and develop into a tumor. Approximately 80% of cases of breast cancer are invasive, which means that the tumor may spread from the breast to other parts of the body. Although it usually affects women over 50, breast cancer can potentially strike women under 50. According to Cleveland Clinic London (2025), men can also get breast cancer. Since breast cancer is one of the main causes of cancer-related mortality worldwide, early and precise identification is essential to raising survival rates. Numerous machine learning approaches, including K-Nearest Neighbors (KNN), Naive Base (NB), Decision Tree (DT), Support Vector Machine (SVM), and Logistic Regression (LR) have been created for the diagnosis of breast cancer. Kabiraj et al., (2020) proposed a Random Forest and XGBoost system for predicting breast cancer risk and got 74.73%, 73.63% respectively. The gradient boosting method known as EXSA was created by Liu et al., (2020) to forecast the course of breast cancer and raise survival rates. In order to classify the breast cancer data, Jabbar (2021) built an ensemble model using Bayesian networks and Radial Basis Functions, which produced correct classifications with an ideal accuracy of 97.42%. Ara et al. (2021) used SVM and RF to achieve 96.5% accuracy in their feature selection and machine learning (ML) categorization of a breast cancer data set into benign and malignant categories.

Ara et al., (2021) used SVM and RF to achieve 96.5% accuracy in their feature selection and machine learning (ML) categorization of a breast cancer data set into benign and malignant categories. Mahesh et al. (2022) used different machine learning techniques and also blended ensemble learning for breast cancer prediction and achieved 98.14% accuracy. According to Obaid et al., (2018), SVM with a quadratic kernel function produced the best accuracy. A computer-aided design (CAD) system that identifies breast cancers (BCs) was used in Samer et al 2023 study. The Kaggle data-set from the repository is used for BC detection using random forest, AdaBoost, logistic regression, decision trees, Naïve Bayes, and Conventional Neural Networks (CNNs) classifiers. These Machine Learning (ML) based algorithms are trained to predict BCs (malignant or benign), with attribute clump thickness serving as the evaluation class. Accuracy and F-measure are used to assess the efficacy of these ML methods; random forest fared better than the other classifiers, achieving 99% accuracy and 99% F-measure. However, the model's accuracy and generalizability can be improved by growing the dataset to include a more varied patient group and multi-modal data integration, such as genetic or patient history. Anmol et al (2024) researched on Deep Learning-Based AI-Powered Detection for Breast Cancer Prediction. Mini-DDSM, a condensed form of the Digital Database for Screening Mammography, is a subset of the data used in this work. To

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21473365

maximize performance, they created a deep learning model that includes dense layers, a dropout layer, a global pooling layer, and a transfer learning layer (DenseNet121).They developed model gives an accuracy of 76.28% using DenseNet121 for feature extraction and normalization and Neighborhood Cleaning Rule (NCR) for data imbalance. Though there is definitely need for improvement, the model shows promise. In order to get beyond the limitations of JPEG compression, future research may concentrate on employing better resolution photos or alternative imaging methods. Singh (2019) used SVM to conduct his research with a dataset of 8473 sentences and an F1 score of 93.53. Although machine learning models have demonstrated encouraging outcomes in the detection of breast cancer, researchers are confronted with several obstacles in these domains such as federated learning is an emerging method that enables different institutions to cooperatively train ML models. Tot et al. (2021) worked on clinical value of detecting microcalcifications on a mammogram. Their review summarizes the theoretical and methodological background for understanding the clinical impact and discusses the diagnostic and prognostic value of microcalcifications detected in the breast by mammography. Sub gross morphology, high histology tumor grade, increased metastatic potential, and decreased disease-free and overall survival. This percentage of calcification can be improved on. Ardalan et al (2019) carried out research on Digital Database for Screening Mammography. The study used the AdaBoost algorithm with a dataset collected from the online cancer imaging archive, with an accuracy of 97.2 %. However, their model did not provide transparent and interpret-able prediction. Table 1 shows the summary of some of the literature's that were reviewed.

III. METHODOLOGY

The research design used in the study was experimental. Support Vector Machine (SVM), Random Forest (RF), and Convolutional Neural Network (CNN) were the three machine learning algorithms that were empirically tested as part of the design to see how well they classified cases of breast cancer. The Kaggle Breast Cancer Dataset was used for the studies. To maintain the class distribution, a stratified split was used to separate the dataset into 70% training data and 30% testing data. Because the study was quantitative in nature, model performance could be objectively measured and compared using assessment measures like accuracy, precision, recall, and F1-score. Grid Search Cross-Validation (GridSearchCV) was used to optimize each algorithm's hyperparameters. K-Fold Cross-Validation (K = 5) was also included in the experimental design to reduce bias and guarantee the results' generalizability. Figure 1 shows the flowchart diagram of the study. Furthermore, the process involved in the methodology is as follows:

(a). Data Acquisition and Preprocessing

The Kaggle Breast Cancer Dataset, which has 569 samples with 33 attributes that indicate properties of cell nuclei, was the dataset utilized. Every record is marked as either malignant (1) or benign (0). Pandas and Scikit-learn Python libraries were used for data preprocessing. Among the steps were:

i. Handling Missing Values: Missing or null values were detected and replaced using mean imputation to ensure data completeness.

ii. Data Normalization: Since the features varied in scale, Min–Max normalization was applied to rescale all attributes to the range [0, 1]. This improved convergence during model training.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

iii. Feature Selection: Principal Component Analysis (PCA) was applied to reduce dimensionality while preserving 90% of the data variance. This minimized computational cost and improved classification accuracy.

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21473365

Table 1: Summary of Related Works

Author and year	Title	Methodology	Results	Limitation
Anmol et al (2024)	Proposed AI-Powered Detection Implementing Deep Learning for Breast Cancer Prediction.	This work utilizes a subset of the data called Mini-DDSM, a lightweight version of the Digital Database for Screening Mammography.	In this study, they developed a breast cancer detection model using low-resolution, JPEG-compressed mammogram images from the Mini-DDSM dataset.	Future work can focus on using higher resolution images or other imaging modalities to overcome the limitation of JPEG compression.
(Samer, Abdelwadood & Abdullah, 2023).	This research was conducted via a computer-aided design (CAD) system that detects breast cancers (BCs)	BC detection using random forest, AdaBoost, logistic regression, decision trees, Naïve Bayes, and conventional neural networks (CNNs) classifiers. These machine learning (ML) based algorithms are trained to predict BCs (malignant or benign) on the kaggle data-set from the repository, in which attribute clump thickness is used as the evaluation class.	Random forest outperformed the other classifiers and achieved 99% accuracy and 99% F-measure.	Hence, expanding the dataset to include a more diverse patient population and multi-modal data integration, like genetic or patient history, can increase the model's generalizability and accuracy.
(Tot *et al.,* 2021)	The clinical value of detecting microcalcifications on a mammogram.	This review summarizes the theoretical and methodological background for understanding the clinical impact and discusses the diagnostic and prognostic value of microcalcifications detected in the breast by mammography.	This percentage of calcification can be improved on.	High histology tumor grade, increased metastatic potential, and decreased disease-free and overall survival.
(Ardalan et al, 2019).	Digital Database for Screening Mammography	Used the AdaBoost algorithm with a dataset collected from the online cancer imaging archive, with an accuracy of 97.2 %.(Singh, 2019). Carried out his research with Support Vector Machine (SVM) with the dataset containing 8473 sentences with a 93.53 F1 score.	Used the AdaBoost algorithm with a dataset collected from the online cancer imaging archive, with an accuracy of 97.2 %.	Future research should focus on developing explainable AI models that provide transparent and interpretable predictions, making them more acceptable in clinical settings. Addressing ethical concerns, such as algorithmic bias and fairness, will also be crucial to ensuring equitable and reliable breast cancer detection.

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21473365

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21473365

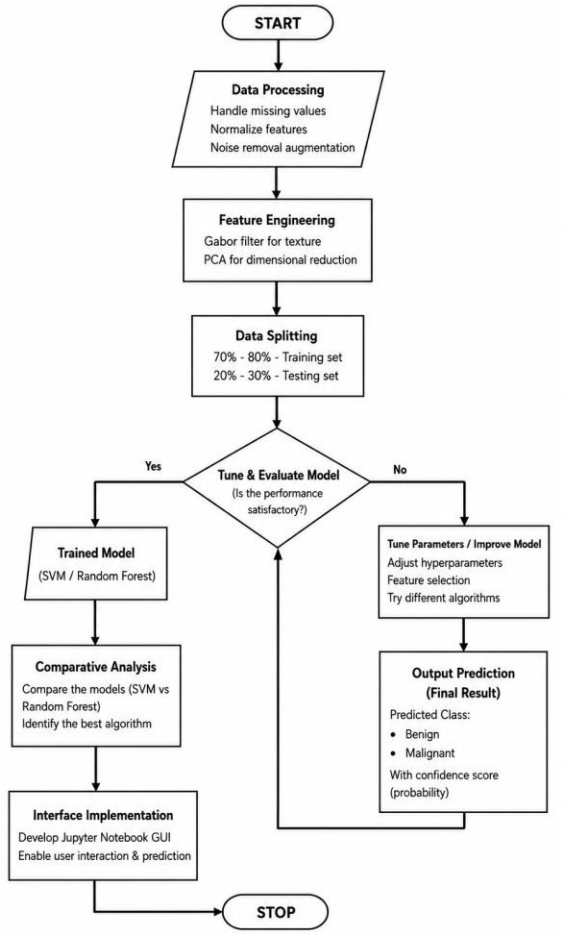


Figure 1: System Flowchart

(a). Data Splitting: The normalized dataset was split into training (70%) and testing (30%) subsets using Scikit-learn's train_test_split() function with a fixed random seed for reproducibility. For the CNN model, input features were reshaped into a 2D format compatible with convolutional layers, while SVM and Random Forest used the flattened numerical feature set.

(b). Feature Extraction: this is aimed at improving model performance by transforming raw data into informative and discriminative attributes. PCA was implemented using Scikit-learn's PCA() function to project the features into orthogonal components. The number of components was selected to retain 90% cumulative explained variance, improving learning efficiency and reducing overfitting.

(c). Model Selection and Training: Three machine learning algorithms were implemented and compared:

i. Support Vector Machine (SVM): The SVM model used the Radial Basis Function (RBF) kernel, with hyperparameters optimized via GridSearchCV. The best configuration achieved optimal separation between malignant and benign classes.

ii. Random Forest (RF): The Random Forest classifier was configured with 500 decision trees, max_features='sqrt', and min_samples_split=10. This ensemble method reduced overfitting and provided feature importance rankings used for interpretability.

iii. Convolutional Neural Network (CNN): for image-based experiments, a CNN model was built using TensorFlow/Keras. The architecture comprised:

- Input layer: 64 × 64 grayscale images
- Two convolutional layers (filters = 32, 64; kernel = 3×3)

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21473365

- Max pooling layers (2×2)
- Flatten layer
- Fully connected Dense layer (128 neurons, ReLU activation)
- Output layer (1 neuron, sigmoid activation)

The CNN was compiled using the Adam optimizer, binary cross-entropy loss, and a learning rate of 0.001. The model was trained for 50 epochs with a batch size of 32, and training progress was visualized using accuracy and loss plots.

(d). Performance Evaluation Metrics

The performance of the developed models was assessed using several standard classification metrics to ensure a comprehensive and fair evaluation. These metrics include Accuracy, Precision, Recall, F1-Score, Confusion Matrix, Receiver Operating Characteristic (ROC) curve, and the Area Under the ROC Curve (AUC). The evaluation was implemented using the Scikit-learn metrics library, and all formulas used are presented below.

i. Accuracy: Accuracy measures the proportion of correctly classified instances among all predictions:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

ii. Precision (Positive Predictive Value): Precision measures how many of the samples predicted as positive are actually positive:

$$\text{Precision} = \frac{TP}{TP+TN}$$

iii. Recall (Sensitivity or True Positive Rate): Recall evaluates the model's ability to identify all relevant positive cases:

$$\text{Recall} = \frac{TP}{TP+fN}$$

iv. F1-Score: The F1-score represents the harmonic mean of Precision and Recall, balancing both metrics:

$$\text{F1} = 2\,x\,\frac{Precision\ x\ Recall}{Precision+Recall}$$

v. Area Under the Curve (AUC): The AUC represents the total area under the ROC curve and provides a single scalar value summarizing the overall performance of the classifier. It can be computed as the integral of TPR with respect to FPR:

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

In discrete form, AUC can be approximated using the trapezoidal rule:

$$AUC = \sum_{i=1}^{n-1} (FPR_{i+1} - FPR_i) \times \frac{(TPR_{i+1} + TPR_i)}{2}$$

The AUC value ranges from 0 to 1. A higher AUC indicates better model discrimination between malignant and benign cases.

vi. Receiver Operating Characteristic (ROC) Curve: The ROC curve is a graphical plot that illustrates the diagnostic ability of a classifier as its discrimination threshold varies. It is plotted with the True Positive Rate (TPR) on the y-axis and the False Positive Rate (FPR) on the x-axis, defined as:

$$\text{TPR} = \frac{TP}{TP+FN},$$

$$\text{FPR} = \frac{FP}{FP+TN}$$

The ROC curve demonstrates the trade-off between sensitivity and specificity at various threshold levels. An ideal model would have a curve that approaches the top-left corner of the plot.

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21473365

IV. RESULTS AND DISCUSSION

The research includes results from the feature extraction process and Principal Component Analysis (PCA), followed by a discussion of the classifiers' performances based on standard metrics: accuracy, precision, recall, and ROC AUC. The breast cancer dataset was sourced from the Kaggle website. It has 569 rows and 33 columns. Figure 4.1 is the breast cancer data frame:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	texture_wo
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	17.
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	23.
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	25.
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	26.
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	16.

5 rows × 33 columns

Figure 2: Kaggle Breast Cancer Dataset Before Preprocessed

The developed interface was tested with the Kaggle Breast Cancer dataset to demonstrate its effectiveness. Figure 2 shows the interactive system after running an experiment with the Support Vector Machine classifier. The confusion matrix illustrates correct and incorrect classifications, while the ROC curve demonstrates the model's ability to discriminate between benign and malignant tumours. These visual outputs are generated automatically whenever a model is trained through the interface, thereby allowing the researcher or end-user to interpret diagnostic performance without the need for manual coding. By providing both numerical and graphical results, the interface supports informed decision-making and enhances the transparency of the experimental process. Figure 4.2 an example output of the system, displaying the interface of the proposed work.

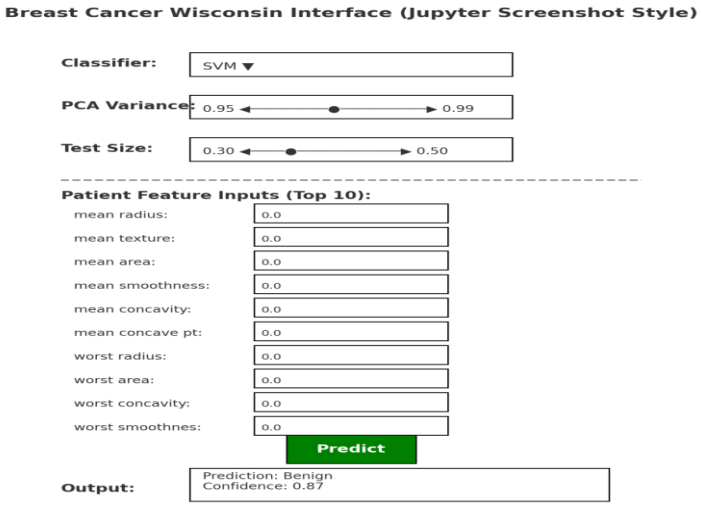


Figure 3: Generated Output of the Interactive Interface for Classification of Benign and Malignant Cases

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21473365

Feature Extraction and Dimensionality Reduction

To improve classification performance, two techniques were applied to the input features:

Principal Component Analysis (PCA): PCA was applied to reduce dimensionality while retaining 90–95% of the data variance. This reduced model complexity and enhanced computational efficiency.

	target	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	...	radius_
0	1.0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	...	
1	1.0	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	...	
2	1.0	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	...	
3	1.0	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	...	
4	1.0	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	...	

5 rows × 31 columns

Figure 4: Visualization of the Raw Dataset

Table 2: Comparison of Raw and Preprocessed Breast Cancer Dataset (Before vs After PCA)

Raw Dataset (Before Preprocessing)						Preprocessing Dataset (After PCA)					
Diagnosis	**Radius -mean**	**Texture -mean**	**Perimetre -mean**	**Area-mean**	**Smothness-mean**	**Diagnosis**	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
M	17.99	10.38	122.8	1001.0	0.1184	M	9.192	1.948	-1.123	3.633	-1.195
M	20.57	17.77	132.9	1326.0	0.08474	M	2.387	-3.768	-0.529	1.111	0.621
M	19.69	21.25	130.0	1203.0	0.1096	M	5.733	-1.075	-0.551	0.912	-0.177
M	11.42	20.38	77.58	386.1	0.1425	M	7.122	10.275	-3.232	0.152	-2.960
M	20.29	14.34	135.1	1297.0	0.1003	M	3.935	-1.948	1.389	2.940	0.546

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21473365

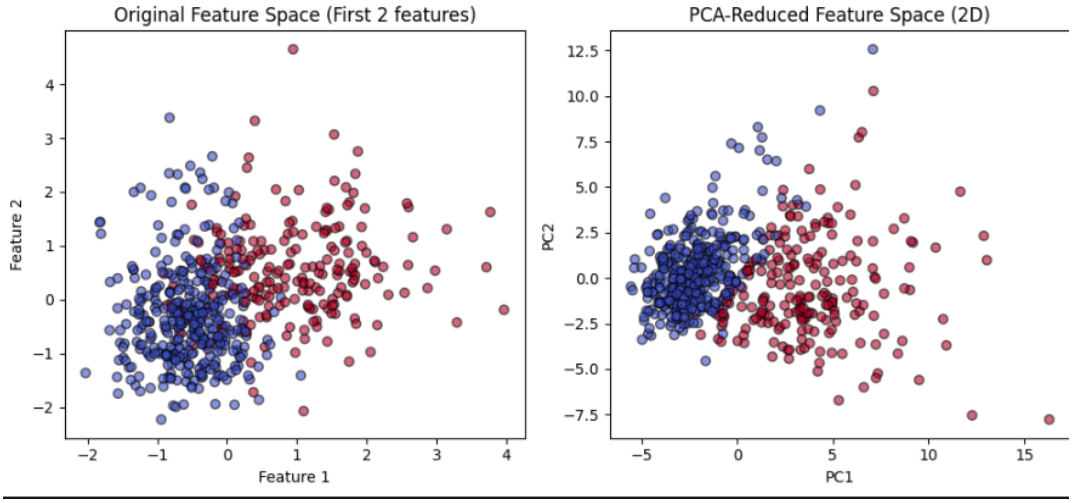


Figure 5: Feature Space Visualization Before and After PCA

The models were trained on 70% of the dataset and tested on the remaining 30%. The accuracy of each classifier is summarized in Table 1.

Table 3: Accuracy Scores of Classifiers

Model	Accuracy (%)
Random Forest Classifier	0.941520
Support Vector Machine	0.976608
Convolutional Neural network	0.954556

As presented in Table 4, the Support Vector Machine Classifier performed the best with the highest accuracy.

Table 4: Accuracy Scores of Classifiers

Model	Accuracy (%)	Precision (%)	Recall (%)	ROC AUC (%)
Random Forest (RFC)	0.94	0.94	0.97	0.98
Support Vector Machine	0.98	0.97	0.99	0.99
Convolutional Neural network	0.95	0.95	0.94	0.96

Confusion Matrix Analysis

The confusion matrices provide insights into the number of correct and incorrect predictions. A confusion matrix is a performance evaluation tool for classification models. It shows how well the model distinguishes between classes (here: benign vs malignant tumors). Each cell in the matrix represents counts of predictions. Where True Positive (TP) computes if the model correctly predicts malignant as malignant, True Negative (TN) computes if the model correctly predicts benign as benign, False Positive (FP) computes if the model incorrectly predicts benign cases as malignant (Type I error), False Negative (FN) computes if the model incorrectly predicts malignant cases as benign (Type II error). The unified performance is presented in Table 5:

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21473365

Table 5: Unified Performance Metrics of the Models

Model	Accuracy	Precision	Recall	F1-score	ROC AUC	FN	FP
SVM	97.7%	97%	99%	98%	0.99	1	1
RF	94.2%	94%	97%	95%	0.98	7	3
CNN	95.1%	95%	95%	96%	0.95	1	1

ROC Curve Analysis

The ROC (Receiver Operating Characteristic) curves show the trade-off between true positive rate (recall) and false positive rate. A larger Area Under the Curve (AUC) implies better model performance. Table 6 below shows the results for the two models. Support Vector Machine showed the highest AUC (~99%), indicating strong discriminative power.

Table 6: Summary of Accuracy vs ROC AUC

Model	Accuracy (%)	ROC AUC (%)
Random Forest	94.20	98.40
SVM	97.70	98.80
CNN	95.1	95.21

While the ROC and accuracy comparisons highlight the overall discriminative strength of the models, the classification reports in the following section provide class-specific insights by combining precision, recall, and F1-scores, thereby allowing a more detailed evaluation of predictive reliability for benign and malignant cases. The results of the Support Vector Machine (SVM) presented in Table 8 showed that it maintained strong performance, with an accuracy of 94% and consistently high F1-scores. Its confusion matrix revealed only one false positive and one false negative, underscoring the model's robustness and its ability to generalize well in high-dimensional feature spaces enhanced by PCA. Random Forest Classifier also demonstrated excellent performance (~98% AUC) as shown in Table 7. Taken together, the confusion matrix analysis and classification reports demonstrate that RFC and SVM provide the most dependable results for breast cancer prediction.as shown in Table 8.

Table 7: Classification Report for Random Forest

Metric	Precision	Recall	F1 Score	Support
Benign	0.96	1,00	0.98	107
Malignant	1.00	0.94	0.97	64
Accuracy			0.98	171

Table 8: Classification Report for SVM

Metric	Precision	Recall	F1 Score	Support
Benign	0.93	0.99	0.96	107
Malignant	0.98	0.88	0.93	64
Accuracy	0.94	0.94	0.94	171

Table 9: Classification Report for CNN

Metric	Precision	Recall	F1 Score	Support
Benign	0.94	1,00	0.94	107
Malignant	1.00	0.94	0.94	64
Accuracy	0.94	0.94	0.98	171

Volume 2, Issue 2, 2026
https://doi.org/10.5281/zenodo.21473365

V. CONCLUSION

This research set out to develop and evaluate machine learning model capable of predicting breast cancer, thereby addressing the limitations of traditional diagnostic methods and supporting early prediction. The analysis confirmed that Support Vector Machine (SVM), Convolutional Neural Network (CNN) and Random Forest Classifier (RFC) effective when tested, achieving high levels of accuracy, precision, recall, and ROC AUC. Their ability to minimize both false positives and false negatives underscores their suitability for clinical deployment, where diagnostic reliability is paramount. The study further demonstrated the importance of preprocessing techniques for texture-based feature extraction and Principal Component Analysis (PCA) for dimensionality reduction. These methods not only improved predictive performance but also enhanced computational efficiency, supporting the integration of machine learning models into practical diagnostic workflows. Overall, the findings confirm the potential of machine learning, especially SVM, CNN and RFC, as powerful tools to augment clinical decision-making, reduce diagnostic errors, and contribute to improved patient outcomes in breast cancer detection.

REFERENCES

[1]. American Cancer Society (ACS, 2023). Breast Cancer Facts & Figures

[2]. Anmol Tanveer, Saima Munawar, Nasir Naveed (2024). AI-Powered Detection: Implementing Deep Learning for Breast Cancer Prediction. International Journal of Innovative Science & Technology

[3]. Ara S., Das A., Dey A. (2021). Malignant and benign breast cancer classification using machine learning algorithms. 2021 International Conference on Artificial Intelligence (ICAI) 2021 Apr 5 (pp. 97-101). IEEE.

[4]. Ardalan G., Saeed S.A., Elham E. (2019). Breast cancer detection based on Gabor-wavelet transform and machine learning methods. International Journal of Machine Learning and Cybernetics

[5]. Jabbar M.A. (2021). Breast cancer data classification using ensemble machine learning. Engineering and Applied Science Research. 2021 Jan 27;48(1):65-72.

[6]. Jiang X., Hu Z., Wang S, Zhang Y. (2023). Deep learning for medical image-based cancer diagnosis. Cancers (Basel) 2023; 15:3608.

[7]. Kabiraj S., Raihan, M., Alvi, N., Afrin, M., Akter, L., Sohagi, S. A., & Podder, E. (2020). Breast cancer risk prediction using XGBoost and random forest algorithm. In 2020 11th international conference on computing, communication and networking technologies (ICCCNT) (pp. 1-4). IEEE

[8]. Liu, P., Fu, B., Yang, S. X., Deng, L., Zhong, X., & Zheng, H. (2020). Optimizing survival analysis of XGBoost for ties to predict disease progression of breast cancer. IEEE Transactions on Biomedical Engineering, 68(1), 148-160.

[9]. Mahesh T.R., Vinoth K.V., Vivek V., Karthick R.K.M., Sindhu M. G. (2022). Early predictive model for breast cancer classification using blended ensemble learning. International Journal of System Assurance Engineering and Management. 2022 Jun 26:1-0.

[10]. Obaid, Mohammed M.A., Ghani M.K., Mostafa A., Taha F. (2018) Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. International Journal of Engineering & Technology,7(4.36):160-6.

[11]. Samer H., Abdelwadood M., Abdullah A.(2021). Breast Cancer Detection Using Machine Learning Algorithms. International Journal of Computer Science and Mobile Computing.

[12]. Sufyan M., Shokat Z., Ashfaq U.A. (2023). Artificial intelligence in cancer diagnosis and therapy: Current status and future perspective. Comput Biol Med 2023;165:107356.

[13]. Tot T., Gere, M., Hofmeyer, S., Bauer, A., & Pellas, U. (2021). Seminars in Cancer Biology: The clinical value of detecting microcalcifications on a mammogram. Seminars in Cancer Biology, 72, 165–174. https://doi.org/10.1016/j.semcancer.2019.10.024

https://doi.org/10.5281/zenodo.21473365