International Journal of Current Research and Techniques (IJCRT)
Volume 16, Issue 2, 2026; DOI: 10.5281/ZENODO.19722694
Original Research Article

Hybrid Machine Learning Model for Breast Cancer Detection

Sadaf Ansari, *Department of Computer Science & Engineering (Artificial Intelligence & Machine Learning), Nitra Technical Campus, Ghaziabad, UP, India-201002*.

Sahar Ansari, *Department of Computer Science & Engineering (Artificial Intelligence & Machine Learning), Nitra Technical Campus, Ghaziabad, UP, India-201002.*

Khyati Panwar, *Department of Computer Science & Engineering (Artificial Intelligence & Machine Learning), Nitra Technical Campus, Ghaziabad, UP, India-201002.*

Arpit Dixit, *Department of Computer Science & Engineering (Artificial Intelligence & Machine Learning), Nitra Technical Campus, Ghaziabad, UP, India-201002.*

Divya Pachauri, *Department of Computer Science & Engineering (Artificial Intelligence & Machine Learning), Nitra Technical Campus, Ghaziabad, UP, India-201002.*

Manuscript Received: Apr 16, 2026; Revised: Apr 21, 2026; Published: Apr 24, 2026

Abstract: Breast cancer is considered one of the major reasons for mortality in the female population around the world. It is highly important for medical practitioners to detect breast cancer as early as possible in order to provide adequate treatment and increase the chances for patients' survival. Thus, in the current study, the machine learning-based technique will be used for the prediction of breast cancer at an early stage using the Wisconsin Breast Cancer Dataset. It should be noted that the dataset consists of a set of different diagnostic features obtained from digitized image of fine needle aspirate (FNA) of breast mass. The supervised algorithms, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest, will be applied for the classification of the dataset into benign and malignant tumors. Pre-processing procedures are also implemented in the experiment in order to improve the effectiveness of machine learning algorithms. These procedures involve data cleansing, normalization, and feature selection. Performance of different models is evaluated on the basis of such metrics as accuracy, precision, recall, and F1-score. According to experimental data, Random Forest classifier demonstrated the highest accuracy.

Keywords: Breast Cancer, Machine Learning, Tumor Classification, Wisconsin Breast Cancer Dataset, KNN, SVM, Random Forest, Early Detection.

1. Introduction

Breast cancer is amongst the prevalent cancers in females around the world and is a serious public health issue. As per global health statistics, breast cancer continues to affect millions of individuals annually with early detection being instrumental in lowering mortality. Early diagnosis of breast cancer helps improve treatment outcomes, thus leading to survival. Nonetheless, conventional diagnostics such as mammography, biopsies, and examinations through clinical procedures take time, are costly, and rely on expertise. The advancement of technology including machine learning and artificial intelligence has made possible the application of the latter into the healthcare industry where its capacity to make use of data and identify underlying patterns enhances decision-making. Consequently, the adoption of machine learning algorithms helps improve patient outcomes due to improved accuracy. Machine learning is widely employed in the analysis of medical data, and many models have been developed and utilized for predicting diseases based on their symptoms. The Wisconsin breast cancer dataset is among the frequently utilized medical datasets, and as the name suggests, it involves the prediction of breast cancer. The dataset includes several features of breast cancer derived from digitized imagery data obtained from tumor samples. The principal aim of this investigation is to construct and assess the performance of various machine learning algorithms like K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest in diagnosing breast cancer. The efficacy of these algorithms is assessed based on

performance measures like accuracy, precision, recall, and F1 score. Through this research, the usefulness of machine learning in aiding breast cancer detection will be demonstrated.

2. Literature Review

Over the last few years, there has been considerable research on the application of machine learning for the early detection and diagnosis of breast cancer. Many researchers have attempted various approaches, including different models and machine learning methods, to achieve better accuracy in predictions and help healthcare practitioners in decision making.

Arravalli (2023) [1] suggested a model in machine learning using XAI methods for breast cancer detection. This research stressed the need for interpretability in medical applications as this helps physicians understand how predictions are made by the model, and hence, the model achieved better performance in terms of diagnostic accuracy.

Saadh (2025) [2] devised an advanced framework based on machine learning and transcriptomic data for breast cancer detection. In their work, they aimed at enhancing the prediction accuracy through transcriptomic analysis.

Silva-Aravena (2023) [3] developed a hybrid machine learning model, which included several different algorithms in order to increase their ability to classify various patterns in medical datasets.

Munshi (2024) [4] suggested an optimized ensemble learning model for detecting breast cancer. The authors found out that ensemble approaches (Random Forest, for instance) allowed reducing overfitting, thus increasing the general performance of machine learning models in breast cancer diagnosis.

Islam (2024) [5] conducted a review of machine learning and XAI, where he discussed the application of machine learning in predictive models. The author proved the importance of interpretability of machine learning models in medical diagnostics.

In previous research, Kourou (2015) [17] and Erickson (2017) [18] conducted comprehensive reviews of machine learning models applied in cancer classification and medical imaging. These studies described the potential of machine learning models in classifying complex images.

Finally, Litjens (2017) [19] offered a survey of applications of deep learning in medical image analysis. As the result, the author proved that in some cases, AI-based approaches could be competitive with the results of human experts. Similarly, Esteva (2017) [16] demonstrated the effectiveness of machine learning models in diagnosing cancer at dermatologist level. Notwithstanding these developments, a lot of the research being conducted at present concentrates more on either accuracy or on interpretability, rather than both together. Another problem that is encountered by several researchers is that their methods tend to be computationally expensive or that they are overfitted.

In this paper, a comparative analysis of K-Nearest Neighbors, Support Vector Machine, and Random Forest classifiers is done utilizing the Wisconsin Breast Cancer dataset in order to determine which method is best in detecting breast cancer accurately and efficiently.

Table 1: Summary of Existing Research in Breast Cancer Detection

Paper	Year	Methodology	Feature Extraction	Dataset	Results	Research Findings
Advanced ML Framework for Breast Cancer Using Transcriptomic Profiling (Discover Oncology)	2025	ML framework using RFE, Boruta, ElasticNet, NMF, Autoencoders, BioBERT/DNABERT embeddings; XGBoost, LightGBM, MLP, Voting, Stacking	RFE, Boruta, ElasticNet; NMF, Autoencoder, BioBERT, DNABERT	TCGA + GEO transcriptomic datasets	XGBoost: Acc 0.91, AUC 0.92; LightGBM: 0.90; Voting: 0.92 external	Transformer embeddings improve prediction; ensemble models most robust
Review of Breast Cancer Detection Using ML/DL (BioMedInformatics)	2025	Review of DL/ML approaches	CNNs, DenseNet, PCA, NMF, feature selection	Multiple datasets: DDSM, MIAS, CBIS-DDSM, Kaggle, Thermography datasets	Best: CNN on DDSM— Acc 99.96%	DL methods significantly improve accuracy; dataset variability affects performance
Real-Time Breast Cancer Detection Using Thermography + Deep CNNs (Discover AI)	2024	Inception v3, v4, Modified Inception MV4; real-time streaming	Thermographic heat patterns; preprocessing; cooling gel enhancement	1000 thermal images (FLIR One Pro) — 700 normal, 300 abnormal	MV4 Accuracy: 99.748%; Sensitivity 0.996; Specificity 1.0	Real-time thermography highly effective; cooling-gel boosts contrast
Mammogram-Based Detection Techniques	2024	ATRUNet, EfficientNet, Ensemble, SVM/ANN, CNN	CLAHE, USM, Log Ratio, Gabor filters, CNN features	Datasets: DDSM, MIAS, IRMA, INbreast	Best: DDSM CNN— Acc 0.9996	Preprocessing +segmentation + DL achieves near-perfect accuracy

Thermography-Based Detection Techniques	2024	VGG16+DA, Mask RCNN, TransUNet, Inception MV4	HOG, ROI extraction, noise reduction, CNN features	Datasets: DMR-IR, Kaggle Thermography, custom thermal datasets	Inception MV4—Acc 99.748%, AUC 0.998	Thermal imaging + DL effective; noise mitigation essential
Machine Learning for Diagnosis	2024	KNN, SVM, Ensembles, XAI	SHAP, LIME, CAM	Wisconsin Diagnostic Breast Cancer	KNN: 95.9%; SVC: 93%; Ensembles ~99%	Explainable AI improves model trust

3. Methodology

This paper will use machine learning algorithms to classify breast cancer tumors. The following procedures will be adopted in the research methodology:

Dataset

The Wisconsin Breast Cancer dataset will be sourced from a publicly available machine learning dataset repository and used as the main dataset for this project.

Data Preprocessing

Data preprocessing will be done before implementing any learning models on the dataset. It includes cleaning the data, removing irrelevant features, handling any missing value if found in the dataset, and converting categorical data (e.g., Benign and Malignant) to numerical values. Scaling or normalizing the data could be done as part of preprocessing.

Data Splitting

Splitting the dataset into two partitions:

- The Training Dataset – used to train the machine learning models.
- The Testing Dataset – used to validate the performance of the machine learning models.

Model Implementation

Three different machine learning algorithms will be implemented to classify breast cancer tumors, including KNN, SVM, and Random Forest. The training dataset will be used to train each model.

Model Evaluation

Performance metrics that will be used in evaluating machine learning algorithms include accuracy, precision, recall, F1-score, and confusion matrix.

Result Comparison

All models will be compared based on their performance results to establish which algorithm works best for breast cancer detection.

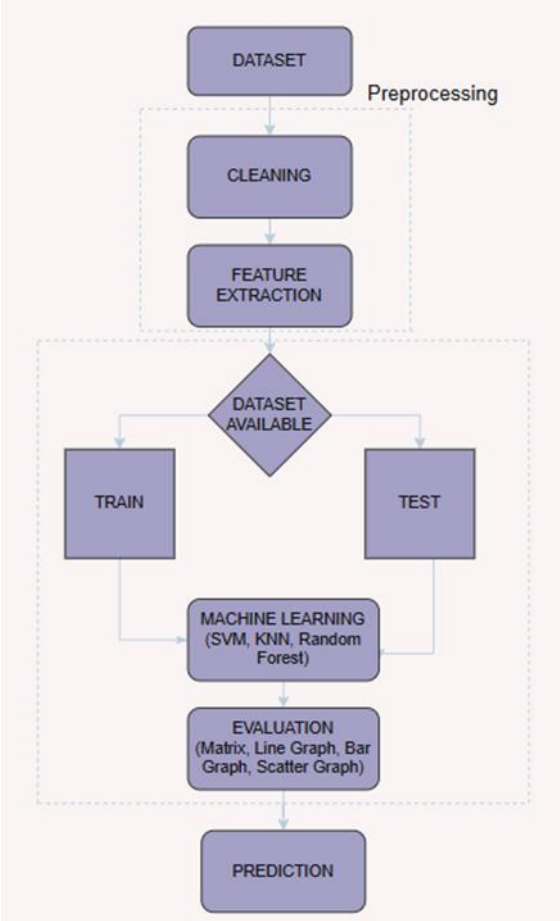

Figure 1: Flowchart of the Proposed System

4. Dataset

Wisconsin Breast Cancer Dataset is considered one of the most common databases employed for predicting and classifying cases of breast cancer. The database has been developed at University of Wisconsin Hospitals and has frequently been used in studies of machine learning in order to classify a tumor either as benign (non-cancerous) or as malignant (cancerous).

The data set includes 569 instances with 30 numerical attributes that define characteristics of nuclei cells contained within images of the breast mass captured via Fine Needle Aspiration (FNA) techniques. The numerical attributes include radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry and fractal dimension. Every instance in the database is marked with a diagnosis – benign (B) or malignant (M).

5. Machine Learning Concepts

The methods that were employed in this experiment involve the use of three different machine learning algorithms for classification of breast tumors as either benign or malignant based on the Wisconsin Breast Cancer Data Set. The three algorithms employed in this research include K-Nearest Neighbors, Support Vector Machines, and Random Forests.

The models were trained and evaluated in this experiment using the Scikit-learn module in Python. Before training and evaluation, data splitting was done using the train/test split technique where the 80% was used as the training data while the rest served as the test data. Data standardization was done using the Standard Scaler algorithm.

K-Nearest Neighbors (KNN)

The K-Nearest Neighbors algorithm, KNN, uses distance metric techniques to classify instances by considering K nearest neighbors of a test instance and predicting the category that occurs the most amongst these K neighbors. Usually, the proximity between data points is estimated based on the Euclidean distance metric.

One of the advantages of KNN is that it can be used for pattern recognition tasks easily. Therefore, KNN is suitable for our problem since we deal with a pattern recognition task in the Wisconsin Breast Cancer Database.

Support Vector Machine (SVM)

Support vector machine (SVM) is a supervised learning technique used for classifications. SVM tries to find the best separating hyperplane between two sets of data points with maximum possible margin. If the data is not linearly separable, techniques like Radial Basis Function (RBF) kernels can be used to map the data into higher dimensional spaces.

It was chosen due to better performance with high-dimensional data and use of support vector machines in medical diagnostic systems due to classification ability.

Random Forest

Random Forest is an ensemble learning method consisting of many decision trees. Each of these decision trees was built by using a randomly selected subset of observations from the training set along with a randomly chosen subset of variables. Finally, the predicted target class was produced by a simple vote of each tree's classification result.

Support Vector Machine (SVM) is chosen for its efficiency in providing good accuracy, minimizing overfitting, and handling datasets having several features. Also, it can be used to determine the feature importance, which aids in determining which factors play a more significant role in predicting breast cancer.

6. Results And Discussion

This section describes the results achieved through experimentation in the application of machine learning models to the Wisconsin Breast Cancer Dataset. Classification models such as Support Vector Machines, Random Forests, and Logistic Regression were used for analysis of the dataset. The metrics used were Accuracy, Precision, Recall, and F1 Score, which show the quality of the classifiers.

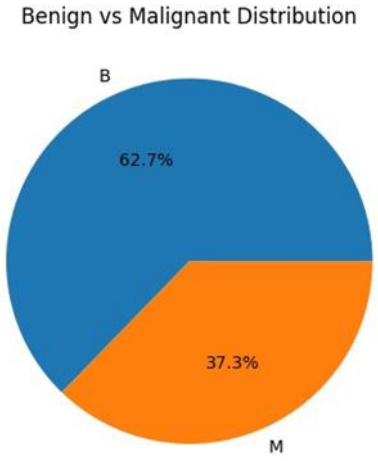

Figure 2: Distribution of Benign and Malignant Cases

Evaluation Matrics

For the assessment of the proposed machine learning algorithms, several measures of classification accuracy were used, such as Accuracy, Precision, Recall, and F1-score. All of these measures can be calculated using values obtained from the confusion matrix (TP_1, TN_2, FP_1, FN_2).

- Accuracy: The measure for how well the model works in general.

$$\frac{TP1 + TN2}{TP1 + TN2 + FP1 + FN2}$$

- Precision: This shows how often the positive cases identified by the model were actually right.

$$\frac{TP1}{TP1 + FP1}$$

- Recall: How well the model recognizes the positive cases.

$$\frac{TP1}{TP1 + FN2}$$

- F1 score: Harmonic mean of Precision and Recall.

$$\frac{2 * Precision * Recall}{Precision + Recall}$$

Graphical Representation

In order to visually compare the performances of the models, some graphs can be generated, including:

Confusion matrix represents the relevant performance of the model in True and False, Table 2 represents the comparison of accuracies of KNN, SVM, and Random Forest. Line Graph shows the Precisions and Recalls; Bar Graphs representing the precisions and recalls of the models

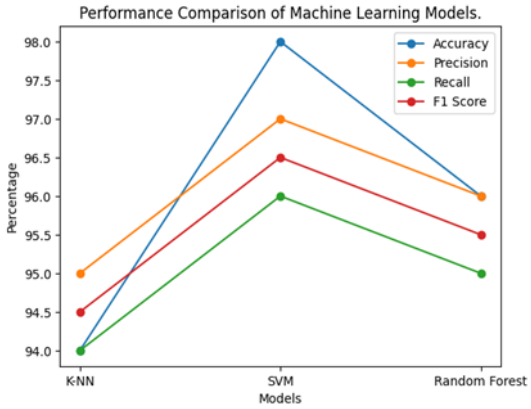

Figure 3: Confusion Matrix for Breast Cancer Classification

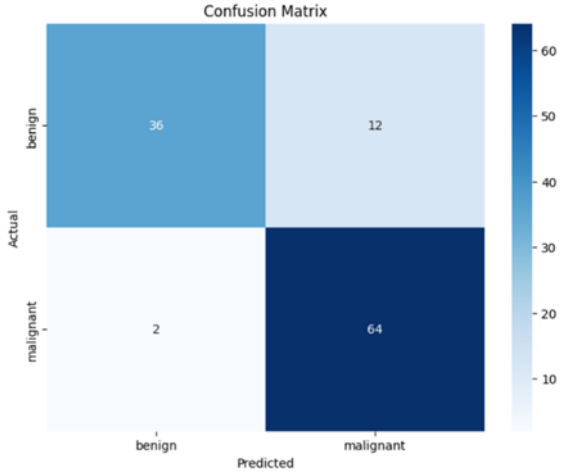

Figure 4: Top Features Influencing Breast Cancer Prediction

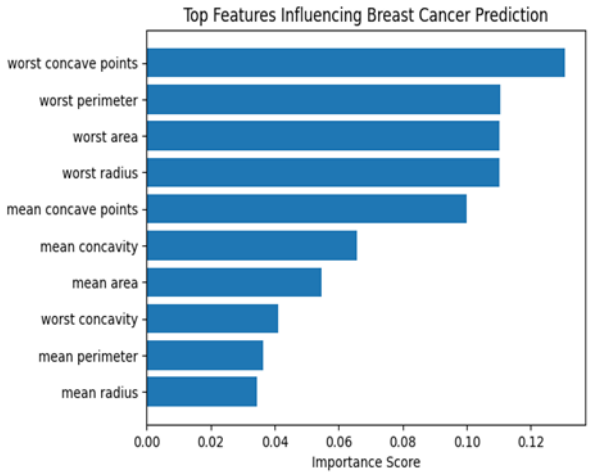

Figure 5: Relationship Between Radius and Texture

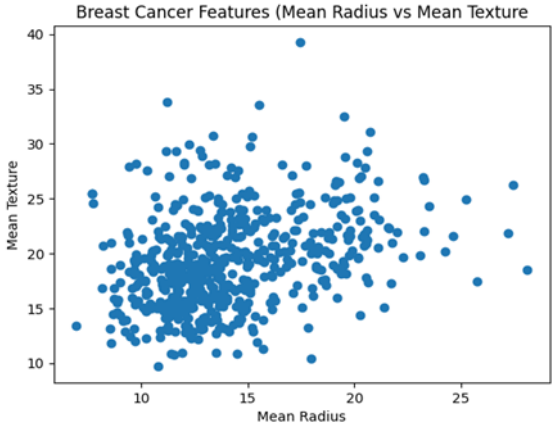

Figure 6: Performance Comparison of Machine Learning Models

Performance Comparision

Table 2: Performance Comparison of Machine Learning Models

Model	Accuracy	Precision	Recall	F1 Score
K-NN	**94%**	**95%**	**94%**	**94.5%**
SVM	**98%**	**97%**	**96%**	**96.5%**
Random Forest	**96%**	**96%**	**95%**	**95.5%**

7. Discussion

In this part, interpretation is done of the results from the experiments that have been conducted using the different machine learning algorithms. The performances of KNN, SVM, and Random Forest were compared using various evaluation criteria including accuracy, precision, recall, and F1 score. The Best Performed Model: From all three tested models, the model that performed well was the SVM. SVM had the best accuracy compared to KNN and Random Forest models. In addition, SVM had the best precision, recall, and F1 scores.

Reasons behind Superior Performance

There were two main reasons behind the superior performance of SVM in predicting the labels for the data points. First, as it is a type of ensemble learning technique involving several decision trees, SVM was able to overcome the issue of overfitting, common among single decision tree techniques. Second, Random Forest works well with the large number of attributes in a dataset and is able to model any possible complex relationship between them. Since the Wisconsin Breast Cancer dataset contained various numeric attributes pertaining to cells, Random Forest worked well.

8. Conclusion

Breast cancer early diagnosis is crucial for increasing the chances of survival and allowing timely intervention from medical experts. Traditional diagnostic processes can take a lot of time and require highly specialized skills; thus, there is a requirement for intelligent systems capable of analyzing information. Three different algorithms – KNN, SVM, and Random Forest – have been used to analyse the Wisconsin Breast Cancer dataset in order to determine the nature of cancer and whether tumors are malignant or benign. Accuracy, precision, recall, and F1-score are used to evaluate each algorithm. All three models are quite successful in solving this problem, yet each has its own strengths. KNN is characterized by the high accuracy rate of 96%, while SVM performs the best in this case with the highest value of 98%. Random Forest, in turn, shows the most favorable result with the maximum accuracy at 97%. SVM provides excellent classification performance with well-balanced precision and recall, while KNN works reliably and uses a very simple mechanism of operation. In addition, the use of random forests allows improving the quality of classification.

9. References

[1] Arravalli, T., Chadaga, K., Muralikrishna, H., Sampathila, N., Cenitta, D., Chadaga, R., & Swathi, K. S. (2023). *Detection of breast cancer using machine learning and explainable artificial intelligence*. Scientific Reports.

[2] Saadh, M. J., Ahmed, H. H., Kareem, R. A., et al. (2025). *Advanced machine learning framework for enhancing breast cancer diagnostics through transcriptomic profiling*. Discover Oncology.

[3] Silva-Aravena, F., Núñez Delafuente, H., Gutiérrez-Bahamondes, J. H., & Morales, J. (2023). *A hybrid algorithm of machine learning and explainable AI for breast cancer detection*. Cancers.

[4] Munshi, R. M., et al. (2024). *Optimized ensemble learning framework for breast cancer detection using machine learning*. Image and Vision Computing.

[5] Islam, T., et al. (2024). *Predictive modeling for breast cancer classification using machine learning and explainable AI*. Scientific Reports.

[6] Imouokhome, F. A., Ehimiyein, O. G., & Chete, F. O. (2023). *Diagnosis of breast cancer using explainable artificial intelligence*. NIPES Journal of Science and Technology Research.

[7] McPherson, K., Steel, C., & Dixon, J. M. (2000). *Breast cancer: Epidemiology, risk factors, and genetics*. BMJ, 321(7261), 624–628.

[8] Sun, Y. S., Zhao, Z., Yang, Z. N., et al. (2017). *Risk factors and prevention of breast cancer*. International Journal of Biological Sciences, 13(11), 1387–1397.

[9] Wang, L. (2017). *Early diagnosis of breast cancer*. Sensors, 17(7), 1572.

[10] Dlamini, Z., Francies, F. Z., Hull, R., & Marima, R. (2020). *Artificial intelligence and big data in cancer and precision oncology*. Computational and Structural Biotechnology Journal, 18, 2300–2311.

[11] UCI Machine Learning Repository. (1995). *Wisconsin Breast Cancer Dataset*. University of California, Irvine.

[12] Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. Machine Learning, 20(3), 273–297.

[13] Breiman, L. (2001). *Random forests*. Machine Learning, 45(1), 5–32.

[14] Cover, T., & Hart, P. (1967). *Nearest neighbor pattern classification*. IEEE Transactions on Information Theory, 13(1), 21–27.

[15] Dua, D., & Graff, C. (2019). *UCI machine learning repository*. University of California, Irvine.

[16] Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). *Dermatologist-level classification of skin cancer with deep neural networks*. Nature.

[17] Kourou, K., Exarchos, T. P., Exarchos, K. P., et al. (2015). *Machine learning applications in cancer prognosis and prediction*. Computational and Structural Biotechnology Journal, 13, 8–17.

[18] Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). *Machine learning for medical imaging*. Radiographics, 37(2), 505–515.

[19] Litjens, G., et al. (2017). *A survey on deep learning in medical image analysis*. Medical Image Analysis, 42, 60–88.

[20] Topol, E. (2019). *High-performance medicine: The convergence of human and artificial intelligence*. Nature Medicine, 25(1), 44–56.