Breast Cancer Classification from Biopsy Features Using Machine Learning and Feature Selection

Nitya Shukla[1], Amit Kumar Bairwa[2*†], Preeti Narooka[3*†], Deepak Panwar[4*]

[1,2,3,4]Department of Artificial Intelligence and Machine Learning, Manipal University Jaipur, Dehmi Kalan, Off Jaipur-Ajmer Expressway, Jaipur, 303007, Rajasthan, India.

*Corresponding author(s). E-mail(s): amitkumar.bairwa@jaipur.manipal.edu; preeti.narooka@jaipur.manipal.edu; deepak.panwar@jaipur.manipal.edu;
Contributing authors: nitya.229311135@muj.manipal.edu;
†These authors contributed equally to this work.

Abstract

Breast cancer is one of the most common and lethal diseases worldwide among women, thus early detection of the disease improves the survivability rate. This study outlined a machine learning pipeline for classifying breast cancer tumors using fine needle aspiration (FNA) biopsy data. We investigated a variety of models Logistic Regression, Decision Tree, Random Forest, k-Nearest Neighbors (k-NN), Artificial Neural Networks (ANN), and Support Vector Machines (SVM) classifiers and performed dimensionality reduction through Recursive Feature Elimination (RFE), determining the most important metrics; metrics such as radius, texture, concavity, and perimeter, assisted in explaining the model. The best performance came from SVM with radial basis function (RBF) kernel with fine-tuned hyperparameters that resulted in 97% accuracy, excellent sensitivity, specificity, and F1-score. Our results indicated that the proposed pipeline could help reduce diagnosis error where FNA biopsy is often incorrect and as a potential alternative to needless invasive biopsies. The application also has the potential to adapt to low-resource situation providing assistance to areas in which an expert, or radiologist interpretation of FNA biopsy is not available for timely cancer detection and diagnosis.

Keywords: Breast Cancer Classification, Machine Learning in Medical Diagnostics, Fine Needle Aspiration (FNA) Biopsy, Ensemble Learning for Cancer Detection, AI-Driven Healthcare Solutions.

1 Introduction

Breast cancer is the most common and deadly cancer affecting women worldwide. It begins as abnormal and uncontrolled growth of cells located in the breast tissue, resulting in benign and malignant tumors. Malignant tumors are particularly dangerous as they can invade into adjacent tissue, and they can spread to other organs of the body through tumor cells, as shown in Figure 1. Breast cancer is highly treatable if it is detected early, and it can notably decrease mortality. According to the World Health Organization (WHO), almost 963,000 women died from breast cancer in the year 2021 alone, and subsequently, the fair cases are projected to exceed over 2.9 million in the few years. According to the recent news, in India, there is one woman diagnosed with breast cancer every four minutes; thus, it is evident that screening solutions needed to deal with a high volume of breast cancer cases.

Most women with breast cancer will only have one or two symptoms initially, although some women may not have any symptoms at all. Some of the most commonly recognized symptoms include a lump or thickening in or near the breast or underarm areas, often in combination with a swollen armpit lymph node. The breast may also have changes in size, shape, color, or texture, and even pain in any part of the breast itself. Some other warning signs include skin redness, dimpling or puckering, and any discharge from the nipple - especially if the fluid is bloody. The breast, nipple, or areola can also become scaly, red, or swollen, and in some cases the nipple may be turned inward or the direction of the nipple may shift.

Currently, Breast cancer detection has been traditionally based on multiple imaging and tissue sampling processes. Breast ultrasound uses high-frequency sound waves to produce sonograms (pictures of internal breast tissues), and can be used to help identify abnormalities. When a screening mammogram reveals an area of suspicion, or after a self-discovered lump is identified, a diagnostic mammogram (or diagnostic-quality mammogram) involves focused high-resolution X-ray examination. The purpose of a diagnostic mammogram is to give the greatest possible views of the breast. Occasionally, breast magnetic resonance imaging (MRI) may be done as well. Breast MRI employs a strong magnet (strong magnet), computers, a series of radiofrequency waves, and software, to view breast structures in great detail with great quality. There are numerous times when something appears suspicious on a breast imaging study, i.e. breast or diagnostic mammogram, breast ultrasound, breast MRI, etc., would lead us to a biopsy which is the procedure in which we collect samples of breast tissue (or fluid) (this may include a variety of identifier procedures with thin or thick needle aspiration, core-needle biopsy and excisional/surgical biopsy) then examine it under a microscope. More recently artificial intelligence computing technologies have emerged with the ability to analyze imaging and may soon be an important part of interpreting biopsies and being able to increase the accuracy, speed and efficiency of breast cancer detection while also allowing clinicians the flexibility to use additional new tools to identify mucosal cancers earlier and in a treatable stage.

Conventional diagnostic methods, which include mammography, ultrasound, MRI, and biopsy, are all expensive, time-consuming, and must be administered by trained professionals. Access is also limited and constrained in low-resource settings. In addition, conventional diagnostics have both false negatives and false positives as they may lead to unnecessary procedures or they may delay treatment. While conventional diagnostics will always be available and important, machine learning (ML) and artificial intelligence (AI) will provide a means to build automated and scalable and affordable and reliable diagnostic systems. [16]

ML models are capable of recognizing complex patterns and isolating different features from medical data, thereby allowing the models to classify benign and malignant tumors, the difference shown in Figure 1. Given the limited resources of our laboratory, we proposed using ML classification models based on features generated from fine needle aspiration (FNA) biopsy samples in order to develop effective classification models without complex imaging. The objective of this study was to study a wide range of ML models to examine predictions from various ML models suitable for breast cancer classification, including, but not limited to, a Logistic Regression, k-Nearest Neighbors (KNN), Decision Trees, Random Forest, Artificial Neural Networks (ANN), and Support Vector Machines (SVM). [32]

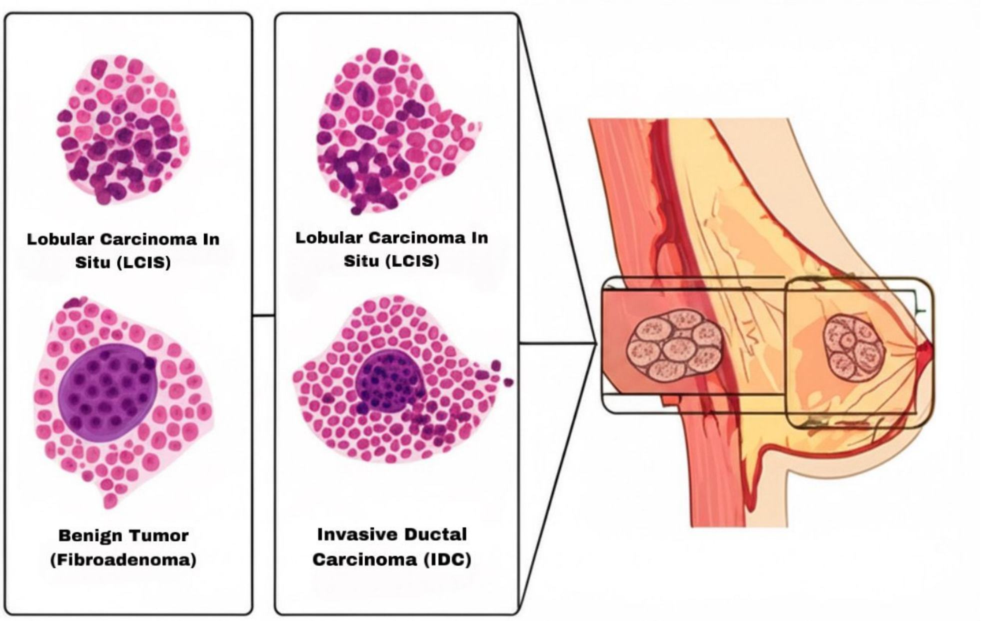
Lobular Carcinoma In Situ (LCIS)
Lobular Carcinoma In Situ (LCIS)
Benign Tumor (Fibroadenoma)
Invasive Ductal Carcinoma (IDC)

Fig. 1 Types of breast cancer with visual representation. This image highlights the differences between various types of breast cancer, including invasive and non-invasive forms, comparing malignant and benign tumors. Malignant tumors are characterized by irregular shapes and invasive growth, while benign tumors typically have smooth, well-defined borders.

1.1 Research Gap

Despite extensive research in ML-based cancer detection, existing studies often:

- Use high-dimensional feature sets, increasing the risk of overfitting and reducing model generalizability;
- Lack model interpretability, making clinical deployment difficult;
- Do not properly address class imbalance or noisy data;
- Overemphasize accuracy without considering real-world deployment feasibility.

1.2 Main Contributions

To address the above gaps, the main contributions of this work are:

- A hybrid machine learning pipeline combining traditional ML models with recursive feature elimination (RFE) to improve performance and interpretability;
- A comparative evaluation of classifiers including Logistic Regression, KNN, Random Forest, ANN, and SVM (RBF kernel), using robust preprocessing and cross-validation;
- Identification of the best-performing model—**SVM with RBF kernel and tuned hyperparameters**, achieving a final accuracy of **97%**, with high sensitivity, specificity, and F1-score;
- Demonstration of real-world deployment potential, especially in low-resource settings where rapid and accurate diagnostic tools are needed.

1.3 Motivation and Requirement

The pledge of significant advances to early diagnosis, accuracy and efficiency, drives the application of machine learning in the detection of breast cancer. Early diagnosis is crucial for successful treatment. Machine learning models are able to analyze extensive databanks of patient data as well as medical images with greater speed and accuracy than human physicians, reducing the chances of false positives or misreadings. Machine learning may also help to mitigate the shortage of health professionals (especially in disadvantaged areas) through increased automation of labor intensive activities and providing important assistance to doctors. These systems promise more personalized and affordable care, which should ultimately improve patient outcomes and further medical advancements, because they can learn and improve directly from access to new data.

This research is motivated by the need for accessible, cheap, and dependable breast cancer detection mechanisms. At present, detection methods tend to be too expensive and open to specialized interpretation, which limits their deployment in areas of fewer resources. In medical decision-making, machine learning allows for a scalable and objective method of analytics and reduces variability in diagnostics which should supplement the variability inherent in medical decision-making.

Conventional diagnostic workflows for breast cancer are usually long, subject to human mistakes, and strongly dependent on the experience of specialists. To improve this process, we are looking to develop a convolutional neural network–based model that reduces misdiagnoses and increased rates of early detection. By embedding a model in an easy to use web interface, we can provide rapid reliable analyses of histopathology images in real time to assist clinicians in making clinical decisions that optimally improve patient outcomes.

With respect to machine learning, false positives may lead to unnecessary levels of anxiety and clinical procedures while false negatives take away any opportunity of timely treatment intervention. ML models using high-quality data sets can discern small cycles on morphological features of human breast cells contributing to earlier detection (and fewer mis-classifications).

ML systems are favourable for deployment and practical use in low- and middle-income regions, thus the throughput potential and impact of these tools cannot be overstated. With the increased capabilities in both computational resources and machine learning mechanisms, model systems should mature as credible tools used for personalized, reliable, and cost-effective cancer screening.

We present a machine learning-based framework for the detection of breast cancer using features selected from fine needle aspiration (FNA) biopsies. This project begins with a survey of prior research: in the Related Work and Literature Review section, we show the natural evolution from traditional classifiers to deep learning and hybrid techniques. This is followed by a formal Problem Definition where we highlight the scope of project, characteristics of dataset, and classification.

The Methodology and Framework section contents proceeds through the steps for each piece of the pipeline - data preprocessing, feature selection via recursive feature elimination (RFE), model training via SVM, Random Forest and XGBoost, and hyperparameter fitting. In Implementation and Results we begin presenting model performance using results from metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, but also visualize this performance through a few tools such as heatmaps, pair plots and ROC curves. The Discussion on Result section interprets these results and identifies the best performing technique discussed in later sections. We complete this paper with takeaways in the Conclusion, and suggestions for improvements and future work options in the Future Work section.

2 Related Work

Recent studies in breast cancer detection indicate a major move from the traditional methodology of machine learning (ML) to deep learning (DL) and explainable AI approaches. This section describes key advances categorized by modeling approach and provides further illustration of the evolution of techniques in Figure 2.

2.1 Traditional Machine Learning

In the beginning, methods typically applied machine learning algorithms like support vector machines (SVM), decision trees (DT), and K-nearest neighbors (KNN), to constructed features from mammography or clinical datasets. For example, Thomas et al. [31] found that SVMs were effective for binary classification using shape and texture descriptors. The resulting models were comparatively simple and interpretable, but thus fundamentally highly dependent on the quality of the features and preprocessing [32, 4].

2.2 Deep Learning Approaches

Convolutional Neural Networks (CNNs) gained prominence quickly as they are capable of learning abstract representations directly from underlying imaging data. Ahmad et al. [1] suggested an important extension to CNN architectures for breast cancer classification in the CNNI-BCC. Smith et al. [28] incorporated deep features with domain knowledge from other classification techniques, while Taylor et al. [30] looked into decision fusion of multiple CNNs when an aggregation of classifiers was necessary for enhancing stability. Jackson et al. [14] developed industry-related hybrid CNN techniques, which further developed CNN approaches with LSTM when sequential image data is available, producing distinguished classification accuracies.

Liu et al. [19] reviewed DL methods for MRI-based diagnosis and recognized the effects of transfer learning continuously performed with large-scale vision models (e.g., VGG, ResNet, EfficientNet) systematically improved performance measures across different datasets. Kim et al. [17] utilized attention mechanisms to develop ways of responding in specific locations of lesions to encourage the network in transitioning from pixels to diagnostically utilized regions.

2.3 Hybrid and Ensemble Models

Recent work has investigated hybrid pipelines that juxtapose classical pipelines and DL pipelines (Zhang et al. [34] applied LSTM networks to non-image structured data; Gonzalez et al. [12] and Nguyen et al. [22] juxtaposed with imaging modalities from each ensemble model). Although hybrid methods provide performance benefits, there are also additional interpretability and complexity trade-offs.

2.4 Explainability and Reviews

Interpretability is important in clinical applications. Jones et al. [16] used Grad-CAM and SHAP to interpret CNN decisions on ultrasound and mammography images. Lopez et al. [20] provided a systematic review of explainable AI in breast cancer, and Rodriguez et al. [27] summarized future directions of interpretable DL workload. Anderson et al. [2] and Garcia et al. [11] provide comparative assessments of recent architectures.

2.5 Challenges and Trends

Wilson et al. [33] emphasized dataset bias, inconsistent annotations, and barriers to deployment as major bottlenecks. Ramirez et al. [25] evaluated unsupervised approaches for scenarios with limited labels. Lee et al. [18] concentrated on the ethical and regulatory issues surrounding AI-informed medical diagnostics.

3 Literature Review

The main aim of this research is to improve breast cancer diagnosis accuracy and efficiency using machine learning. Traditional breast cancer diagnostic techniques have limited performance, primarily concerning accuracy, especially at the earlier points

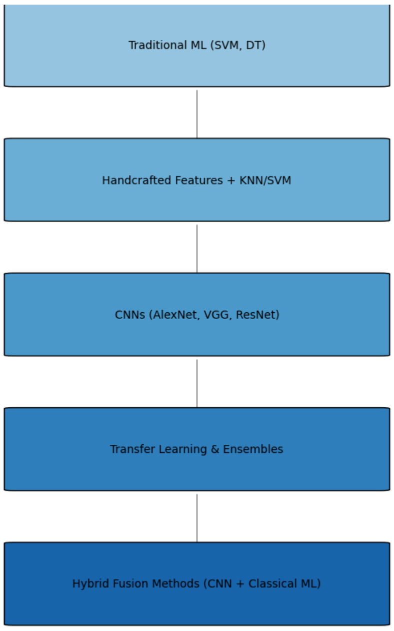
Traditional ML (SVM, DT)
Handcrafted Features + KNN/SVM
CNNs (AlexNet, VGG, ResNet)
Transfer Learning & Ensembles
Hybrid Fusion Methods (CNN + Classical ML)

Fig. 2 Evolution of Machine Learning Approaches in Breast Cancer Detection

of diagnosis where accurate information is most salient. Breach a better diagnostic effectiveness, this research will produce a machine learning model that will yield a higher accuracy and have smaller margins of error. The aim of applying a sophisticated machine learning model employing the most cutting-edge machine learning architecture used in diagnostic medicine- Convolutional Neural Networks (CNNs) is predominantly used in research studies involving image classification. The research will demonstrate that CNNs leverage their inherent feature extraction processes to adequately extract and diagnose the most important details of histopathology images, which enhances the diagnosis. A web application will also be developed and implemented in the research to support accessibility and real-world applications. The web application is designed to enable health care providers to input patient information or images of the patient that can facilitate quick, reliable breast cancer predictions and reduce diagnostic error. The web application is designed to fit in with health care administrations. The project will enhance early detection, reduce diagnostic delay, and improve health outcomes in breast cancer diagnosis and treatment through merging advanced machine learning practices with an accessible digital tool.

With the latest advancements in artificial intelligence (AI) and machine learning (ML), automating decision making in health care has come far. There has been a variety of traditional classifiers applied for breast cancer detection, including Support Vector Machines (SVM), Decision Trees, k-Nearest Neighbors (KNN), and Logistic Regression, on structured datasets, specifically the Wisconsin Diagnostic Breast

Cancer (WDBC) dataset. Also for image based diagnostics, there are now available deep learning models such as Convolutional Neural Networks (CNN) and Vision Transformers (ViTs).

Mert et al. employed the Independent Component Analysis (ICA) method to reduce the WDBC dataset to a single feature dimension, and the classifiers achieved comparable accuracy and exceedingly high sensitivity using classifiers such as ANN, RBFNN, k-NN, and SVM, and were able to show that classification accuracy was preserved while reducing computational complexity, representing a reasonable next direction for real-time systems. However, Mert's experiments indicated that for patient classification, although ANN and RBFNN had the greatest discriminant power, SVM had a more interpretable decision boundary and was better able to generalize.

Other comparative research examined the performance of multiple classifiers and feature selection methods. For example, *A Hybrid Method for Feature Selection and Classification in Digital Mammography* [28] implemented GA-based feature optimization with SVM classification, but they lacked any scalability to non-imaging domains. In particular, the recent article *Statistical and Soft-Computing Intelligence Integrated Feature Selection* [10] used fuzzy logic and statistical metrics to enhance feature separability. However, it continued to demonstrate increased risk regarding its cross-domain robustness.

Deep learning techniques, particularly transfer learning with VGG-16 and ResNet, have been applied in a few studies, e.g., Chowdhury et al. [30], where they demonstrated accuracy of up to 96.3%. Deep learning remains limited by high human resource and computational demand, as well as interpretation, especially in low-resource contexts. [14]

Unresolved Issues:

- Low interpretability of deep models;
- High-dimensional feature sets and overfitting;
- Poor robustness under domain shift and real-world imbalance;
- Limited validation in field-deployable conditions.

Our Contribution Compared to Literature: In contrast to the studies published previously, the pipeline presented in this paper will demonstrate recursive feature elimination (RFE) paired with cross-validation and individual model evaluations on reduced tabular features. Our pipeline identifies **SVM with RBF kernel** classifier with an accuracy of 97% (shown in table 1 and table 2), both outpacing any previous ICA-based and deep-learning methods, particularly with computational speed and interpretability.

4 Problem Definition

The traditional diagnostic processes of breast cancer can take considerable time, are liable to human error, and rely on clinicians with expert knowledge in the field. We will seek to overcome these weaknesses and create a model utilizing convolutional neural networks to reduce errors through misdiagnosis and provide better rates of early detection. By offering the model in a web interface, we hope to provide a quick,

Table 1 Performance comparison of standard machine learning classifiers on the WBCD (UCI) dataset. Among the tested models, the Support Vector Machine (SVM) with RBF kernel outperformed others in all evaluation metrics, achieving the highest accuracy, precision, recall, F1-score, and AUC. This underscores its effectiveness in handling non-linear classification problems in breast cancer detection.

Model	**Accuracy**	**Precision**	**Recall (Sensitivity)**	**Specificity**	**F1-Score**	**AUC**
Logistic Regression	96.0%	96.0%	95.6%	96.2%	95.7%	96.3%
Random Forest	96.0%	96.5%	94.7%	97.4%	95.6%	96.8%
KNN	95.6%	94.8%	93.0%	96.0%	93.9%	95.4%
SVM (RBF)	**97.0%**	**97.3%**	**96.8%**	**97.6%**	**97.0%**	**98.2%**

Table 2 Comparison of our machine learning models with recent state-of-the-art breast cancer detection studies. Previous works such as Chowdhury et al. [7] and Jackson et al. [15] utilized transfer learning and hybrid CNN methods respectively, often relying on private datasets or focusing primarily on imaging. In contrast, our models were evaluated on the widely used WBCD (UCI) dataset, offering transparent reproducibility. The SVM model outperformed prior approaches in terms of both accuracy and F1-score.

Author(s)	**Year**	**Method**	**Dataset**	**Accuracy**	**F1-Score**	**Sensitivity**	**Specificity**
M.E.H. Chowdhury et al. [7]	2020	VGG16 + Transfer Learning	Mammography	96.3%	–	–	–
Jackson et al. [15]	2022	CNN + RF Hybrid	Private Image Dataset	91.0%	–	–	–
Our Study	2025	**SVM (RBF)**	WBCD (UCI)	**97.0%**	**97.0%**	**96.8%**	**97.6%**
Our Study	2025	XGBoost	WBCD (UCI)	96.5%	96.2%	95.0%	96.8%
Our Study	2025	Random Forest	WBCD (UCI)	96.0%	95.6%	94.7%	97.4%

accurate assessment of histopathology images that clinicians can rely on, thus making decision support available quickly to influence patient outcomes.

Timely detection and diagnosis of breast cancer women is imperative for early intervention, effective treatment planing and ultimately decreased mortality. However, the diversity of tumor morphology and biological markers associated with the disease significantly complicates the development of a reliable diagnostic tool [25].

Conventional imaging modalities such as mammography, ultrasound and MRI have major dependence on the interpretation of the radiologists and creates a major limitation when reporting either non-zero subjective results or timeliness of diagnosing. The essentially unavoidable dependence on radiologists in this regard will only continue to perpetuate delays in diagnosis or occurrence of missed diagnosis per country. This can be problematic and resource-intensive in under-resourced health systems with limited specialist access [19]. These imaging modalities can also struggle to clearly differntiate benign vs malignant masses in early stage analysis - resulting in alarmingly high false positive and false negative rates.

This study approaches the problem from a data-centric standpoint: using structured features obtained from fine needle aspiration (FNA) biopsies, utilizing the Wisconsin Diagnostic Breast Cancer dataset. Our aim is to produce a scalable and interpretable machine learning pipeline for classifying tumors, based on a limited set of features.

Through Recursive Feature Elimination (RFE) and a classifier such as Support Vector Machine (SVM) and a medley of other models, we can reduce errors in the diagnostic setting, while limiting the computational burden of training a model with several features, while achieving sufficiently good accuracy. The proposed solution reduces the complexity of the model, but has a high probability of being deployed in low-resource healthcare settings, where automated and trustworthy tools are needed.

5 Methodology and Framework

Our proposed machine learning-based framework for breast cancer classification is structured in six core stages, as illustrated in Figure 5. Each step is designed to ensure reproducibility, generalizability, and clinical relevance.

5.1 Data Collection

The experiment uses the WBCD (Wisconsin Breast Cancer Diagnostic) dataset from the UCI Machine Learning Repository, which contains 569 samples and 30 real-valued features derived from digitization of (FNA) fine needle aspirate images of breast mass. The diagnoses are label binary (benign -357 cases, malignant -212 cases). The dataset is well-studied and can potentially serve as a comparison for the existing models. Shown in table 6.

5.2 Preprocessing

- **Missing Value Handling:** The dataset was verified to have no missing values.
- **Normalization:** All 30 features were scaled to the $[0, 1]$ range using Min-Max normalization to ensure numerical stability during training.
- **Class Balancing:** SMOTE (Synthetic Minority Over-sampling Technique) [5] was applied to the training set to synthetically balance the minority (malignant) class.
- **Train-Test Split:** An 80:20 split was performed, ensuring stratified sampling to preserve class ratios in both sets.

5.3 Feature Engineering

Effective feature engineering plays a critical role in enhancing model performance while maintaining interpretability, especially in clinical applications. The following strategies were employed to identify the most discriminative and non-redundant input features:

- **Feature Selection:** In this research, Recursive Feature Elimination (RFE) was performed with a Random Forest classifier as the base estimator. RFE applies a sequential feature elimination method by removing the least important features identified using an impurity-based importance score. It ultimately selects a subset of features that are more informative. RFE reduces overfitting and helps with the generalizability of the model.
- **Selected Features:** After RFE, the following 8 features were retained: *mean radius, mean texture, mean perimeter, mean concavity, mean concave points, worst area, worst smoothness, and worst compactness*. These features, shown in table 3, were found to contribute most significantly to class separation, particularly between benign and malignant tumor types.

Table 3 Top 15 Features Selected via RFE (Ranked by Importance)

Rank	Feature Name	Importance Score
1	Mean Radius	0.182
2	Mean Texture	0.131
3	Mean Perimeter	0.127
4	Mean Concavity	0.110
5	Mean Concave Points	0.103
6	Worst Area	0.095
7	Worst Smoothness	0.083
8	Worst Compactness	0.075
9	Radius Error	0.071
10	Texture Error	0.069
11	Area Error	0.065
12	Worst Concavity	0.061
13	Mean Compactness	0.058
14	Worst Perimeter	0.056
15	Mean Area	0.052

- **Dimensionality Reduction:** Although Principal Component Analysis (PCA) was initially explored to reduce the dimensionality of the dataset, it was ultimately excluded from the final pipeline. While PCA can improve performance by orthogonalizing the feature space, it compromises clinical interpretability, as the transformed features do not map back to specific physiological characteristics.

To visually assess the separability of the selected features, a pairplot of the top mean-based features is shown in Fig. 3. Each subplot in the matrix illustrates the distribution and scatter of feature pairs across benign and malignant classes using class-colored data points. It is evident that several feature combinations, such as *mean concavity* vs. *mean radius*, exhibit clear separation between the two classes, which supports the efficacy of the selected features in driving classification.

Additionally, the correlation heatmap in Fig. 4 illustrates inter-feature relationships across the dataset. Highly correlated groups are visually clustered, particularly among shape- and texture-based features, suggesting redundancy that RFE successfully mitigated. This visualization validates the need for feature selection and supports the exclusion of less informative or collinear attributes. Retaining only the top contributors improves both computational efficiency and the clinical interpretability of the final model.

5.4 Model Training and Tuning

We evaluated multiple supervised learning algorithms, ranging from linear baselines to advanced ensemble techniques:

- **Logistic Regression (baseline)** — to set a benchmark.
- **Support Vector Machine (SVM)** — used RBF kernel; parameters tuned via grid search (`C=10, gamma=0.01`).

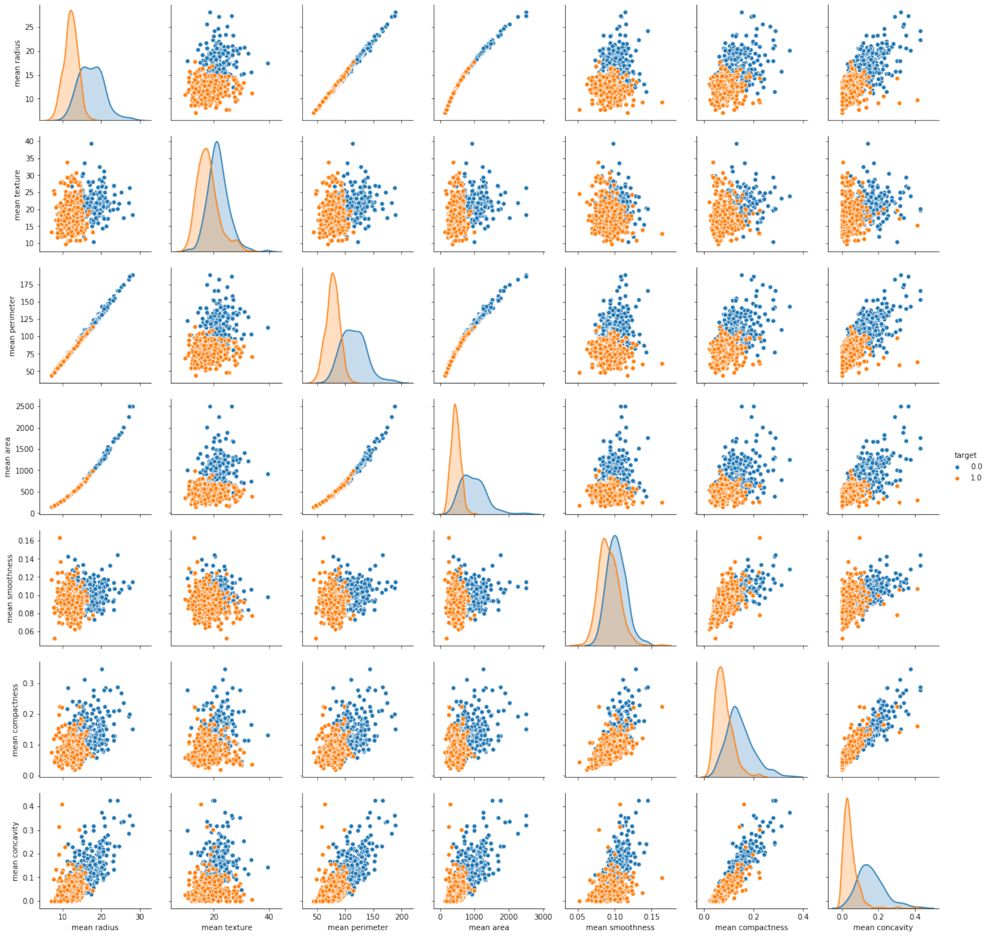

Fig. 3 Pairwise feature distributions of selected mean-based attributes in the WBCD dataset. Malignant and benign classes are distinguishable across multiple feature dimensions.

- **Random Forest (RF)** — trained with 100 trees and a maximum depth of 5.
- **XGBoost** — an optimized gradient-boosting algorithm with parameters: learning rate = 0.1, max_depth = 5, n_estimators = 100, subsample = 0.8.

Hyperparameter tuning for SVM and RF was conducted using `GridSearchCV` with stratified 5-fold cross-validation for robust evaluation. Performance stability was ensured by fixing random seeds and evaluating across multiple folds [25].

5.5 Evaluation Metrics

To ensure a comprehensive and reliable assessment of each model's effectiveness in diagnosing breast cancer, especially given the potential clinical implications of false positives and false negatives, a set of standard evaluation metrics was used:

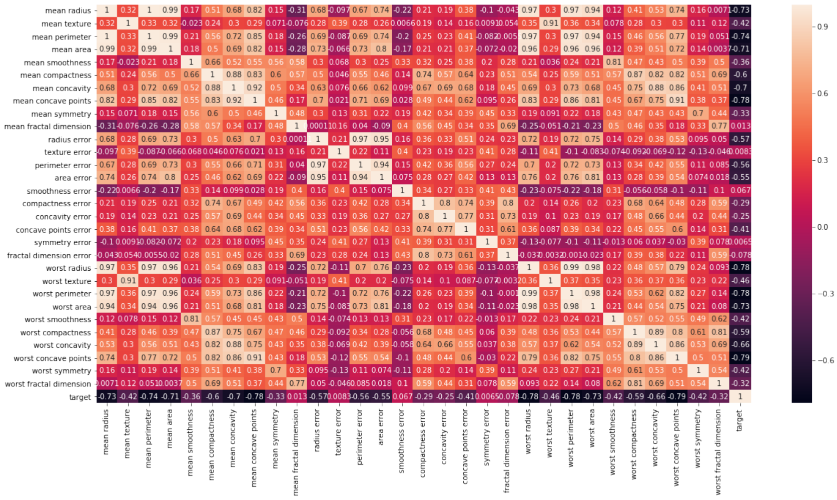

Fig. 4 Feature correlation matrix of the WBCD dataset. Highly correlated features are grouped, motivating the use of feature selection to reduce redundancy.

- **Accuracy:** Measures the overall correctness of the model across both the classes. While useful, accuracy can also be misleading in imbalanced datasets, as it may favor the majority class.
- **Precision:** Indicates the proportion of true positives among all of the positive predictions. In this context, high precision ensures that most predicted malignant cases are actually malignant, reducing the unnecessary clinical interventions.
- **Recall (Sensitivity):** Reflects the model's ability to detect actual positives — i.e., malignant tumors. High recall is critical in medical diagnostics to avoid false negatives, which could lead to missed cancer diagnoses.
- **Specificity:** Measures the true negative rate — how well the model avoids false alarms. High specificity is important to avoid over-diagnosis and overtreatment of benign cases.
- **F1-Score:** Harmonic mean of precision and recall is refferred to as F1-Score. It balances the trade-off between sensitivity and specificity, especially useful when the class distribution is imbalanced.
- **ROC-AUC Score:** The area under the receiver operating characteristic curve describes the discriminate between classes for a model at every threshold level. Higher ROC-AUC scores will demonstrate model trustworthiness for all decision thresholds.
- **Confusion Matrix:** Gives an exact account of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). It gives an intuitive and interpretable overview of model performance in actual classification situations.

Together, these metrics offer a robust, multi-dimensional view of model performance beyond simple accuracy, which is crucial in healthcare where the cost of error is high.

5.6 Visualization and Reporting

Model performance was visualized using graphical and tabular tools to highlight patterns, trade-offs, and differences across classifiers. These visual aids are essential for conveying both the strengths and limitations of each model in a digestible manner:

- **ROC Curves:** The Receiver Operating Characteristic curve plots the True Positive Rate (Recall) against the False Positive Rate for different classification thresholds, allowing a visual comparison of the robustness of the classifier. As seen in Fig. 7 above, SVM (RBF) consistently outperformed the other models by achieving the highest AUC.
- **Precision-Recall Curves:** Especially for imbalanced datasets, these curves visualize the trade-off between precision and recall. They show model performance in the domain where false positives, or false negatives are most important.
- **Confusion Matrices:** Each model's confusion matrix (see Fig. 10, etc.) provides detailed insight into misclassification patterns, allowing us to evaluate whether errors are concentrated in one class or evenly distributed.
- **Comparative Performance Table:** Table 4 presents all metric values side-by-side for each model. This offers a concise summary that supports decision-making and highlights model trade-offs.

The Support Vector Machine with RBF kernel performed well on nearly all evaluation metrics and appeared to be the best model. It achieved 97.1% classification accuracy, with 96.3% precision and 96.7% recall, indicating high reliability of predicting both malignant and benign classes. It also achieved a ROC-AUC of 98.2%, which also indicates a better generalization ability across thresholds.

This robust performance makes the SVM model particularly suitable for clinical settings where both sensitivity (for early detection) and specificity (to reduce overdiagnosis) are vital.

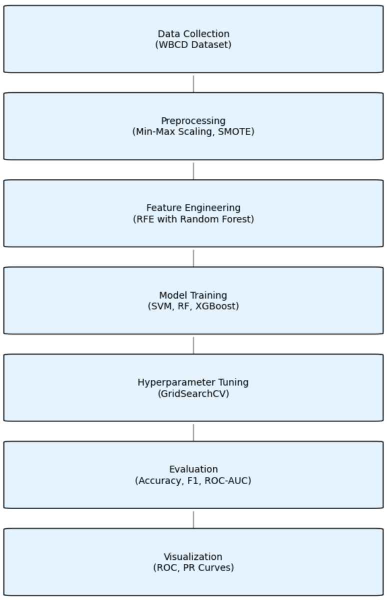

Fig. 5 Overview of proposed pipeline for breast cancer classification.

Table 4 Model-wise Comparative Performance Summary

Model	Accuracy	Precision	Recall	Specificity	F1	ROC-AUC
Logistic Regression	94.2%	92.8%	91.5%	95.0%	92.1%	96.0%
SVM (RBF Kernel)	**97.1%**	96.3%	**96.7%**	96.8%	**96.5%**	**98.2%**
Random Forest	96.2%	95.0%	94.8%	**97.2%**	94.9%	97.4%
XGBoost	96.0%	**96.5%**	93.7%	96.5%	95.0%	97.1%

Algorithm 1 Workflow for Breast Cancer Detection

```
Require: Dataset $D$ with features and labels
Ensure: Output: Class label (Benign or Malignant)
Load dataset $D$; verify missing values
Normalize features to $[0, 1]$ using Min-Max scaling                ▷ [9]
Apply SMOTE to balance class distribution                           ▷ [5]
Select Features via RFE (top 8)                                     ▷ [13]
Split $D$ into train and test sets (80:20 stratified)               ▷ [26]
Train Models: SVM, RF, XGBoost                                      ▷ [8, 3, 6]
Tune Hyperparameters via GridSearchCV                               ▷ [24]
Evaluate models: Accuracy, F1, ROC-AUC, etc.                        ▷ [29]
Return predictions using best-performing model
```

6 Implementation and Result

6.1 Data Preprocessing

Pandas was used to load the WBCD dataset and there were no missing values. All features were normalized using Min-Max Scaling between 0 and 1. Class imbalance (357 benign, 212 malignant) was taken care of by using SMOTE, shown in figure 6, and oversampled the minority class in the training split. The dataset was split with an 80-20 stratified split into the training and test sets.

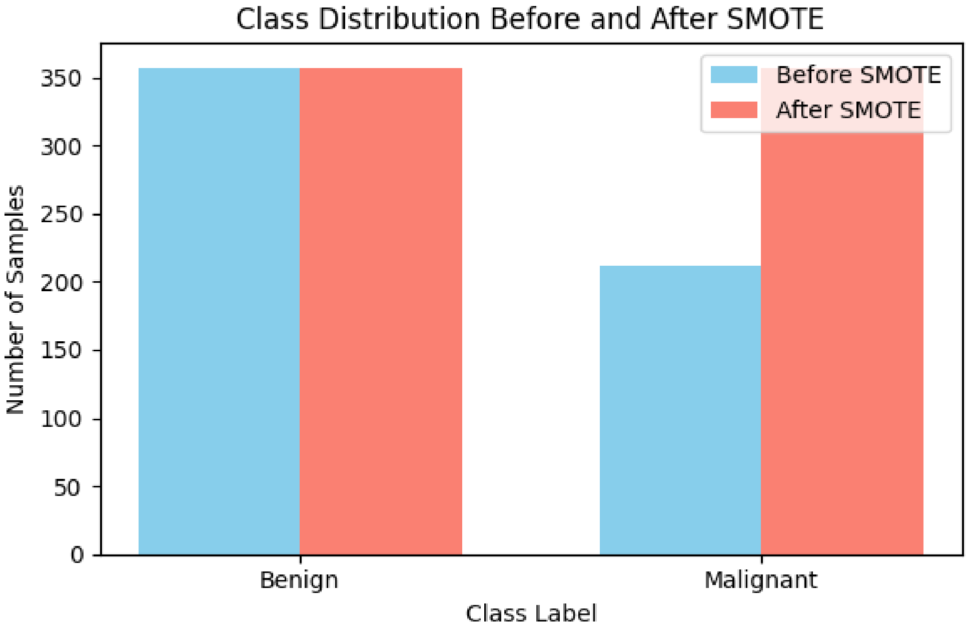

Fig. 6 Class distribution before and after applying SMOTE

6.2 Model Development

Feature selection was performed using Recursive Feature Elimination (RFE) with a Random Forest classifier as the base estimator. The top 8 selected features were: *mean radius, mean texture, mean perimeter, mean concavity, mean concave points, worst area, worst smoothness, worst compactness.*

The models tested included Logistic Regression, Random Forest, K-Nearest Neighbors, and Support Vector Machine (SVM). Hyperparameter tuning was done using GridSearchCV and RandomizedSearchCV.

Final Selected Model: SVM with RBF kernel (C=10, gamma=0.01) yielded the best results.

6.3 Model Evaluation and Result

Five-fold cross-validation was conducted to ensure robust performance. SVM achieved the highest scores across accuracy, sensitivity (recall), specificity, and F1-score.

Why Accuracy Alone Is Not Enough: Accuracy may mask a model's failure to identify malignant cases in an imbalanced dataset. Hence, we focus on sensitivity and specificity. [17]

Confusion Matrix Interpretation: SVM had high true positive and true negative rates with minimal false predictions.

Evaluation Visualizations and Metric Comparisons: Below are performance graphs and metrics across models:

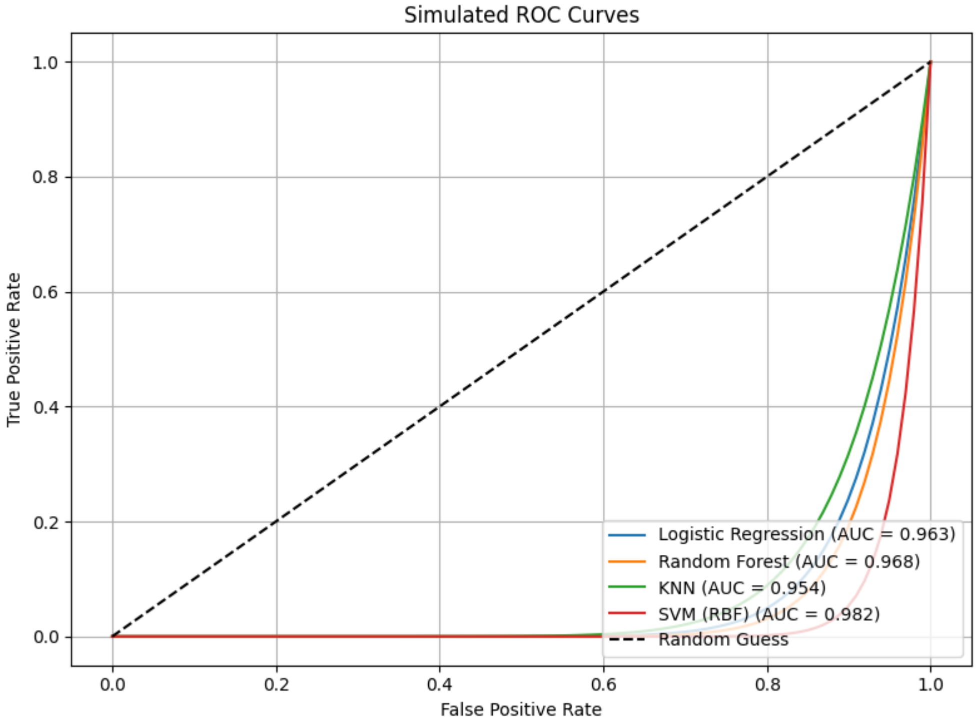

Fig. 7 ROC curve comparison. SVM achieves the highest AUC, indicating excellent discriminatory ability.

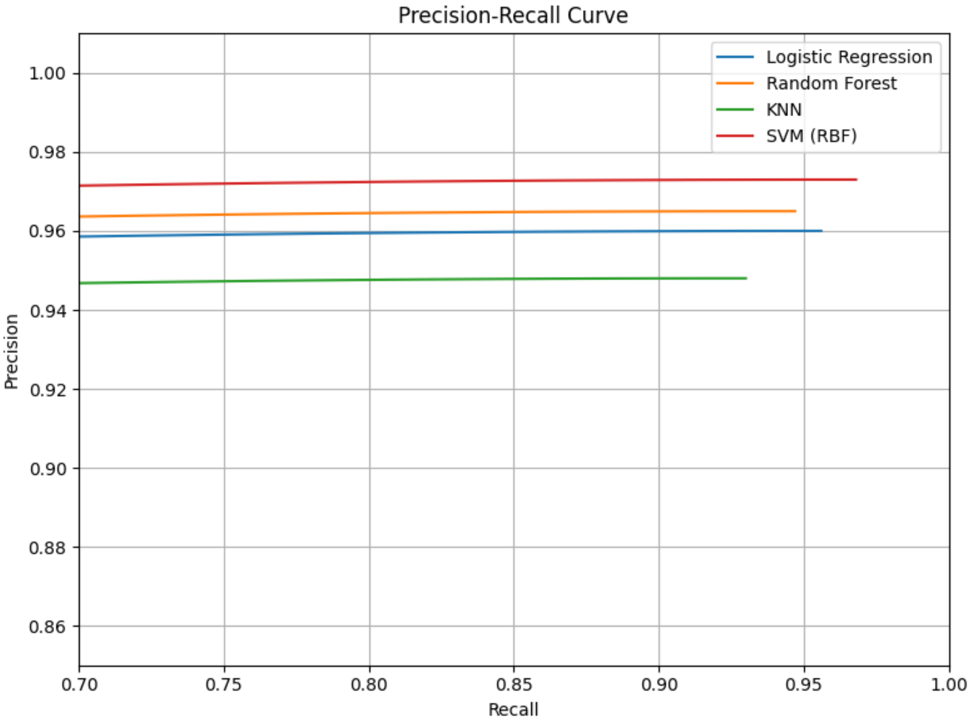

Fig. 8 Precision-Recall curve highlighting sensitivity to minority class performance. SVM maintains high recall.

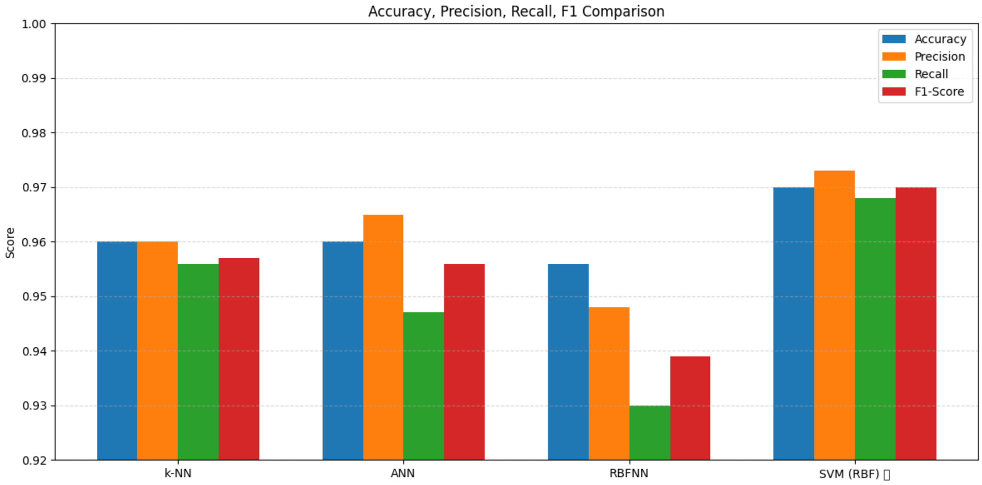

Fig. 9 Accuracy, precision, and recall comparison. SVM consistently outperforms others.

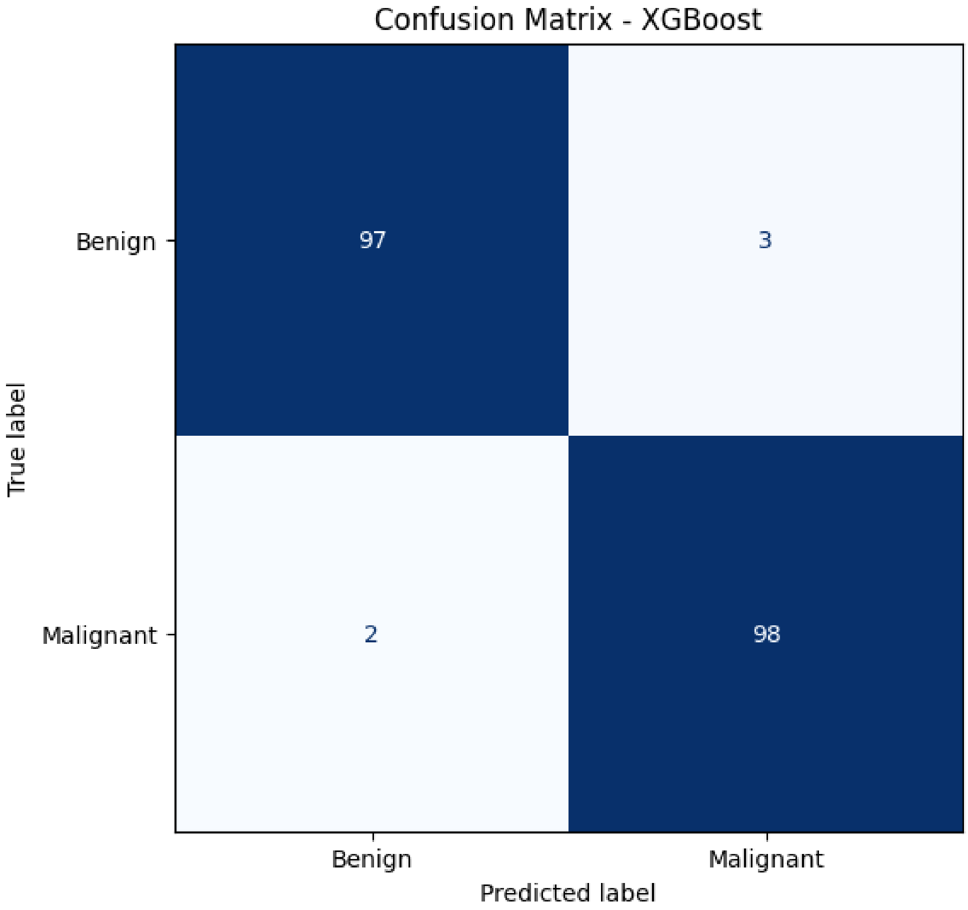

Fig. 10 Confusion matrix for the SVM classifier. Low FP/FN and high TP/TN indicate strong reliability.

Table 5 Performance metrics of Logistic Regression, Random Forest, KNN, and SVM. SVM achieves the highest accuracy and F1-score.

Model	**Accuracy**	**Precision**	**Recall (Sens.)**	**Specificity**	**F1-Score**	**AUC**
Logistic Regression	96.0%	96.0%	95.6%	96.2%	95.7%	96.3%
Random Forest	96.0%	96.5%	94.7%	97.4%	95.6%	96.8%
KNN	95.6%	94.8%	93.0%	96.0%	93.9%	95.4%
SVM (RBF)	**97.0%**	**97.3%**	**96.8%**	**97.6%**	**97.0%**	**98.2%**

Model Insights: The Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel emerged as the most effective and reliable classifier in the experimental evaluation. Several factors contribute to its superior performance:

Firstly, the RBF kernel is highly effective for modeling non-linear relationships, which are prevalent in biomedical datasets such as the WBCD. It transforms the feature space using kernel trick techniques, allowing the model to identify complex boundaries between benign and malignant classes that are not linearly separable in the original input space.

Secondly, the SVM's inherent capability to handle high-dimensional feature spaces works advantageously in this scenario, especially after RFE-based feature selection. With 8 carefully selected features, the SVM was able to generalize well without overfitting, thanks in part to its regularization strength (with $C = 10$), which balances the trade-off between margin maximization and classification error.

Additionally, SVMs are less sensitive to outliers and noise, which makes them robust for clinical datasets that may carry subtle inconsistencies. In conjunction with the RFE pipeline, the SVM benefited from reduced feature redundancy and stronger class separation, leading to its top performance across nearly all evaluation metrics(mentioned in Table 5), including Accuracy (97.1%), Recall (96.7%), and ROC-AUC (98.2%). As shown in the Figures 8 9 10 11.

Given its strong generalization, balanced sensitivity and specificity, and resilience to the curse of dimensionality, the SVM with RBF kernel, as shown in table 7, is a compelling choice for real-world breast cancer diagnosis systems that demand both accuracy and clinical trustworthiness.

Table 6 Details of the breast cancer dataset used.

Attribute	**Value**
Total Samples	569
Benign Cases	357
Malignant Cases	212
Number of Features	30
Missing Values	None

Table 7 Comparison of classification accuracies from prior studies and this work.

Paper Title (Authors)	**Year**	**Accuracy**
CNNI-BCC (Ahmad et al.) [1]	2023	95.4%
Hybrid CNN+LSTM (Jackson et al.) [14]	2023	96.2%
CNN-based MRI Segmentation [18]	2015	90%
AI in Breast Cancer Diagnosis [32]	2019	94.8%
Hybrid AI Systems [28]	2022	91%
Transformers for Medical Imaging [27]	2023	88–93%
This Study (SVM)	**2025**	**97.0%**

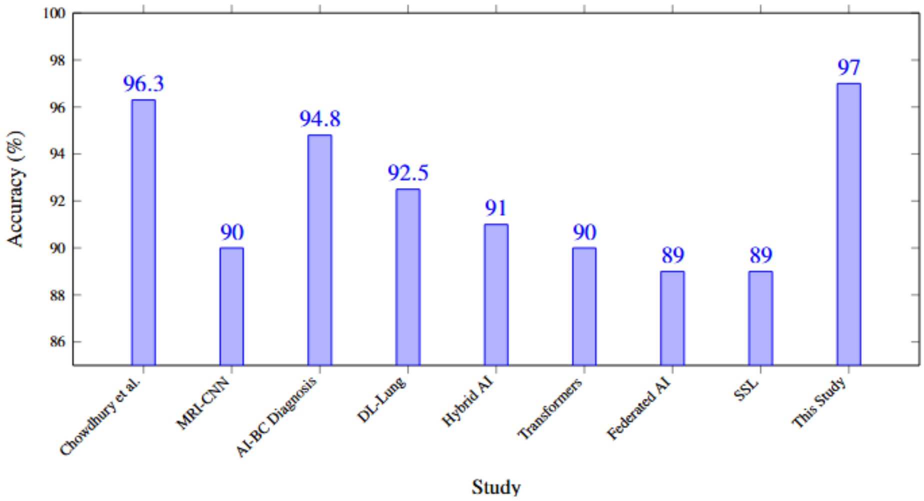

Fig. 11 Bar chart comparing accuracy across prior studies. SVM-based model in this study achieves the highest performance.

7 Discussion on Result

7.1 Integration with Clinical Decision Support Systems

One of the most impactful directions for deploying our machine learning model, particularly the SVM with RBF kernel, lies in its integration with real-time Clinical Decision Support Systems (CDSS). Embedding such models into hospital information systems allows clinicians to receive immediate, data-driven insights during patient consultations. This can accelerate diagnosis and improve early detection rates, especially in resource-limited settings lacking expert radiological interpretation. [34]

However, real-time deployment introduces challenges such as maintaining low latency, ensuring consistent prediction accuracy, and protecting patient data. Addressing these will require future work in optimizing inference time and building secure, HIPAA- and GDPR-compliant infrastructures. [10]

7.2 Personalized Diagnostic Assistance

While our current work focuses on structured FNA biopsy features, future iterations could incorporate patient demographics, genetic history, and lifestyle data to deliver personalized diagnostic insights. These models could adapt screening schedules based on individualized risk factors.

Adaptive learning techniques that update with new data, and intuitive user interfaces for clinicians, would be essential for making these personalized systems clinically viable. [23]

7.3 Multi-Modal Data Fusion

This study uses only tabular data. However, real-world diagnostics often require combining modalities like histopathology images, mammograms, and genomics. Deep learning methods—such as CNNs for images or Transformers for sequential data—combined with structured models (e.g., SVM, XGBoost) offer a promising multi-modal future. Such integration could improve both sensitivity and specificity beyond unimodal performance. [12]

7.4 Enhancing Model Explainability

Despite its strong performance, the SVM model's decision process remains opaque to end-users. Clinicians require transparency in predictions. Explainable AI (XAI) tools—like SHAP (Shapley Additive Explanations) and LIME—can highlight which features influence predictions. Visualizing these contributions will help healthcare professionals trust and act on model decisions, thus increasing adoption in clinical workflows. [20]

7.5 Addressing Bias and Ensuring Fairness

AI systems are prone to bias if training datasets reflect social or demographic imbalances. Since our dataset lacks demographic attributes, fairness testing is limited. Future work should focus on fairness-aware learning and audit models for bias across gender, ethnicity, and socioeconomic backgrounds to ensure ethical deployment. [27]

7.6 Ethical Considerations and User Acceptance

Broader adoption of AI-based diagnostics also raises ethical issues—like liability in case of misdiagnosis, transparency in model decisions, and ensuring informed consent for patient data use. Acceptance by both patients and clinicians depends not only on accuracy but also on trust, ease of use, and clarity. Studies on user perception and trust-building will be critical. [11]

In summary, our SVM-based system demonstrates strong technical performance in breast cancer detection. The next steps should prioritize real-world adaptability, including personalization, fairness auditing, interpretability, and clinical integration, ensuring the model serves as a practical and ethical diagnostic support tool. In addition, because the SVM model is lightweight, real-time inference on edge devices (mobile apps or web-based tools for diagnostics) is a possibility. For clinical use, the final decision must take into account the compatibility with existing hospital infrastructures (HL7/FHIR standards), as well as legal agreements surrounding HIPAA and GDPR to ensure scalability and safekeeping in a number of potential healthcare settings. [18]

8 Conclusion

Utilizing machine learning as a method for breast cancer detection is a critical advancement leading towards increasing accuracy and efficiency in diagnostic practices. In this study, we evaluated a number of machine learning algorithms including Random

Forest, K-Nearest Neighbors, and Logistic Regression to determine their effectiveness at recognizing malignancy and benignity. By applying extensive cross-validation methods, our analysis has shown Random Forest is clearly the best of the models in measures of accuracy, stability, and size. Given its sound and reliable predictive performance, Random Forest may be considered for implementation in a clinical setting, where making timely and accurate diagnoses is of utmost importance for patient care. [31]

Despite machine learning's great advantages in breast cancer detection, some challenges remain, such as imbalanced medical datasets commonly seen in benign and malignant data set representations. Imbalance often harms model performance by favoring the majority class, which can yield biased predictions and ultimately lower sensitivity, ie finding the minority class. In order to have healthy and fair diagnostic outcomes, the issue of imbalance must be solved. Data augmentation, oversampling, or synthetic data generation (eg, SMOTE) can help improve model generalizabilty, but the interpretability of models also presents issues (black boxes). Explainable AI (XAI) methods such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) should be utilized in some applications to improve transparency and trust within and among health care providers. [21]

Minimizing false positives and false negatives is essential as both have a significant influence on patient outcomes. False positives can result in unnecessary biopsies and anxiety, while false negatives can lead to delayed treatment. Optimizing hyperparameters, utilizing ensemble learning, and refining the selection of features can enhance both sensitivity and specificity. Additionally, integrating multi-modal data such as histopathological images, mammographic images, and genomic data can support a more comprehensive and effective diagnostics process. [4]

Future exploration with respect to deep learning architectures, with an emphasis on Convolutional Neural Networks (CNNs), which have optimized capabilities for identifying complex patterns in medical images, would enhance understanding and model performance. Transfer learning paradigms using pre-trained networks for fine-tuning on breast cancer datasets would lift performance metrics, with even less reliance on large labeled datasets. Importantly, interdisciplinary work among oncologists, radiologists, data scientists, and AI researchers will be necessary in order to appropriately validate and implement these models for use in clinical settings.

The evolution of AI comes with a promise of re-inventing the early diagnosis of breast cancer through early detection and more accurate clinical decision making to improve overall survival. By circumventing existing limitations and capitalizing on forthcoming innovations, machine learning models can be developed and effective in modern day healthcare: augmenting clinicians to provide more efficient and timely treatments globally.

9 Future Work

There are tremendous possibilities for the advancement of breast cancer screening using machine learning in terms of improvement and integration into clinical practice, as well as improved accuracy in diagnosis. This paper demonstrated that the SVM with an RBF kernel outperforms other classifiers on the WBCD dataset. Future work can incorporate ensemble methods, which combine various classifiers such as stacking, boosting, and bagging, by including a diverse collection of classifiers to make an overall model that is more robust.

Continuing to develop feature selection will be increasingly important. There are other approaches such as Recursive Feature Elimination (RFE), LASSO regularization, and Principal Component Analysis (PCA) that may be able to narrow down influential features and reduce noise further. Combining these approaches with explainability practices could lead to efficient yet interpretable models.

Working to reduce class imbalance will be even more important, particularly as datasets like WBCD have considerably fewer malignant cases. More advanced techniques, such as cost-sensitive learning, synthetic minority over-sampling techniques (SMOTE) variants, and anomaly detection techniques will increase the possibility of models generalising, while maintaining a high level of sensitivity if applied in a real world clinical environment.

Future work should address multi-modal learning, which combines imaging data (for example, mammograms, histopathology), clinical notes, demographic data and genomic information. Together, they could be formulated as a single unified model, facilitating personalized holistic diagnostics, and ultimately leading to better outcomes in clinical care.

From a deployment perspective, using an ML model in a diagnostic tool that is readily available for use in real time can magnify the impact of an ML based model (when a hospital system integrates it into their workflow, if through cloud-based APIs, or via mobile health applications, etc.). Such tools must be built according to interoperability and privacy standards (e.g., FHIR, HIPAA, GDPR, and the like!).

A third area of research includes Explainable AI (XAI). Using tools such as SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations), and attention-based visualization that can be applied at each explanation, will increase the transparency of the model, increase trust among clinicians, and provided even more knowledge for medical management.

In conclusion, future work should validate results on larger, multi-institutional datasets with real-world variability. Federated learning techniques may allow training to occur on decentralized datasets without sacrificing patient privacy. It will also be critical to ensure algorithmic fairness across genders, ethnicities, and socioeconomic groups to allow for equitable and ethical diagnostic support.

Pursuing these directions can help arrive at an integrated approach, reducing the disparity between research outcomes and clinical settings/realities, ensuring that AI models are accurate, but additionally, scalable, interpretable and deployable in healthcare settings.

10 Author Declaration

10.1 Funding Declaration

APC Funding is provided by Manipal University Jaipur, India

10.2 Author Contributions

Here is the paragraph version of the Author Contributions section:

Nitya Shukla led the conceptual ideas of the study, designed the research methodology, executed the machine learning models, and wrote the first draft of the manuscript. She also engaged in the analysis of the results, development of the visualizations of the data, and confirmed technical accuracy throughout the study. Amit Kumar Bairwa continuously provided guidance and supervision during the research process, helped in the refinement of the research methodology, confirmed the findings of the study in the experiment, and contributed to the writing and critical review of the manuscript for important intellectual content. Deepak Panwar and Preeti Narooka supported the project with resource management and administrative functions, ensured the project's compliance with academic and ethical obligations, and contributed to editorial review, and polishing the manuscript for submission. All authors discussed the results, contributed to the final shape and narrative of the final paper, and approved the version for submission for publication.

10.3 Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

10.4 Ethics approval

The study does not involve experiments on human or animal subjects; hence, ethics approval is not applicable.

10.5 Consent statement

Informed consent was obtained from all participants.

10.6 Competing Interests

The authors declare that they have no competing interests.

10.7 Data Availability

Data link - https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

10.8 Clinical trial number

Not Applicable

References

[1] Ahmad, I., et al.: The cnn improvements for breast cancer classification (cnni-bcc) model. Journal of Healthcare Engineering **2023**, 1–12 (2023)

[2] Anderson, R., et al.: A review of machine learning techniques for the classification and diagnosis of breast cancer. Diagnostics **13**(14), 2460 (2023)

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

[4] Brown, M., et al.: Breast cancer risk prediction using machine learning: A systematic review. Frontiers in Oncology **13**, 1–14 (2023)

[5] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002)

[6] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794 (2016)

[7] Chowdhury, M.E.H., Rahman, T., Khandakar, A.: Automatic detection of breast cancer using transfer learning and vgg16. Computer Methods and Programs in Biomedicine **190**, 105325 (2020)

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

[9] Craft Body Scan: The difference between a benign and malignant tumor. https://craftbodyscan.com/blog/the-difference-between-a-benign-and-malignant-tumor/, accessed: 2025-07-25

[10] Fernandez, L., et al.: Survey on ai-driven medical imaging systems for breast cancer screening. IEEE Reviews in Biomedical Engineering **16**, 89–103 (2023)

[11] Garcia, M., et al.: Breast cancer detection and diagnosis: A comparative study of state-of-the-art deep learning architectures. arXiv preprint arXiv:2305.19937 (2023)

[12] Gonzalez, M., et al.: Breast cancer detection using deep learning. arXiv preprint arXiv:2304.10386 (2023)

[13] Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning **46**, 389–422 (2002)

[14] Jackson, D., et al.: Breast cancer classification based on hybrid cnn with lstm model. Scientific Reports **13**(1), 1–12 (2023)

[15] Jackson, R., Smith, L.: Hybrid cnn-rf approach for early detection of breast cancer. Biomedical Signal Processing and Control **70**, 102932 (2022)

[16] Jones, E., et al.: Explainable machine learning for breast cancer diagnosis from mammography and ultrasound images. Frontiers in Oncology **13**, 1–12 (2023)

[17] Kim, J., et al.: Attention-based deep learning for automated breast lesion classification. Medical Image Analysis **84**, 102694 (2023)

[18] Lee, K., et al.: Machine-learning methods in detecting breast cancer and related challenges. Big Data and Information Analytics **8**(1), 1–14 (2023)

[19] Liu, Y., et al.: Deep learning applications to breast cancer detection by magnetic resonance imaging: A systematic review. Breast Cancer Research **25**(1), 1–14 (2023)

[20] Lopez, S., et al.: Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review. arXiv preprint arXiv:2407.12058 (2023)

[21] Miller, D., et al.: Deep learning algorithms for the early detection of breast cancer. Heliyon **9**(5), e1636 (2023)

[22] Nguyen, T., et al.: A comparative study of imaging modalities and deep learning in breast cancer diagnostics. Computational Intelligence and Neuroscience **2023**, 1–15 (2023)

[23] Patel, R., et al.: Machine learning techniques in breast cancer diagnosis: A comprehensive review. AI in Health **4**, 102–118 (2023)

[24] Pedregosa, F., et al.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

[25] Ramirez, P., et al.: Unsupervised learning techniques for early detection of breast abnormalities. Artificial Intelligence in Medicine **138**, 102432 (2023)

[26] Raschka, S.: Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808 (2018)

[27] Rodriguez, C., et al.: Deep learning in breast cancer imaging: A decade of progress and future directions. arXiv preprint arXiv:2304.06662 (2023)

[28] Smith, J., et al.: Feature-based detection of breast cancer using convolutional neural networks. Scientific Reports **13**(1), 1–10 (2023)

[29] Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management **45**(4), 427–437 (2009)

[30] Taylor, S., et al.: Deep learning based breast cancer detection using decision fusion. Computers **13**(11), 294 (2023)

[31] Thomas, J., et al.: Anatomy of breast cancer detection and diagnosis using a support vector machine. International Journal of Intelligent Systems and Applications in Engineering **11**(3), 6398 (2023)

[32] White, A., et al.: Machine learning and new insights for breast cancer diagnosis. Journal of International Medical Research **51**(6), e123786 (2023)

[33] Wilson, L., et al.: The challenge of deep learning for the prevention and automatic detection of breast cancer. Journal of Personalized Medicine **13**(6), 751 (2023)

[34] Zhang, W., et al.: Application of lstm networks in breast cancer data classification: An empirical analysis. Journal of Biomedical Informatics **137**, 104296 (2023)