Cureus Journal of Computer Science

Open Access Original Article

Comparative Evaluation of Support Vector Classifier, Random Forest, and XGBoost for Early Breast Cancer Prediction With Feature Importance and Class Balancing

Received 04/28/2025
Review began 06/03/2025
Review ended 06/18/2025
Published 06/19/2025

DOI: https://doi.org/10.7759/s44389-025-05794-5

Deepti Sharda [1] , Ritika Bansal [1] , Prachi Gumber [1] , Swati Kandari [1] , Indu Arora [1]

1. Department of Computer Science and Applications, Mehr Chand Mahajan DAV College for Women, Chandigarh, IND

Corresponding author: Ritika Bansal, ritika.bansal@mcmdavcwchd.in

Abstract

Timely detection of breast cancer is crucial for improving patient outcomes. While machine learning (ML) models have shown promise in supporting diagnosis, there is still a need for frameworks that combine predictive accuracy with model interpretability and balanced evaluation.

This study compares three widely used supervised ML algorithms, Support Vector Classifier (SVC), Random Forest, and XGBoost, using the Wisconsin Diagnostic Breast Cancer dataset. To address class imbalance, the dataset was balanced using the Synthetic Minority Oversampling Technique, and hyperparameter tuning was also performed through GridSearchCV. Each model was evaluated using accuracy, precision, recall, F1-score, and receiver operating characteristic area under the curve.

XGBoost exhibited the best results, achieving 98.60% accuracy and perfect recall for malignant cases, which are crucial for reducing false negatives in clinical settings. Random Forest also demonstrated better results, while SVC showed slightly lower accuracy, but it still maintained balanced class-wise performance. Feature importance analysis was applied to these models to enhance clarity and highlight key diagnostic features. SVC could not generate feature importance output due to its kernel-based approach.

This study presents a reproducible and interpretable ML-based framework for breast cancer classification. By integrating model optimization, class balancing, and explainability, the proposed approach provides a robust foundation for future clinical decision-support systems.

Categories: AI applications, Algorithm Analysis, Predictive Analytics
Keywords: breast cancer prediction, machine learning, feature selection, xgboost and random forest, predictive analytics, medical diagnostics, ai-driven disease detection

Introduction

Breast cancer is among the most prevalent and life-threatening diseases globally, posing a significant public health challenge, and exerts a substantial burden on healthcare systems. According to the World Health Organization (WHO), breast cancer was the most commonly diagnosed cancer in 2020, accounting for approximately 2.3 million new cases and 0.68 million deaths worldwide [1]. Early and accurate detection of breast cancer is critical to improve survival rates and to enable timely treatment. The conventional diagnostic techniques such as mammography, physical examinations and biopsies are widely adopted and clinically effective, but these methods, although essential, can occasionally be time-consuming and resource-intensive, and there still remains a possibility of human interpretation variability. In particular, distinguishing between malignant and benign lesions, especially in early-stage breast cancer, can be challenging and may lead to occasional diagnostic uncertainty [2].

Integrating artificial intelligence (AI) and machine learning (ML) techniques into the diagnostic workflow holds promise in complementing existing practices by enhancing accuracy, reducing variability and assisting clinicians in making more informed decisions. ML techniques can identify patterns present in complex medical dataset, which might not be detected by traditional methods, which can help in early diagnosis. It has been observed that combination of AI and ML techniques can achieve equivalent or in some cases better performance than an expert, as shown by Esteva et al. [3], where deep neural networks achieved performance comparable to or exceeding that of expert radiologists.

ML-based models have a potential to improve the diagnostic reliability by reducing interpretation time and minimizing human error [4,5]. Feature selection is one of the key factors for the success of ML-based models in the field of medical. Medical datasets are often high-dimensional and not all features contribute meaningfully to model performance. By detecting the relevant features, accuracy as well as efficiency of

How to cite this article
Sharda D, Bansal R, Gumber P, et al. (June 19, 2025) Comparative Evaluation of Support Vector Classifier, Random Forest, and XGBoost for Early Breast Cancer Prediction With Feature Importance and Class Balancing. Cureus J Comput Sci 2 : es44389-025-05794-5. DOI https://doi.org/10.7759/s44389-025-05794-5

the model can be improved, thereby reducing the overall computational costs. In breast cancer diagnosis, identifying critical predictive features can also support transparency and trust in AI-driven systems, which is vital for clinical integration.

Many ML techniques are available; therefore, choosing the effective technique out of all is crucial. This study focuses on three widely used classifiers - Support Vector Classifier (SVC), Random Forest (RF) and XGBoost - to study their complementary strengths in handling non-linear patterns, ensemble learning and feature prioritization. The study evaluates and compares these three models using metrics such as accuracy, precision, recall and F1-score. These models are evaluated in the context of breast cancer classification using a well-known diagnostic dataset. The novelty of this study lies in presenting a well-structured, comparative evaluation of these three ML models for breast cancer detection, aimed at improving both predictive accuracy and actionable insights for clinical use.

Materials And Methods

Literature review

The application of ML techniques in breast cancer detection has been widely explored. Many studies achieved high classification accuracy using models such as SVM, Artificial Neural Networks (ANN), k-Nearest Neighbors (k-NN), Decision Trees (DT), RF and XGBoost. Table *1* summarizes some notable research that has applied these algorithms across various breast cancer datasets.

Reference	Dataset Used	Feature Selection Applied?	Feature Selection Technique	ML Algorithm Used	Results & Findings
[6]	UCI Machine Learning Repository	No	RapidMiner 5.0	Discriminant Analysis, ANN, DT, LR, SVM, NB, k-NN	SVM achieved the best accuracy (96.5%) for breast cancer diagnosis.
[7]	Wisconsin Original Breast Cancer Dataset	No	-	SVM, RF, BN	Random Forest had the best discriminatory power for malignant vs. benign cases.
[8]	Wisconsin Breast Cancer Dataset	Yes	Feature Selection	SVM, ANN, LR, DT, k-NN, BN	SVM outperformed all classifiers (97.28%).
[9]	Wisconsin Breast Cancer Dataset	No	-	ANN, SVM	SVM outperformed ANN with 99.51% accuracy
[10]	BreaKHis dataset	Yes	DenseNet Feature Extractor	Deep Learning XGBoost	Achieved 97% accuracy
[11]	Wisconsin Breast Cancer Dataset	Yes	Box Plots	XGBoost	Achieved 94.74% accuracy
[12]	Wisconsin Breast Cancer Dataset	No	-	SVM, k-NN, RF, AdaBoost, XGBoost	XGBoost achieved highest accuracy (98%), followed by SVM (96%)
[13]	Wisconsin Breast Cancer Dataset	No	-	RF assembled with Adaboost	Accuracy of 98.57%
[14]	MIAS 2012 dataset	No	-	CNN-XGBoost	XGBoost improved 7-class accuracy
[15]	Wisconsin Breast Cancer Dataset	Yes	SMOTE	XGBoost with combination of NV, RF, DT	XGBoost + RF give best accuracy (98.2%)
[16]	Wisconsin Breast Cancer Dataset	Yes	Pearson Correlation Coefficient, Recursive Feature Elimination, LR, and Univariate Selection	RF, SVM, k-NN, DT, Multilayer Perception	RF gives best accuracy (99%)
[17]	Dataset from Internet	No	-	LR, DT, k-NN, NB, SVM, RF	RF gives best accuracy (98%)

TABLE 1: Literature Review of Machine Learning Application for Breast Cancer Detection

AdaBoost, Adaptive Boosting; ANN, Artificial Neural Network; BN, Bayesian Network; CNN, Convolutional Neural Network; DT, Decision Tree; k-NN, K-Nearest Neighbors; LR, Logistic Regression; ML, Machine Learning; NB, Naïve Bayes; RF, Random Forest; SMOTE, Synthetic Minority Over-sampling Technique; SVM, Support Vector Machine; XGBoost, Extreme Gradient Boosting

Among these studies, SVM can be seen as a strong performer by achieving accuracy rates above 96% in classifying malignant and benign tumors [6,8,9,11]. From this table, it can be seen that RF also delivered excellent results because of its ability to leverage multiple decision trees, which in turn reduce overfitting and improve predictive reliability [7,13,16,17]. XGBoost, a gradient boosting framework, is known for its speed and precision because of which it can be helpful for medical diagnosis tasks. Studies that included XGBoost [10-12,14,15] exhibited accuracy up to 97% or more. XGBoost when combined with deep learning led to even greater improvements such as a 7% accuracy boost in multi-class classification using mammogram images.

Although existing studies have shown high accuracy, there are certain concerns that need to be addressed. Feature selection methods that are essential for reducing the model complexity have not been implemented by many studies. In some studies [10,11,16], techniques like Pearson correlation, recursive feature elimination and information gain were used, but they are not universally adopted. Apart from reducing model complexity, feature selection methods also provide clarity and confidence to clinicians by highlighting which features the model uses for its predictions.

Another concern that was observed is that the datasets used to carry on the research are not balanced. They contain larger number of benign cases and lesser number of malignant cases. When model is trained

using imbalanced data, it tends to become biased towards majority class and as a result might miss the malignant cases, which is dangerous in medical diagnosis because it could mean failing to detect cancer in a patient. Class imbalance is not being addressed in most of the studies. It has been handled by only few studies such as that by Mahesh et al. [15] where they have employed balancing techniques like SMOTE to improve fairness in classification performance.

Yet another important aspect that remains underexplored is hyperparameter tuning. Many of the studies have overlooked it and have used the default settings of the algorithms, which might not be the best for the specific dataset or problem. While methods like Grid Search or Bayesian Optimization have shown effectiveness in other domains, their adoption in the reviewed literature is sparse [18-20]. This inconsistency in parameter tuning makes it hard to fairly compare the results from different studies. One model may look better than another just because it was better tuned - not because it is inherently better. So, proper hyperparameter tuning is important for both performance and fair comparison.

In addition, many previous studies, as shown in Table *1*, have compared these ML models using different datasets and feature selection methods as well as experimental settings, making it difficult to fairly compare their performance and determine which model is most suitable for breast cancer diagnosis. In response to these gaps, the present study proposes a structured and comprehensive comparison of three widely used classifiers, SVC, RF, and XGBoost, using the Wisconsin Diagnostic Breast Cancer dataset. To ensure a fair and clinically relevant evaluation, the study incorporates SMOTE for class balancing, GridSearchCV for systematic hyperparameter tuning and feature importance analysis for interpretability in tree-based models. Each model is assessed using a consistent set of evaluation metrics, including accuracy, precision, recall, F1-score and area under the curve (AUC). By consolidating these elements into a single experimental pipeline, this research aims to provide a reproducible and interpretable ML framework tailored for breast cancer prediction and clinical decision support.

Dataset description

The dataset employed in this research is the Breast Cancer Wisconsin (Diagnostic) dataset, curated by Wolberg et al. [21]. It is a benchmark dataset frequently used in ML research for breast cancer diagnosis. It contains features extracted from digitized images of fine needle aspirate (FNA) samples of breast tissue, offering a rich source of information to distinguish between malignant (cancerous) and benign (non-cancerous) cases [21]. The key characteristics of the dataset are as follows:

Number of Instances: 569 samples, each representing a unique case of breast tissue analysis.

Number of Features: 30 numerical attributes that describe various properties of cell nuclei.

Class Distribution: Malignant - 212 instances (representing cancerous cases). Benign - 357 instances (representing non-cancerous cases).

The dataset comprises 30 numerical features extracted from FNA images of breast masses, representing critical geometrical and textural properties of cell nuclei. These features are grouped into three categories, each capturing different statistical aspects of the measurements:

Mean Values

These represent the average measurement for each property, providing an overall estimate of the cell nuclei's shape, size and texture.

Features: radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean.

Standard Errors

These measure the variability or spread of each property, offering insights into the consistency and reliability of the observed measurements.

Features: radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, fractal_dimension_se.

Worst Values

These indicate the largest measurement observed among the three largest cell nuclei in each image, highlighting potential extremes associated with malignancy.

Features: radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst.

Methodology

The following methodology was implemented to develop an optimized ML-based Breast Cancer Prediction System. Figure *1* provides a visual representation of the workflow.

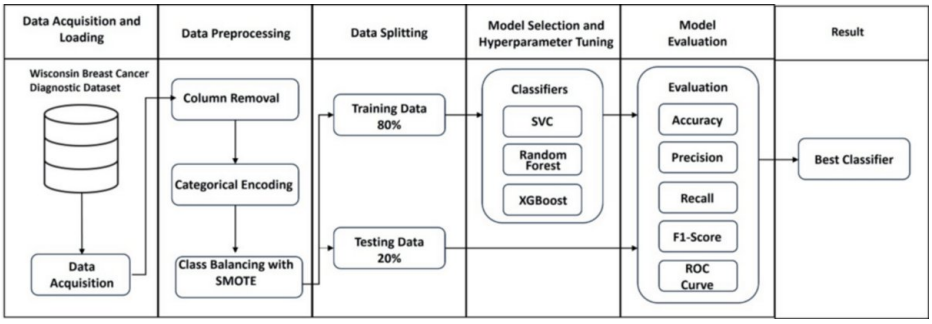

FIGURE 1: Stages for Breast Cancer Prediction

Data Acquisition and Loading

The breast cancer dataset was acquired in CSV format and loaded into the Python environment using the Pandas library. This step ensured that all data attributes and records were correctly read into a structured DataFrame for subsequent processing and analysis. The use of CSV format ensures compatibility and simplicity during import, while Pandas provides efficient and flexible tools for data manipulation, making it a standard in data science workflows [22].

Data Preprocessing

Preprocessing was performed to prepare the dataset for effective ML model training. The following sub-steps were executed:

Column removal: The id column was removed from the dataset as it serves as a unique identifier without any contribution to prediction. Including such non-informative features may introduce noise and unnecessarily increase dimensionality [23].

Categorical encoding: The target variable diagnosis, originally labeled as 'M' (Malignant) and 'B' (Benign), was encoded into binary values using Label Encoding, where 1 denotes Malignant and 0 denotes Benign. Binary encoding is appropriate for categorical variables in binary classification tasks and ensures compatibility with ML algorithms that require numerical input [24].

Class balancing with SMOTE: The dataset exhibited slight class imbalance, with benign samples outnumbering malignant ones. Class imbalance can bias models toward the majority class, leading to poor recall on the minority class [25]. To address this, SMOTE was employed. SMOTE generates new synthetic examples of the minority class rather than duplicating existing ones, effectively improving classifier performance on the underrepresented class [26].

Data Splitting

To evaluate the model's generalization capabilities, the preprocessed dataset was split into 80% for training, used to fit the models and learn patterns, and 20% for testing, used to evaluate the models' performance on unseen data. This stratified split was performed using the train_test_split() function from the scikit-learn library, which maintains the original class distribution across subsets, ensuring a balanced and fair model evaluation [27].

Feature Selection and Hyperparameter Tuning

These two processes were critical to optimize model performance and interpretability:

Feature selection: For tree-based models such as RF and XGBoost, feature_importances_ was used to identify and visualize the most significant features. Feature selection reduces overfitting, improves accuracy and enhances interpretability by removing irrelevant or redundant features [24]. Feature

importance was used as a post-training interpretability tool, rather than a pre-training feature selection filter. This approach allowed for better insight into model decision-making while retaining all features during training to maintain consistency across models. For models like SVC that do not directly provide feature importance, this step was not applied.

Hyperparameter tuning: A GridSearchCV strategy was implemented to identify the optimal combination of hyperparameters for each classifier. This approach performs an exhaustive search over a predefined parameter grid using 5-fold cross-validation, ensuring robust performance across different training/validation splits. Hyperparameter tuning is crucial for avoiding underfitting/overfitting and improving model generalization [18].

Model Selection and Training

Three ML classifiers were selected for comparative analysis, representing diverse learning paradigms:

SVC: A kernel-based margin classifier known for effectiveness in high-dimensional, low-sample-size datasets [28].

RF: An ensemble method that combines multiple decision trees and aggregates their outputs to reduce overfitting and improve accuracy [29].

XGBoost: A gradient boosting algorithm that builds sequential trees to minimize classification error, optimized for speed and predictive power [30]. It is widely used in medical ML applications due to its superior handling of complex patterns and feature interactions.

All models were trained using the training data with the best hyperparameter combinations obtained via GridSearchCV.

Model Evaluation

Each trained model was evaluated using the following metrics:

Accuracy: The proportion of correctly predicted instances across all predictions.

Precision: The proportion of true positive predictions among all predicted positives, critical in clinical scenarios to avoid false alarms.

Recall (Sensitivity): The ability to correctly identify actual malignant cases; a high recall is crucial in medical diagnosis to minimize false negatives.

F1-Score: A harmonic mean of precision and recall, offering a balanced assessment when class distribution is imbalanced.

In addition to the above standard performance metrics, this study incorporated visual tools such as confusion matrices and receiver operating characteristic curves to better compare the models. The confusion matrices offered a clear view of how each model classified benign and malignant cases, while the ROC curves provided insight into how well the models distinguished between the two classes at different threshold levels. Feature importance graphs were made for both the tree-based models, i.e. RF and XGBoost, to highlight which features had the most influence in making predictions.

Results

All experiments were executed on Python 3.11 (64-bit) on a Windows 10 machine with an Intel Core i5 (11th Gen) processor and 8 GB of RAM. Model building and visualization were done in a Jupyter Notebook setup.

The Wisconsin Breast Cancer Diagnostic dataset was imported in CSV format with the help of the pandas library. The uninformative id column was dropped, and the target variable diagnosis was label-encoded as malignant with value 1 and benign with value 0. Mild class imbalance was present in the dataset, which was handled by utilizing the SMOTE method from the imblearn library.

The data was divided into 80% train and 20% test sets via train_test_split() with a fixed random state for reproducibility. Three supervised classifiers were tested: SVC, RF and XGBoost. Each model was hyperparameter tuned via GridSearchCV with 5-fold cross-validation. Models were tested on the test set with accuracy, precision, recall, F1-score and AUC-ROC. For tree models (RF and XGBoost), feature importance scores were calculated to rank important predictor variables. Visualizations such as confusion

matrices, ROC curves and feature importance plots were created using matplotlib and seaborn.

This part deals with the performance of these three ML models, SVC, RF, and XGBoost model, in predicting the breast cancer cases. Table 2 enlists the optimal hyperparameters chosen for each model. The SVC performed best with a regularization parameter of C = 1 and a gamma of 0.001, which made it sensitive to fine patterns in the feature space. RF model was better with 200 decision trees, presumably gaining more diversity in the ensemble to enhance generalizability. XGBoost performed optimally with 100 estimators, striking a nice balance between obtaining high accuracy and preventing overfitting. The final accuracy figures indicate how good each model is at generalizing and performing well on previously unseen data.

Model	Best Parameters	Accuracy
Support Vector Classifier	C = 1, gamma = 0.001	0.9091 (90.91%)
Random Forest	n_estimators = 200	0.9720 (97.20%)
XGBoost	n_estimators = 100	0.9860 (98.60%)

TABLE 2: Model Hyperparameters and Accuracy

Table 3 presents class-wise performance, which is particularly relevant in healthcare applications where both precision and recall for malignant cases are critical. While SVC showed reasonably balanced performance (F1 = 0.91 for both classes), its lower recall for malignant cases (0.89) could translate to a higher false negative rate. RF demonstrated strong overall performance and higher recall for malignant cases (0.99), indicating its robustness in clinical screening scenarios. XGBoost achieved a perfect recall of 1.00 for malignant tumors while maintaining high precision, making it ideal for minimizing the risk of undetected cancer cases.

Model	Class	Precision	Recall	F1-Score	Support
Support Vector Classifier	0 (Benign)	0.89	0.93	0.91	69
	1 (Malignant)	0.93	0.89	0.91	74
Random Forest	0 (Benign)	0.99	0.96	0.97	69
	1 (Malignant)	0.96	0.99	0.97	74
XGBoost	0 (Benign)	1	0.97	0.99	69
	1 (Malignant)	0.97	1	0.99	74

TABLE 3: Classification Report – Class-Wise Metrics

The confusion matrices, shown in Figure 2, provide a detailed view of the classification performance of the three models. The SVC model correctly identified 64 benign and 66 malignant cases, but misclassified 5 benign as malignant and 8 malignant as benign. The RF model demonstrated improved accuracy, misclassifying only three benign and one malignant case. Notably, the XGBoost model achieved the highest classification accuracy, correctly identifying all 74 malignant cases and misclassifying only 2 benign cases.

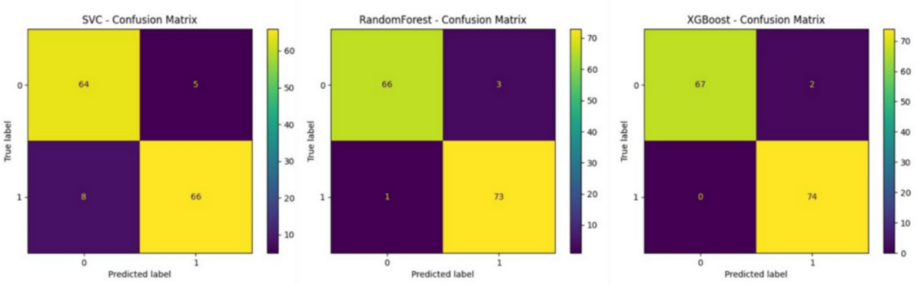

FIGURE 2: Confusion Matrices of Support Vector Classifier (SVC), Random Forest and XGBoost Models, Illustrating the Classification Performance Across Benign (Label 0) and Malignant (Label 1) Tumor Classes

These results highlight XGBoost's superior capability in distinguishing between the two classes, especially in minimizing false negatives - a critical metric in medical diagnosis where failing to detect a malignant tumor could lead to severe consequences. RF also performed well with minimal errors, whereas SVC, while still effective, showed relatively higher misclassification.

Figure *3* displays ROC curves for the SVC, RF and XGBoost classifiers. The ROC curve illustrates how each model balances sensitivity (true positive rate) and specificity (1 - false positive rate) across various classification thresholds. It serves as a valuable tool for evaluating a model's ability to discriminate between classes.

In this study, SVC achieved an AUC score of 0.97, which reflects that it can distinguish between malignant and benign cases well, but its ROC curve showed slightly more distance from the ideal top-left corner compared to the other models, suggesting that there were still some instances where it could not predict. AUC score of RF and XGBoost is coming out to be exact 1.00, and their ROC curves closely followed the top-left boundary of the plot, indicating better predictive power than SVC. These results align with the findings from the confusion matrix as discussed above.

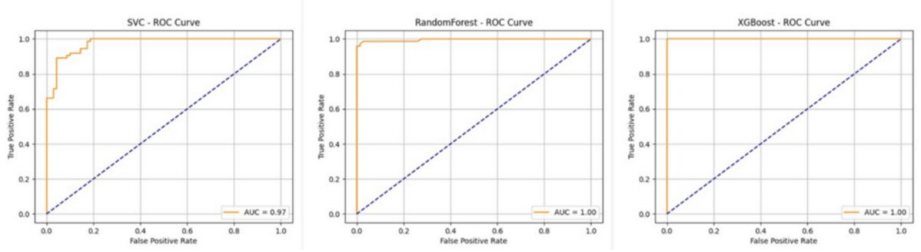

FIGURE 3: ROC Curves of SVC, Random Forest and XGBoost Models, Showing Their Ability to Distinguish Between Benign and Malignant Tumor Classes. AUC Values Are Also Indicated for Each Model

AUC, Area Under the Curve; ROC, Receiver Operating Characteristic; SVC, Support Vector Classifier

Following the ROC analysis, it is also important to examine performance metrics that encapsulate model behavior across both classes. These include macro and weighted averages for precision, recall and F1-score, as summarized in Table *4*.

Model	Macro Precision	Macro Recall	Macro F1	Accuracy
Support Vector Classifier	0.91	0.91	0.91	0.91
Random Forest	0.97	0.97	0.97	0.97
XGBoost	0.99	0.99	0.99	0.99

TABLE 4: Macro and Weighted Averages for All Models

Macro and weighted averages, shown in Table *4*, provide a holistic view of model performance across imbalanced datasets. Macro averaging ensures both classes are given equal importance, while weighted averaging reflects class frequency. XGBoost achieved near-perfect scores across all metrics, reinforcing its effectiveness in addressing both class sensitivity and balanced performance.

This consolidated summary in Table *5* highlights that while all three models performed admirably, XGBoost consistently outperformed SVC and RF in every metric. Its perfect recall for malignant cases and highest F1-score makes it a highly suitable model for critical applications such as cancer detection, where the cost of false negatives is high.

Model	Accuracy	Precision	Recall	F1-Score
Support Vector Classifier	0.9091	0.9296	0.8919	0.9103
Random Forest	0.972	0.9605	0.9865	0.9733
XGBoost	0.986	0.9737	1	0.9867

TABLE 5: Model Performance Summary

Discussion

This section presents experimental results obtained by comparing SVC, RF and XGBoost on the task of classifying breast cancer. Overall accuracy of 90.91% and F1-score of 0.91 were observed for the SVC model. It was noted that it had equitable performance for both benign and malignant cases, and its recall for malignant tumors was 0.89, which reflected an increased false negative rate. In clinical diagnosis, this can be serious as it signifies a chance of missing cancer-positive cases. The confusion matrix validated these results, revealing more misclassifications in the malignant class compared to the other models.

RF surpassed SVC in that it attained 97.20% accuracy and a recall of 0.99 for malignant samples. RF classified only one malignant sample incorrectly as can be seen from its confusion matrix. The matching ROC curve was near perfect, and its AUC value was 1.00, demonstrating high discriminatory power between benign and malignant cases.

XGBoost surpassed both these models by having the highest accuracy of 98.60% as well as a recall of 1.00 for malignant cases. XGBoost identified all the positive cancer cases effectively and, therefore, is extremely well-suited for use in diagnostic purposes where false negatives need to be minimized. The ROC curve of XGBoost was closest to the top-left corner with an AUC value of 1.00, thus validating its best classification performance.

To further understand the model decision-making process, feature importance plots were generated for RF and XGBoost, as shown in Figures *4* and 5, respectively. In both models, perimeter_worst, radius_worst and concave points_worst emerged as the most influential features in classifying breast cancer cases. These attributes are related to tumor size, shape and margin irregularities - factors known to correlate strongly with malignancy in clinical diagnostics. While RF distributed importance across a broader set of features, XGBoost concentrated more weight on just a few key predictors, particularly perimeter_worst and radius_worst, suggesting its effectiveness in narrowing down the most critical features for accurate classification.

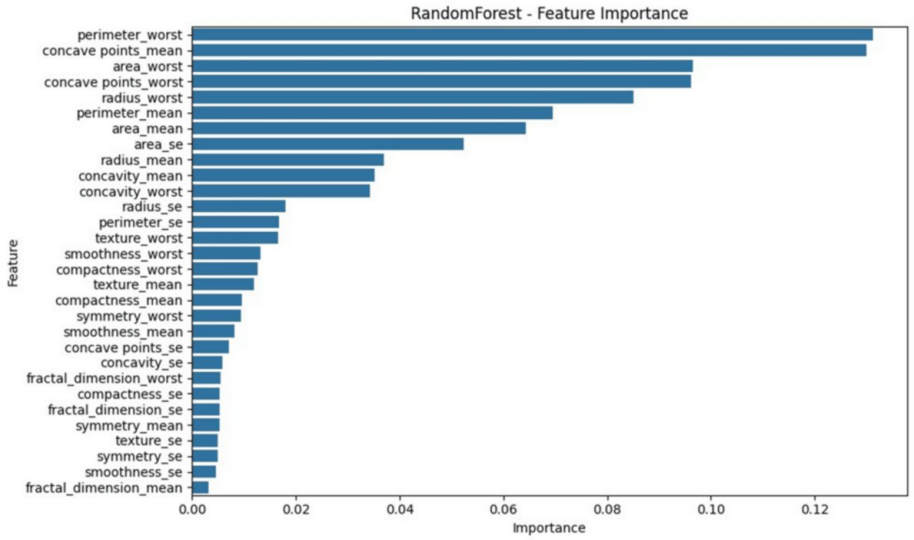

FIGURE 4: Feature Importance Scores Generated by Random Forest

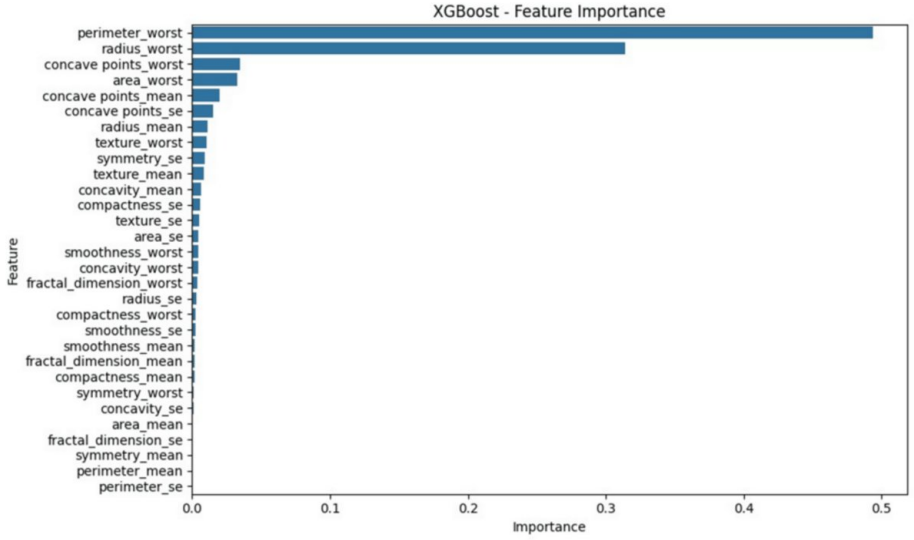

FIGURE 5: Feature Importance Scores Generated by XGBoost Models

It is important to note that SVC does not offer feature importance visualization due to its kernel-based mechanism. SVC constructs a decision boundary in a higher-dimensional space, which while effective makes it difficult to interpret the contribution of individual features in a manner similar to tree-based models [31].

The experimental findings clearly demonstrate that XGBoost is the most effective model in this study, both in terms of classification accuracy and clinical reliability. The use of SMOTE played an important role in ensuring fair evaluation across classes particularly for the malignant class.

This research has few limitations, which need to be noted. The analysis was conducted with a comparatively small dataset (Wisconsin Diagnostic Breast Cancer dataset), and its impact might reduce the generalizability of the results. The research did not encompass external verification using independent datasets. Future research will concentrate on remedying these points by using larger and more diverse datasets and conducting external verification.

Conclusions

This study conducted a comparative evaluation of three supervised ML algorithms - SVC, RF and XGBoost - for classifying breast cancer using the Wisconsin Diagnostic Breast Cancer dataset. The dataset was

balanced using SMOTE, and each model was fine-tuned through GridSearchCV to optimize performance. Among the classifiers, XGBoost delivered the best results with an accuracy of 98.60% and a perfect recall for malignant cases, making it particularly effective in reducing false negatives - a critical aspect in clinical diagnostics. ROC curve and confusion matrix analyses further supported the strong performance of both XGBoost and RF in accurately distinguishing between benign and malignant tumors. Feature importance plots from the tree-based models indicated that attributes such as perimeter_worst and concave points_worst were most influential in classification. While SVC lacks built-in feature importance due to its kernel-based design, it still maintained competitive accuracy above 90%.

To conclude, this study presents an interpretable and optimized ML framework for breast cancer prediction, with XGBoost emerging as the most promising model. Future research will focus on validation with larger and diverse clinical datasets and real-world integration into diagnostic support tools.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Ritika Bansal, Deepti Sharda, Prachi Gumber, Swati Kandari, Indu Arora

Acquisition, analysis, or interpretation of data: Ritika Bansal, Deepti Sharda, Prachi Gumber, Swati Kandari, Indu Arora

Drafting of the manuscript: Ritika Bansal, Deepti Sharda, Prachi Gumber, Swati Kandari, Indu Arora

Critical review of the manuscript for important intellectual content: Ritika Bansal, Deepti Sharda, Prachi Gumber, Swati Kandari, Indu Arora

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. World Health Organization: Breast cancer. (2024). Accessed: May 19, 2025: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.
2. Duffy SW, Tabar L, Olsen AH, et al.: Absolute numbers of lives saved and overdiagnosis in breast cancer screeninq, from a randomized trial and from the breast screening programme in England. Journal of Medical Screening. 2010, 17:25-30. 10.1258/jms.2009.009094
3. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S: Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017, 542:115-118. 10.1038/nature21056
4. Litjens G, Kooi T, Ehteshami Bejnordi B, et al.: A survey on deep learning in medical image analysis. Medical Image Analysis. 2017, 42:60-88. 10.1016/j.media.2017.07.005
5. Araújo T, Aresta G, Castro E, et al.: Classification of breast cancer histology images using convolutional neural networks. PLoS ONE. 2017, 12:e0177544. 10.1371/journal.pone.0177544
6. Senturk ZK, Kara R: Breast cancer diagnosis via data mining: Performance analysis of seven different algorithms. Computer Science & Engineering: An International Journal. 2014, 4:35-46.
7. Bazazeh D, Shubair R: Comparative study of machine learning algorithms for breast cancer detection and diagnosis. 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, United Arab Emirates. 2016, 1-4. 10.1109/ICEDSA.2016.7818560
8. Naji MA, El Filali S, Aarika K, Benlahmar EH, Ait Abdelouhahid R, Debauche O: Machine learning algorithms for breast cancer prediction and diagnosis. Procedia Computer Science. 2021, 191:487-492. 10.1016/j.procs.2021.07.062
9. Ahmad Ubaidillah SHS, Sallehuddin R, Ali NA: Cancer detection using artificial neural network and support vector machine: A comparative study. Jurnal Teknologi. 2013, 65:33-38. 10.11113/jt.v65.1788
10. Liew XY, Hameed N, Clos J: An investigation of XGBoost-based algorithm for breast cancer classification. Machine Learning with Applications. 2021, 6:100154. 10.1016/j.mlwa.2021.100154
11. Hoque R, Das S, Hoque M, Haque E: Breast cancer classification using XGBoost. World Journal of Advanced Research and Reviews. 2024, 21:1985-1994. 10.30574/wjarr.2024.21.2.0625
12. Sinha NK, Khulal M, Gurung M, Lal A: Developing a web based system for breast cancer prediction using XGboost classifier. International Journal of Engineering Research & Technology. 2020, 9:852-856. 10.17577/IJERTV9IS060612

13. Rohan TI, Awan-Ur-Rahman, Siddik AB, Islam M, Yusuf MSU: A precise breast cancer detection approach using ensemble of random forest with AdaBoost. 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh. 2019, IEEE:1-4. 10.1109/IC4ME247184.2019.9036697
14. Sugiharti E, Arifudin R, Wiyanti DT, Susilo AB: Convolutional neural network-XGBoost for accuracy enhancement of breast cancer detection. Journal of Physics: Conference Series. 2021, 1918:042016. 10.1088/1742-6596/1918/4/042016
15. Mahesh TR, Kumar VV, Muthukumaran V, Shashikala HK, Swapna B, Guluwadi S: Performance analysis of XGBoost ensemble methods for survivability with the classification of breast cancer. Journal of Sensors. 2022, 2022:4649510. 10.1155/2022/4649510
16. Minnoor M, Baths V: Diagnosis of breast cancer using random forests. Procedia Computer Science. 2023, 218:429-437. 10.1016/j.procs.2023.01.025
17. Anisha PR, Reddy CKK, Apoorva K, Mangipudi CM: Early diagnosis of breast cancer prediction using random forest classifier. IOP Conference Series: Materials Science and Engineering. 2021, 1116:012187. 10.1088/1757-899X/1116/1/012187
18. Bergstra J, Bengio Y: Random search for hyper-parameter optimization. Journal of Machine Learning Research. 2012, 13:281-305.
19. Snoek J, Larochelle H, Adams RP: Practical Bayesian optimization of machine learning algorithms. NIPS'12: Proceedings of the 26th International Conference on Neural Information Processing Systems. 2:2951-2959.
20. Hutter F, Kotthoff L, Vanschoren J: Automated Machine Learning: Methods, Systems, Challenges. Springer, Cham; 2019. 10.1007/978-3-030-05318-5
21. Wolberg W, Mangasarian O, Street N, Street W: Breast Cancer Wisconsin (Diagnostic) [Dataset]. UCI Machine Learning Repository. (2019). https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic.
22. McKinney W: Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference. 2010, 445:56-61. 10.25080/Majora-92bf1922-00a
23. Guyon I, Elisseeff A: An introduction to variable and feature selection. Journal of Machine Learning Research. 2003, 3:1157-1182.
24. Han J, Kamber M, Pei J: Data Mining: Concepts and Techniques. Elsevier, Cambridge, MA; 2011.
25. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002, 16:321-357. 10.1613/jair.953
26. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F: Learning from Imbalanced Data Sets. Springer, Switzerland; 2018. 10.1007/978-3-319-98074-4
27. Pedregosa F, Varoquaux G, Gramfort A, et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011, 12:2825-2830.
28. Cortes C, Vapnik V: Support-vector networks. Machine Learning. 1995, 20:273-297. 10.1007/BF00994018
29. Breiman L: Random forests. Machine Learning. 2001, 45:5-32. 10.1023/A:1010933404324
30. Chen T, Guestrin C: XGBoost: A scalable tree boosting system. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016, 785-794. 10.1145/2939672.2939785
31. Hastie T, Tibshirani R, Friedman J: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer, New York, NY; 2009. 10.1007/978-0-387-84858-7