Science in Information Technology Letters
Vol. 6., No. 1, May 2025, pp. 16-33
ISSN 2722-4139
http://pubs2.ascee.org/index.php/sitech

Optimizing breast cancer classification using SMOTE, Boruta, and XGBoost

Cicin Hardiyanti P [a,1,*]

[a] Informatics Department, Universitas Alma Ata, Daerah Istimewa Yogyakarta, Indonesia
[1] cicinhardiyanti@almaata.ac.id
* Corresponding Author

ARTICLE INFO

Article history
Received April 22, 2025
Revised May 17, 2025
Accepted May 28, 2025

Keywords
Breast cancer
SMOTE
Boruta
XGBoost
Imbalanced data handling

ABSTRACT

Breast cancer remains one of the leading causes of death among women worldwide. This study aims to develop a clinical data-based breast cancer classification framework by integrating the Synthetic Minority Oversampling Technique (SMOTE), the Boruta feature selection algorithm, and the XGBoost classifier. The proposed approach is tested using the Wisconsin Breast Cancer Diagnostic (WBCD) dataset, consisting of 569 samples and 30 numerical features. SMOTE addresses class imbalance, Boruta selects the most relevant diagnostic features, and XGBoost is the main classification algorithm due to its tabular and imbalanced data robustness. Model validation is conducted through Repeated Stratified K-Fold Cross Validation with 30 repetitions to ensure statistical stability. The resulting model achieves excellent classification performance, with an average accuracy of 0.9608 ± 0.0274, precision of 0.9465 ± 0.0481, Recall of 0.9512 ± 0.0524, and F1-score of 0.9475 ± 0.0374. The ROC-AUC value reaches 0.9926 ± 0.0094, the PR-AUC is 0.9906 ± 0.0113, and the Matthews Correlation Coefficient (MCC) is 0.9179 ± 0.0575, indicating a well-balanced model. Clinically, this model can aid early diagnosis by effectively reducing irrelevant diagnostic attributes, retaining only 10 key features without compromising accuracy, thereby offering a lightweight yet reliable diagnostic tool. However, limitations include the relatively small dataset and the absence of hyperparameter tuning. Future research should explore larger datasets, advanced ensemble methods, and interpretability techniques such as SHAP or LIME to improve clinical transparency and adoption.

1. Introduction

Breast cancer is one of the most common types of cancer affecting women worldwide and remains a leading cause of cancer-related deaths. According to the Global Cancer Observatory report from the World Health Organization (WHO), breast cancer caused 670,000 deaths globally in 2022 [1]. About half of all breast cancer cases occur in women without specific risk factors other than gender and age. In 2022, breast cancer was the most common type of cancer among women in 157 out of 185 countries. It occurs in every country around the world. In the United States, according to the American Cancer Society, breast cancer is the most common cancer among women and a leading cause of cancer-related

deaths. A similar situation exists in Indonesia. Globocan reported that in 2022, there were 66,271 new cases of breast cancer, with a total of 209,748 deaths over the past five years [2].

Breast cancer occurs when abnormal cells develop and spread uncontrollably throughout the body [3]–[5]. Cancer cells first appear in the milk ducts of the breast and spread through other breast tissues, eventually leading to tumors and can spread to other organ tissues [6]. In general, two types of abnormal cells can be detected, namely benign cells and cancer cells. Benign cells are characteristic of not spreading and not damaging the surrounding tissues [7], [8] while cancer cells can spread rapidly and damage the tissues present in the affected individual [9]. Many patients realise they have this disease when it has reached an advanced stage, resulting in a low success rate for treatment. The main contributing factors are low public awareness of early detection and limited diagnostic facilities, especially in remote areas.

Early detection is essential to increase recovery chances and reduce mortality risks. Recently, machine learning (ML) technology based on clinical data has been introduced as an alternative approach to assist the diagnostic process. ML algorithms can analyse patterns in patient data, such as blood test results, anthropometric data, or biopsy findings, to classify whether a tumour is benign or malignant. One commonly used dataset is the Wisconsin Breast Cancer Dataset (WBCD) from the UCI Machine Learning Repository or Kaggle, which provides information from 569 samples with diagnostic features.

Although various breast cancer diagnostic approaches have been developed, many challenges remain unaddressed, particularly when using real-world clinical data. Class imbalance, where benign tumor samples significantly outnumber malignant ones, often leads to misclassification, making models less accurate in detecting actual cancer cases. This misclassification may cause delayed treatment and pose serious risks to patient safety. Moreover, not all attributes in clinical data contribute meaningfully to classification. Including less relevant features may degrade model performance and increase the risk of overfitting. Thus, effective feature selection methods are required to filter out unimportant features and preserve model performance.

Ensemble-based classification algorithms like XGBoost have performed excellently in various competitions and research studies. However, their implementation in real-world breast cancer cases requires systematic validation, particularly due to the complexities of imbalanced data. Combining data balancing techniques, optimal feature selection, and appropriate algorithms may offer a more robust approach for data-driven clinical diagnosis. ML-based methods have now become a prominent choice in breast cancer classification. Several algorithms have been applied, including Naïve Bayes, K-Nearest Neighbour (KNN), Decision Tree, Support Vector Machine (SVM), and ensemble methods like Random Forest and XGBoost.

Several studies using Naïve Bayes algorithms for breast cancer classification have shown high accuracy, such as those conducted by Oktavianto and Handri [10] with an average accuracy of 96.9%. Research by Muntiari & Hanif [11] compared seven algorithms, including SVM, Random Forest, and Neural Network, concluding that some classical methods, such as Decision Tree and KNN, provide high accuracy. However, the evaluation only used one metric (accuracy), and there was still a lack of in-depth analysis, such as ROC-AUC, PR-AUC, or MCC.

The ensemble classifier approach has also been studied to combine the strengths of various models. Khadijah & Kusumaningrum [12] combine SVM, ELM, and KNN through majority voting, increasing accuracy in several scenarios. However, it is limited to a small dataset (116 samples) and has not yet utilised data balancing techniques. Meanwhile, Jamaludin et al. (2024) [13] compared Random Forest

and Neural Network, with Random Forest performing better with an accuracy of 98.86%. However, the modelling is limited to basic evaluation and does not include ROC or PR curve visualisation.

Deep learning-based approaches also demonstrate high performance. Erwandi & Suyanto [14] used the ResNet50 architecture and achieved an accuracy of 99.3% for binary classification. However, this study focuses on histopathological image data and does not use interpretability metrics such as feature importance, which are important for medical applications. The study by Supriyanto et al. (2022) [15] using the Inception-V3 architecture combined with various machine learning algorithms shows that the combination of CNN and Logistic Regression provides the best accuracy of 93% for images magnified 40x. Nevertheless, this study has not utilised feature selection techniques and feature weighting..

Other studies also showed the success of ensemble and SVM models in breast cancer classification. Mohammed Amine Naji et al. [16]. SVM has the highest accuracy (97.2%), while Sharmin Ara et al [17]. Show SVM and Random Forest as the best methods with an accuracy of 96.5%. The research by Naufal Cahya Ramadhan et al. (2024) [18] integrates SMOTE and feature selection XGBoost into the KNN, Naïve Bayes, and Random Forest models, and achieves performance improvements, particularly in Random Forest (accuracy 98%, AUC 94%). Although many approaches have been developed, integrating data balancing, stable feature selection, and cross-fold performance evaluation remains underexplored within a single framework.

This study aims to integrate SMOTE, Boruta, and XGBoost into a unified framework to address class imbalance and feature redundancy in breast cancer classification. The proposed model is validated using Repeated Stratified K-Fold Cross Validation (30 folds) and evaluated using multiple performance metrics (accuracy, precision, Recall, F1-score, ROC-AUC, PR-AUC, and MCC). Feature importance analysis is performed using Boruta and XGBoost to enhance clinical relevance, providing insights into the most influential clinical features.

This research makes several contributions:

- Proposing an integrated classification framework combining SMOTE, Boruta, and XGBoost for breast cancer classification based on clinical data..
- Conducting a comprehensive evaluation using robust statistical validation and diverse performance metrics, while enhancing clinical interpretability through feature importance analysis to support informed medical decision-making.

2. Method

2.1. Research stages

The stages in this research broadly consist of data collection, preprocessing, correlation analysis and multicollinearity reduction, feature selection, handling imbalanced data, classification, and evaluation. The stages can be seen in Fig. 1. Research stages.

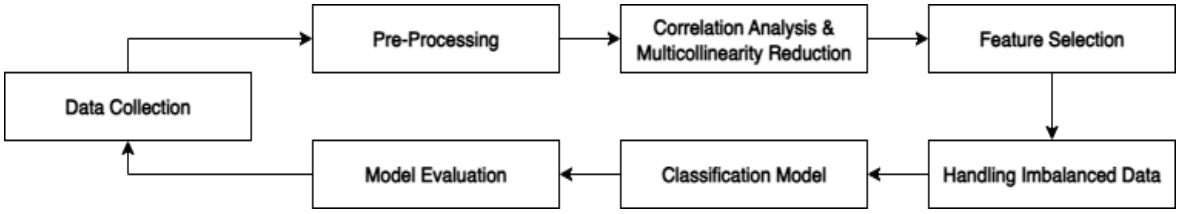

Fig. 1. Research stages

2.1.1. Data Collection

This research uses public data from kaggle.com (https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data). This dataset contains 569 samples of breast tissue examination results. Each sample consists of 30 numerical features that describe the morphological characteristics of cell nuclei from microscopic images. These features are generated from digital image analysis of fine-needle aspiration (FNA) procedures on breast tissue masses. There are ten main cell characteristics measured, namely: radius (the average distance from the center to the edge of the cell), texture (standard deviation of grayscale values), perimeter (circumference of the cell), area (surface area of the cell), smoothness (local variation in radius length), compactness (ratio of the square of the perimeter to the area, minus 1), concavity (the degree of indentation in the contour), concave points (the number of concave points on the contour), symmetry (the degree of symmetry of the cell), and fractal dimension (the complexity of the cell contour resembling the length of a coastline).

2.1.2. Preprocessing

Preprocessing is the initial step in data processing, carried out to clean the data, prepare the data, and align the data for optimal use [19], [20]. The first step in preprocessing is the removal of irrelevant columns, such as id and Unnamed: 32, as shown in Table 1, which do not provide information for the classification process and only serve as administrative identification. Removing these columns is important to avoid noise disrupting the model's performance.

Table 1. Initial data before column deletion

id	diagnosis	radius_mean	texture_mean	...	Unnamed: 32
842302	M	17.99	10.38	...	NaN
842517	M	20.57	17.77	...	NaN
84300903	M	19.69	21.25	...	NaN
84348301	M	11.42	20.38	...	NaN

The next step is the encoding process for the diagnosis labels, where the categories M (Malignant) and B (Benign) are converted into binary numeric form, namely 1 for M and 0 for B, as shown in Table 2.

Table 2. Data before conversion

diagnosis	radius_mean	texture_mean
M	17.99	10.38
...	...	...
B	13.540	14.36

This conversion allows the target data to be recognized and processed by machine learning algorithms requiring a numeric representation, data after conversion, as shown in Table 3.

Table 3. Data after conversion

diagnosis	radius_mean	texture_mean
1	17.99	10.38
...	...	...
0	13.540	14.36

Next, handling of missing values is carried out using a simple imputation method with SimpleImputer from the Scikit-learn library, using the mean strategy. Each numeric feature that contains missing values will be filled with the average value of that feature, thus avoiding errors during model training due to incomplete data.

2.1.3. Correlation Analysis and Multicollinearity Reduction

After the data cleaning phase, the next step is to conduct a correlation analysis between features to evaluate potentially strong linear relationships between pairs of features, as shown in Table 4. High correlation between features can lead to multicollinearity issues, which can ultimately disrupt model interpretation and cause information redundancy [21].

Table 4. Features with high correlation (>0.9)

	Feature 1	Feature 2	Correlation
0	perimeter_mean	radius_mean	0.997855
1	perimeter_worst	radius_worst	0.993708
2	area_mean	radius_mean	0.987357
3	area_mean	perimeter_mean	0.986507
4	area_worst	radius_worst	0.984015
5	area_worst	perimeter_worst	0.977578
6	perimeter_se	radius_se	0.972794
7	perimeter_worst	perimeter_mean	0.970387
8	radius_worst	radius_mean	0.969539
9	radius_worst	perimeter_mean	0.969476
10	perimeter_worst	radius_mean	0.965137
11	radius_worst	area_mean	0.962746
12	area_worst	area_mean	0.959213
13	perimeter_worst	area_mean	0.959120
14	area_se	radius_se	0.951830
15	area_worst	perimeter_mean	0.941550
16	area_worst	radius_mean	0.941082
17	area_se	perimeter_se	0.937655
18	concave points_mean	concavity_mean	0.921391
19	texture_worst	texture_mean	0.912045
20	concave points_worst	concave points_mean	0.910155

Correlation analysis was performed using Pearson's correlation coefficient to understand the structure of relationships between numeric features in the dataset. The correlation between features was visualised as a heatmap to identify potential multicollinearity before further feature selection with the Boruta algorithm. Dark red indicates a high positive correlation, while blue indicates a high negative correlation. Zero or near-zero correlation is visualised in white or light grey. Due to the symmetry of the correlation matrix, this heatmap is symmetric with respect to the main diagonal, displaying a value of 1 as the correlation of a feature with itself. Based on Fig. 2, it was found that several pairs of features have very high correlations ($r > 0.9$), including: 1) radius_mean, perimeter_mean, and area_mean; 2) concavity_mean and concave_points_mean; 3) radius_worst, perimeter_worst and area_worst.

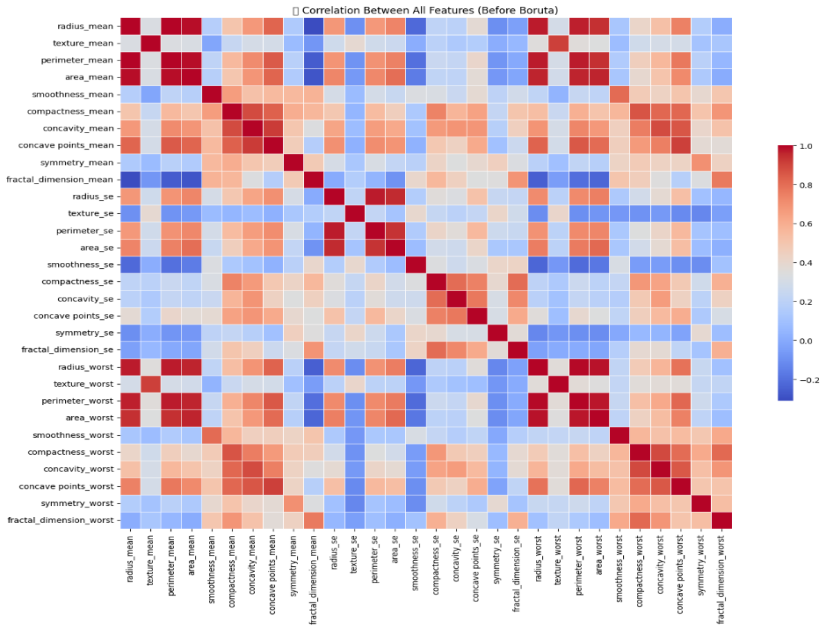

Fig. 2. Correlation Between All Features (Before Boruta)

2.1.4. Feature Selection with Boruta

To filter features relevant to the prediction target, feature selection is performed using the Boruta algorithm, a wrapper method based on Random Forest designed to identify all important features in a dataset [22]–[25]. This algorithm works by comparing the importance of the original features against the 'shadow' features (random features generated through permutation) and retaining the original features that are statistically more relevant than the random features. The Boruta algorithm uses the Random Forest Classifier as the main estimator in this study. Boruta is conservative and iterative; thus, it can preserve important features and eliminate features that do not contribute significantly [26], [27]. This process results in a more concise and meaningful feature subset, which is then used as the classification model's main input, as shown in Fig. 3.

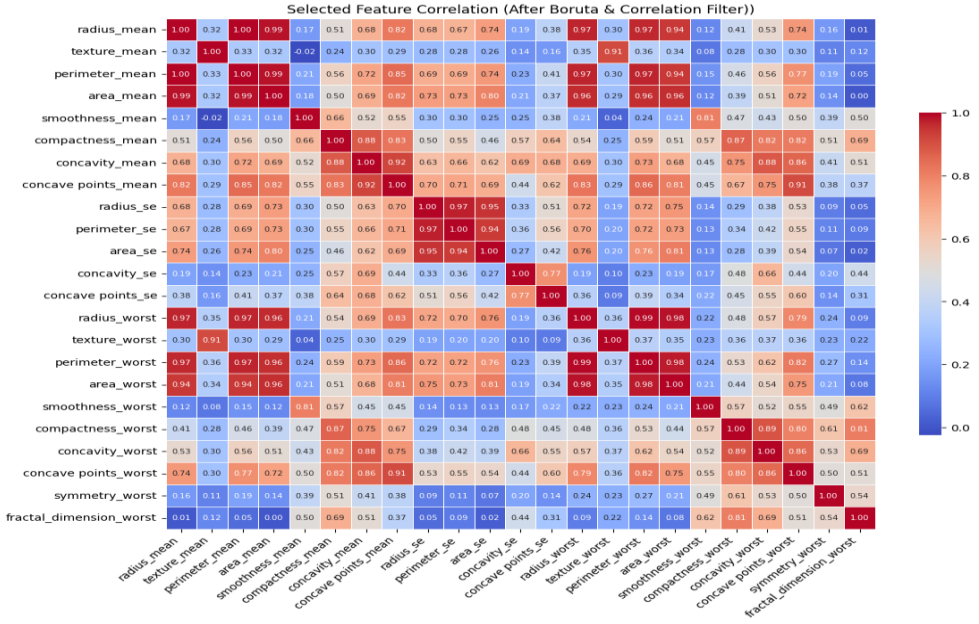

Fig. 3. Selected Feature Correlation (After Boruta & Correlation Filter)

2.1.5. Handling Imbalanced Data

A common challenge in disease diagnosis classification is class imbalance, where the number of samples in one class (usually benign) is significantly greater than in the other class (malignant) [28], [29]. This imbalance can lead to a model being biased towards the majority class and neglecting detection in the minority class, which in the medical context is the most crucial case. To address this issue, this study applies the Synthetic Minority Oversampling Technique (SMOTE). In this research, SMOTE is specifically applied to the training data in each iteration of cross-validation (Repeated Stratified K-Fold), and is not applied to the test data to maintain the accuracy of evaluation. This strategy is implemented so the test data's distribution reflects the dataset's original distribution, while the model is trained on balanced data. Thus, the model's performance is measured to reflect the true generalisation capability of new data. The implementation of SMOTE is carried out using the imblearn library with default parameters, and it is applied after the training and testing data split in each fold. This technique allows for an increase in the model's sensitivity (Recall) towards cancer cases without decreasing overall accuracy and produces a more balanced model in detecting both classes.

2.1.6. Classification Model

The classification model used in this study is XGBoost (Extreme Gradient Boosting), an efficient and accurate boosting algorithm, especially for tabular data. XGBoost can handle small to large datasets and address overfitting through regularisation and model complexity management [30]–[33]. The main parameters of the XGBoost model used include the number of trees (n_estimators) set to 100, maximum tree depth (max_depth) of 5, and learning rate (learning_rate) of 0.1. The scale_pos_weight parameter is set to a default value of 1 with no adjustments to the class distribution, and the evaluation function during training is adjusted to logloss as the primary metric. The primary metric maintains result reproducibility, with the random_state value set to 42.

To evaluate the impact of features that have high correlation on model performance, a comparison is made between two classification approaches: the model trained using all original features and the model filtered based on the correlation between features before feature selection using the Boruta algorithm. This experiment aims to assess the extent to which the initial filtering process of highly correlated features can influence the results of feature selection as well as the final performance of the classification model.

The final stage of the optimisation process, the best classification model is formed by combining three main components: handling class imbalance using Synthetic Minority Over-sampling Technique (SMOTE) on the training data of each fold, relevant feature selection using the Boruta algorithm, and classification using XGBoost with the specified parameter configuration. This strategy is designed to produce a cancer classification model optimal for breasts, considering common challenges in medical data such as imbalanced class distribution and multicollinearity among input features.

2.1.7. Model Evaluation

To ensure the validity and stability of the model's performance, a Repeated Stratified K-Fold Cross-Validation technique was used, which is a layered cross-validation method that divides the data into 10 folds with a consistent class proportion in each fold, and repeats the process 3 times. Thus, there are a total of 30 evaluation scenarios that provide a more robust and reliable estimate of the model's performance against variations in the training data.

The model's performance is evaluated using various metrics that reflect the accuracy and quality of classification from different aspects. The first metric used is accuracy, which represents the proportion of correct predictions to the total amount of data, and is formulated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision measures the accuracy of the model in predicting the positive class, and is defined as:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall or sensitivity measures the model's ability to detect all actual positive data, with the formula:

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

To balance precision and Recall, the F1-score is used, which is the harmonic mean between the two, formulated as:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

Furthermore, an evaluation was performed using Receiver Operating Characteristic - Area Under Curve (ROC-AUC), which measures the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR), with TPR and FPR calculated as:

$$PR = \frac{TP}{TP+FN}, \; FPR = \frac{FP}{FP+TN} \quad (5)$$

An evaluation that considers all elements in the confusion matrix more evenly uses the Matthews Correlation Coefficient (MCC), which is formulated as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (6)$$

In addition to numerical evaluation, visualisation provides a more comprehensive understanding of the model's behaviour. The visualisation includes the confusion matrix for the first fold as an illustration of the distribution of correct and incorrect predictions, a boxplot showing the variation of performance metrics across all 30 cross-validation folds, and ROC and Precision-Recall curves for each fold and in aggregate form (mean curve).

3. Results and Discussion

3.1. Initial Experiment: Impact of Removing Highly Correlated Features

This initial experiment was conducted to evaluate the impact of removing features that have a high correlation on the performance of the classification model. The goal is to identify whether reducing features based on correlation can improve predictive accuracy without sacrificing important information. In this case, features with an absolute Pearson correlation value greater than 0.9 were identified as redundant and removed from the data subset. This testing was carried out by comparing two scenarios. The XGBoost model was trained using all features (X_imputed), 30 original features from the dataset and the XGBoost model was trained after removing high correlation features (X_no_high_corr).

Both models were trained using XGBClassifier from the xgboost library with default parameters, including eval_metric=logloss, and consistent data splitting using train_test_split from scikit-learn with an 80:20 ratio and random_state=42 to maintain reproducibility. Evaluation was conducted by calculating the accuracy of the test data. The comparison results are visualised using bar graphs as shown in Fig. 4.

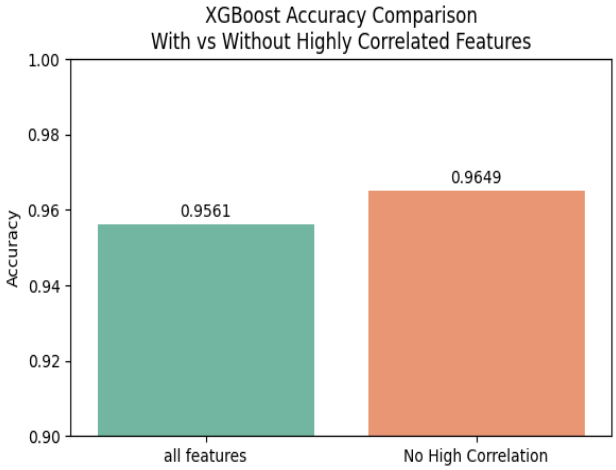

Fig. 4. Comparison of the accuracy of the XGBoost model between the use of all features and after the removal of highly correlated features

The results show that the model trained with all features achieved an accuracy of 0.9561, while the model with reduced features achieved a higher accuracy of 0.9649. Removing highly correlated features can enhance model performance, possibly because the model becomes simpler and free from features that carry redundant information. However, it is important to note that during the advanced feature selection stage using the Boruta algorithm, several features that statistically had a high correlation were still retained. This indicates that even though these features correlate, they are still considered predictively relevant to the target label by decision tree-based algorithms. In other words, correlation among features does not necessarily mean that one of those features is not helpful in the context of ensemble models like XGBoost; redundant information can still have contributions.

Therefore, the results of this preliminary experiment provide an initial justification for cleaning features before further selection stages, while still highlighting the importance of model-based feature selection techniques like Boruta, so that important features are not prematurely removed just because of statistical relationships between features.

3.2. Final Model Evaluation: SMOTE + Boruta + XGBoost

Per-fold model evaluation is an important component in validating the performance of machine learning models, especially in the medical context, which requires a high level of trust in prediction results [34]. In this experiment, a 30-fold Repeated Stratified K-fold cross-validation approach was used, which allows for a comprehensive measurement of the variability of model performance on varied training data. This provides a more accurate estimate of generalisation compared to single testing. Based on Table 5, the model shows consistently high performance in almost all folds, with some folds even achieving perfect performance (such as Fold 3, 10, and 21) with Accuracy, Precision, Recall, F1-Score, ROC-AUC, MCC, and PR-AUC values of 1.0000. This indicates that on specific subsets of data, the XGBoost model can completely distinguish between benign and malignant classes without prediction errors. There are also some folds with relatively lower performance, such as: Fold 4: Recall = 0.7619 and F1-score =

0.8205, Fold 8: Precision = 0.8077, although Recall = 1.0000, Fold 22: Recall = 0.8636 and F1-score = 0.9048.

Table 5. Evaluation of each fold

	Fold	Accuracy	Precision	Recall	F1-Score	ROC-AUC	MCC	PR-AUC
0	1	0.9474	0.8800	1.0000	0.9362	1.0000	0.8970	1.0000
1	2	0.9123	0.8696	0.9091	0.8889	0.9909	0.8170	0.9867
2	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	4	0.8772	0.8889	0.7619	0.8205	0.9762	0.7330	0.9643
4	5	0.9649	1.0000	0.9048	0.9500	0.9696	0.9258	0.9694
5	6	0.9649	0.9524	0.9524	0.9524	0.9987	0.9246	0.9978
6	7	0.9825	1.0000	0.9524	0.9756	1.0000	0.9626	1.0000
7	8	0.9123	0.8077	1.0000	0.8936	0.9987	0.8340	0.9978
8	9	0.9474	0.9500	0.9048	0.9268	0.9735	0.8864	0.9715
9	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	11	0.9825	0.9565	1.0000	0.9778	1.0000	0.9639	1.0000
11	12	0.9649	0.9167	1.0000	0.9565	0.9974	0.9297	0.9962
12	13	0.9649	0.9524	0.9524	0.9524	0.9974	0.9246	0.9959
13	14	0.9649	1.0000	0.9048	0.9500	1.0000	0.9258	1.0000
14	15	0.9825	0.9545	1.0000	0.9767	0.9974	0.9633	0.9956
15	16	0.9474	0.9500	0.9048	0.9268	0.9788	0.8864	0.9763
16	17	0.9825	1.0000	0.9524	0.9756	0.9854	0.9626	0.9836
17	18	0.9649	0.9524	0.9524	0.9524	0.9974	0.9246	0.9959
18	19	0.9474	0.9500	0.9048	0.9268	0.9894	0.8864	0.9851
19	20	0.9464	0.9091	0.9524	0.9302	0.9932	0.8874	0.9887
20	21	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
21	22	0.9298	0.9500	0.8636	0.9048	0.9818	0.8518	0.9763
22	23	0.9474	0.9091	0.9524	0.9302	0.9960	0.8886	0.9936
23	24	0.9474	0.8750	1.0000	0.9333	0.9987	0.8956	0.9978
24	25	0.9825	0.9545	1.0000	0.9767	1.0000	0.9633	1.0000
25	26	0.9649	0.9524	0.9524	0.9524	0.9987	0.9246	0.9978
26	27	0.9825	1.0000	0.9524	0.9756	0.9987	0.9626	0.997
27	28	0.9825	1.0000	0.9524	0.9756	0.9868	0.9626	0.9846
28	29	0.9474	0.9091	0.9524	0.9302	0.9749	0.8886	0.9659
29	30	0.9821	0.9545	1.0000	0.9767	1.0000	0.9630	1.0000

These folds show a momentary imbalance between the model's ability to recognise positive and negative cases. For example, in Fold 4, although the precision is high (0.8889), the lower Recall indicates the presence of false negatives, which are cancer cases that were not detected, a condition that must be avoided in medical practice. Meanwhile, Fold 8 shows the opposite situation: perfect Recall but low precision, meaning the model identifies all cancer patients and misclassifies some healthy patients as positive (false positives). The F1-score value, as a harmonic metric between precision and Recall, has a relatively narrow range, from 0.8205 (Fold 4) to 1.0000 (Fold 3, 10, 21). This indicates that the model can consistently balance prediction performance across both target classes.

The Matthews Correlation Coefficient (MCC) values are also high and consistent, mostly above 0.88 with an average of around 0.92, indicating symmetric performance in imbalanced binary classification. ROC-AUC and PR-AUC reached the maximum value (1.0) in most folds, indicating excellent class discrimination ability and strong resilience to imbalanced data.

The results of this evaluation indicate that after applying the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance and the Boruta algorithm for feature selection, the performance of the XGBoost model significantly improved compared to the initial baseline. The model became more accurate and more stable when handling data with complex and variable distributions. Integrating SMOTE and Boruta has been shown to improve generalisation and reduce the potential for overfitting.

3.2.1. Analysis of Confusion Matrix Fold-1

An analysis was conducted on the confusion matrix in Fold-1 to evaluate the model's performance in more depth. This analysis aims to identify the model's ability to distinguish between breast cancer cases (malignant) and non-cancer cases (benign) based on the predictions produced by the combination of SMOTE, Boruta, and XGBoost. Fig. 5 Confusion Matrix Fold-1.

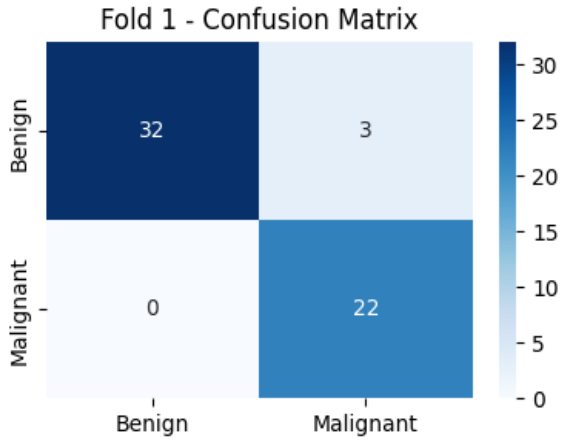

Fig. 5. Confusion Matrix Fold -1

Based on Fig. 5, the True Negative (TN) value is 32, the number of benign cases classified correctly as benign. False Positive (FP) is 3, the number of benign cases incorrectly classified as malignant. False Negative (FN) is 0, meaning no malignant cases went undetected. True Positive (TP) is 22, the number of malignant cases correctly classified as malignant. Based on these values, performance metrics can be calculated, accuracy 0.9474, precision 0.8800, recall 1.0000 and F1-Score 0.9362.

The confusion matrix results in Fold-1 indicate that the model has perfect Recall (1.0), signifying that all cancer cases were correctly identified, which is an important advantage in the medical field to avoid life-threatening false negatives. However, three false positive cases reduced the precision to about 0.88, indicating that approximately 12% of cancer predictions were incorrect. Although this may cause anxiety or require additional examinations, it is still tolerable in the context of early detection. The model demonstrates excellent performance with high sensitivity and a low error rate.

3.2.2. Average and Standard Deviation of Matrix

The average and standard deviation of various evaluation metrics were obtained after evaluating the model on all folds (30 folds). Table 6 shows a 30-fold cross-validation evaluation; the model shows strong and stable predictive performance. The average accuracy is 96.08% with a standard deviation of 2.74%, indicating consistent classification ability and good generalisation across various data subsets. On the precision side, it is 94.65% with a standard deviation of 4.81%. Most positive predictions are indeed cancer cases, resulting in a low risk of overdiagnosis. The Recall is 95.12% ± 5.24% indicating low false

negatives. This is very important in the medical context, as errors in detecting cancer cases can have fatal consequences. The F1-score (94.75% ± 3.74%) shows a balance between precision and sensitivity, reinforcing the model's stability. The high ROC-AUC value (99.26% ± 0.94%) demonstrates excellent ability to distinguish between cancer and non-cancer across various classification thresholds. The Matthews Correlation Coefficient (91.79% ± 5.75%) underscores the overall prediction balance. Finally, a PR-AUC of 99.06% ± 1.13% indicates the model's robustness in maintaining the quality of positive detection, especially for the minority class.

Table 6. Average and Standard Deviation Matrix

	Average	Standard Deviation
Accuracy	0.9608	0.0274
Precision	0.9465	0.0481
Recall	0.9512	0.0524
F1-Score	0.9475	0.0374
ROC-AUC	0.9926	0.0094
MCC	0.9179	0.0575
PR-AUC	0.9906	0.0113

Overall, the SMOTE, Boruta, and XGBoost combination produces an accurate, robust, and reliable model for detecting breast cancer. The model combining SMOTE, Boruta, and XGBoost demonstrates superior and consistent predictive performance across 30 cross-validation scenarios. The average values of all metrics are high with slight variation, which means the model is accurate but also stable and reliable for implementation in real clinical contexts, particularly for early and accurate breast cancer detection.

3.2.3. Matrix Distribution Visualisation

To evaluate the performance of the breast cancer classification model, Repeated Stratified K-Fold was run 30 times, measuring seven key metrics: accuracy, precision, Recall, F1 score, ROC-AUC, MCC, and PR-AUC, as shown in Fig. 6.

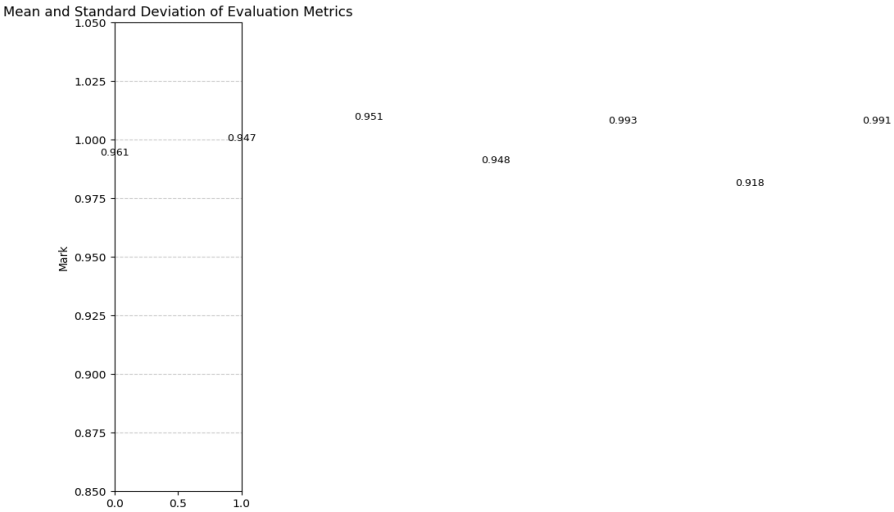

Fig. 6. Matrix Distribution Visualisation

The results show excellent and stable model performance, with a ROC-AUC of 0.993 and a PR-AUC of 0.991, reflecting a very high ability to distinguish between classes, even under conditions of class imbalance. The recall value (0.951) and F1-score (0.947) indicate a balance between sensitivity and

precision. In contrast, the precision (0.946) is slightly lower but still acceptable in a medical context, as the main priority is to reduce Type II errors (false negatives). The MCC value of 0.918 indicates a strong correlation between predictions and actual labels, which is important for evaluating imbalanced data. The low standard deviation across all metrics indicates the stability and generalizability of the model across different folds.

The bar chat visualization shown in Fig. 7 summarizes descriptive statistics in the form of a bar chart showing the model's average evaluation metric values, accompanied by error bars indicating the standard deviation across folds. This presentation aims to quantitatively assess the model's performance stability within a framework of 30-fold cross-validation.

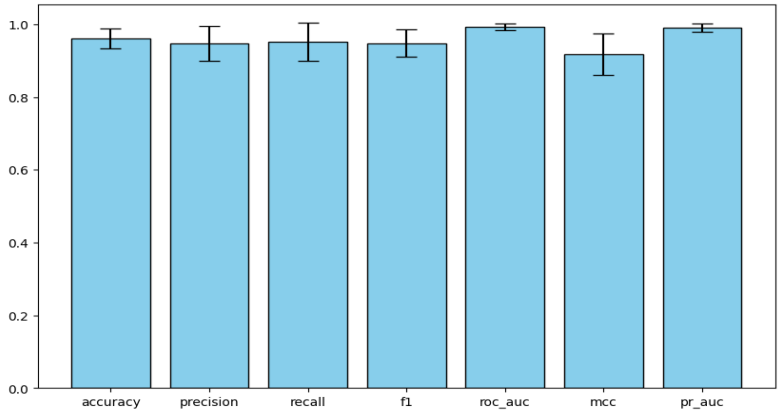

Fig. 7. Bar Chat Visualisation

Figure 7 shows that all model evaluation metrics are at a high performance level, with an average score above 0.90 indicating excellent classification ability. Here are some important points:

- ROC-AUC and PR-AUC are the highest metrics with the smallest error bars, reflecting the model's strong and stable ability to differentiate between benign and malignant classes. This is crucial for medical diagnosis applications where detection accuracy is critical.
- Accuracy, Precision, Recall, and F1-Score also demonstrate high and consistent performance. Although Precision and Recall have slightly larger fluctuations, their values remain within reasonable limits, indicating the model's capability to detect cancer cases while effectively minimising false positives.
- The Matthews Correlation Coefficient (MCC) is slightly lower and more variable, but still shows good values, indicating that the model is not biased towards either class despite the imbalanced data.

Overall, this graph confirms that the combination of SMOTE, Boruta, and XGBoost produces a strong, stable model with small inter-fold variation that is feasible to apply in medical classification.

3.2.4. Combined ROC and PR Curve Fold

The initial visual evaluation through the Precision-Recall (PR) curve on fold 1 provides an overview of the model's ability to balance precision and Recall. As shown in Fig. 8, the PR curve indicates very high performance, with precision nearing 1.0 across almost the entire recall range. This indicates the model can identify most cancer cases with very low error. The sharp drop in precision as Recall approaches 1.0 is a common phenomenon, occurring when the model attempts to classify all samples as

positive for maximum sensitivity. These results indicate that the SMOTE-based approach and XGBoost with selected features can provide precise detection even in a single data split.

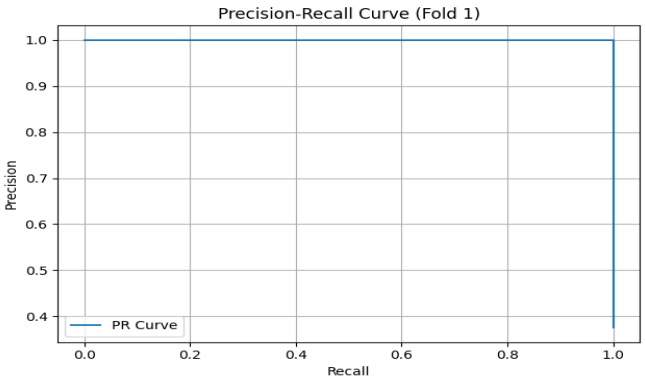

Fig. 8. PR-Curve

Furthermore, to obtain an overall picture of the model's stability and generalisation, the results from 30-fold cross-validation were aggregated. Fig. 9 shows ROC (left) and PR (right) curves, each representing the average results of all folds.

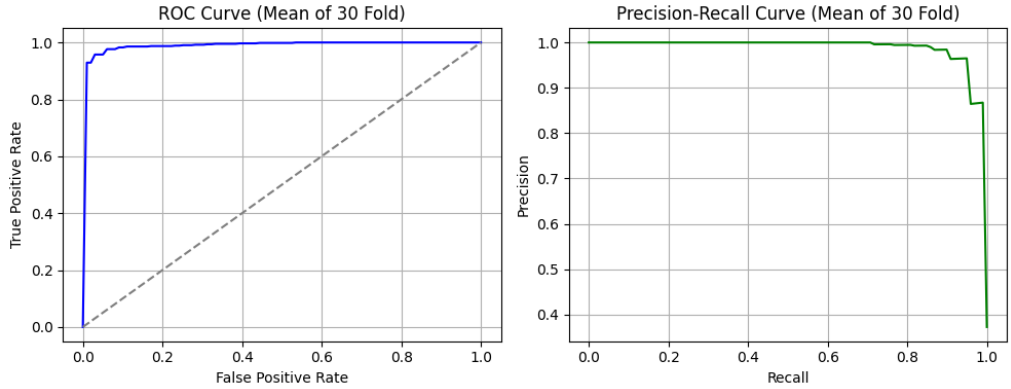

Fig. 9. ROC and PR-Curve

The average ROC curve shows a concave shape approaching the top left point, reflecting a high True Positive Rate despite a still low False Positive Rate, with an AUC ROC of 0.993. This indicates the model can distinguish between benign and malignant cases with minimal error. Meanwhile, the aggregate PR curve also performs well, with precision remaining high across the recall range. The decline in precision as Recall approaches 1.0 is gradual and insignificant, indicating that the model remains precise even as sensitivity increases. With an AUC PR of 0.991, the model proves reliable in dealing with class imbalance. Both curves reinforce the previous numerical evaluation results and demonstrate that the SMOTE, Boruta, and XGBoost-based classification pipeline can deliver consistent and reliable predictive performance on complex medical data.

3.3. Feature Importance

After the data balancing was performed using the Synthetic Minority Over-sampling Technique (SMOTE), it was followed by feature selection based on Boruta, and the final modelling was done using the XGBoost algorithm, resulting in ten important features that made the most significant contribution to breast cancer classification. Feature selection was conducted to eliminate less relevant or redundant attributes, while XGBoost provided feature importance scores based on the contribution to the model's

performance. Figure 10 presents the Top 10 Feature Importance visualisation based on integrating the Boruta selection results and the XGBoost gain scores.

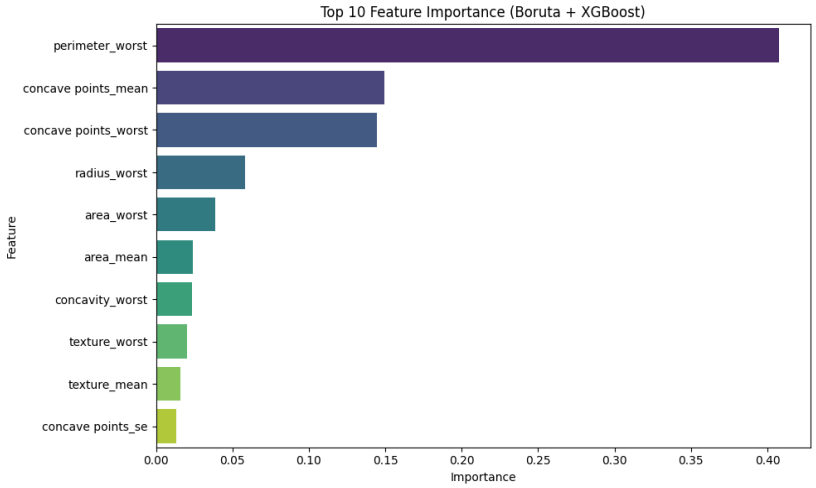

Fig. 10. Top 10 Feature Importance

The perimeter_worst feature emerges as the most influential attribute in classification, reflecting the circumference length of the tumour mass under the most extreme conditions. Along with other features such as concave points_mean, concave points_worst, and concave points_se, the model utilises the edge shape information of the tumour to detect malignancy, where deeper and more varied concavities are often associated with aggressive tumours. The radius_worst, area_worst, and area_mean features describe the size and area of the tumour mass, which correlate with growth and malignancy levels. Concavity_worst measures the depth of the tumour boundary's concavity, indicating the potential for invasion into surrounding tissue. Meanwhile, texture_worst and texture_mean reflect variations in pixel intensity, where heterogeneous textures tend to indicate more dangerous tumour tissue. Overall, the most important features in classification are dominated by the morphological characteristics of tumours in worst-case scenarios, which are relevant clinically in assessing malignancy. Thus, these important features improve the classification performance over the combination of SMOTE, Boruta, and XGBoost but also strengthen the clinical relevance in identifying the degree of tumour malignancy, making the prediction results more relevant and medically interpretable.

3.4. Discussion

The model evaluation results show that the combination of SMOTE, Boruta, and XGBoost provides excellent classification performance in distinguishing between benign and malignant tumours in the Wisconsin Breast Cancer Diagnosis (WBCD) dataset. Based on Repeated Stratified K-fold validation, the model can achieve high and consistent average values for accuracy, precision, Recall, and F1-score across all folds. This indicates that the model is good at classifying the training data and can maintain performance on different test data, thereby reducing the risk of overfitting.

From the feature importance analysis produced by the XGBoost model, features such as worst concave points, worst perimeter, and mean concave points are the most influential variables in the classification process. This finding supports several previous studies that indicate that the characteristics of the shape and contour of tumour tissue have a strong relationship with the degree of malignancy. Feature selection using the Boruta algorithm also plays a role in simplifying the model's complexity by discarding less

relevant features, thus not only increasing processing efficiency but also improving the performance of the model's results.

Some advantages of this combined approach can be seen from three aspects. First, using SMOTE successfully addresses the class imbalance in the data, a common challenge in cancer diagnosis, by increasing the model's sensitivity to the minority class (malignant tumours). Second, the Boruta algorithm effectively filters important features, thus reducing noise and potential multicollinearity that can affect model performance. Third, XGBoost, as a strong classification algorithm, shows advantages in handling tabular data, providing high accuracy and tolerance to missing values and outliers.

4. Conclusion

This study contributes significantly to developing breast cancer classification systems by proposing an integrated framework combining SMOTE for data balancing, Boruta for feature selection, and XGBoost as the primary classification model. This approach successfully addresses the problem of clinical data imbalance while maintaining efficiency by using only about 10 important features. Evaluation using Repeated Stratified K-Fold (30 folds) showed excellent results, with an accuracy of 96.08% ± 2.74%, ROC-AUC of 99.26% ± 0.94%, and PR-AUC of 99.06% ± 1.13%. Morphological features such as perimeter_worst and concave points_mean proved to be dominant, thus improving the clinical relevance and interpretability of the model. However, this approach has limitations, such as the limited amount of data, potential model inaccuracy on more complex data, the use of default hyperparameters in XGBoost that may not be optimal, the risk of SMOTE in generating synthetic samples that may not accurately represent the original distribution, and this study has not implemented advanced model interpretability methods such as SHAP and LIME, which are recommended for future research to improve model transparency in the medical context. As future development directions, it is recommended to: 1) Optimize hyperparameters using Grid Search, Random Search, or Bayesian Optimization; 2) Explore alternative models such as LightGBM and CatBoost; 3) Combine multimodal data (radiology images, genetics, and medical records) and utilize deep learning to build a more comprehensive and accurate diagnostic system.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
Funding statement. None of the authors has received funding or grants from any institution or funding body for the research.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] "Breast cancer," *World Health Organisation*, 2024. [Online]. Available at: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[2] "Global Cancer Observatory," *International Agency for Research on Cancer*, 2022. [Online]. Available at: https://gco.iarc.who.int/media/globocan/factsheets/populations/360-indonesia-fact-sheet.pdf.

[3] K. Shaikh, S. Krishnan, and R. Thanki, "An Introduction to Breast Cancer," in *Artificial Intelligence in Breast Cancer Early Detection and Diagnosis*, Cham: Springer International Publishing, 2021, pp. 1–20, doi: 10.1007/978-3-030-59208-0_1.

[4] S. Sriharikrishnaa, P. S. Suresh, and S. Prasada K., "An Introduction to Fundamentals of Cancer Biology," Springer, Cham, 2023, pp. 307–330, doi: 10.1007/978-3-031-31852-8_11.

[5] E. Bassey, B. Chinemelum, and A. Huygens, "Review Paper Breast Cancer," *J. Glob. Biosci.*, vol. 11, no. 3, pp. 9248–9257, 2022, [Online]. Available at: https://www.mutagens.co.in/jgb/vol.11/110304.pdf.

[6] A. K. Das, S. K. Biswas, A. Bhattacharya, and E. Alam, "Introduction to Breast Cancer and Awareness," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Mar. 2021, no. March, pp. 227–232, doi: 10.1109/ICACCS51430.2021.9441686.

[7] P. Bisoyi, "Malignant tumors – as cancer," in *Understanding Cancer*, Elsevier, 2022, pp. 21–36, doi: 10.1016/B978-0-323-99883-3.00011-1.

[8] H. T. Nia, L. L. Munn, and R. K. Jain, "Physical traits of cancer," *Science (80-)*, vol. 370, no. 6516, p. 12, Oct. 2020, doi: 10.1126/science.aaz0868.

[9] P. Bisoyi, "A brief tour guide to cancer disease," in *Understanding Cancer*, Elsevier, 2022, pp. 1–20, doi: 10.1016/B978-0-323-99883-3.00006-8.

[10] H. Oktavianto and R. P. Handri, "Breast Cancer Classification Analysis Using Naïve Bayes Algorithm," *INFORMAL Informatics J.*, vol. 4, no. 3, p. 117, Jan. 2020, doi: 10.19184/isj.v4i3.14170.

[11] N. R. Muntiari and K. H. Hanif, "Classification of Breast Cancer Disease Using Comparison of Machine Learning Algorithms," *J. Ilmu Komput. dan Teknol.*, vol. 3, no. 1, pp. 1–6, May 2022, doi: 10.35960/ikomti.v3i1.766.

[12] K. Khadijah and R. Kusumaningrum, "Ensemble Classifier for Breast Cancer Classification," *IT J. Res. Dev.*, vol. 4, no. 1, pp. 61–71, Aug. 2019, doi: 10.25299/itjrd.2019.vol4(1).3540.

[13] jamaluddin, A. Kholiq Fajar, M. Zaenal Mutaqin, M. Malik Mutoffar, and D. Setiyadi, "Breast Cancer Classification Using Neural Network and Random Forest Algorithms," *J. Manaj. Inform. Sist. Inf.*, vol. 7, no. 1, p. 77, 2024, [Online]. Available at: https://e-journal.stmiklombok.ac.id/index.php/misi/article/view/1082.

[14] R. Erwandi and Suyanto, "Breast Cancer Classification Using Residual Neural Network," *J. Comput.*, vol. 5, no. 1, pp. 45–52, 2020, [Online]. Available at: https://socjs.telkomuniversity.ac.id/ojs/index.php/indojc/article/download/373/170/1691.

[15] A. Supriyanto, W. A. Kusuma, and H. Rahmawan, "Breast Cancer Tumor Classification Using Inception-V3 Architecture and Machine Learning Algorithms," *J. Al-AZHAR Indones. SERI SAINS DAN Teknol.*, vol. 7, no. 3, p. 187, Sep. 2022, doi: 10.36722/sst.v7i3.1284.

[16] M. A. Naji, S. El Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis," *Procedia Comput. Sci.*, vol. 191, pp. 487–492, 2021, doi: 10.1016/j.procs.2021.07.062.

[17] S. Ara, A. Das, and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," in *2021 International Conference on Artificial Intelligence (ICAI)*, Apr. 2021, no. June, pp. 97–101, doi: 10.1109/ICAI52203.2021.9445249.

[18] N. C. Ramadhan, H. H. H, T. Rohana, and A. M. Siregar, "Machine Learning Algorithm Optimization Using Xgboost Feature Selection for Breast Cancer Classification," *TIN Terap. Inform. Nusant.*, vol. 5, no. 2, pp. 162–171, 2024, doi: 10.47065/tin.v5i2.5408.

[19] K. Mallikharjuna Rao, G. Saikrishna, and K. Supriya, "Data preprocessing techniques: emergence and selection towards machine learning models - a practical review using HPA dataset," *Multimed. Tools Appl.*, vol. 82, no. 24, pp. 37177–37196, Oct. 2023, doi: 10.1007/s11042-023-15087-5.

[20] B. L. Ortiz *et al.*, "Data Preprocessing Techniques for AI and Machine Learning Readiness: Scoping Review of Wearable Sensor Data in Cancer Care," *JMIR mHealth uHealth*, vol. 12, no. 1, p. e59587, Sep. 2024, doi: 10.2196/59587.

[21] J. Y.-L. Chan *et al.*, "Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review," *Mathematics*, vol. 10, no. 8, p. 1283, Apr. 2022, doi: 10.3390/math10081283.

[22] S. Subbiah, K. S. M. Anbananthen, S. Thangaraj, S. Kannan, and D. Chelliah, "Intrusion detection technique in wireless sensor network using grid search random forest with Boruta feature selection algorithm," *J. Commun. Networks*, vol. 24, no. 2, pp. 264–273, Apr. 2022, doi: 10.23919/JCN.2022.000002.

[23] M. B. Kursa, "Robustness of Random Forest-based gene selection methods," *BMC Bioinformatics*, vol. 15, no. 1, p. 8, Dec. 2014, doi: 10.1186/1471-2105-15-8.

[24] R. Iranzad and X. Liu, "A review of random forest-based feature selection methods for data science education and applications," *Int. J. Data Sci. Anal.*, pp. 1–15, Feb. 2024, doi: 10.1007/s41060-024-00509-w.

[25] M. Galih Pradana, K. Palilingan, Y. Vanli Akay, D. Puspasari Wijaya, and P. Hari Saputro, "Comparison of Multi Layer Perceptron, Random Forest & Logistic Regression on Students Performance Test," in *2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, Nov. 2022, pp. 462–466, doi: 10.1109/ICIMCIS56303.2022.10017501.

[26] H. Zhou, Y. Xin, and S. Li, "A diabetes prediction model based on Boruta feature selection and ensemble learning," *BMC Bioinformatics*, vol. 24, no. 1, p. 224, Jun. 2023, doi: 10.1186/s12859-023-05300-5.

[27] H. Gharoun, N. Yazdanjue, M. S. Khorshidi, F. Chen, and A. H. Gandomi, "Leveraging Neural Networks and Calibration Measures for Confident Feature Selection," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 9, no. 3, pp. 2179–2193, Jun. 2025, doi: 10.1109/TETCI.2025.3535659.

[28] H. Matsuo *et al.*, "Diagnostic accuracy of deep-learning with anomaly detection for a small amount of imbalanced data: discriminating malignant parotid tumors in MRI," *Sci. Rep.*, vol. 10, no. 1, p. 19388, Nov. 2020, doi: 10.1038/s41598-020-76389-4.

[29] A. Ali, S. Shamsuddin, and A. Ralescu, "Classification with class imbalance problem: A review," *Int. J. Adv. Soft Comput.*, vol. 5, no. 3, pp. 1–30, 2013. [Online]. Available at: https://www.researchgate.net/profile/Aida-Ali-4/publication/288228469.

[30] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *Int. J. Distrib. Sens. Networks*, vol. 18, no. 6, p. 155013292211069, Jun. 2022, doi: 10.1177/15501329221106935.

[31] S. Fatima, A. Hussain, S. Bin Amir, S. H. Ahmed, and S. M. H. Aslam, "XGBoost and Random Forest Algorithms: An in Depth Analysis," *Pakistan J. Sci. Res.*, vol. 3, no. 1, pp. 26–31, Oct. 2023, doi: 10.57041/pjosr.v3i1.946.

[32] J. Pasaribu, N. Yudistira, and W. F. Mahmudy, "Tabular Data Classification and Regression : XGBoost or Deep Learning with Retrieval-Augmented Generation," *IEEE Access*, vol. 12, pp. 1–1, 2024, doi: 10.1109/ACCESS.2024.3518205.

[33] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," *Technologies*, vol. 13, no. 3, p. 88, Feb. 2025, doi: 10.3390/technologies13030088.

[34] D. Wilimitis and C. G. Walsh, "Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial," *JMIR AI*, vol. 2, no. 1, p. e49023, Dec. 2023, doi: 10.2196/49023.