

# Improving Breast Cancer Classification with Adaptive Synthetic Sampling, Feature Selection, and Hyperparameter Optimization

Hayder Nsaif Jasim<sup>†</sup>, Wesam M. Jasim and Mohammed S. Ibrahim

College of Computer Science and Information Technology, University of Anbar,  
Anbar, Iraq

**Abstract**—Breast cancer is a major global health concern, highlighting the need for accurate and efficient diagnostic solutions rather than persistent issues with detection accuracy. This study presents an enhanced machine learning framework to improve breast cancer classification by addressing key limitations: Class imbalance, irrelevant features, and suboptimal hyperparameters. Adaptive synthetic sampling (ADASYN) was used to balance class distribution and various feature selection techniques. Univariate Selection and recursive feature elimination improved feature relevance, and arctic puffin optimization (APO) was applied for hyperparameter tuning. Multiple classifiers were evaluated using the Wisconsin Diagnostic Breast Cancer dataset. The random forest (RF) with ADASYN approach, optimized using APO, achieved outstanding results – 99.53% accuracy, 100% precision, 99.07% recall, and 99.53% F1-score – with only one misclassification out of 569 samples. This framework, while not modifying ADASYN or RF algorithms themselves, significantly enhances diagnostic performance and serves as a robust foundation for clinical decision support systems.

**Index Terms**—Adaptive synthetic sampling, Arctic puffin optimization, Breast cancer detection, Feature selection, Hyperparameter optimization, Machine learning.

## I. INTRODUCTION

Breast cancer is one of the most frequent and deadly diseases worldwide, which is a leading death, especially for women. The uncontrolled proliferation of atypical breast cells causes benign or malignant tumors. Benign tumors are localized and pose little health problems, whereas malignant tumors are aggressive and can metastasize if not detected and treated. Treatment efficacy increases with a quick and accurate

diagnosis, improving survival rates (Lukong, 2017). Machine learning (ML) methods in breast cancer diagnosis automate and categorize tumors using data. ML methods, including logistic regression (LR), support vector machines (SVM), decision trees (DT), random forest (RF), k-nearest neighbors (k-NN), multi-layer perceptrons (MLP), XGBoost, and Naïve Bayes (NB), excel in predicting performance (Obaid, et al., 2018); (Yadav, Singh and Kashtriya, 2022). However, various barriers hinder classification efficacy, affecting model resilience and generalizability. Class imbalance, which under-represents malignant instances, is a big issue. This mismatch distorts model training, favoring the majority class and reducing malignant tumor detection sensitivity. Adaptive synthetic sampling (ADASYN) treats minority-class synthetic samples to improve class representation and classifier efficacy (He, et al., 2008). Feature selection enhances model performance by reducing noise, overfitting, and computation time. This study employs Univariate Selection and recursive feature elimination (RFE) with cross-validation (Yadav, Singh and Kashtriya, 2022).

Improving model performance requires hyperparameter tweaking. Incorrect hyperparameter selection can overfit or underfit a model, limiting its generalization. Arctic puffin optimization (APO) methodically finds the optimal hyperparameter setups for many classifiers (Wang, et al., 2024). We will evaluate the proposed technique using accuracy, precision, recall, and F1-score. In addition to that, confusion matrices, and receiver operating characteristic-area under the curve (ROC-AUC) are used to show models performance. Different feature selection, oversampling, and hyperparameter-optimized models will be compared. This comprehensive study will provide the best ML methods for breast cancer categorization with balanced accuracy and better resource utilization.

## II. RESEARCH OBJECTIVES

This research focuses on creating a more effective machine-learning diagnostic system for breast cancer classification by tackling class imbalance, enhancing feature selection, and meticulously adjusting model parameters to achieve optimal performance.

ARO-The Scientific Journal of Koya University  
Vol. XIV, No. 1 (2026), Article ID: ARO.12386. 10 pages  
DOI: 10.14500/aro.12386

Received: 27 June 2025; Accepted: 01 February 2026

Regular research paper; Published: 15 March 2026

<sup>†</sup>Corresponding author's e-mail: hay23c1016@uoanbar.edu.iq

Copyright © 2026 Hayder Nsaif Jasim, Wesam Mohammed Jasim and Mohammed Salah Ibrahim. This is an open-access article distributed under the Creative Commons Attribution License (CC BY-NC-SA 4.0).



This study highlights several important contributions, which encompass:

- Implementation of ADASYN to address class imbalance and enhance classifier sensitivity.
- Assessment of various feature selection methodologies, including Univariate selection and RFE and cross validation (RFECV), to ascertain the most pertinent characteristics.
- Meticulous hyperparameter optimization with APO to enhance the efficacy of diverse machine-learning models.
- Improves computer-aided diagnosis (CAD) systems for breast cancer detection by integrating class balance, feature selection, and hyperparameter tuning, leading to enhanced accuracy and interoperability.

This document is organized as follows. Section 2 presents an examination of the present literature on ML methods for classifying breast cancer. Section 3 provides a comprehensive analysis of dataset preparation, feature selection, oversampling methods using ADASYN, and the procedure for hyperparameter optimization. Section 4 reports on the findings from experiments. In-depth examination of various ML classifiers utilizing a range of performance metrics. Section 5 represents the examination and final reflections overview of significant findings, obstacles faced, and possible avenues for further exploration.

### III. RELATED WORKS

This section aims to show the research that has been done on using ML to find breast cancer through numerical data analysis.

(Gupta and Garg, 2020) analyzed the performance of six ML models – k-NN, LR, DT, RF, SVM with radial basis function kernels, and deep learning using Adam gradient descent – for breast cancer prediction using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The study emphasizes hyperparameter optimization for model performance. At 98.24%, deep learning using Adam Gradient Descent is the most accurate, followed by SVM and RF at 97.2%. However, the investigation has limitations. We prioritize hyperparameter tuning over feature selection and class imbalance algorithms, both of which are crucial for medical diagnoses. Deep learning is accurate, but computational complexity and overfitting remain unexplored, limiting its applicability in resource-limited contexts.

The study by (Naji, et al., 2021) uses the WDBC dataset to test five ML algorithms to see how well they can diagnose breast cancer. The algorithms are SVM, RF, LR, C4.5 DT, and k-NN. SVM had the best accuracy (97.2%), precision, sensitivity, and F1 score. SVM can discriminate benign from malignant instances, proving its medical diagnostic value. The analysis' focus on WDBC may limit its applicability to other breast cancer datasets with various feature distributions. Clinical application requires accurate characteristics, class imbalance adjustment, and calculation efficiency assessment, which it lacks.

(Magboo and Magboo, 2021) used the Wisconsin Prognostic Breast Cancer dataset to evaluate LR, NB, KNN,

and SVM for breast cancer recurrence prediction. It examines four scenarios: Scaling alone, scaling with principal component analysis (PCA), scaling with PCA and minority class oversampling, and picking features by removing closely connected variables. The results show that LR is always better than other models. It has the highest accuracy in all configurations (80%, 80%, 72%, and 75%), as well as the highest F1-score (0.76–0.74), AUROC (0.81–0.78), and Cohen's Kappa statistic (0.34–0.29). The performance of SVM remains consistent, achieving a stable accuracy of 75% across various configurations. KNN recall improves with oversampling, but accuracy stays at 60–74%. NB struggles with PCA-based frameworks, resulting in a 60% accuracy. However, feature selection improves accuracy to 65–69%. Nonetheless, the study presents certain limitations. The small size of the dataset, despite its high-dimensional structure, may impact on the generalizability of the models. Reduce class imbalance but distort synthetic data by oversampling. The feature selection technique only considers correlation-based deletion, even if other selection strategies increase model understanding and performance.

Gopal, et al., 2021 used IoT and ML to identify breast cancer by categorizing data and choosing key factors. This study selects and classifies Wisconsin Breast Cancer Dataset characteristics using RF, LR, and MLP. MLP classifies better than RF and LR. It had an F1 score of 96%, accuracy of 98%, precision of 98%, recall of 97%, and only 10 errors in 569 cases. MLP can distinguish benign from malignant patients better than RF and LR (MAR: 34.21%, RMSE: 45.82%, and RAE: 64.47%). However, the investigation shows limitations. The analysis focuses on Wisconsin Breast Cancer Diagnostic (WBCD), which may affect its relevance to other datasets. Features are selected using PCA and correlation functions, without considering other methods. The computational complexity of the model can affect real-time clinical applications, but the study ignores it.

(Allam and Nandhini, 2022) The Binary teaching learning-based optimization (BTLBO) breast cancer feature selection method obtained 98.43% accuracy on the WDBC dataset. Limiting processing and selecting relevant feature groups improves classification accuracy. BTLBO improves Genetic Algorithm results. The investigation is limited to conventional classifiers, such as SVM, k-NN, NB, DT, and Discriminant Analysis. Practical applications need class imbalance, scalability, and processing complexity. This study ignores these difficulties with large datasets.

Uddin, et al., 2023 employed feature optimization on the Wisconsin Breast Cancer Dataset to improve breast cancer diagnosis with ML. They analyze eleven classifiers, including SVM, RF, KNN, DT, NB, LR, AdaBoost, GB, MLP, NCC, and VC. The voting classifier, which combines LR and SVM, gets an amazing 98.77% accuracy after feature optimization, which is better than the performance of the separate models. Fine-tuning hyperparameters, implementing cross-validation, and applying PCA for dimensionality reduction significantly improve model performance. They created a website for real-time breast cancer prediction using Flask and React. Multiple study limitations exist. They limit the analysis to the WBCD

dataset, which may not apply to other datasets with different feature distributions. The study detects features using PCA and ignores other optimization methods. The ensemble model's untested computational complexity, notably the voting classifier, may limit real-time clinical use.

(Nemade and Fegade, 2022) examined different machine-learning breast cancer classifiers using the WDBC dataset. Tests, such as NB, LR, SVM, KNN, DT, RF, AdaBoost, and X, are used to check how accurate, precise, recalling, and AUC-ROC a model is. XGBoost classifiers outperform other models with 97% accuracy. XGBoost categorizes the best with an area under the ROC curve (AUCROC) of 0.999. LR has an AUC-ROC of 0.993, while NB has the lowest accuracy at 90%. The study has flaws. Using only the WDBC dataset may limit its application to other breast cancer datasets. The study did not explore advanced feature selection methods or ensemble model computational complexity, which could affect real-time medical applications.

Minnoor and Baths (2023), RFE and LR determine feature importance, and the analysis of variance (ANOVA) F-test determines the optimal variable. This will help the RF model find breast cancer more accurately using the WBCD dataset. First, the Pearson correlation coefficient eliminates strongly associated features, leaving 16. ANOVA F-test, LR, and RFE restrict the study to 8 features with two positive scores. RF outperforms SVM, DT, Multilayer Perceptron, and k-NN, with 100% and 99.30% accuracy for 16 and 8 features. We only study correlational and statistical feature selections, which may complicate model optimization on larger datasets.

Aiyeniko, (2023) Metaheuristic methods for selecting breast cancer features have been explored using the Wisconsin Breast Cancer Dataset. ACO, Bat, and Wolf algorithms choose attributes before categorization. Trained models include SVM, DT (C4.5), K-Nearest Neighbors (KNN), Neural Network, and NB with unique attributes. Bat and Wolf feature options yield SVM accuracy to 96.9957%. In combination with the ACO, Bat, and Wolf algorithms, SVM yields the highest recall of 0.9740 and an F1 score of 0.9770, while NB consistently yields the highest precision of 0.9860. Metaheuristic-based feature selection improves SVM

accuracy and efficiency, beating other classifiers. The study is flawed. The study only supports the WBCD dataset, thereby restricting its use. We do not analyze the computational complexity of the metaheuristic algorithm, which potentially limits its real-time clinical use.

Assegie, et al., 2024 tested ADASYN resampling for class imbalances in breast cancer diagnoses using the Wisconsin Breast Cancer Dataset. LR, SVM, KNN, GB, and AB assess unfair and ADASYN-resampled data. After ADASYN oversampling, LR classifies best at 99.46%. Resampling lowers false negatives, as shown by one confusion matrix of malignant misclassification. SVM is 97.87% accurate, and RF is 95.74%. SHAP considers texture poor, concave points, and radius influential. Bad research. Since ADASYN resamples SMOTE and Borderline-SMOTE, they are not compared. Overfitting and unanalyzed computational complexity may limit ADASYN's real-world use.

Ahamed, 2024 detected early breast cancer using the Wisconsin Breast Cancer Dataset. Classifiers, such as LR and DT distinguish malignant and benign tumors. To improve the model, remove unnecessary attributes, missing values, one-hot categorical feature encoding, and Min-Max numerical value scaling while preparing data. LR has 100% precision, 93.65% recall, 96.72% F1, and 97.66% accuracy. However, the DT has 92.40% accuracy, 87.88% precision, 92.06% recall, and 89.92% F1-score. The study links tumor size, shape, and margins to classification. LR outperforms DT in ROC curve analysis with AUC 1.00 versus 0.92. Bad research. The research utilized two ML models – not SVM, RF, or ensembles. No feature selection may influence model efficiency and interpretability. Finally, the study does not address class imbalance issues that may limit diagnostic use.

#### IV. METHODOLOGY

For the methodology of this research, Fig. 1. shows the main steps of our proposed model.

##### A. Dataset

The dataset, developed by Dr. William H. Wolberg at the University of Wisconsin Hospital, is widely employed in

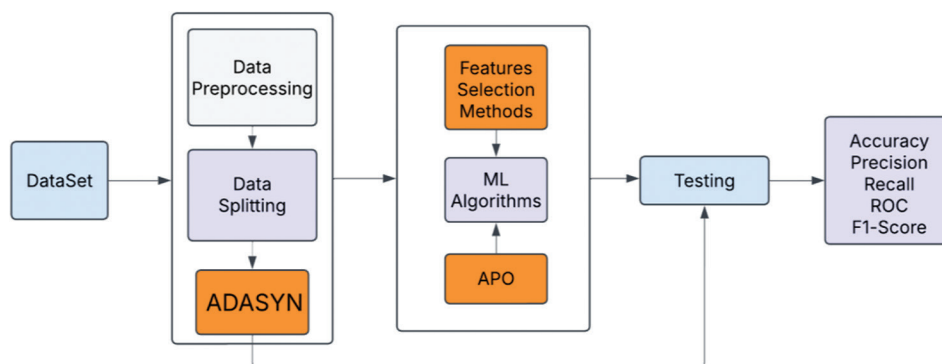


Fig. 1. The main steps of our proposed system.

medical diagnostics and ML research. The dataset, available in the UCI ML Repository, there are a total of 569 occurrences, with 357 classified as benign (non-cancerous) and 212 as malignant (cancerous), as shown in Fig. 2. The features derive from fine-needle aspiration biopsies of breast tumors and comprise 30 numerical attributes, derived from ten fundamental characteristics of cell nuclei, which encompass radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. Each attribute is recorded utilizing three statistical metrics: Mean, standard error, and worst value (the mean of the three greatest values) (Wolberg, et al., 1993), (Mangasarian, Street and Wolberg, 1994).

*B. Data Pre-processing*

Before model training, we pre-processed the WDBC dataset to improve data quality and predictive accuracy. We removed missing data to ensure consistency and prevent model performance concerns. The ID column is deleted because it only serves as a unique identifier and does not help categorize, simplifying the structure. Normalization uniformizes feature scales to prevent any property from significantly affecting the model. This maintains feature uniformity. This stage improves model convergence and efficacy. The goal labels for the categories are encoded as 0 for malignant tumors and 1 for benign tumors to help the ML system read them. These pre-processing procedures greatly improve classification model robustness and efficacy.

*C. ADASYN*

The ADASYN method was selected due to its effectiveness with tabular data and its ability to generate data adaptively based on local distribution, making it well-suited for the WDBC dataset. It addresses class imbalance by generating synthetic samples for the minority class based on the learning difficulty of each instance. The distribution of the majority class surrounding each instance of the minority class

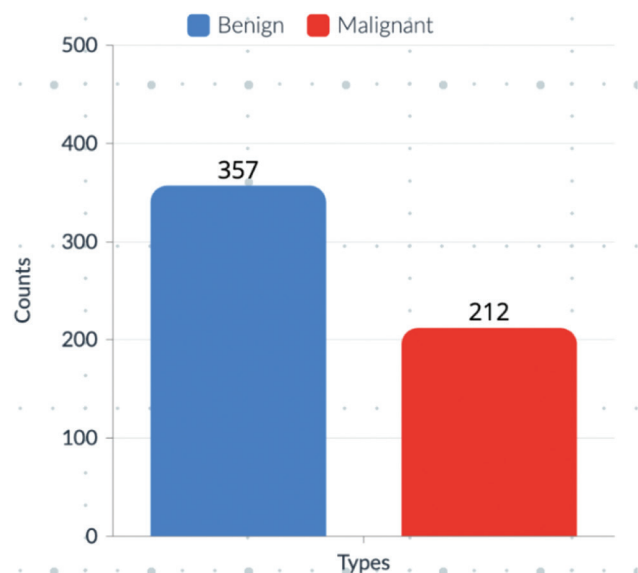


Fig. 2. Illustrates the distribution of benign and malignant cases. M: Malignant tumors, B: Benign tumors

determines the difficulty. The imbalance ratio, expressed as  $d = \frac{ms}{ml}$ , where ms and ml signify the counts of minority and majority class samples, respectively, determines the need for dataset resampling. If d is below a defined threshold dth, the technique computes the total number of synthetic samples generated using  $G=(ml-ms) \times \beta$ , where  $\beta$  modulates the desired balancing level. Each minority instance  $x_i$  is evaluated to ascertain its difficulty ratio,  $r_i = \frac{\Delta_i}{k}$  where  $\Delta_i$  represents the count of majority samples among its k-NN. The normalized ratio  $r^i$  guarantees an equitable allocation of synthetic samples among minority instances, resulting in the computation  $g_i=r^i \times G$  for the quantity of synthetic samples per instance. The equation  $s_i=x_i+(x_{zi}-x_i) \times \lambda$ , produces synthetic samples, with  $x_{zi}$  representing a randomly chosen neighbor from the minority class and  $\lambda$  being a stochastic variable that ranges from 0 to 1. This adaptive approach improves the performance of classifiers by minimizing bias, adjusting decision thresholds, and enhancing the detection of challenging events in the dataset (He, et al., 2008), (Kulkarni, Chong and Batarseh, 2020), (Krawczyk, 2016). The effect of ADASYN is demonstrated in Fig. 3. and Table I.

*D. Features Selection Methods*

Feature significance methods rank input features to determine their impact on a predictive model’s decision-making. Analyzing feature importance reveals which attributes most affect model performance and provides useful dataset insights. This analysis explains how features affect breast cancer detection.

Our work used trained models to calculate feature importance scores and determine the most important features for accurate categorization. Fig. 4. shows how several mammographic characteristics correlate, revealing their importance in breast cancer detection.

ML pre-processing selects attributes to reduce dimensionality, increase model performance, and improve interpretability. The top breast cancer categorization features were selected using two ways.

- Univariate Feature Selection: Using SelectKBest is a univariate selection method that ranks features based on statistical tests (such as the Chi-square test) and retains the top k features that show the highest relevance to the target variable. Early feature elimination and computational efficiency are gained with this method (Gárate-Escamila, Hajjam El Hassani and Andrés, 2020).
- RFECV using RF RFECV extends RFE by incorporating 5-fold cross-validation to determine the optimal number of features. Rather than selecting a fixed number of features, RFECV automatically identifies the subset that maximizes

TABLE I  
THE DATA BEFORE AND AFTER THE APPLICATION OF ADASYN

Data	Original (train-test)	ADASYN (train-test)
Benign	249–108	249–108
Malignant	149–63	250–105

ADASYN: Adaptive synthetic sampling

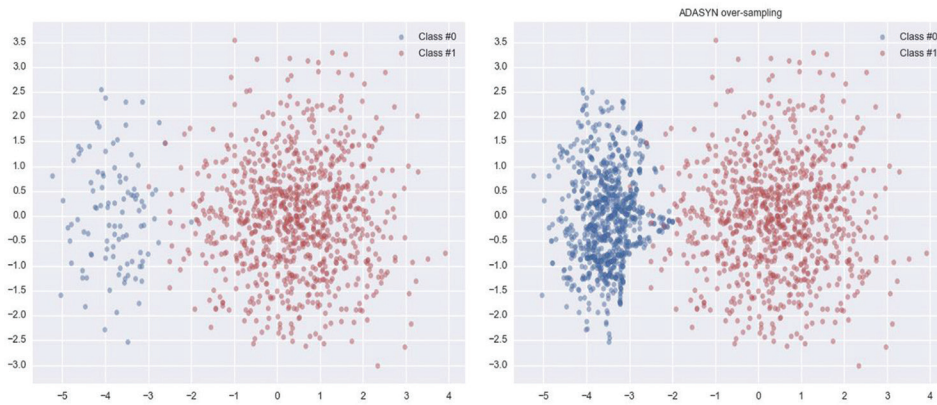


Fig. 3. The dataset before and after the application of adaptive synthetic sampling.

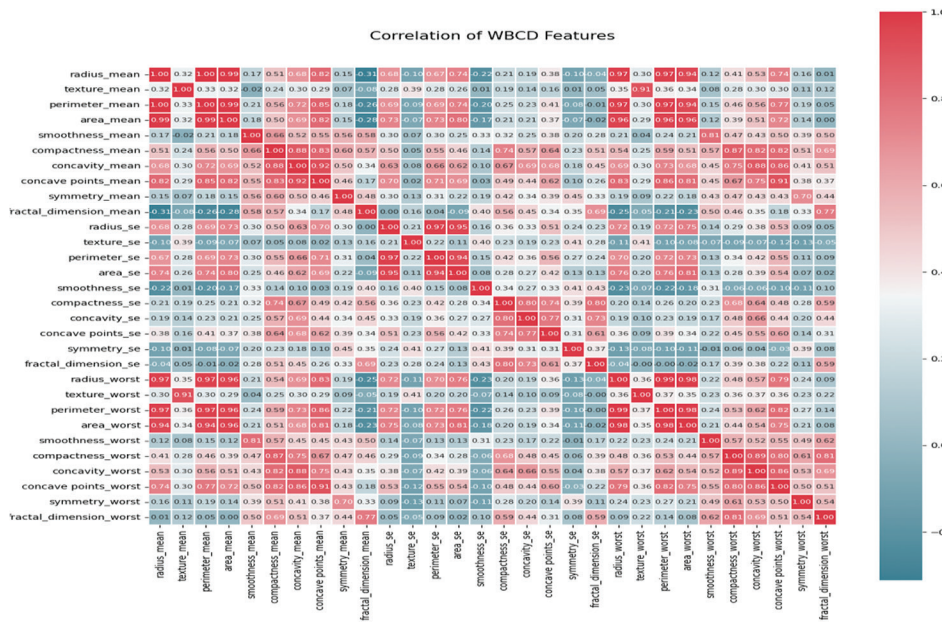


Fig. 4. Correlation between the features of breast cancer.

classification accuracy. RFECV determined that 18 features were optimal, balancing performance and model complexity (Mowri, Siddula and Roy, 2023).

**E. ML Algorithms**

Different classification methods in ML algorithms increase accuracy and generalization. This investigation uses several optimized predictive models.

- LR: Is a simple and effective way to make clear decision boundaries for binary classification models that use the logistic function to estimate the chances of each class (Cox, 1958).
- KNN: A non-parametric model that classifies data points by distance metrics and closest neighbor majority votes (Cover and Hart, 1967).
- DT: A hierarchical model that categorizes data by feature values for efficiency and interpretability (Quinlan, 1986).
- Supervised learning approach support vector classifier (SVC): Efficiently constructs a hyperplane to partition

classes in high-dimensional domains (Cortes and Vapnik, 1995).

- The RF algorithm: Uses many DT to avoid overfitting and enhance accuracy (Breiman, 2001).

These algorithms use model strengths and ensemble methods to develop a comprehensive categorization framework that improves forecast accuracy.

**F. APO**

The APO technique (Wang, et al., 2024) was employed to optimize the hyperparameters for RF, KNN, SVC, and LR, among others. APO is a metaheuristic optimization technique that efficiently navigates a hyperparameter search space to enhance classification accuracy.

APO was initialized with a population of 10 solutions and iterated over 50 generations to search for optimal hyperparameters. For each classifier, the search space was broken down into numerical bounds. Categorical parameters were stored as discrete indices, and continuous values

were limited to ranges that had already been set. We used 5-fold cross-validation accuracy as the fitness function to test each solution in the population against a different set of hyperparameters. The objectives of APO were to maximize classification performance by iteratively refining hyperparameters using two key mechanisms:

- Aerial flight stage: Broad search using random exploration and perturbations
- Underwater foraging stage: Fine-tuning solutions by leveraging differential evolution techniques.

After testing different hyperparameter setups, the best ones were accurate across modeling. APO optimized classification by balancing model complexity with generalization. This improved the ML breast cancer detection models. The search space range and optimal parameters discovered for each classifier are summarized in Table II. The table presents the range of values explored for each hyperparameter during optimization and the corresponding best-performing configuration identified by APO.

*G. Model Performance and Results*

This section presents the results of various classification models applied to the WDBC dataset, each utilizing different feature selection techniques. The models were

trained on 70% of the data and evaluated on the remaining 30%, using metrics, such as accuracy, ROC AUC, precision, recall, and F1-score to assess performance. The comparison focuses on two feature selection methods. Univariate Feature Selection and RFECV. The performance of each classifier is assessed using multiple assessment metrics. The ROC curves for each classification model are depicted in Fig. 5. This figure highlights the equilibrium between sensitivity and specificity applicable to all feature selection methods.

A comparison of the precision-recall curves for each of the models is shown in Fig. 6, which highlights how these models are effective in addressing class imbalance.

Bar charts are used to highlight the accuracy and ROC AUC values for each model that utilizes the four different feature selection procedures. These bar charts are displayed in Figs. 7 and 8.

An assortment of confusion matrices for every classifier is illustrated in Figs. 9 and 10, which can be found here. The goal labels for the categories are encoded as 0 for malignant tumors and 1 for benign. The distribution of cases that were properly identified will be displayed in these matrices, along with the distribution of examples that were incorrectly labeled.

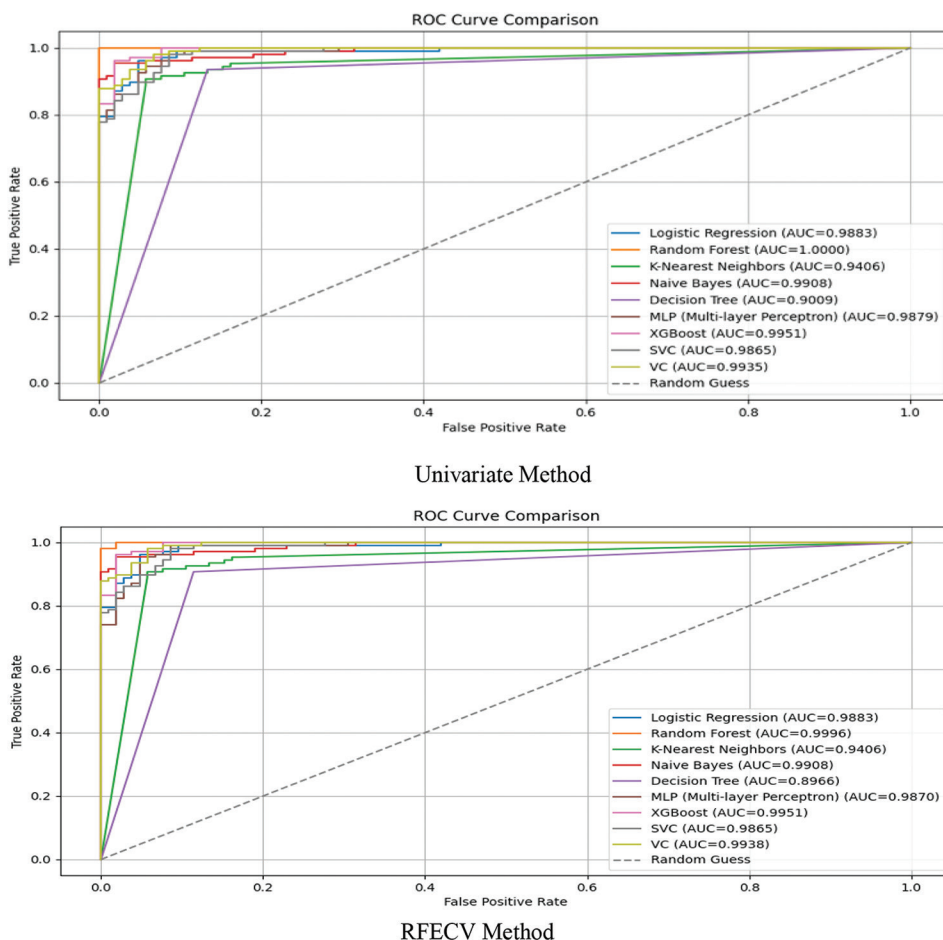


Fig. 5. Hyperparameter search space and optimal selection.

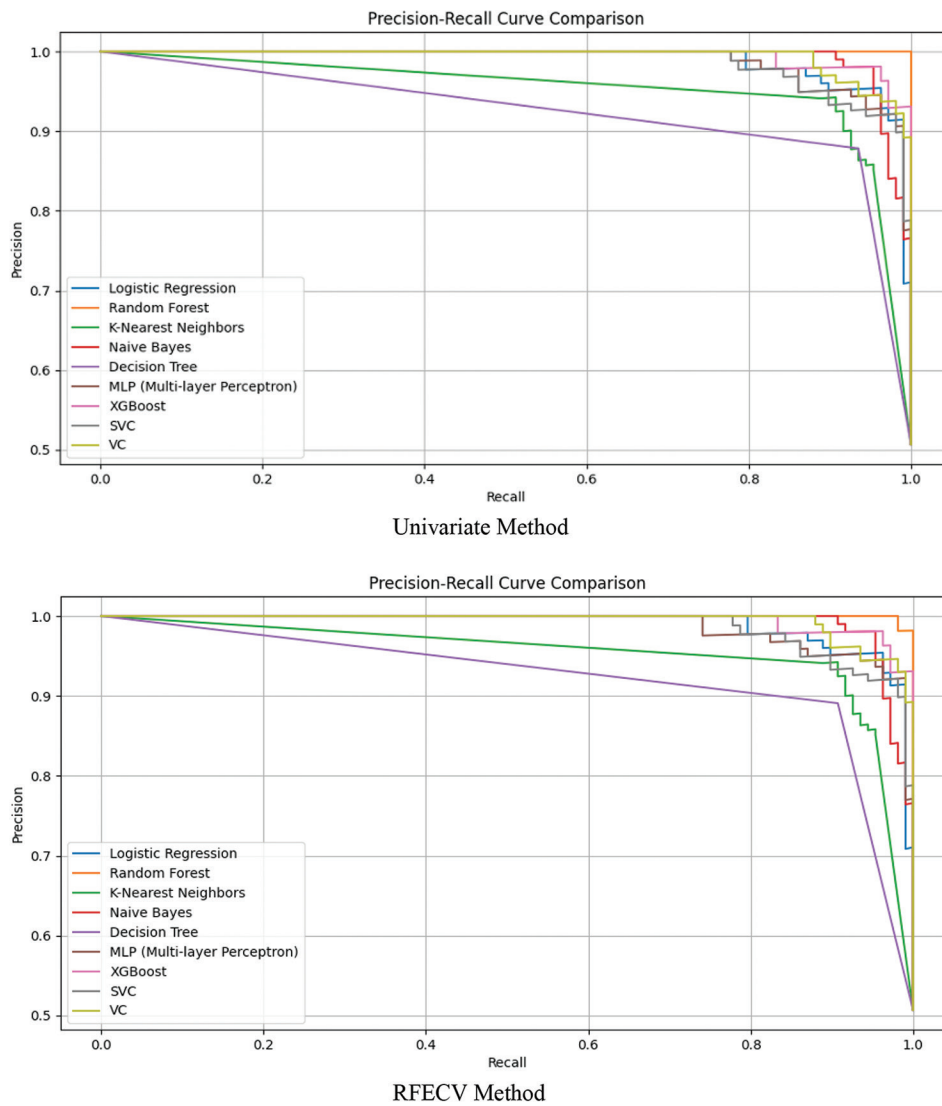


Fig. 6. Precision-Recall curve comparison.

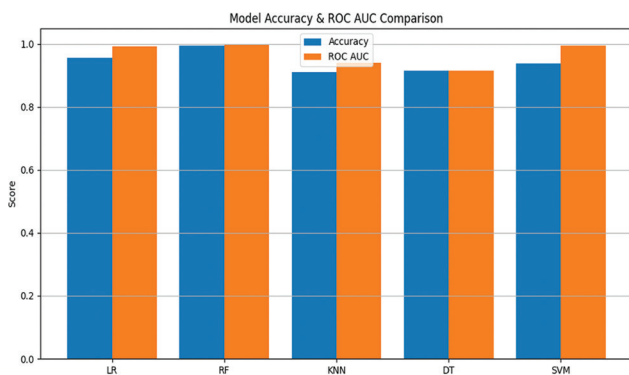


Fig. 7. Model accuracy and receiver operating characteristic-area under the curve comparison for univariate method.

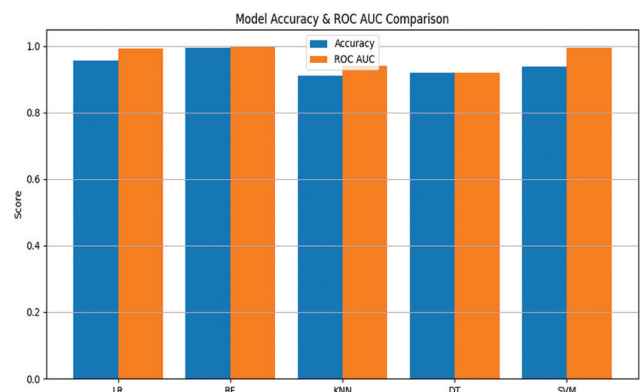


Fig. 8. Model accuracy and receiver operating characteristic-area under the curve comparison for recursive feature elimination and cross-validation method.

The findings of the evaluation reveal that the RF classifier was the most effective model. This is since it exhibited the highest accuracy and ROC AUC when compared to all of

the other feature selection methods taken into consideration. With greater specificity:

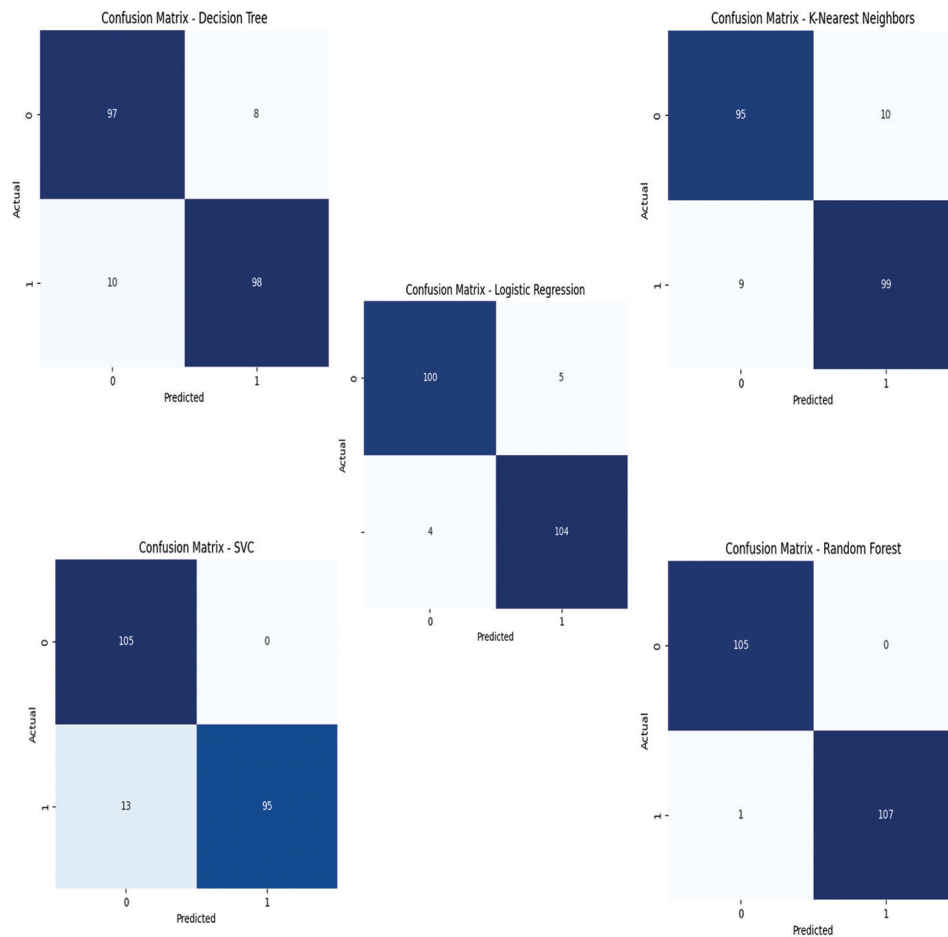


Fig. 9. Confusion matrices for the univariate method. 0 = Malignant (cancerous tumor), 1 = Benign (non-cancerous tumor)

TABLE II  
HYPERPARAMETER SEARCH SPACE AND OPTIMAL SELECTION

Model	Hyperparameter with search space	Best parameter
LR	'penalty': ['l2', 'none'], 'solver': ['lbfgs', 'liblinear', 'saga'], 'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000], 'max_iter': [10, 100, 200, 500, 1000, 10000]	C=1000, penalty=l2, max_iter=10, solver=liblinear
SVM	'degree': [2, 3, 4], 'gamma': ['scale', 'auto'], 'C': [0.001, 0.01, 0.1, 1, 10], 'kernel': ['linear', 'poly', 'rbf', 'sigmoid']	C=10, degree=2, kernel=poly, gamma=scale,
KNN	'p': [1, 2] 'leaf_size': [30, 40, 50], 'n_neighbors': [3, 5, 7, 10, 15], 'weights': ['uniform', 'distance'], 'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute']	p=2, leaf_size=30, algorithm=auto, n_neighbors=3, weights=distance
RF	'bootstrap': [True, False] 'min_samples_leaf': [1, 2, 4], 'n_estimators': [50, 100, 200], 'min_samples_split': [2, 5, 10], 'max_depth': [None, 10, 20, 30], 'max_features': ['auto', 'sqrt', 'log2'], 'criterion': ['gini', 'entropy', 'log_loss']	criterion=gini, bootstrap=False, max_depth=None, n_estimators=100, max_features=sqrt, min_samples_leaf=1, min_samples_split=2,

LR: Logistic regression, SVM: Support vector machines, RF: Random forest, K-NN: k-nearest neighbors

With univariate feature selection, the RF classifier achieved an outstanding accuracy of 99.53% and an ROC AUC of 0.9998. Similarly, using RFECV, the RF model maintained the same accuracy and ROC AUC, confirming its robustness across different feature selection methods. Several other models fared well. LR performed well with 95.77% accuracy and 0.9934 ROC AUC. The SVC had good ROC AUC (0.9957) but lower accuracy (93.90%). However, KNN and DT models lagged with 91.08% and 91.55–92.02% accuracy, respectively. RF is the best classification model for this problem due to its accuracy and stability across feature selection methods.

Our method's results will be compared to the best breast cancer detection studies. Table III compares machine and deep learning methods by accuracy, ROC AUC, precision, recall, F1 score, and error situations. We found that APO-RF had the highest accuracy at 99.53%. ROC AUC and F1 scores of 99.87% and 99.53% are excellent. Our performance beats ANN, SVM, LR, and ensemble-based approaches in earlier research. Our technique improves breast cancer detection by improving classification accuracy and reducing mistakes.

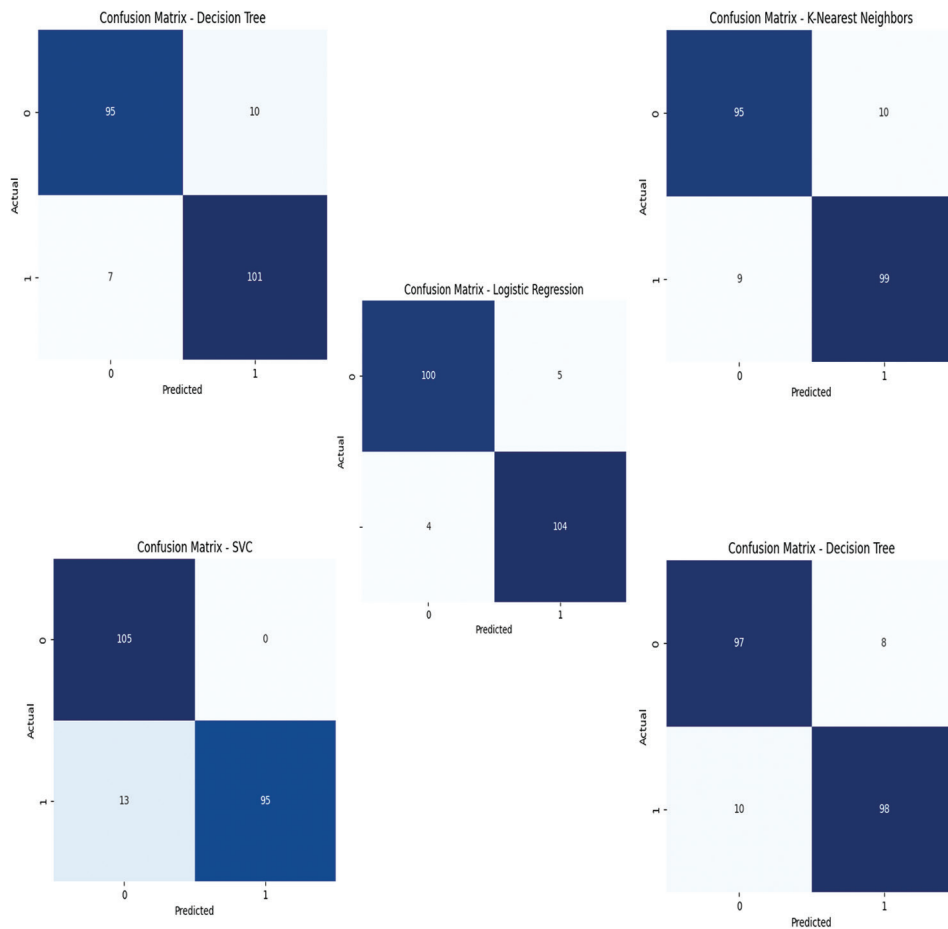


Fig. 10. Confusion matrices for recursive feature elimination and cross-validation method.

TABLE III  
COMPARISON OF THE PROPOSED MODEL WITH THE MOST RELEVANT RESEARCH

Ref	Method	Accuracy (%)	ROC AUC	Precision	Recall	F1 score	Error cases
[6–20] 2020	ANN	98.24	-	-	-	-	-
[7–6] 2021	SVM	97.2	-	-	-	-	-
[8–8] 2021	LR	80	0.81	-	-	76%	-
[9–9] 2021	MLP	98	-	98%	97%	96%	10–569
[10–5] 2022	BTLBO-SVM	98.43	-	-	-	-	-
[11–10] 2023	VC (LT+SVM)	98.77	-	-	-	-	-
[12–12] 2023	XGBoost	97	99%	-	-	-	-
[13–16] 2023	RF	99.30	-	-	-	-	-
[14–27] 2023	ACO+Bat+Wolf (SVM)	96.99	-	-	97.40%	97.70%	-
[15–19] 2024	ADASYN (LR)	99.46	-	-	-	-	-
[16–26] 2024	LR	97.66	100%	100%	93.65%	96.72%	-
Proposed-model	APO-RF	99.53	99.98%	100%	99.07%	99.53	1-569

ROC: Receiver operating characteristic, ADASYN: Adaptive synthetic sampling, APO: Arctic puffin optimization, BTLBO: Binary teaching learning-based optimization, LR: Logistic regression, SVM: Support vector machines, DT: Decision trees, RF: Random forest, K-NN: k-nearest neighbors, MLP: Multi-layer perceptrons

## V. CONCLUSION AND FUTURE WORKS

This study introduces an improved framework for breast cancer classification that employs ADASYN to rectify class imbalances, integrates multiple feature selection techniques (such as univariate selection and RFECV), and utilizes hyperparameter tuning through APO to enhance the efficacy of ML models. According to the results of the evaluation of numerous models, the RF classifier, which was optimized with APO, achieved the highest accuracy,

which was 99.53%, and an AUCROC of 99.98%, exceeding a variety of conventional and deep learning approaches. The application of ensemble learning, feature selection, and hyperparameter adjustment markedly enhanced the accuracy and reliability of the classification process. This technique is a strong candidate for a CAD system capable of detecting breast cancer. The findings indicate that the new technology improves the ability to identify malignant tumors while avoiding the detection of non-cancerous ones.

This subsequently aids in the early diagnosis process and the development of treatment options.

The paradigm shows promising results, yet it has severe drawbacks. The model performed well when tested on specific datasets, but we are not sure how well it will work with different types of images, features, or patient groups. We need larger, more diverse datasets to evaluate resilience. GridSearchCV and ensemble learning hyperparameter optimizations are computationally demanding for clinical usage. Feature selection impacts model performance. Feature selection strategies differ for each dataset; therefore, a universal strategy may not work. Clinicians may not understand the model's decision-making process despite its high classification accuracy. Explainable AI, such as SHAP or LIME, could assist doctors in understanding models.

#### REFERENCES

- Ahamed, M.R.H., 2024. Early detection: Machine learning for breast cancer prediction. In: *Conference: Early Detection: Machine Learning for Breast Cancer Prediction At: Galle, Sri Lanka*.
- Aiyeniko, O., 2023. Performance evaluation of metaheuristic algorithms for feature selection in breast cancer predictive model. *American Journal of Computer Sciences and Applications*, 11(2), pp.1-10.
- Allam, M., and Nandhini, M., 2022. Optimal feature selection using binary teaching learning based optimization algorithm. *Journal of King Saud University - Computer and Information Sciences*, 34, pp.329-341.
- Assegie, T.A., Salau, A.O., Sampath, K., Govindarajan, R., Murugan, S., and Lakshmi, B., 2024. Evaluation of adaptive synthetic resampling technique for imbalanced breast cancer identification. *Procedia Computer Science*, 235, pp.1000-1007.
- Breiman, L., 2001. Random forests. *Machine Learning*, 45, pp.5-32.
- Cortes, C., Vapnik, V., 1995. Support-vector networks. *Machine Learning*, 20, pp.273-297.
- Cover, T.M., and Hart, P.E., 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), pp.21-27.
- Cox, D.R., 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20, pp.215-232.
- Gárate-Escamila, A.K., Hajjam El Hassani, A., and Andrés, E., 2020. Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked*, 19, 100330.
- Gopal, V.N., Al-Turjman, F., Kumar, R., Anand, L., and Rajesh, M., 2021. Feature selection and classification in breast cancer prediction using IoT and machine learning. *Measurement*, 178, p.109442.
- Gupta, P., and Garg, S., 2020. Breast cancer prediction using varying parameters of machine learning models. *Procedia Computer Science*, 171, pp.593-601.
- He, H., Bai, Y., Garcia, E.A., and Li, S., 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: *Proceedings of the International Joint Conference on Neural Networks*, IEEE. pp.1322-1328.
- Krawczyk, B., 2016. Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5, pp.221-232.
- Kulkarni, A., Chong, D., and Batarseh, F.A., 2020. Foundations of data imbalance and solutions for a data democracy. In: *Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering*, Academic Press, United States, pp.83-106.
- Lukong, K.E., 2017. Understanding breast cancer – The long and winding road. *BBA Clinical*, 7, pp.64-77.
- Magboo, V.P.C., and Magboo, M.S., 2021. Machine learning classifiers on breast cancer recurrences. *Procedia Computer Science*, 192, pp.2742-2752.
- Mangasarian, O.L., Street, W.N., and Wolberg, W.H., 1994. Breast cancer diagnosis and prognosis via linear programming. *AAAI Spring Symposium – Technical Report*, SS-94-01, pp.83-86.
- Minnoor, M., and Baths, V., 2022. Diagnosis of breast cancer using random forests. *Procedia Computer Science*, 218, pp.429-437.
- Mowri, R.A., Siddula, M., and Roy, K., 2023. Is iterative feature selection technique efficient enough? A comparative performance analysis of RFECV feature selection technique in ransomware classification using SHAP. *Discover Internet of Things*, 3, p.21.
- Naji, M.A., Filali, S.E., Aarika, K., Benlahmar, E.H., Abdelouhahid, R.A., and Debauche, O., 2021. Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Computer Science*, 191, pp.487-492.
- Nemade, V., and Fegade, V., 2022. Machine learning techniques for breast cancer prediction. *Procedia Computer Science*, 218, pp.1314-1320.
- Obaid, O.I., Mohammed, M.A., Abd Ghani, M.K., Mostafa, S.A., and Al-Dhief, F.T., 2018. Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer. *International Journal of Engineering and Technology*, 7, pp.160-166.
- Quinlan, J.R., 1986. Induction of decision trees. *Machine Learning*, 1(1), pp.81-106.
- Uddin, K.M.M., Biswas, N., Rikta, S.T., and Dey, S.K., 2023. Machine learning-based diagnosis of breast cancer utilizing feature optimization technique. *Computer Methods and Programs in Biomedicine Updates*, 3, p.100098.
- Wang, W.C., Tian, W.C., Xu, D.M., and Zang, H.F., 2024. Arctic puffin optimization: A bio-inspired metaheuristic algorithm for solving engineering design optimization. *Advances in Engineering Software*, 195, p.103694.
- Wolberg, W., Mangasarian, O., Street, N., and Street, W., 1993. *Breast cancer Wisconsin (Diagnostic) [Dataset]*. *UCI Machine Learning Repository*.
- Yadav, R.K., Singh, P., and Kashtriya, P., 2022. Diagnosis of breast cancer using machine learning techniques - a survey. *Procedia Computer Science*, 218, pp.1434-1443.