Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 9 (2027) 3495

Journal of the Nigerian Society of Physical Sciences

An improved SMOTE-based hybrid resampling (pipelining) framework for imbalance-aware breast cancer diagnosis

Yusuf Abubakar Kutigi[a,*], Abdullahi Muhammad Bashir[b], Mohammed Abdulmalik Danlami[a], Adabara Nasiru Usman[c]

[a]*Department of Computer Science, Federal University of Technology, Minna, Nigeria*
[b]*Department of Data Science, Federal University of Technology, Minna, Nigeria*
[c]*Department of Microbiology, Federal University of Technology, Minna, Nigeria*

Abstract

Class imbalance remains a persistent challenge in medical diagnosis, particularly in breast cancer datasets, which often leads to models that overlook malignant cases. While the Synthetic Minority Oversampling Technique (SMOTE) and its variants address this issue, they frequently introduce noise and fail to preserve the intrinsic data structure. This study proposes a sequentially integrated resampling framework combining Adaptive Synthetic Sampling (ADASYN), Cluster-based oversampling and density-based filtering to enhance data quality and classification performance. The framework was evaluated using extreme gradient boosting (XGBoost) on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which contains 569 observations and 30 real-valued diagnostic features derived from digitized images of fine-needle aspirates (FNA) of breast masses. The original data also contains an identification variable and a diagnosis variable. The identification variable was excluded from model development because it has no diagnostic meaning, while the diagnosis variable was retained as the target variable and encoded as malignant = 1 and benign = 0. Thus, only the 30 continuous diagnostic features were used as predictor variables. The dataset was split into training and testing sets. The proposed improved SMOTE achieved an accuracy of 96.49%, recall of 97.22%, precision of 97.22%, F1 score of 97.22% and area under the receiver operating characteristic curve (ROC-AUC) of 96.23%, achieving higher threshold-dependent classification performance while yielding a lower ROC-AUC than the evaluated baseline resampling approaches using the same XGBoost classifier.

DOI: 10.46481/jnsps.2027.3495

Keywords: SMOTE, ADASYN, Class imbalance, Breast cancer diagnosis, Data resampling

Article History:
Received: 08 May 2026
Received in revised form: 06 August 2026
Accepted for publication: 08 October 2026
Available online: 18 November 2026

© 2027 The Author(s). Published by the Nigerian Society of Physical Sciences under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

Communicated by: S. Folorunso

1. Introduction

Breast cancer affects approximately 2.4 million women annually, with early detection being critical for survival rates. However, in the WDBC dataset used in this study, 212 of the 569 observations are malignant and 357 are benign, corresponding to approximately 37.3% malignant and 62.7% benign cases. Although this imbalance is less severe than that encountered in some clinical screening datasets, it remains relevant because the malignant class represents the clinically important minority class [1]. Most medical datasets exhibit class imbalance where the number of benign cases far exceeds malignant cases. This imbalance often leads to biased models that favour the majority class, resulting in poor detection of critical minority cases. The Synthetic Minority Oversampling Technique (SMOTE) and its variants have been widely used to address this issue. However, standalone SMOTE has limitations, including sample overlap,

*Corresponding Author Tel. No.: +234-803-064-2825.
Email address: yusufkutigi1970@unimaid.edu.ng (Yusuf Abubakar Kutigi)

noise amplification and boundary distortion.

Traditional machine learning algorithms assume balanced class distributions and therefore struggle in such scenarios. Oversampling techniques have been widely adopted to address this issue by artificially increasing the number of minority samples. Consequently, recent research has focused on improved SMOTE frameworks, including ADASYN-enhanced methods and other SMOTE variants, but these have been studied separately to improve sample realism and class separability [2, 3].

This study introduces an improved SMOTE framework that combines adaptive sampling, clustering and density-based filtering to generate high-quality synthetic data while preserving the structure of the original dataset to improve minority-class detection. The contributions of this study include (1) adaptively generating minority malignant samples through ADASYN, (2) preserving intrinsic tumour pattern structures using cluster-based oversampling and (3) eliminating noisy synthetic observations through density-based filtering. The integration of these three mechanisms, namely, ADASYN, Cluster-Based SMOTE and DBF produces a balanced, structured and noise-reduced dataset, thereby improving classifier sensitivity, precision, robustness and overall diagnostic performance to produce quality synthetic instances while keeping original dataset robust compared with conventional oversampling approaches [1, 4].

2. Literature review of related works

2.1. Basic SMOTE and its limitations

Chawla *et al.* [5] introduced SMOTE by generating synthetic minority instances through linear interpolation between existing minority samples and their k nearest neighbours. SMOTE effectively expanded the minority decision region and significantly reduced the overfitting associated with random oversampling. However, standard SMOTE operates blindly, often generating noisy synthetic samples in overlapping regions between the minority and majority classes, which can obscure the true decision boundary.

To address the limitations of basic interpolation in complex medical domains, Gurcan and Soylu [6] explored advanced generative architectures. For instance, these authors utilized Deep Generative Adversarial Networks (GANs) for synthetic boosted resampling to improve cancer prediction from highly imbalanced datasets. While deep generative models can capture complex, non-linear data distributions, they are often computationally prohibitive, prone to mode collapse, and lack interpretability. Therefore, there is a strong motivation to utilize optimized, density-aware variants of the SMOTE family, which offer a computationally efficient and highly effective alternative for tabular medical datasets like the WDBC dataset.

The integration of density-awareness into oversampling mechanisms has proven critical for preserving decision boundaries in complex diagnostic tasks. Irfan *et al.* [2] demonstrated that incorporating density-aware filtering into the SMOTE framework drastically improves diagnostic accuracy in complex fault diagnosis. By evaluating the local data density, their approach successfully filtered out synthetic samples generated in noisy, overlapping regions, ensuring that the augmented data strictly reinforced the minority class manifold. This principle of density-based filtering is paramount in medical diagnosis, where the cost of a false negative (misclassifying a malignant tumour as benign) is critically high, and the decision boundary must remain pristine.

However, empirical evaluations by Wongvorachan *et al.* [7] demonstrated that while standard SMOTE outperforms basic resampling techniques, its blind interpolation mechanism is fundamentally restricted to the convex hull of the minority class. Consequently, standard SMOTE frequently generates noisy synthetic instances in the overlapping regions between minority and majority classes, obscuring the true decision boundary and degrading classifier performance in complex diagnostic tasks [8].

2.2. Justifying ADASYN and cluster-SMOTE

To overcome the strict geometric limitations of standard SMOTE, recent literature emphasised the necessity of expanding the sample generation space. Li *et al.* [9] proposed an Improved SMOTE algorithm that explicitly relaxes interpolation constraints, proving that expanding the generation space is critical for capturing the true, extended manifold of the minority class. This principle directly justifies the first two stages of the proposed Improved SMOTE. The authors demonstrated that incorporating density –awareness into the SMOTE framework drastically improves diagnostic accuracy by focusing on synthetic generation on complex, underrepresented regions of the feature space. Also, the cluster SMOTE stage further refines this expansion by partitioning the minority class into localized sub-clusters before interpolation. By expanding the generation space within these localized clusters rather than the global minority space, Cluster SMOTE preserves the underlying sub structures of the data, ensuring a more representative synthetic distribution.

2.3. Preserving the decision boundary: Justifying DBF

While expanding the generation space (via ADASYN and Cluster SMOTE) improves minority representation, it inherently increases the risk of generating synthetic samples in ambiguous, overlapping class boundaries. To mitigate this, the third stage of the proposed framework, DBF is strictly necessitated by recent advancements in local density estimation. Taccaliti and Aguilar-Ruiz [10] demonstrated the efficacy of Kernel Density Estimation (KDE) based synthetic sampling for imbalanced genomic data. By modeling the underlying probability density function, their approach ensures that synthetic samples are retained only in high density, safe regions while strictly avoiding low density areas indicative of class overlap. Building upon this, Lyu *et al.* [11] introduced LD-SMOTE which utilizes local density estimation in overlapping regions. These studies conclusively proved that integrating a local density threshold into the oversampling pipeline is paramount.

In the proposed framework, the DBF stage operationalizes these findings by acting as a strict post generation gatekeeper, discarding any synthetic samples that fall below a minimum

local density threshold, thereby preserving a pristine, noise-free decision boundary.

2.4. Improved sample generation beyond traditional SMOTE

Li *et al.* [9] proposed an improved SMOTE algorithm designed to expand the sample generation space during oversampling. The study noted that traditional SMOTE and several of its variants often fail to adequately consider local data distribution characteristics, resulting in synthetic samples that may not accurately represent the underlying minority class structure. To address this situation, the authors introduced a modified oversampling strategy that generates more representative minority class instances while preserving distributional properties. Experimental evaluation across multiple benchmark datasets showed improved classification performance compared with conventional SMOTE-based methods.

ADASYN focuses on learning harder regions but may still introduce noise in highly overlapping areas [2].

Cluster –based SMOTE applies clustering (for example, k-means) before oversampling to preserve local data structure. Synthetic samples are generated within clusters rather than globally, improving class separability [12]. Demonstrated that clustering improves the quality of generated samples. Additionally, density-based filtering methods have been used to remove noisy or outlier samples after oversampling. These approaches improve model generalization by eliminating unreliable synthetic points.

2.5. Density-aware oversampling for imbalanced classification

Lyu *et al.* [11] introduced LD-SMOTE, a local density estimation based oversampling technique developed to overcome limitations associated with traditional interpolation-based methods. Unlike SMOTE, LD-SMOTE estimates local data density and uses this information to control both the number and location of generated samples. The authors reported improvements in accuracy, F-measure, G-mean and AUC across nineteen benchmark datasets. Their findings suggest that density-aware oversampling can generate more representative minority class samples while reducing the risk of producing noisy or overlapping observations. This work provides strong evidence for incorporating density information into oversampling frameworks, particularly for medical datasets where data distributions are often heterogeneous.

2.6. Hybrid resampling and classification performance

Irfan *et al.* [2] proposed a hybrid oversampling and undersampling framework and demonstrated that combining complementary resampling operations can outperform standalone balancing techniques.

Khan *et al.* [3] reviewed the integration of data augmentation and ensemble learning methods for imbalanced classification and reported that combining resampling techniques with advanced learning models generally produces more robust predictive performance than using individual methods in isolation. These findings support the argument that hybrid frameworks are more capable of addressing the complex characteristics of imbalanced datasets than single method approaches.

2.7. Research gap leading to the proposed study

The reviewed literature established a clear evolutionary trajectory in imbalance mitigation, from the foundational but spatially restricted standard SMOTE [5, 7] to the necessity of expanding the generation space to capture complex minority manifolds [2, 11] and finally to the critical requirement of local density estimation to prevent noise injection at class boundaries [10, 11]. However, most existing studies focus on improving only one aspect of the oversampling process, such as adaptive sample generation. For instance, a study might employ ADASYN or DBF, but rarely both in coordinated, sequential manner tailored for clinical dataset. Therefore, there remains a need for a unified imbalance-aware framework that combines ADASYN based adaptive sampling, clustering based structural preservation, density-based filtering to improve breast cancer classification performance.

There is limited work on integrating adaptive sampling, clustering and filtering into a single pipelining framework to address the class imbalance, especially in medical datasets which form the motivation for the study.

3. Methodology of the study

To ensure an unbiased evaluation and prevent data leakage, the WDBC dataset was first partitioned into training (80%) and testing (20%). All resampling operations were performed independently within the training portion to prevent information leakage. The 20% test partition was kept completely untouched throughout model development and was used only for final performance evaluation.

Figure 1 illustrates the proposed pipeline.

The proposed method integrates three components:

Adaptive sampling (ADASYN).

ADASYN generates synthetic samples based on local data difficulty:

$$x_{\text{new}} = x_i + \lambda(x_{zi} - x_i), \tag{1}$$

where $\lambda \in [0, 1]$, x_i is a minority sample, and x_{zi} is its neighbour.

This ensures that more samples are generated in difficult regions.

Cluster-based SMOTE.

The dataset is partitioned into clusters:

$$C = \{C_i, C_2, \ldots, C_k\}. \tag{2}$$

Clustering ensures that synthetic samples follow the natural structure of the data, reducing overlap between classes.

Density-based filtering.

To remove noise, Density-based filtering (DBF) is applied after synthetic minority observations have been generated by the adaptive and cluster-based oversampling stages. Unlike a centroid-distance criterion, this study uses a local inverse-k-nearest-neighbour density measure throughout

$$D(x_{\text{new}}) = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{\|x_{\text{new}} - x_j\| + \varepsilon}. \tag{3}$$

Table 1: Summary of related literature.

Improved SMOTE stage	Core functions	Supporting literature
Baseline critique	Identifies the need to move beyond standard SMOTE	Chawla *et al.* (2002); Wongvorachan *et al.* (2023)
ADASYN	Adaptively expands the generation space, focusing on synthetic creation in low-density areas	Li *et al.* (2025); Latif *et al.* (2023)
Cluster-based SMOTE	Preserves local data substructures by interpolating within localized minority clusters rather than globally	Li *et al.* (2025); Latif *et al.* (2023)
DBF	Removes synthetic samples in overlapping or low-density regions	Taccaliti & Aguilar-Ruiz (2025); Lyu *et al.* (2025)
Methodological choice	Justifies using SMOTE variants over GANs for clinical data	Gurcan & Soylu (2024)

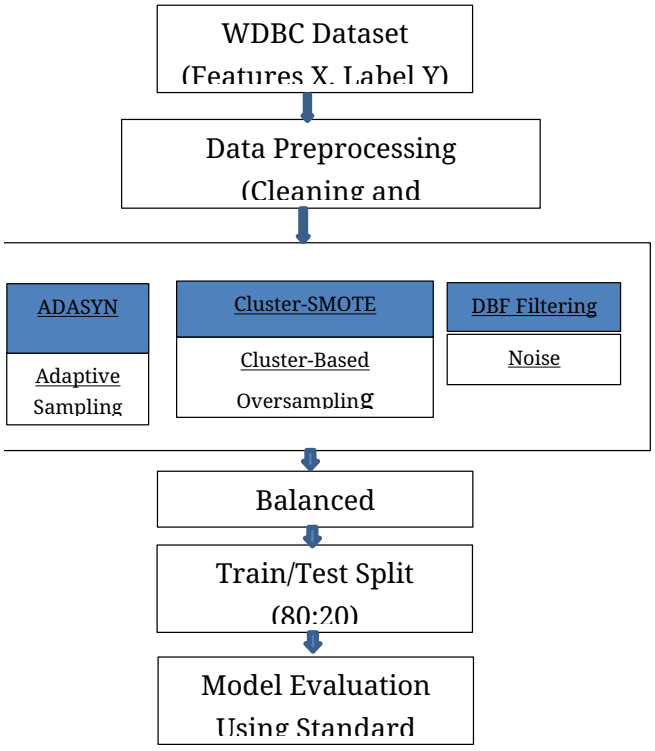


Figure 1: Improved SMOTE pipeline framework.

3.1. Improved SMOTE technique

The overall goal is to improve synthetic sampling for breast cancer diagnosis, by (1) Focusing more on hard-to-learn areas (ADASYN), (2) Organizing similar samples (Clustering) (3) Removing noisy points (DBF).

Step 1: Measure local imbalance.

To quantify how difficult each minority sample is to learn based on its local neighbourhood composition, then allocate synthetic generation effort proportionality.

For each $x_i \in C_j$, find the result of

$$r_i = \frac{\Delta_i}{k_{\text{nn}}}, \tag{4}$$

where Δ_i represents the number of majority-class neighbours around x_i, and k_{nn} represents the number of neighbours considered (typically, $k_{\text{nn}} = 5$). The difficulty score r_i is bounded in $[0, 1]$; higher scores indicate samples that are harder to learn.

Then, we normalize the difficulty score so that the data will be adjusted to a common scale

$$r'_i = \frac{r_i}{\sum_{x_j \in C_j} r_j}, \tag{5}$$

where r'_i is the normalized difficulty score for x_i, and $\sum_{x_j \in C_j} r_j$ is the sum of difficulty scores for all minority samples within cluster C_j.

The normalization converts the raw difficulty scores into a probability distribution within each cluster.

The normalized scores satisfy:

$$\sum_{x_i \in c_j} r'_i = 1. \tag{6}$$

This ensures that the total synthetic budget for each cluster is distributed proportionally among its members based on their relative difficulty.

Allocate Synthetic Samples per instance,

then, we have,

$$g_i = r'_i \cdot G_j, \tag{7}$$

where G_j represents the total synthetic samples to generate in cluster c_j ($G_j = [(n_{\text{maj}} - n_{\text{min}})/k]$ for balanced allocation across clusters), g_i is the number of new samples to create for x_i, and r'_i is the normalized difficulty ratio from Eq. (5).

This equation distributes the cluster level synthetic budget among individual minority samples based on their normalized difficulty. Samples with higher receive more synthetic neighbours, while samples in safe regions receive fewer.

The total number of synthetic samples generated in a cluster should satisfy

$$\sum_{x_i \in c_j} g_i = G_j. \tag{8}$$

Step 2: Generate synthetic samples.

A new minority-class sample is created by interpolating between existing minority instances within the same cluster.

For each sample $x_i \in C_j$, select a random neighbour $x_{nn} \in C_j$,

$$x_{new} = x_i + \lambda \cdot (x_{nn} - x_i), \tag{9}$$

where λ is a random scalar drawn uniformly from 0 to 1, x_{new} is the newly generated synthetic sample, x_i is the base minority sample, and x_{nn} is a randomly selected minority neighbour of x_i within the same cluster. The difference $x_{nn} - x_i$ is the directional vector from x_i to x_{nn}.

Step 3: Cluster the minority samples.

From,

$$X_{minority} = \{x_1, x_2, \ldots, x_m\}, \tag{10}$$

where $X_{minority}$ denotes the complete set of minority-class samples (malignant samples), x_i is the ith minority sample for $i \in \{1, 2, \ldots, m\}$, and m is the total number of minority-class samples.

Clustering is then applied

$$C = \{c_1, c_2, \ldots, c_k\}, \tag{11}$$

where C is the set of clusters formed from $X_{minority}$, c_i is the ith cluster containing similar minority samples, and k is the number of clusters.

A clustering algorithm, k-means, partitions $X_{minority}$ into k disjoint subsets such that $\bigcup_{i=1}^{k} c_i = X_{minority}$ and $c_i \cap c_j = \varnothing$ for $i \neq j$.

This ensures that every minority sample belongs to exactly one cluster.

Step 4: Apply density-based filtering.

The local density is computed by Eq. (12):

$$D(x_{new}) = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{\|x_{new} - x_j\| + \varepsilon}, \tag{12}$$

subject to $x_j \neq x$. Here $D(x_{new})$ is the local density score of the synthetic sample, k_{nn} denotes the number of nearest neighbours, x_j is the jth nearest neighbour of x_{new}, and $\|x_{new} - x_j\|$ is the Euclidean distance between the synthetic sample and that neighbour. The inverse-distance weight $1/(1 + \|x_{new} - x_j\|)$ ensures greater weight for closer neighbours.

Eq. (12) computes a density score based on inverse distance. Unlike a simple count based density, this formulation is continuous and differentiable.

This establishes a baseline density threshold. Synthetic samples with density below this average are considered "less representative" of the minority class manifold.

Apply Density Threshold for Filtering

The density multiplier is defined as

$$\theta = \frac{\gamma}{\overline{D}}, \tag{13}$$

Table 2: DBF reproducibility parameters.

DBF parameters	Experimental specifications
Density definition	Mean inverse k-nearest-neighbour distance
Reference observations	Original minority training samples
Number of neighbours (k_d)	5
Distance metric	Euclidean
Feature space	Standardized training feature space
Self-neighbour rule	Self-neighbour explicitly excluded for original training observations
Zero-distance safeguard	$\varepsilon = 10^{-12}$
Threshold	$\tau = \alpha\bar{\rho}_{train}$
Density multiplier (α)	1.0
Retention rule	Retain a synthetic observation if $\rho_s \geq \tau$; discard otherwise
Samples subject to DBF	Synthetic samples only
Test-set involvement	None

where γ is the density threshold and $\overline{D}$ is the average local density, expressed as

$$\overline{D} = \left(\frac{1}{n} \sum \text{Density}(x_{new})\right). \tag{14}$$

The synthetic sample x_{new} is retained if

$$\text{Density(new)} \geq \theta \cdot \overline{D}, \tag{15}$$

where θ represents the density-threshold multiplier, $\overline{D}$ is the average density, and Density(new) is the local density of the synthetic sample.

This filters out low-density (noisy) synthetic samples.

3.2. Improved SMOTE equations put together

Putting all the three equations together, we have

$$\begin{aligned} D(x_{new}) &= \frac{1}{k} \sum_{j=1}^{k} \frac{1}{1 + \|x_{new} - x_j\|} \\ &\geq \theta \cdot = \left(\frac{1}{n} \sum \text{Density}(x_{new})\right). \end{aligned} \tag{16}$$

3.3. Classification model

The effectiveness of Improved SMOTE method was evaluated using XGBoost classifier due to its strong performance in structured datasets [13]. To avoid information leakage, particle swarm optimization (PSO) was conducted using only the model-development training data. The independent hold-out test partition was excluded from the optimization, resampling, and cross-validation procedures and was used only for the final evaluation of the optimized model. PSO algorithmic description of optimization procedure is as follows:

Table 3: Dataset column encoding.

Target	Count	Class name	Percent
0	212	Malignant	37.26
1	357	Benign	62.74

1. Define the PSO search range as $N_{est} \in [100, 1000]$.
2. Initialize a swarm of 20 candidate values.
3. For each candidate, train XGBoost using five-fold stratified cross-validation.
4. Compute mean validation accuracy.
5. Define fitness as the negative mean accuracy because PSO minimizes the objective.
6. Update particle positions according to their personal and global best solutions.
7. Repeat for a maximum of 10 iterations.
8. Select the value of $n_{\text{estimators}}$ with the highest cross-validated accuracy.
9. Use the optimized XGBoost model for recursive feature elimination (RFE) and subsequent final model training.
10. Evaluate the final model only on the untouched 30% hold-out test set.

4. Results and discussion

4.1. Dataset description

The Wisconsin Diagnostic Breast Cancer dataset (https://www.kaggle.com/datasets/erdemtaha/cancer-data/) was collected from Kaggle. For the present study, a Kaggle-hosted copy of the dataset was used. The downloaded file was checked against the canonical WDBC schema to verify the identifier, diagnosis variable and 30 diagnostic features before processing. The dataset was originally introduced by Wolberg *et al.* [14] and is distributed through the UCI Machine Learning Repository. It contains 569 observations and the identification variable was excluded from model development because it has no diagnostic meaning while the diagnostic variable was retained as the target variable and encoded as shown in Table 3

Table 3 showed the malignant (212) and Benign (357) counts and their target and percentages.

Figure 2 compares model performance.

Figure 3 shows the class distributions before and after resampling.

Figure 2 showed a comparison of the various Improved SMOTE model performances which is a direct graphical representation of the model performance. The accuracy has improved to 96.49% over the Cluster SMOTE and DBF which attain a 94.12% accuracy.

From Figure 3, (Table 4 values), is the class distribution of the WDBC training data before and after resampling. Class 0 represents benign observation and class 1 represents malignant observations. The improved resampling was applied to the original dataset (569, 30) which was split into training distribution

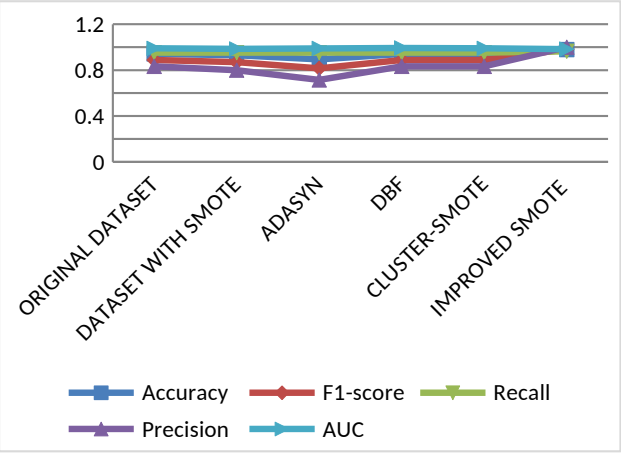


Figure 2: Comparison of the various improved SMOTE model performances.

(398, 30) and testing hold out (171, 30) from its original distribution.

Figure 4 presents the confusion matrix.

The confusion matrix demonstrates that the model achieved very strong classification performance and the classifier correctly classified 159 out of 171 observations. Out of the 71 actual malignant cases, 61 were correctly identified while 10 were incorrectly classified as benign.

Table 5 presented state-of-the-art studies which specifically use WDBC to prove that their improved oversampling techniques can handle complex, overlapping medical data without generating noise. However, the proposed model because of its more robust techniques outperformed most of them in standard metric measurements.

4.2. Discussion of results

From Figure 2 and Table 4 above, the accuracy has improved to 96.49% over the Cluster SMOTE and DBF which attain a 94.12% accuracy, F1 score of 97.22%, Recall and precision of 97.22 indicating that the model improved the class imbalance. The table and figures indicated that, there are no major shifts when DBF and cluster-based SMOTE was applied on the datasets. Figure 2 presented a line chart of the breast cancer classification before and after SMOTE techniques, the Improved SMOTE achieved the highest reported accuracy, precision, recall and F1 scores among the evaluated approaches, while its ROC-AUC was lower than that of the baseline approaches. The integration of ADASYN ensures better representation of difficult regions, while clustering preserves structural integrity. Density-based filtering removes noise, resulting in cleaner datasets.

From Table 4, it is evident that the Improved SMOTE yields substantial improvements in Accuracy, Precision and F1 score compared to the original dataset. Also the complete Improved SMOTE pipeline produced the best reported threshold dependent metrics, but it produced a lower ROC-AUC. The lower

Table 4: Breast cancer dataset classification outcomes before and after SMOTE techniques.

Technique	Accuracy	F1 score	Recall	Precision	AUC
ORIGINAL DATASET	0.9412	0.8889	0.9524	0.8333	0.9903
DATASET WITH SMOTE	0.9294	0.8696	0.9524	0.8000	0.9851
ADASYN	0.8941	0.8163	0.9524	0.7143	0.9896
DBF	0.9412	0.8889	0.9524	0.8333	0.9926
CLUSTER-SMOTE	0.9412	0.8889	0.9524	0.8333	0.9903
IMPROVED SMOTE	0.9649	0.9722	0.9722	0.9722	0.9623

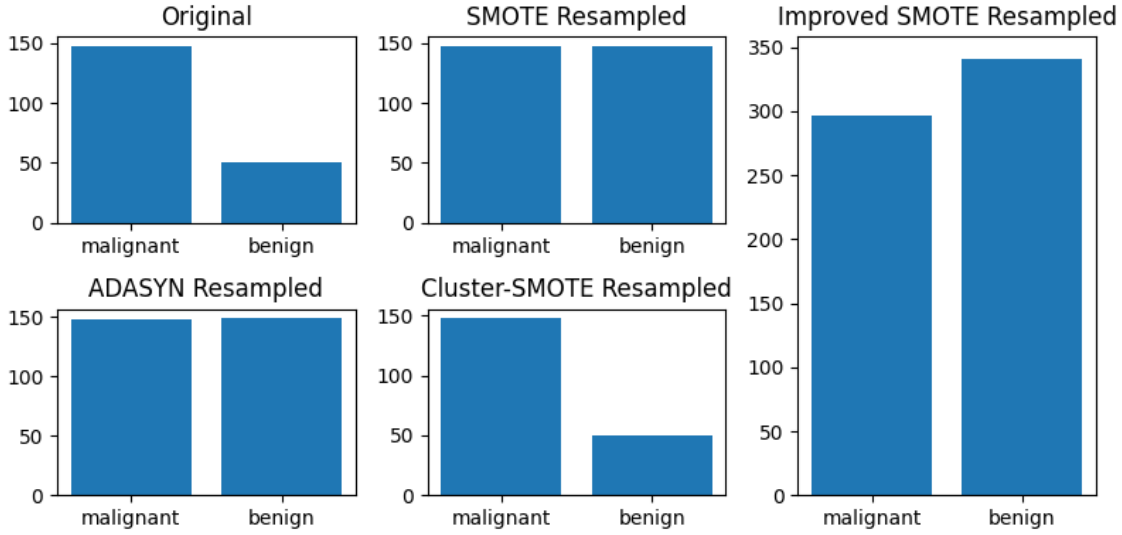


Figure 3: Class distributions before and after resampling.

Table 5: Model comparison with state-of-the-art models.

Authors and year	SMOTE variant evaluated	Verified DOI	Specific application to WDBC
Gurcan & Soylu (2024)	Advanced Resampling (ADASYN, BSMOTE, SVM-SMOTE)	10.3390/cancers16234046	Evaluated 10 classification algorithms on the WDBC
Assegie *et al.* (2024) [15]	ADASYN (Adaptive Synthetic Sampling)	10.1016/j.procs.2024.04.095	Evaluated on WDBC and demonstrated that using ADASYN, with SVM Logistic Regression is more effective
Hardiyanti P. (2025) [16]	Standard SMOTE + Boruta Feature Selection	10.31763/sitech.v6i1.2109	Tested on WDBC combining SMOTE for balancing with Boruta for feature optimization
Proposed Model	ADASYN, Cluster Base SMOTE and DBF		Tested on WDBC using XGBoost classifier

AUC is an observed result; its specific cause cannot be established from the present experiment. DBF related explanations are hypotheses requiring ablation and score distribution analysis.

In breast cancer diagnosis, false negatives are particularly consequential because failure to identify a malignant case may delay further diagnostic investigation and treatment. Therefore, sensitivity or recall for the malignant class is a critical performance measure. False positives are also clinically relevant because they may lead to unnecessary follow-up procedures, consequently precision and specificity should also be considered when evaluating the diagnostic model.

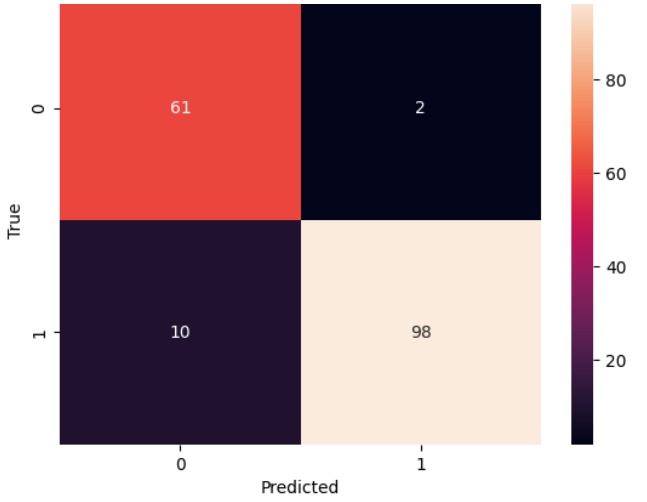


Figure 4: Confusion matrix for the improved SMOTE.

5. Conclusion, limitations and recommendations for future work

This study proposed an improved SMOTE-based hybrid (pipelining) resampling framework for breast cancer diagnosis. By sequentially integrating of adaptive sampling, clustering and density-based filtering, the method addresses the limitations of traditional oversampling techniques. The experimental results demonstrate that the proposed approach significantly improves classification performance, making it suitable for imbalance-aware medical diagnosis systems.

A limitation for the improved SMOTE is a slight reduction in the AUC metric compared to the baseline. One possible explanation is that density-based filtering alters the distribution of synthetic training observations, which may affect the ranking of model prediction scores. However, the current experiment does not establish that DBF necessarily reduces ROC-AUC. Further ablation and score distribution analyses would be required to establish the mechanism. Future work should look at applying the framework to larger and more complex datasets. Future studies may also investigate probability calibration methods such as Platt scaling to assess and improve the calibration of predicted probabilities using calibration curves and the Brier score. Because ROC-AUC evaluates ranking performance, improvement in calibration alone should not be assumed to increase AUC.

Data availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

Funding

The authors received no specific funding from any public, commercial, or not-for-profit organisation for this research.

References

[1] S. Susan & A. Kumar, "The balancing trick: Optimized sampling of imbalanced datasets—a brief survey of the recent state of the art", Engineering Reports **3** (2021) e12298. https://doi.org/10.1002/eng2.12298.

[2] M. Irfan, Z. Mushtaq, N. A. Khan, S. N. F. Mursal, S. Rahman, M. A. Magzoub, M. A. Latif, F. Althobiani, I. Khan & G. Abbas, "A scalo gram-based CNN ensemble method with density-aware SMOTE oversampling for improving bearing fault diagnosis", IEEE Access **11** (2023) 127783. https://doi.org/10.1109/ACCESS.2023.3332243.

[3] A. A. Khan, O. Chaudhari & R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation", Expert Systems with Applications **244** (2024) 122778. https://doi.org/10.1016/j.eswa.2023.122778.

[4] T. Chen & C. Guestrin, *XGBoost: A scalable tree boosting system*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, USA, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall & W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique", Journal of Artificial Intelligence Research **16** (2002) 321. https://doi.org/10.1613/jair.953.

[6] F. Gurcan & A. Soylu, "Synthetic boosted resampling using deep generative adversarial networks: A novel approach to improve cancer prediction from imbalanced datasets", Cancers **16** (2024) 4046. https://doi.org/10.3390/cancers16234046.

[7] T. Wongvorachan, S. He & O. Bulut, "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining", Information **14** (2023) 54. https://doi.org/10.3390/info14010054.

[8] O. Njama-Abang, D. U. Ashishie & P. T. Bukie, "Addressing class imbalance in lassa fever epidemic data, using machine learning: A case study with SMOTE and random forest", Journal of the Nigerian Society of Physical Sciences **7** (2025) 2586. https://doi.org/10.46481/jnsps.2025.2586.

[9] Y. Li, Y. Yang, P. Song, L. Duan & R. Ren, "An improved SMOTE algorithm for enhanced imbalanced data classification by expanding sample generation space", Scientific Reports **15** (2025) 23521. https://doi.org/10.1038/s41598-025-09506-w.

[10] E. Taccaliti & J. S. Aguilar-Ruiz, "Improving classification on imbalanced genomic data via KDE-based synthetic sampling", BioData Mining **18** (2025) 60. https://doi.org/10.1186/s13040-025-00474-5.

[11] J. Lyu, J. Yang, Z. Su & Z. Zhu, "LD-SMOTE: A novel local density estimation-based oversampling method for imbalanced datasets", Symmetry **17** (2025) 160. https://doi.org/10.3390/sym17020160.

[12] H. He, Y. Bai, E. A. Garcia & S. Li, *ADASYN: Adaptive synthetic sampling approach for imbalanced learning*, 2008 IEEE International Joint Conference on Neural Networks, Hong Kong, China, 2008, pp. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969.

[13] G. Douzas & F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks", Expert Systems with Applications **91** (2018) 464. https://doi.org/10.1016/j.eswa.2017.09.030.

[14] W. H. Wolberg, W. N. Street & O. L. Mangasarian, "Image analysis and machine learning applied to breast cancer diagnosis and prognosis", Analytical and Quantitative Cytology and Histology **17** (1995) 77. https://pubmed.ncbi.nlm.nih.gov/7612134/.

[15] T. A. Assegie, A. O. Salau, K. Sampath, R. Govindarajan, S. Murugan & B. Lakshmi, "Evaluation of adaptive synthetic resampling technique for imbalanced breast cancer identification", Procedia Computer Science **235** (2024) 1000. https://doi.org/10.1016/j.procs.2024.04.095.

[16] C. P. Hardiyanti, & S. Sarina, "Optimizing breast cancer classification using SMOTE, Boruta, and XGBoost", Science in Information Technology Letters **6** (2025) 16. https://doi.org/10.31763/sitech.v6i1.2109.