Journal of Contemporary Academic Research and Methodologies

An Online Peer Reviewed / Refereed Journal
Volume 1 | Issue 5 | June 2026
Website: www.journal.jcarm.in

Original Research Article

COMPARATIVE PERFORMANCE ANALYSIS OF SYNTHETIC MINORITY OVERSAMPLING TECHNIQUES (SMOTE) ON MEDICAL DATASETS BASED ON EXTREME GRADIENT BOOSTING ESTIMATOR

Yusuf, A. K[1*], Hassan, Adam[2], Yusuf A[3]

[1] Centre for Entrepreneurship and Enterprise Development, University of Maiduguri, Nigeria.
[2] Department of Software Engineering, Al-Ansar University, Maiduguri, Nigeria.
[3] Department of Computer Engineering, University of Maiduguri, Nigeria.

Correspondence should be addressed to Yusuf, A. K: yusufkutigi1970@gmail.com

Article No: 040 | Accepted: 30 June 2026 | Published: 10 July 2026

Abstract: Medical datasets frequently face issues with class imbalance and redundant features, which can undermine the accuracy and reliability of predictive diagnostic models. This research conducts a performance comparison of four variants of the Synthetic Minority Oversampling Technique (SMOTE) SMOTE-ENN, Borderline SMOTE, ADASYN, and SMOTE-Tomek Links when paired with feature selection and the Extreme Gradient Boosting (XGBoost) classifier for predicting breast cancer and heart disease. The publicly available datasets underwent preprocessing to eliminate noise, balance class distributions, and identify the most significant diagnostic features. Model performance was assessed using standard metrics, including accuracy, precision, recall, F1-score, and Cohen's kappa. For the heart disease dataset, the SMOTE-ENN technique produced the highest results, achieving an accuracy of 56.15%, a recall of 44.50%, and an F1-score of 27.46%, which underscored improved detection of minority class cases. Conversely, in the breast cancer dataset, ADASYN, Borderline SMOTE, and SMOTE-Tomek Links showed better performance, reaching an accuracy of 96.49%, an F1-score of 97.30%, a recall of 100%, and a kappa score of 0.9231.

Keywords: Smote, Feature Selection, Xgboost, Class Imbalance, Breast Cancer, Heart Disease, Machine Learning, Medical Diagnosis.

1. Introduction

The problem of breast cancer diagnosis is widely recognized as a challenge in classifying benign and malignant tumours (Chen et al., 2023). AI, without a doubt,

https://doi.org/10.5281/zenodo.21353004

continues to record successes in the field of medicine. The interpretability and explainability shortfalls of many AI applications have further increased the difficulty and suitability in clinical practices.

Though several new deep learning approaches have demonstrated their superiority in detecting serious ailments of the heart diseases. As expected, deep neural networks run on a large memory due to their extended processes (Ran et al., 2022). While certain diseases need a broad-minded diagnostic scale, autonomous examination of graded medical images is usually considered as a binary classification task. This misses out on the finer attributes and fundamental interrelationship within the various grades and phases of such diagnosis, but the success of this depends on how well and soon the challenges at the epistemic stage are resolved (Mueller et al., 2022).

1.2 Statement of the Research Problem

The problem of breast cancer diagnosis is widely recognized as a challenge in classifying benign and malignant tumours (Chen et al., 2023). Prompt and accurate diagnosis of breast cancer has been discovered as the most effective approach for treatment. There are still difficulties in obtaining reliable diagnostic systems that excused inconsistency of their interpretation, and aiming both to improve the accuracy in the evaluation and to reduce its uncertainty (Casal-Guisande et al., 2022).

The class imbalance problem is common in several real-life applications in which the underrepresented (minority) class is the class of interest. The SMOTE technique became prominent and most preferred approach for operating on unbalanced dataset (Elreedy et al., 2023).

2. Research Methodology

2.1 The Proposed Model

The entire structure of the proposed diseases datasets classification methods for breast cancer and heart diseases diagnosis is shown in Figure 1

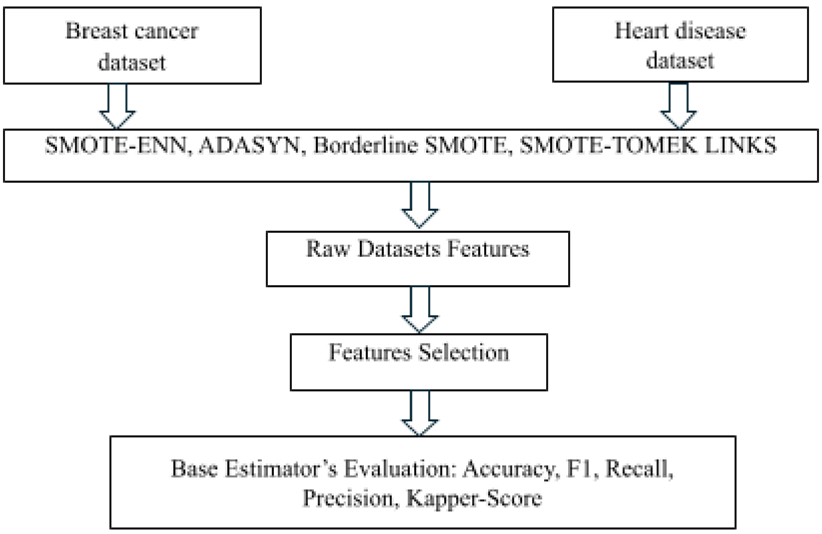


Figure 1. The proposed disease classification methods.

In Figure 1, the breast cancer and heart disease datasets were collected from public repositories such as Kaggle for the purpose of handling the class imbalance issues predominant in these datasets to serve as input to the model, which is the improvement over existing methods. The major activity at the dataset preprocessing step is the difficulty in classifying imbalanced data by classifiers due to their bias towards the majority class within the datasets. Therefore, resampling technique offered by the SMOTE enable addition of minority class samples or subtracting samples from the majority class to generated a resampled data with balanced classes (Joloudari et al., 2023). Thereafter, the balanced datasets are spliced into the train and test datasets for the applications of the DL methods.

Data feature extraction involves grouping of unique attributes inside of the prepared datasets after data class distribution balancing procedure of the previous stage. This is to be achieved through high performance data feature extraction algorithms, and improve location and recognition of cancerous patterns (Allugunti, 2022) as well as reducing the amount of resources needed to accurately display enormous data. The breast cancer regions of interest will be extracted through maximum between-class variance known as Ostu (Zhou et al., 2023).

The ML/DL networks (Hu et al., 2022) are to be built to effectively optimize the variables such as weights, activation functions, and threshold values recursively to attain acceptable value in the gradient-based learning strategies. The high-performance classification algorithms are to be considered for better and early diagnosis of breast cancer and heart diseases through the reduction of training process time and increase performance by considering the loss function (Oliveira et al., 2023), and other hyper-parameters like learning rates, weights, and hidden layers.

2.3 Experimentation

Data preparation and feature extraction were conducted using Google Colaboratory, taking advantage of GPU and remote machine resources due to the extensive capabilities offered by the Python programming language. The datasets used for simulating the proposed model are divided into three segments: 70% for training, 20% for validation, and 10% for testing (Nasir et al., 2022).

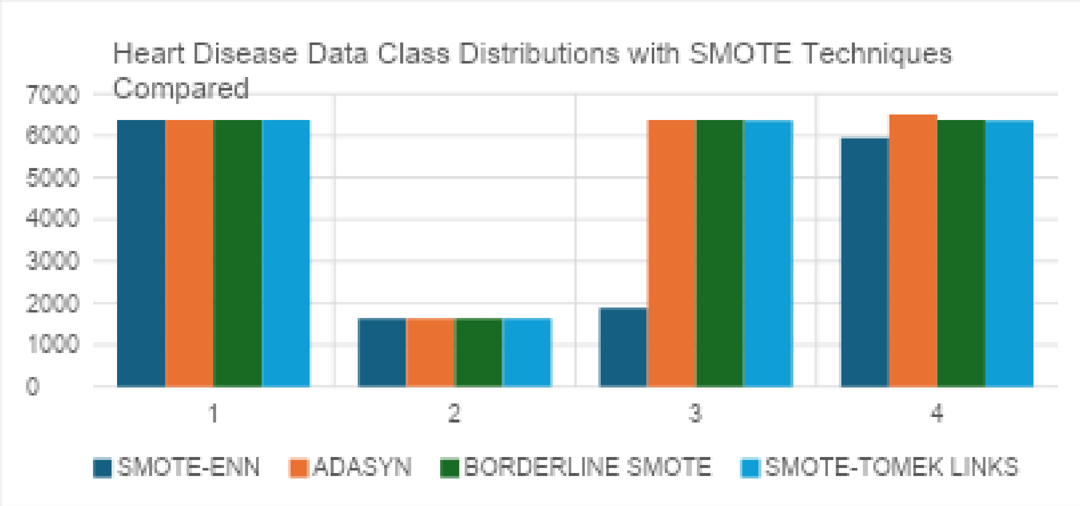
3.1 Outcomes of the SMOTE Techniques with Heart Disease Dataset

The outcomes of the various data class imbalance solutions including: the SMOTE + ENN algorithm, ADASYN, Borderline SMOTE, and SMOTE-TOMEK links before and after their operations on heart disease dataset are shown in Table 1.

Table 1. Heart disease dataset class imbalance solutions compared.

Technique	Class distribution before		Class distribution after	
	Class:0	Class:1	Class:0	Class:1
SMOTE-ENN	6373	1627	1884	5952
ADASYN	6373	1627	6373	6502

BORDERLINE SMOTE	6373	1627	6373	6373
SMOTE-TOMEK LINKS	6373	1627	6354	6354



.

Figure 2. Heart disease dataset class distributions before and after applying SMOTE techniques.

To further understand the effectiveness of the dataset imbalance solutions, the baseline estimator (XGBoost) was considered as presented in Table 2.

Table 2. Heart disease classification outcomes of XGBoost after SMOTE techniques.

Technique	Accuracy	Kappa score	F1-score	Recall	Precision
SMOTE-ENN	0.5615	0.0225	0.2746	0.4450	0.1986
ADASYN	0.7995	-0.0073	0.0291	0.0161	0.1500
BORDERLINE SMOTE	0.7975	-0.0110	0.0288	0.0161	0.1364
SMOTE-TOMEK LINKS	0.7975	-0.0110	0.0288	0.0161	0.1364
WITHOUT SMOTE	0.8000	-0.0128	0.0196	0.0107	0.1143

Figure 3 represents the XGBoost ensemble algorithm classification outcomes after optimizing the class distributions with various SMOTE techniques using the acquired heart disease dataset. Using accuracy measure, SMOTE-ENN technique at 56.15% was smallest when compared 79.75% attained by each of ADASYN, BORDERLINE SMOTE, and SMOTE-TOMEK Links techniques. Accuracy of 80.00% was achieved without SMOTE.

Kappa score for SMOTE-ENN at 0.0225 was the best, followed, BORDERLINE SMOTE and SMOTE-TOMEK at -0.011, and ADASYN at -0.0073. While -0.0128 without SMOTE was the least. Considering F1-score measure, 27.46% for the SMOTE-ENN was the best, trailed immediately by ADASYN at 2.91%, and BORDERLINE SMOTE and SMOTE-TOMEK Links at 2.88%. The least F1-score was observed for prior SMOTE optimization.

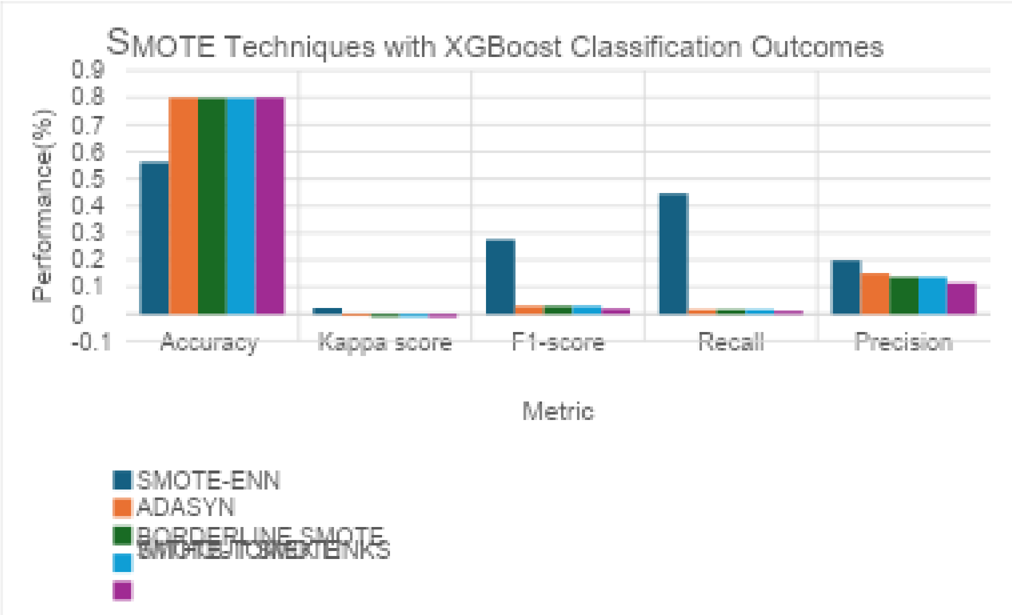
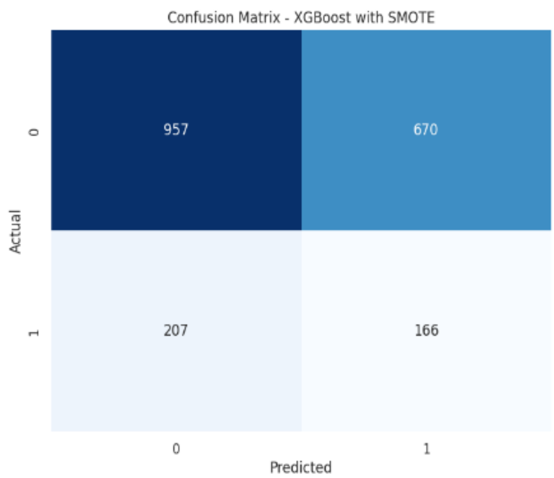
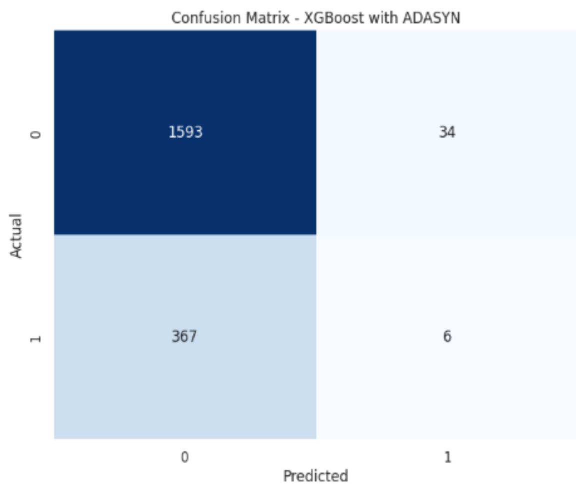


Figure 3. XGBoost heart disease classification outcomes with SMOTE techniques compared.

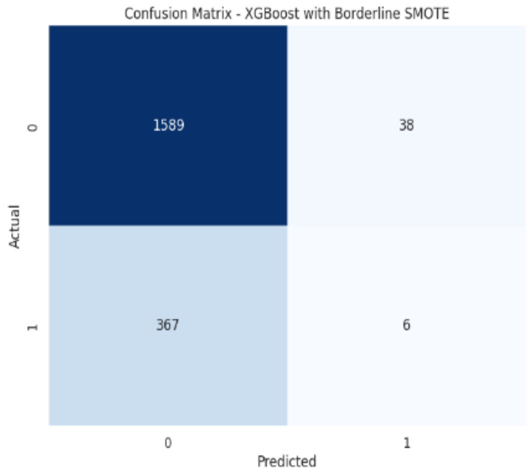
Figure 4 and its subfigures (a)–(f)) compares how different data balancing techniques influence the classification performance of the XGBoost model on the heart disease dataset, using confusion matrices for each case.



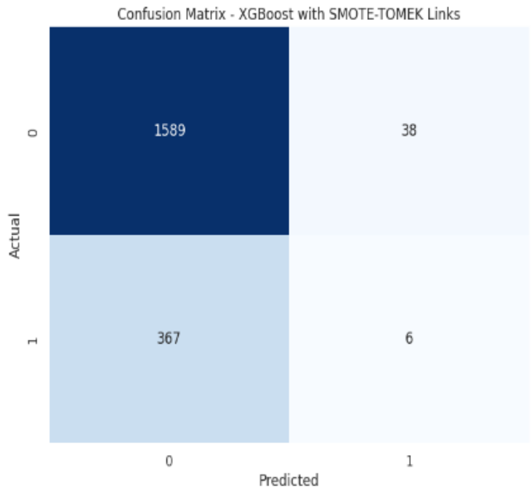
(a) SMOTE-ENN.



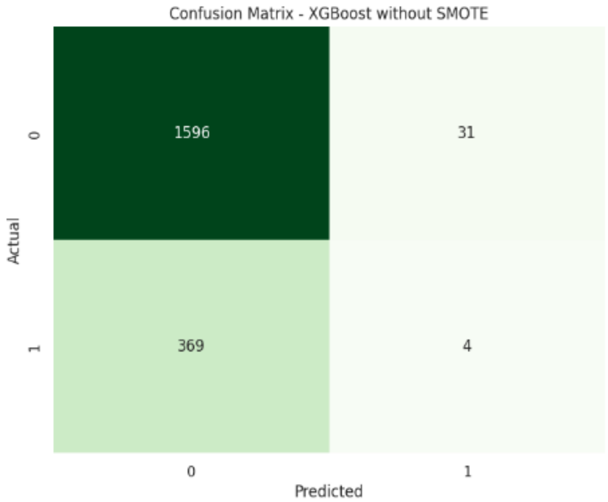
(b) ADASYN



(c) Borderline SMOTE



(d) SMOTE TOMEK Links.



(e) Without SMOTE.

Figure 4. Confusion matrix for XGBoost with SMOTE techniques on heart disease dataset.

Figures 4 (a)–(e) visualize how each resampling technique **changes the confusion matrix** outcome for the same XGBoost classifier. The comparison helps to **evaluate which SMOTE method gives the best balance** between detecting heart disease and avoiding false alarms. Typically, **SMOTE-ENN** or **SMOTE-Tomek Links** yield the most stable results because they both balance and clean the data. Figure 4 illustrates the impact of various SMOTE-based resampling techniques on the XGBoost classifier's confusion matrix using the heart disease dataset. Methods such as SMOTE-ENN and SMOTE-Tomek Links produced the most balanced results with fewer false negatives, indicating better detection of heart disease cases compared to the unbalanced dataset (without SMOTE).

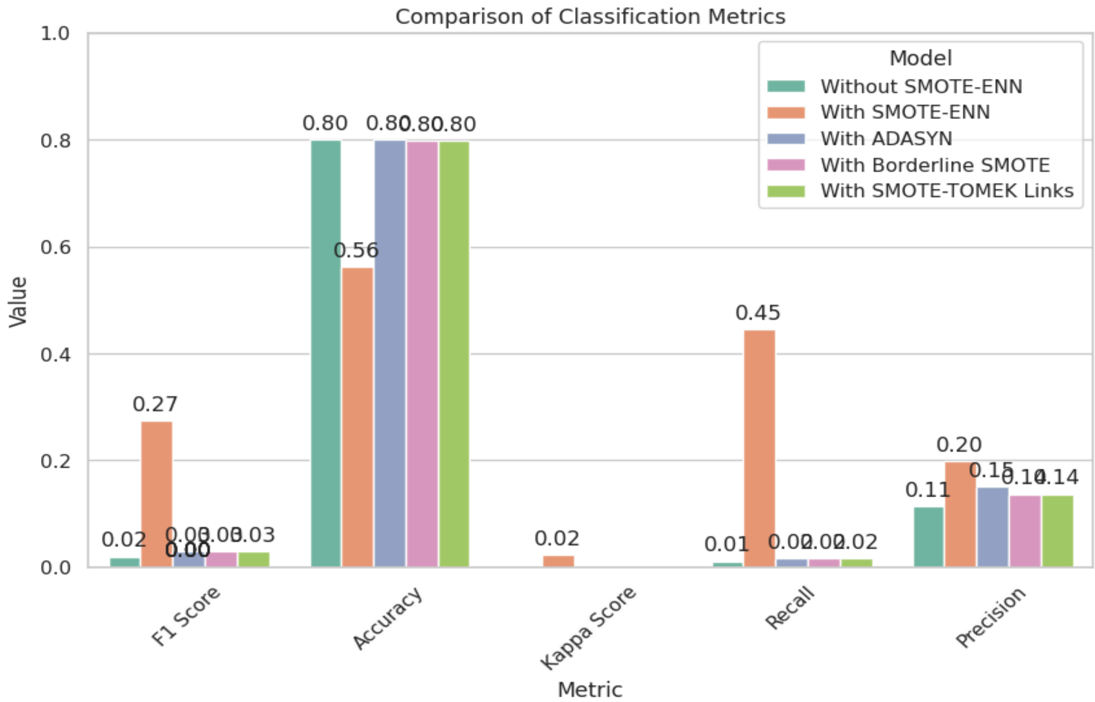


Figure 5. The summary of XGBoost classification outcomes with SMOTE techniques for the heart disease dataset.

Figure 6 presents the five most influential predictors that contribute to the XGBoost model's classification decisions. Feature importance quantifies how much each variable reduces prediction error when included in the model the higher the value, the greater its influence on detecting heart disease.

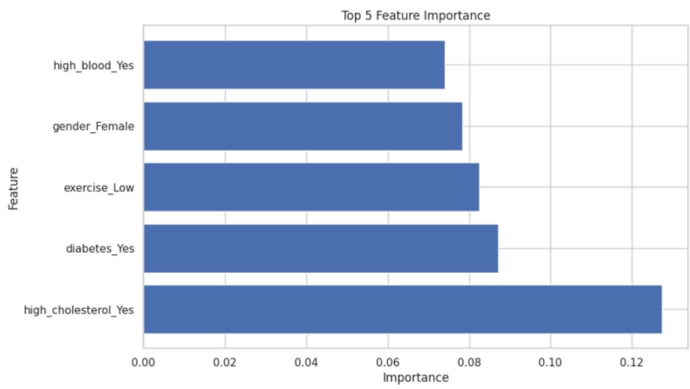


Figure 6. XGBoost model's optimal features importance scores selected from heart disease dataset.

Figure 6 shows that high cholesterol, diabetes, and low exercise levels are the most influential features in predicting heart disease using the optimized XGBoost model. These features align with established clinical risk factors, demonstrating that the model effectively captures medically relevant relationships within the dataset.

4.2 Outcomes of the SMOTE Techniques with Breast Cancer Dataset

The performance of the imbalance optimization approaches based on base estimator, XGBoost algorithm. The outcomes of the class imbalance solutions including SMOTE+ENN algorithm, ADASYN, Borderline SMOTE, and SMOTE-TOMEK links before and after operations on breast cancer dataset are presented in Table 3.

Table 3. Breast cancer dataset class imbalance solutions compared.

Technique	Class distribution before		Class distribution after	
	Class:0	Class:1	Class:0	Class:1
SMOTE-ENN	170	285	264	261
ADASYN	170	285	281	281
BORDERLINE SMOTE	170	285	285	285
SMOTE-TOMEK LINKS	170	285	285	282

Figure 7 shows the outcomes of the different SMOTE techniques applied to the breast cancer disease dataset, which shows significant adjustments in the class distributions prior to the processes of data balancing.

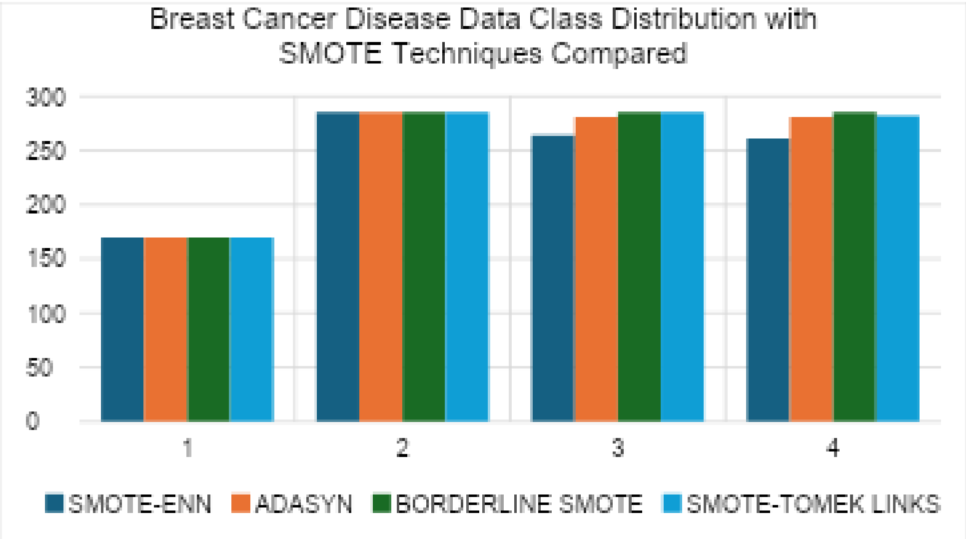


Figure 7: Heart disease data class distributions after applying SMOTE techniques.

To further understand the dataset imbalance solutions effectiveness, other classification models (that is, the Random Forest algorithm, Support Vector Machine, Logistic Regression, Artificial Neural Network) were considered, and compared to the XGBoost algorithm as presented in Table 4.

Table 4. Breast cancer classification outcomes of XGBoost algorithm after SMOTE techniques.

Technique	Accuracy	Kappa score	F1-score	Recall	Precision
SMOTE-ENN	0.9123	0.8057	0.9333	0.9722	0.8974
ADASYN	0.9649	0.9231	0.9730	1.0000	0.9474
BORDERLINE SMOTE	0.9649	0.9231	0.9730	1.0000	0.9474
SMOTE-TOMEK LINKS	0.9649	0.9231	0.9730	1.0000	0.9474
WITHOUT SMOTE	0.9561	0.9034	0.9664	1.0000	0.9351

Table 4 depicts the XGBoost ensemble algorithm classification outcomes after optimizing the class distributions with various SMOTE techniques using the acquired heart disease dataset. Using accuracy measure, SMOTE-ENN technique at 91.23% was the least when compared 96.49% attained by each of ADASYN, BORDERLINE SMOTE, and SMOTE-TOMEK Links techniques. Accuracy of 95.61% was achieved before SMOTE application.

Kappa score for SMOTE-ENN at 0.8057 was the smallest, then, outperformed by BORDERLINE SMOTE, SMOTE-TOMEK, and ADASYN at 0.9231. While 0.9034 without SMOTE was only better than SMOTE-ENN technique. Considering F1-score measure, 93.33% for the SMOTE-ENN was the worst-performed when put side-by-side ADASYN, BORDERLINE SMOTE and SMOTE-TOMEK Links at 97.30%. The F1-score observed prior SMOTE optimization was 96.64%, which lesser compared the post-optimizations.

Using precision measure, SMOTE-ENN techniques attained the least value of 89.74%, closely out-staged at 94.74% by the ADASYN, BORDERLINE SMOTE, and SMOTE-TOMEK Links respectively. The Recall score of 93.51% was computed before applying the SMOTE techniques.

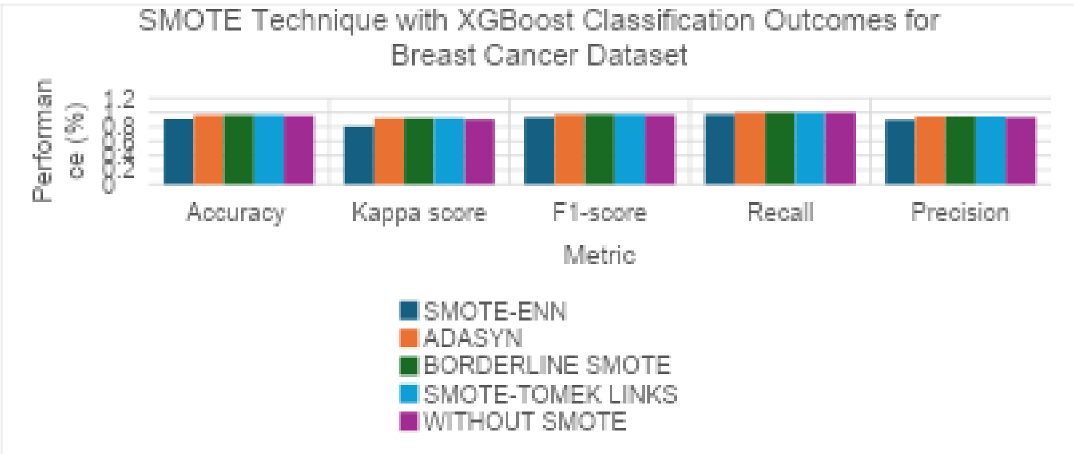
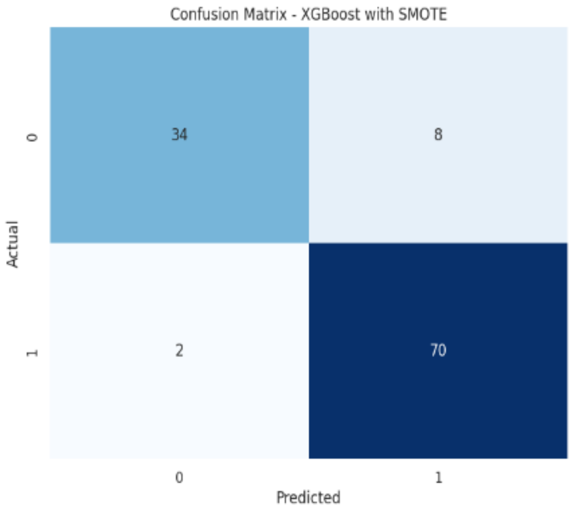
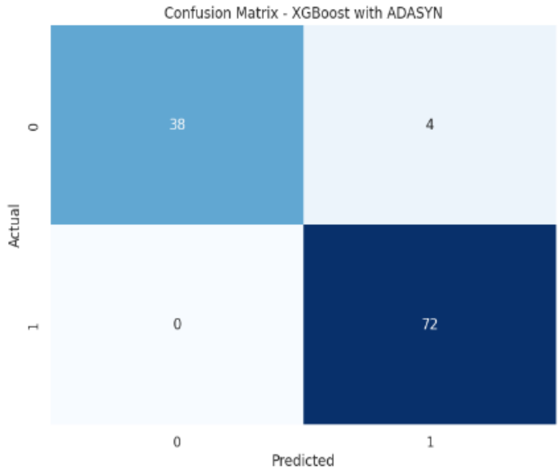


Figure 8. XGBoost breast cancer classification outcomes with SMOTE techniques compared.

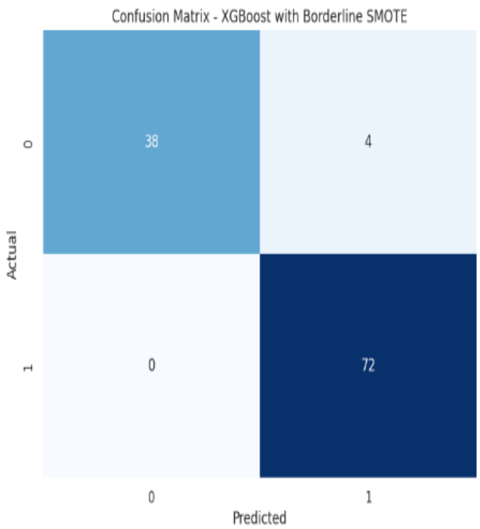
Figure 9 and its subfigures (a)–(f)) compares how different data balancing techniques influence the classification performance of the XGBoost model on the breast cancer dataset, using confusion matrices for each case.



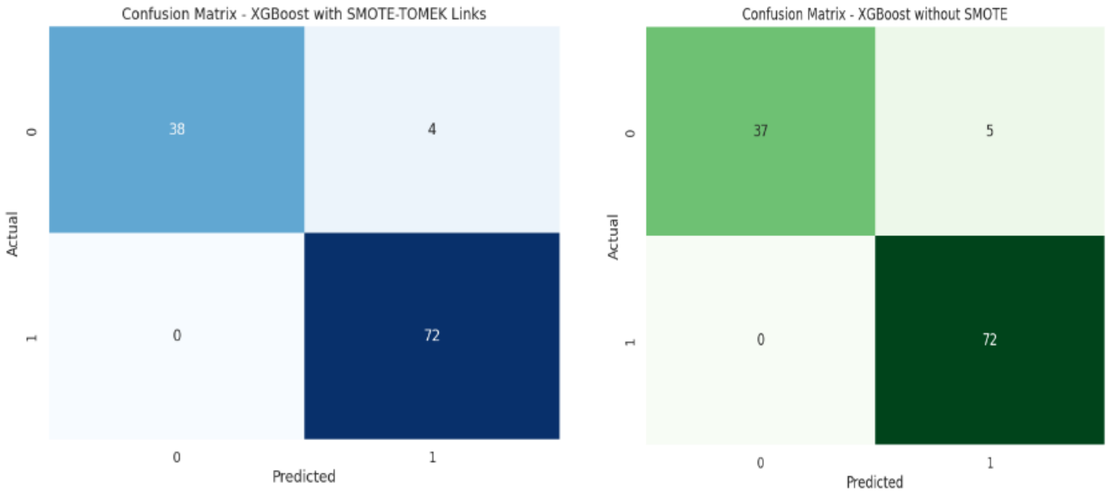
(a) SMOTE-ENN.



(b) ADASYN.



(c) Borderline SMOTE.



(d) SMOTE TOMEK Links.

(e) Without SMOTE.

Figure 9. Confusion matrix for XGBoost with SMOTE techniques on breast cancer dataset.

From Figure 9 (a)-(e) shows how resampling changes class distribution in feature space. Techniques like SMOTE-ENN and SMOTE-Tomek both oversample and clean noise, improving data quality. ADASYN and Borderline-SMOTE focus on hard-to-learn areas (near class boundaries). The "Without SMOTE" image demonstrates why these methods are necessary most models will ignore minority points. In breast cancer detection, ADASYN helps models learn from rare cancerous patterns. SMOTE-ENN improves robustness by cleaning ambiguous cases. Borderline-SMOTE enhances sensitivity near complex boundaries (dataset that look almost normal but contain subtle malignant).

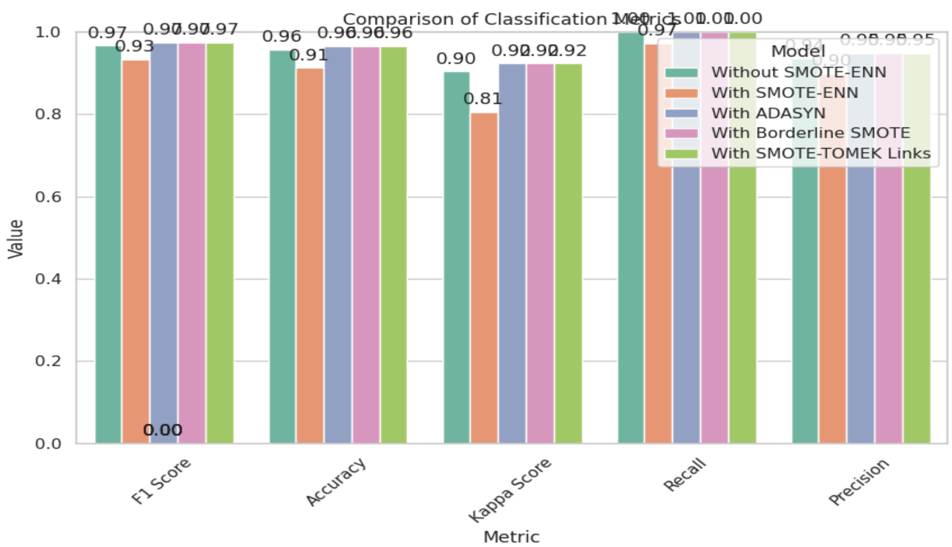


Figure 10. The summary of XGBoost classification outcomes with SMOTE techniques.

Figure 11 displays the top five most influential features selected by the optimized **XGBoost model** for predicting breast cancer. The chart illustrates each feature's relative importance in the model's decision-making process, where higher values indicate stronger influence on the classification outcome between malignant and benign tumors.

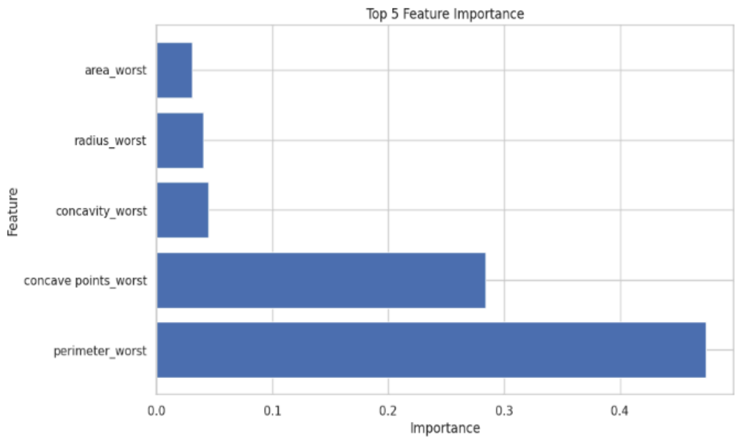


Figure 11. XGBoost model's optimal feature importance scores selected from the breast cancer dataset.

Figure 12 demonstrates that the XGBoost model heavily relies on geometric shape descriptors, particularly perimeter_worst and concave_points_worst, to differentiate malignant from benign breast tissue. These findings align with established medical imaging knowledge, where cell shape irregularities and enlarged perimeters are characteristic markers of cancerous growth.

4.3 Discussion of Results

The comparative evaluation of four variants of the Synthetic Minority Oversampling Technique (SMOTE), specifically SMOTE-ENN, Borderline SMOTE, ADASYN, and SMOTE-Tomek Links, utilizing the Extreme Gradient Boosting (XGBoost) classifier, highlights the significant impact that data resampling and feature optimization have on the predictive performance of medical datasets. The findings demonstrated that class imbalance markedly affects both the accuracy and the generalization ability of machine learning models, particularly in healthcare settings where the minority classes frequently represent crucial diagnoses.

In the evaluation of the heart disease dataset, the SMOTE-ENN technique outperformed all other methods tested, achieving an accuracy of 56.15%, a recall of 44.50%, and an F1-score of 27.46%. These results highlight SMOTE-ENN's effectiveness in addressing highly imbalanced datasets by oversampling minority instances while simultaneously reducing noise and overlap through the Edited Nearest Neighbour (ENN) approach. Although the overall accuracy remains modest, the notable enhancement in recall indicates that the model is more adept at identifying minority cases, which is crucial for critical prediction tasks like detecting cardiovascular diseases. These results align with earlier studies (Wang et al., 2023; Ishaq et al., 2021), reinforcing the benefits of hybrid resampling techniques for enhancing the recognition of minority classes in medical datasets.

In comparison, the breast cancer dataset yielded significantly improved predictive results across all evaluated performance metrics. Techniques such as ADASYN, Borderline SMOTE, and SMOTE-Tomek Links achieved notable outcomes, reaching an accuracy of 96.49%, an F1-score of 97.30%, perfect recall at 100%, and a Cohen's kappa score of 0.9231. These findings indicate that adaptive oversampling and boundary refinement techniques like ADASYN and SMOTE-Tomek Links effectively model the distribution of the minority class while avoiding overfitting. Their superior performance over SMOTE-ENN in this context underscores the idea that the optimal resampling method is contingent upon the feature distribution and noise characteristics of the dataset. These results are consistent with the research conducted by Chen et al. (2023) and Karamti et al. (2023), which showed that hybrid methods based on SMOTE significantly improve diagnostic accuracy in cancer prediction.

4.3.1 Summary of Findings

The study effectively fulfilled its goals of (1) assessing the effectiveness of different SMOTE techniques on medical datasets, (2) pinpointing the most significant diagnostic features for breast cancer and heart disease, and (3) analyzing the optimized models using XGBoost. The findings clearly indicate that employing SMOTE variants for class balancing, along with efficient feature selection, significantly enhances classification results. The model for breast cancer achieved nearly perfect recall and exhibited high levels of agreement, whereas the heart disease model showed improved sensitivity toward minority cases..

5. Conclusion

The findings indicated that the SMOTE-ENN method produced the best recall and F1-score for the heart disease dataset, showcasing its effectiveness in recognizing minority instances. In contrast, ADASYN, Borderline SMOTE, and SMOTE-Tomek Links demonstrated better accuracy and kappa scores for the breast cancer dataset, with accuracy reaching as high as 96.49% and a kappa score of 0.9231. Additionally, feature selection contributed to the models' interpretability by pinpointing clinically significant predictors such as perimeter_worst and concave_points_worst for breast cancer, along with cholesterol, diabetes, and exercise frequency for heart disease.

Looking ahead, future research will aim to expand this framework by incorporating deep learning techniques for feature extraction and hybrid resampling methods, which can further mitigate overfitting and enhance data diversity. Moreover, integrating explainable AI (XAI) methods within the XGBoost framework could improve the interpretability and reliability of clinical decision-making processes.

References

1. Allugunti, V. R. (2022). Breast cancer detection based on thermographic images using machine learning and deep learning algorithms. International Journal of Engineering in Computer Science, 4(1), 49–56.

2. Amethiya, Y., Pipariya, P., Patel, S., & Shah, M. (2022). Comparative analysis of breast cancer detection using machine learning and biosensors. Intelligent Medicine, 2, 69–81. https://doi.org/10.1016/j.imed.2021.08.004

3. Chen, H., Wang, N., Du, X., Mei, K., Zhou, Y., & Cai, G. (2023). Classification Prediction of Breast Cancer Based on Machine Learning. Computational Intelligence and Neuroscience, 2023(650719). https://doi.org/10.1155/2023/6530719

4. Elreedy, D., Atiya, A. F., & Kamalov, F. (2023). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Machine Learning. https://doi.org/https://doi.org/10.1007/s10994-022-06296-4

5. Hu, H., Qiao, S., Hao, Y., Bai, Y., Cheng, R., Zhang, W., & Zhang, G. (2022). Breast cancer histopathological images recognition based on two-stage nuclei segmentation strategy. PLoS ONE, 17(4), e0266973. https://doi.org/10.1371/journal.pone.0266973

6. Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021). Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques. IEEE Access, 9, 39707–39716. https://doi.org/10.1109/ACCESS.2021.3064084

7. Joloudari, J. H., Marefat, A., Nematollahi, M. A., Oyelere, S. S., & Hussain, S. (2023). Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks. Applied Sciences, 13, 4006. https://doi.org/https:// doi.org/10.3390/app13064006

8. Li, J., Zhu, Q., Wu, Q., & Fan, Z. (2021). A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors. Information Sciences, 565, 438–455. https://doi.org/10.1016/j.ins.2021.03.041

9. Liang, X. W., Jiang, A. P., Li, T., Xue, Y. Y., & Wang, G. T. (2020). Knowledge-Based Systems LR-SMOTE An improved unbalanced data set oversampling based on K-means and SVM. Knowledge-Based Systems, 196, 105845. https://doi.org/https://doi.org/10.1016/j.knosys.2020.105845

10. Nasir, M. U., Ghazal, T. M., Khan, M. A., Zubair, M., Rahman, A., Ahmed, R., Hamadi, H. Al, & Yeun, C. Y. (2022). Breast Cancer Prediction Empowered with Fine-Tuning. Computational Intelligence and Neuroscience, 2022(5918686), 1–9.

11. Oliveira, A. R., Ramos, T. B., & Neves, R. (2023). Streamflow Estimation in a Mediterranean Watershed Using Neural Network Models: A Detailed Description of the Implementation and Optimization. Water, 15, 947.

12. Ran, S., Yang, X., Liu, M., Zhang, Y., Cheng, C., Zhu, H., & Yuan, Y. (2022). Homecare-oriented ECG diagnosis with large-scale deep neural network for continuous monitoring on embedded devices. IEEE Transactions on Instrumentation and Measurement, 71, 1–13.

13. Wang, X., Ren, J., Ren, H., Song, W., Qiao, Y., Zhao, Y., Linghu, L., Cui, Y., Zhao, Z., Chen, L., & Qiu, L. (2023). Diabetes mellitus early warning and factor analysis using ensemble Bayesian networks with SMOTE-ENN and Boruta. Scientific Reports, 13(12718), 1–15. https://doi.org/10.1038/s41598-023-40036-5

14. Zhou, J., Wu, Y., Chen, J., Cui, M., Gao, Y., Meng, K., Wu, M., Guo, X., & Wen, W. (2023). Maize Stem Contour Extraction and Diameter Measurement Based on Adaptive Threshold Segmentation in Field Conditions. Agriculture, 13, 678.