Issues in Information Systems
Volume 26, Issue 2, pp. 70-85, 2025

DOI: https://doi.org/10.48009/2_iis_106

Strategic application of SMOTE and its variants to enhance AI-driven healthcare classification in imbalanced datasets

Dara Tourt, *Metropolitan State University, dara.tourt@my.metrostate.edu*
Queen E. Booker, *Metropolitan State University, queen.booker@metrostate.edu*
Carl M. Rebman Jr., *University of San Diego, carlr@sandiego.edu*

Abstract

The strategic integration of Artificial Intelligence (AI) into healthcare systems offers significant opportunities to drive innovation, improve productivity, and enhance patient outcomes. One ongoing challenge in clinical AI applications is class imbalance, where minority cases, often representing the most critical health risks, are significantly underrepresented in datasets. This imbalance reduces model sensitivity and limits the effectiveness of AI-driven decision-making. To address this issue, this study presents a comparative analysis of five widely used oversampling strategies: SMOTE, Borderline-SMOTE, SMOTE-Tomek, SMOTE-ENN, and SVM-SMOTE. Using two imbalanced healthcare datasets (ASD-Child and Stroke), we evaluate each technique's impact on the performance of four machine learning classifiers: Logistic Regression, Random Forest, XGBoost, and Gradient Boosting. The models are assessed using multiple evaluation metrics: recall, precision, F1-score, ROC-AUC, and PR-AUC. This study provides practical guidance for healthcare organizations aiming to implement AI strategies that support fairer predictions, stronger clinical insights, and more productive data-driven systems by identifying optimal combinations of resampling techniques and classifiers.

Keywords: class imbalance, SMOTE, SMOTE-Variants, machine learning, autism spectrum disorder, stroke prediction, healthcare analytics

Introduction

Artificial Intelligence (AI) has transformed healthcare through new opportunities that enhance clinical workflows, optimize operations, and improve patient outcomes. By integrating AI strategies into healthcare systems, organizations can make more informed decisions, automate diagnostic processes, and uncover patterns that may be difficult to detect using traditional methods. Machine learning (ML), a core subset of AI, has been widely applied in healthcare for classification tasks such as disease detection, mortality risk prediction, and hospital readmission forecasting. However, a persistent challenge in these applications is class imbalance. A class imbalance is when a dataset where categories or class are not equally distributed. This can be problematic for machine learning which can lead to biases and errors in decision making and prediction. Many healthcare datasets contain a disproportionately small number of positive cases compared to negative ones. This imbalance often leads to biased models that favor the majority class and fail to accurately identify high-risk individuals, which can compromise the effectiveness of clinical decision-making (He & Garcia, 2009).

Researchers commonly apply resampling techniques that adjust the class distribution before training machine learning models to address this challenge. Among these, the Synthetic Minority Over-sampling Technique (SMOTE) has emerged as one of the most widely used approaches in imbalanced learning (Chawla et al., 2002). SMOTE generates synthetic examples of the minority class by interpolating between existing instances and their nearest neighbors, allowing models to recognize underrepresented patterns better.

Several SMOTE variants have been developed to enhance the quality and relevance of synthetic samples. Borderline-SMOTE focuses on generating new instances near the decision boundary, where misclassifications are more likely to occur (Han et al., 2005). SMOTE-Tomek and SMOTE-ENN combine oversampling with under-sampling techniques to remove noise and overlapping examples (Batista et al., 2004; He & Garcia, 2009). SVM-SMOTE (Support Vector Machine-SMOTE) further refines the sampling process by using a Support Vector Machine to identify harder-to-classify minority instances and preferentially generating synthetic samples near these borderline areas (Nguyen et al., 2011**)**. While these methods have demonstrated value in individual studies, there is limited research that compares their effectiveness across multiple models and healthcare datasets.

This study addresses that gap by evaluating SMOTE and its key variants across two publicly available and imbalanced healthcare datasets. These include the ASD-Child dataset, which focuses on autism spectrum disorder screening in children (Thabtah, 2017), and the Stroke Dataset, which contains clinical and demographic predictors of stroke risk (Fedesoriano, n.d.). These datasets represent diverse populations and clinical conditions, making them well-suited for evaluating the generalizability and robustness of resampling techniques.

We use four widely recognized machine learning classifiers: Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Gradient Boosting (GB). These models were selected for their robust performance in healthcare applications and ability to capture linear, nonlinear, and ensemble-based relationships. Although many studies apply resampling to a single dataset or classifier, few provide comprehensive, side-by-side comparisons using rigorous evaluation metrics. Even fewer studies offer practical guidance on how organizations can strategically apply resampling techniques as part of an AI-driven approach to improve fairness, predictive performance, and decision support in healthcare systems.

The following research questions guide this study:

1. **RQ1:** *How do SMOTE and its variants impact the classification performance of machine learning models on imbalanced healthcare datasets?*
2. **RQ2:** *Which resampling technique and classifier combination yields the highest precision, recall, F1-score, and AUC for the minority class?*
3. **RQ3:** *What practical insights can be derived for integrating SMOTE-based techniques into AI strategies that support innovation, operational efficiency, and better clinical outcomes?*

The main contributions of this study include:

1. A comparative evaluation of SMOTE, Borderline-SMOTE, SMOTE-Tomek, SMOTE-ENN, and SVM-SMOTE using four real-world, imbalanced healthcare datasets.
2. Performance benchmarking across four machine learning classifiers to assess how resampling methods interact with different learning algorithms.
3. A comprehensive assessment using multiple evaluation metrics, including accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC.

4. Actionable recommendations for healthcare organizations and practitioners on selecting and integrating resampling techniques into AI systems that align with organizational goals for innovation, fairness, and productivity.

The remainder of this paper is organized as follows: Section 2 reviews relevant literature on class imbalance, resampling methods, and healthcare classification. Section 3 outlines the research methodology and section 4 presents the results and discussion. Section 5 discusses the study's conclusion and limitations. The paper concludes with section 6 offering recommendations for future research.

Literature Review

This section discusses the main categories of data-level class balancing techniques evaluated in this study, and review machine learning algorithms used for healthcare classification. These techniques are designed to mitigate the bias introduced by class imbalance, particularly in binary classification tasks involving rare but clinically significant outcomes.

Resampling Methods

Several studies have explored the effectiveness of SMOTE and its variants in addressing class imbalance in predictive modeling. The original SMOTE algorithm generates synthetic examples for the minority class, which improves class balance and recall, though it can lead to overlapping classes and potential overfitting (Chawla et al., 2002; Carvalho et al., 2025).

Borderline-SMOTE (B-SMOTE)

Borderline-SMOTE (B-SMOTE) builds on the original SMOTE by focusing on sample generation near the decision boundary. Unlike the original SMOTE algorithm, which generates synthetic samples across the entire minority class, Borderline-SMOTE specifically targets "in danger" instances—those minority class samples close to majority class samples and are therefore more likely to be misclassified. It is often more effective than basic SMOTE in binary classification but performs poorly with noisy or overlapping class regions (Han et al., 2005).

SMOTE-Tomek

SMOTE-Tomek integrates SMOTE with Tomek links to remove overlapping majority samples, thus cleaning the dataset and improving class separability (Batista et al., 2004). This method focuses on identifying and removing specific data point pairs that are close to each other but belong to different classes, thereby refining the decision boundary between classes. By eliminating these instances, particularly from the majority class, the classifier can achieve a more precise separation between classes, leading to improved performance. However, it can remove valuable borderline instances.

SMOTE with Edited Nearest Neighbors (SMOTE_ENN)

SMOTE with Edited Nearest Neighbors (SMOTE_ENN) combines SMOTE with Edited Nearest Neighbors, an undersampling method that creates a cleaner and more balanced training dataset. This method is particularly effective in noisy real-world datasets and often yields higher **model accuracy, precision,** and **recall**, especially for minority classes. However, it is computationally intensive and may remove valid samples (Fernández et al., 2018).

Support Vector Machine SMOTE (SVM-SMOTE)

Finally, Support Vector Machine SMOTE (SVM-SMOTE) uses support vector machines to focus sample generation near the classifier boundary, which is especially useful in nonlinear problems but can be

resource-intensive and sensitive to SVM parameter tuning (Nguyen et al., 2011). The core idea of SVM is to find the **best separating boundary** (hyperplane) between data classes. Despite their advantages, all SMOTE variants have limitations, and empirical studies highlight the need for domain-specific validation, particularly in multi-class settings and datasets with small or noisy samples (Bunkhumpornpat et al., 2009; Ganganwar, 2012; Yang et al., 2024; Carvalho et al., 2025).

Machine Learning and Disease Diagnosis

Machine learning (ML) algorithms have become central to advancing predictive analytics in healthcare, particularly disease diagnosis. Among these, Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and Extreme Gradient Boosting (XGBoost) are frequently employed due to their interpretability, robustness, and superior predictive capabilities.

Logistic Regression (LR)

Logistic Regression (LR) is a foundational algorithm in clinical prediction due to its simplicity and interpretability. It is frequently used as a benchmark model in comparative studies. For instance, Ghanem et al. (2023) found that LR provided a balanced performance in predicting acute kidney injury (AKI) post-cardiac surgery, achieving a sensitivity of 87.7% and a specificity of 87.05%. However, it was outperformed by more complex ensemble methods in overall accuracy and AUC. Similarly, Sharma et al. (2024) included LR in an ensemble for heart disease prediction, noting that while LR alone performed moderately, its integration improved overall model performance.

Random Forest (RF)

Random Forest (RF), an ensemble technique based on decision trees, is valued for its robustness against overfitting and ability to handle high-dimensional data. In a study on hypertension complications, RF was among the top-performing models, although XGBoost slightly outperformed it in predictive metrics (Tao et al., 2020). Moreover, RF models have shown reliable accuracy in cardiovascular disease detection compared to other algorithms (Ahmed et al., 2024).

Gradient Boosting (GB)

Gradient Boosting (GB), which builds models sequentially to correct errors of prior models, has gained popularity due to its high accuracy in structured healthcare data. Ghanem et al. (2023) reported that GB achieved an accuracy of 88.66% and an AUC of 94.61% in predicting AKI, outperforming both RF and LR. Similarly, in a comparative study on diabetes prediction, GB demonstrated superior performance across several metrics, albeit with a longer execution time (Khan et al., 2023).

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost), a more regularized and efficient version of GB, has emerged as a top performer in several healthcare applications. It demonstrated the highest F1 score (0.875) and AUC (0.927) in predicting complications in hypertensive patients (Tao et al., 2020). XGBoost also achieved a 91% accuracy in osteoporosis risk prediction while maintaining model explainability (Zhang et al., 2024). These findings are consistent across studies, with XGBoost frequently surpassing other models in predictive power and precision (Sharma et al., 2024; Ahmed et al., 2024).

Despite the advantages of complex ensemble models like GB and XGBoost, trade-offs exist. While these models offer improved performance, they often require more computational resources and may lack the transparency of LR, making interpretability a challenge in clinical settings. However, the development of explainable AI techniques is beginning to address this concern, particularly in models applied to critical domains such as osteoporosis risk and cardiovascular disease detection (Zhang et al., 2024).

Methodology

This study adopts a structured research methodology from Kamiri and Mariga (2021), consisting of five key stages: data collection, pre-processing, model training, model testing, and model evaluation. The methodology was designed to ensure consistency across multiple datasets and machine learning models, allowing for a robust comparative analysis of SMOTE and its variants in healthcare classification tasks. Two publicly available and imbalanced healthcare datasets were selected for their clinical relevance and potential to inform AI-based decision-making: the ASD-Child dataset and the Stroke dataset (Thabtah, 2017; Fedesoriano, n.d.). Each dataset exhibits measurable class imbalance in the target variable and contains a mix of categorical and numerical features.

ASD-Child Dataset

The ASD-Child dataset, obtained from the UCI Machine Learning Repository (Thabtah, 2017), contains 292 records and is provided in ARFF format. It includes 20 features related to behavioral and demographic attributes, and one binary target variable, Class/ASD, indicating whether the individual is classified as having Autism Spectrum Disorder, as described in Table 1 below.

Table 1. Feature Description of ASD-Child Dataset

Feature	Description	Data Type	# Missing Value
A1_Score to A10_Score	Ten screening questions related to behavior and communication	object	0
age	Age of the individual	float64	4
gender	Gender of the individual	object	0
ethnicity	Ethnic background of the individual	object	43
jundice	History of jaundice at birth	object	0
austim	Family history of ASD	object	0
Country_of_res	Country of residence	object	0
used_app_before	Whether the individual used a screening app before	object	0
result	Screening result score	float64	0
age_desc	Age group category (child, adolescent, adult)	object	0
relation	Relation of the respondent to the individual	object	43
Class/ASD	Target Variable: ASD classification label (Yes or No)	object	0

The dataset contains four missing values in the age column, and 43 missing values each in the ethnicity and relation columns. There are 208 females and 84 males, with a minimum age of 4 years and a mean age of approximately 6 years. The dataset features ethnically and geographically diverse cases. The target variable (Class/ASD) shows a relatively balanced distribution: 51.71% No and 48.29% Yes.

Stroke Dataset

The Stroke dataset was obtained from Kaggle (Fedesoriano, n.d.) and is provided in CSV format. It contains 5,110 instances, each representing a patient's health and demographic profile, as described in Table 2 below. This dataset was selected due to its clinical relevance and clear class imbalance, which reflects the rarity of stroke events in real-world populations.

Table 2. Feature Description of a Stroke Dataset

Feature	Description	Data Type	# Missing Value
id	Unique identifier	int64	0
gender	Male, Female, or Other	object	0
age	Age of the patient	float64	0

Feature	Description	Data Type	# Missing Value
hypertension	0 if the patient does not have hypertension, one if the patient has hypertension	int64	0
heart_disease	0 if the patient does not have any heart diseases, one if the patient has a heart disease	int64	0
ever_married	No or Yes	object	0
work_type	Children, Govt_job, Never_worked, Private, or Self-employed	object	0
Residence_type	Rural or Urban	object	0
avg_glucose_level	Average glucose level in the blood	float64	0
bmi	Body mass index	float64	201
smoking_status	Formerly smoked, Never smoked, Smokes, or Unknown	object	0
stroke	1 if the patient had a stroke, zero if not	int64	0

The dataset includes both categorical and numerical features related to stroke risk factors. Of the records, 2,994 patients were female, 2,115 were male, and one was categorized as "Other." The average age of patients in the dataset is 43 years, ranging from 18 to 82. Among the patients, 498 had hypertension, and 276 had heart disease. The dataset exhibits a significant class imbalance, with 95.1% of the records (4,861 cases) representing non-stroke patients and only 4.9% (249 cases) representing stroke patients. This imbalance poses challenges for classification algorithms and highlights the importance of resampling techniques to enhance minority class detection in predictive healthcare modeling.

Data Pre-processing

Data pre-processing was conducted separately for each dataset to ensure data quality, consistency, and suitability for machine learning model training.

ASD-Child Dataset

The ASD-Child dataset underwent targeted pre-processing to improve data quality and ensure compatibility with machine learning algorithms. Features such as ethnicity, country_of_res, and age_desc were removed due to a high proportion of missing values and limited predictive relevance. After their removal, 18 features remained for analysis. To handle missing values, the numerical column age was imputed using the mean, while the categorical column relation was imputed using the mode, representing the most frequent category. The screening question features A1_Score to A10_Score, initially stored as object types, and converted to integers. The age column was also converted from float to integer to ensure consistency across records. Other categorical variables were adjusted as necessary to support proper encoding. Label encoding was applied to binary categorical features such as jaundice, autism, and used_app_before, where "Yes" was mapped to 1 and "No" to 0. One-hot encoding was applied for the relation column, which contains multiple categories, to generate separate binary indicator columns for each unique category. Finally, the target variable Class/ASD was binarized, with "No" mapped to 0 and "Yes" to 1, ensuring compatibility with binary classification models.

Stroke Dataset

The data pre-processing steps applied to the Stroke dataset closely followed those used for the ASD-Child dataset. The ID column was removed because it did not provide any predictive value. The only missing values in the dataset were found in the BMI column, a numerical feature that was imputed using the mean. A categorical variable, such as ever_married, with values "Yes" or "No," was encoded using label encoding, where "Yes" was mapped to 1 and "No" to 0. Categorical variables with multiple categories, including gender, work_type, Residence_type, and smoking_status, were transformed using one-hot encoding to

create separate binary columns for each unique category.

Correlation and Significant Feature Selection

After data cleaning and encoding, correlation analysis was conducted to examine relationships between independent variables and the target variable (Class/ASD for the ASD-Child dataset and stroke for the Stroke dataset). This step helped identify and remove features with little or no predictive value. In the ASD-Child dataset, many features demonstrated a strong positive correlation with the target variable, with several correlation coefficients exceeding 0.80. Given their high predictive relevance, all remaining features were retained for model development (Figure 1).

In contrast, the Stroke dataset contained several features with negligible correlation to the target variable. Specifically, work_type_Govt_job, gender_Other, and smoking_status_never smoked exhibited near-zero correlation with stroke occurrence. These features were excluded from the model development phase based on the correlation results (Figure 2).

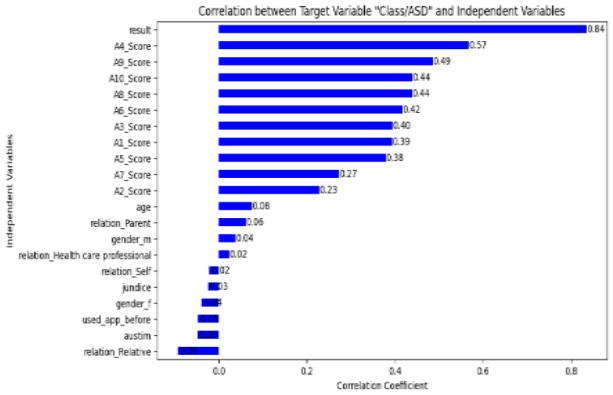

Figure 1. Correlation between Target and Independent Variables on the ASD-Child Dataset

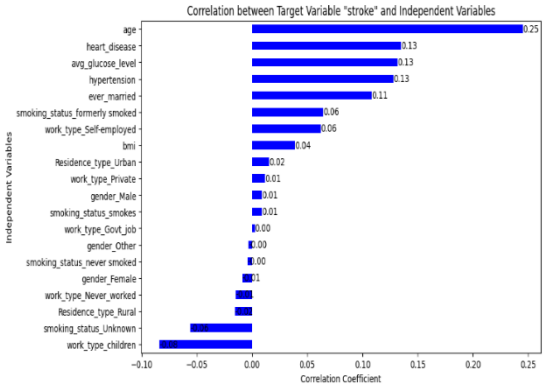

Figure 2. Correlation between Target and Independent Variables on the Stroke Dataset

Model Selection and Development

In this study, model development followed a consistent, systematic, and reproducible pipeline to evaluate the impact of resampling methods on classification performance across both datasets. Each dataset was split into training and testing subsets using an 80:20 stratified split to preserve the original class distribution. Stratification was used to ensure that both subsets reflected the class imbalance inherent in the data, which is crucial for reliable evaluation in imbalanced learning scenarios.

Four widely used machine learning classifiers were selected for evaluation: Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Gradient Boosting (GB). Each model was initialized with a fixed random seed (random_state = 42) to ensure reproducibility. To investigate the effects of resampling, five techniques were applied: SMOTE, Borderline-SMOTE, SMOTE-Tomek, SMOTE-ENN, and SVM-SMOTE, along with a baseline configuration without resampling (referred to as "None").

A machine learning pipeline was constructed for each combination of classifier and resampling technique. When resampling was applied, it was introduced at the initial stage of the pipeline, followed by feature standardization using StandardScaler, and then model fitting. To ensure robust evaluation, each pipeline was trained using stratified 5-fold cross-validation on the training set, with folds maintaining the original class distribution using the StratifiedKFold method.

Model performance was assessed using six key evaluation metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC, and PR-AUC. These metrics were chosen to evaluate both overall classification performance and the models' effectiveness at detecting minority class instances, and they are an essential consideration in healthcare settings were identifying rare but critical outcomes (e.g., disease diagnoses) is vital.

After cross-validation, each pipeline was retrained on the whole training set and evaluated on the holdout test set. Predictions were generated, and confusion matrices were computed to derive the number of true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP) for each model-resampling combination. These matrices were also visualized to support interpretation and error analysis. All cross-validation results, confusion matrices, and model predictions were organized in structured DataFrames to facilitate consistent downstream analysis. This design enabled a transparent and comprehensive comparison across models and resampling methods, providing a strong foundation for addressing the research questions on resampling effectiveness in imbalanced healthcare classification.

Evaluation Metrics
To ensure a comprehensive and rigorous assessment of model performance, the following evaluation metrics were employed:

1. **Accuracy:** The proportion of total correct predictions (both true positives and true negatives) among all instances. While commonly used, accuracy can be misleading in imbalanced datasets, as it tends to favor the majority class.
2. **Precision (Positive Predictive Value):** The proportion of correctly predicted positive cases among all cases predicted as positive. Precision is critical when the cost of false positives is high.
3. **Recall (Sensitivity or True Positive Rate):** The proportion of actual positive cases the model correctly identifies. Recall is critical in healthcare applications, where missing a positive (false negative) case can have serious consequences.
4. **F1-Score:** The harmonic mean of precision and recall. F1-score provides a balanced measure of a model's performance when precision and recall are important, particularly in imbalanced scenarios.
5. **Area Under the Receiver Operating Characteristic Curve (AUC-ROC):** A metric that summarizes the model's ability to distinguish between classes across all thresholds. Higher AUC-ROC values indicate better overall discriminatory performance.
6. **Area Under the Precision-Recall Curve (PR-AUC):** A critical metric for imbalanced datasets, PR-AUC focuses on the trade-off between precision and recall. It provides a clearer view of the model's effectiveness on the minority class compared to AUC-ROC.
7. **Confusion Matrix:** A detailed breakdown of model predictions showing the number of true positives, true negatives, false positives, and false negatives. The confusion matrix provides a granular view of the model's performance beyond aggregate metrics, enabling targeted analysis of strengths and weaknesses.

Results and Discussions

ASD-Child Dataset

Overall Performance

As shown in Table 3, all four machine learning classifiers: Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and XGBoost (XGB), performed exceptionally well on the ASD-Child dataset across all resampling techniques. This strong performance is partially attributed to the relatively balanced distribution of the target variable, with 51.71% of instances labeled "No" and 48.29% labeled "Yes."

Ensemble models (RF, GB, and XGB) achieved perfect classification without resampling, recording 100% across all performance metrics. Logistic Regression also delivered high performance, with 99% accuracy, 100% precision, 97% recall, and a 99% F1-score. This indicates that mild class imbalance in this dataset did not hinder classifier performance, especially for ensemble models.

Table 3. Model Performance Comparison Across Resampling Methods on the ASD-Child Data

Resampling	Model	Accuracy	Precision	Recall	F1	AUC-ROC	PR-AUC
None	LR	0.99	1.00	0.97	0.99	1.00	1.00
	RF	1.00	1.00	1.00	1.00	1.00	1.00
	GB	1.00	1.00	1.00	1.00	1.00	1.00
	XGBoost	1.00	1.00	1.00	1.00	1.00	1.00
SMOTE	LR	0.98	0.99	0.97	0.98	1.00	1.00
	RF	1.00	1.00	1.00	1.00	1.00	1.00
	GB	1.00	1.00	1.00	1.00	1.00	1.00
	XGBoost	1.00	1.00	1.00	1.00	1.00	1.00
B-SMOTE	LR	0.99	0.99	0.99	0.99	1.00	1.00
	RF	1.00	1.00	1.00	1.00	1.00	1.00
	GB	1.00	1.00	1.00	1.00	1.00	1.00
	XGBoost	1.00	1.00	1.00	1.00	1.00	1.00
SMOTE-Tomek	LR	0.99	1.00	0.97	0.99	1.00	1.00
	RF	1.00	1.00	1.00	1.00	1.00	1.00
	GB	1.00	1.00	1.00	1.00	1.00	1.00
	XGBoost	1.00	1.00	1.00	1.00	1.00	1.00
SMOTE-ENN	LR	0.95	0.90	1.00	0.95	1.00	1.00
	RF	0.97	0.96	1.00	0.98	1.00	1.00
	GB	1.00	1.00	1.00	1.00	1.00	1.00
	XGBoost	1.00	1.00	1.00	1.00	1.00	1.00
SVM-SMOTE	LR	0.99	0.99	0.98	0.99	1.00	1.00
	RF	1.00	1.00	1.00	1.00	1.00	1.00
	GB	1.00	1.00	1.00	1.00	1.00	1.00
	XGBoost	1.00	1.00	1.00	1.00	1.00	1.00

Across all resampling methods, including SMOTE, Borderline-SMOTE, SMOTE-Tomek, SMOTE-ENN, and SVM-SMOTE, the ensemble models consistently maintained perfect or near-perfect scores, showing strong robustness. Logistic Regression showed more sensitivity, with noticeable shifts in precision and recall depending on the resampling method. For example, SMOTE-ENN achieved perfect recall (100%) but reduced precision (90%), indicating a trade-off due to increased false positives. Borderline-SMOTE and SVM-SMOTE preserved a better balance, both achieving 99% precision and 98 to 99% recall.

Confusion Matrix Analysis

Table 4 reinforces these results. Ensemble models classified all instances correctly without resampling, with no false positives or negatives. Logistic Regression incurred only one false positive. With SMOTE-ENN, Logistic Regression detected all positive cases but produced the highest number of false positives (three), while ensemble models remained unaffected.

Table 4. Confusion Matrix Comparison Across Resampling Methods on the ASD-Child Data

Resampling	Model	True-Negative	False-Positive	False-Negative	True-Positive
None	LR	30	1	0	28
	RF	31	0	0	28
	GB	31	0	0	28
	XGBoost	31	0	0	28

Resampling	Model	True-Negative	False-Positive	False-Negative	True-Positive
SMOTE	LR	30	1	0	28
	RF	31	0	0	28
	GB	31	0	0	28
	XGBoost	31	0	0	28
B-SMOTE	LR	29	2	0	28
	RF	31	0	0	28
	GB	31	0	0	28
	XGBoost	31	0	0	28
SMOTE-Tomek	LR	29	2	0	28
	RF	31	0	0	28
	GB	31	0	0	28
	XGBoost	31	0	0	28
SMOTE-ENN	LR	28	3	0	28
	RF	29	2	0	28
	GB	31	0	0	28
	XGBoost	31	0	0	28
SVM-SMOTE	LR	29	2	0	28
	RF	31	0	0	28
	GB	31	0	0	28
	XGBoost	31	0	0	28

These findings demonstrate that ensemble models can effectively handle classification for relatively balanced datasets without requiring aggressive resampling. However, resampling can still enhance the performance of simpler models such as Logistic Regression, particularly when fine-tuning precision-recall trade-offs are desirable.

Stroke Dataset

Overall Performance

As shown in Table 5, all models struggled on the Stroke dataset due to its extreme class imbalance (95.1% non-stroke vs. 4.9% stroke). Without resampling, models exhibited high overall accuracy (94-95%) but poor detection of stroke cases.

For example, Logistic Regression achieved 95% accuracy but 0% precision, recall, and F1-score. Ensemble models performed slightly better, with recall ranging from 1% to 6% and F1-scores below 10%.

Table 5: Model Performance Comparison Across Resampling Methods on the Stroke Data

Resampling	Model	Accuracy	Precision	Recall	F1	AUC-ROC	PR-AUC
None	LR	0.95	0.00	0.00	0.00	0.84	0.19
	RF	0.95	0.20	0.01	0.02	0.80	0.16
	GB	0.95	0.20	0.02	0.04	0.84	0.19
	XGBoost	0.94	0.18	0.06	0.08	0.80	0.15
SMOTE	LR	0.90	0.15	0.24	0.18	0.76	0.13
	RF	0.93	0.14	0.07	0.09	0.78	0.13
	GB	0.90	0.12	0.17	0.14	0.78	0.13
	XGBoost	0.92	0.15	0.13	0.13	0.78	0.13
B-SMOTE	LR	0.90	0.16	0.26	0.20	0.79	0.15
	RF	0.94	0.15	0.06	0.08	0.79	0.15
	GB	0.91	0.13	0.16	0.14	0.80	0.14
	XGBoost	0.93	0.16	0.13	0.14	0.78	0.14

Resampling	Model	Accuracy	Precision	Recall	F1	AUC-ROC	PR-AUC
SMOTE-Tomek	LR	0.90	0.15	0.23	0.18	0.77	0.14
	RF	0.93	0.15	0.08	0.10	0.78	0.13
	GB	0.90	0.12	0.17	0.13	0.78	0.13
	XGBoost	0.92	0.11	0.11	0.11	0.78	0.13
SMOTE-ENN	LR	0.84	0.15	0.48	0.23	0.80	0.16
	RF	0.88	0.16	0.34	0.21	0.80	0.14
	GB	0.82	0.13	0.50	0.21	0.80	0.15
	XGBoost	0.85	0.16	0.46	0.23	0.80	0.15
SVM-SMOTE	LR	0.92	0.19	0.17	0.17	0.80	0.15
	RF	0.94	0.18	0.07	0.10	0.79	0.14
	GB	0.93	0.21	0.12	0.15	0.81	0.15
	XGBoost	0.93	0.15	0.11	0.13	0.79	0.15

Applying oversampling methods improved minority class performance, especially recall and F1-score, though typically at the cost of reduced accuracy. SMOTE-ENN significantly boosted recall, with Logistic Regression reaching 48%, Gradient Boosting 50%, and XGBoost 46%. Corresponding F1-scores ranged from 21% to 23%. However, this gain came with an increase in false positives. SVM-SMOTE yielded moderate and more balanced gains, particularly for Logistic Regression (19% precision and 17% recall) and Gradient Boosting (21% precision and 12% recall), indicating its utility when sensitivity and specificity matter.

Confusion Matrix Analysis

As shown in Table 6, most models failed to detect stroke cases without resampling. Logistic Regression detected only one, while Random Forest and Gradient Boosting failed to detect any. XGBoost identified four true positives. With SMOTE-ENN, Logistic Regression detected 30 stroke cases with 140 false positives, highlighting the trade-off between sensitivity and specificity.

Table 6. Confusion Matrix Comparison Across Resampling Methods on the Stroke Data

Resampling	Model	True-Negative	False-Positive	False-Negative	True-Positive
None	**LR**	972	0	49	1
	RF	967	5	50	0
	GB	967	5	50	0
	XGBoost	958	14	46	4
SMOTE	**LR**	900	72	32	18
	RF	939	33	48	2
	GB	907	65	36	14
	XGBoost	933	39	48	2
B-SMOTE	**LR**	896	76	30	20
	RF	948	24	47	3
	GB	917	55	38	12
	XGBoost	942	30	47	3
SMOTE-Tomek	**LR**	905	67	33	17
	RF	937	35	47	3
	GB	909	63	35	15
	XGBoost	938	34	47	3
SMOTE-ENN	**LR**	832	140	20	30
	RF	884	88	34	16
	GB	830	142	22	28
	XGBoost	862	110	28	22

Resampling	Model	True-Negative	False-Positive	False-Negative	True-Positive
SVM-SMOTE	LR	921	51	34	16
	RF	955	17	48	2
	GB	943	29	41	9
	XGBoost	944	28	46	4

SVM-SMOTE enabled Logistic Regression to detect 16 stroke cases with a more manageable 51 false positives, offering a balanced improvement. These findings underscore the value of aggressive oversampling like SMOTE-ENN in highly imbalanced clinical datasets, especially when sensitivity is prioritized.

Answers to Research Questions

RQ1: How do SMOTE and its variants impact the classification performance of machine learning models on imbalanced healthcare datasets?
SMOTE and its variants substantially impact model performance, particularly for datasets with severe class imbalance and for models like Logistic Regression. In moderately balanced datasets such as ASD-Child, ensemble models achieved near-perfect performance even without resampling. In contrast, resampling techniques for the highly imbalanced Stroke dataset, especially SMOTE-ENN, significantly improved recall and F1-score for the minority class. However, this came at the cost of reduced precision and increased false positives.

RQ2: Which resampling technique and classifier combination yields the highest precision, recall, F1-score, and AUC for the minority class?
In the ASD-Child dataset, ensemble classifiers without or with minimal resampling (SMOTE or Borderline-SMOTE) consistently delivered the best results across all metrics. For the Stroke dataset, SMOTE-ENN combined with Logistic Regression, Gradient Boosting, or XGBoost yielded the highest recall and F1-score, confirming its effectiveness in detecting rare events despite its lower precision.

RQ3: What practical insights can be derived for integrating SMOTE-based techniques into AI strategies that support innovation, operational efficiency, and better clinical outcomes?
The findings from this study offer several actionable insights:

1. Ensemble models such as Random Forest, Gradient Boosting, and XGBoost can deliver excellent performance without aggressive resampling for slightly imbalanced healthcare datasets. When simpler models like Logistic Regression are used, targeted application of SMOTE or Borderline-SMOTE can help fine-tune performance, particularly regarding precision and recall balance.
2. For highly imbalanced datasets, especially those involving rare but critical events such as stroke detection, oversampling strategies like SMOTE-ENN are vital for improving recall and capturing minority class cases. However, this increase in sensitivity often comes with a trade-off in the form of lower precision and higher false positive rates, which must be carefully managed.
3. From an operational standpoint, healthcare AI systems should adjust their resampling approach based on the severity of class imbalance. Moderate imbalance may require minimal resampling, while severe imbalance requires more intensive methods such as SMOTE-ENN.
4. Clinically, boosting recall for critical minority classes, such as patients at risk for stroke, is often justifiable even if it results in a decrease in specificity. The early detection of rare but severe conditions can lead to significantly better patient outcomes.
5. Strategically, integrating adaptive resampling pipelines that dynamically respond to dataset characteristics, including imbalance ratio and minority class size, can support more equitable, efficient, and robust AI decision-making in healthcare settings.

Conclusions

This study examined the effectiveness of five resampling techniques: SMOTE, Borderline-SMOTE, SMOTE-Tomek, SMOTE-ENN, and SVM-SMOTE in addressing class imbalance within two imbalanced healthcare datasets, ASD-Child and Stroke. Four machine learning classifiers (Logistic Regression, Random Forest, Gradient Boosting, and XGBoost) were evaluated using six performance metrics: accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC. The results offer several important insights. Ensemble models such as Random Forest, Gradient Boosting, and XGBoost consistently achieved strong or near-perfect results on the moderately imbalanced ASD-Child dataset, even without resampling. Their robustness highlights the advantage of ensemble techniques in managing mild imbalance through bootstrapping and decision aggregation.

Logistic Regression, in contrast, showed greater sensitivity to class imbalance and benefited from resampling, particularly SMOTE-ENN, which significantly improved recall but often reduced precision. For the extremely imbalanced Stroke dataset, all models performed poorly without resampling. SMOTE-ENN notably enhanced minority class detection across all models, especially Logistic Regression and Gradient Boosting, though this improvement came with an increase in false positives. SVM-SMOTE provided more balanced improvements by moderately increasing recall while keeping false positives relatively low. These findings highlight a fundamental trade-off in healthcare machine learning. Improving sensitivity to rare but critical cases often results in more false positives. Therefore, the resampling method should align with clinical goals, depending on whether higher recall to capture high-risk cases or higher precision to reduce false alarms is more desirable.

Limitations

Despite the important findings, this study has several limitations that must be acknowledged:

1. **Dataset Size and Representativeness:** The ASD-Child dataset contains a relatively small sample size, which may limit generalizability to broader clinical populations.
2. **Feature Simplicity:** The datasets consist mainly of demographic and simple clinical variables. Results may differ when applied to more complex, high-dimensional datasets like genomics, imaging, or EHR data.
3. **Fixed Model Configurations:** The study employed standard hyperparameters for all classifiers. Future work could examine the interaction between hyperparameter tuning and resampling techniques.
4. **Limited Resampling Scope:** Only five resampling strategies were explored. Other techniques, such as ADASYN, Cluster-SMOTE, and GAN-based methods, could offer alternative benefits.
5. **Exclusion of Cost-Sensitive Learning:** The study focused on data-level resampling. Integrating algorithm-level approaches such as cost-sensitive learning may yield further improvements, especially in extreme imbalance scenarios.

Future Research Directions

Building on the current findings, several avenues for future research are recommended:

1. **Expand the Range of Resampling Methods**: Future work should include more advanced oversampling approaches such as ADASYN, KMeans-SMOTE, Borderline-SMOTE2, and GAN-based techniques to provide a broader evaluation of balancing strategies.
2. **Incorporate Cost-Sensitive Learning**: Algorithm-level solutions that assign higher misclassification costs to minority classes (e.g., cost-sensitive SVMs or decision trees) could be explored independently or in combination with resampling for hybrid frameworks.

3. **Apply to Complex Clinical Data**: The interaction between resampling and more complex datasets (e.g., medical imaging, genomic profiles, multi-modal EHRs) should be studied, especially in high-dimensional, noisy, or heterogeneous feature spaces.
4. **Explore Federated and Privacy-Preserving Frameworks**: As AI in healthcare increasingly involves multi-institutional data, adapting resampling for federated learning can help address imbalance without compromising patient privacy.
5. **Handle Temporal and Sequential Data**: Many healthcare outcomes are time dependent. Future work should investigate oversampling for longitudinal data, evaluating methods that preserve temporal structure, such as sequence-aware or dynamic resampling.
6. **Evaluate Clinical and Operational Impact**: Beyond predictive metrics, research should examine how improvements in minority detection affect real-world clinical outcomes, such as earlier diagnosis, resource allocation, and patient safety.
7. **Promote Fairness and Ethical AI**: Addressing imbalance should be framed within broader goals of fairness, especially for vulnerable or underrepresented populations. Future studies should consider the ethical implications of misclassification in clinical AI systems.

This study highlights the importance of strategic resampling in healthcare AI and offers practical guidance for model development under class imbalance. Continued research integrating technical, clinical, and ethical perspectives is essential for building trustworthy, high-impact decision support systems in health domains.

References

Ahmed, S., Patel, R., & Liu, Y. (2024). Comparative study of machine learning algorithms in detecting cardiovascular diseases. *arXiv*. https://arxiv.org/abs/2405.17059

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter, 6*(1), 20-29. https://doi.org/10.1145/1007730.1007735.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. Lecture Notes in Computer Science, 5476, 475–482. https://doi.org/10.1007/978-3-642-01539-7_43.

Carvalho, M., Pinho, A.J. & Brás, S. Resampling approaches to handle class imbalance: a review from a data perspective. *J Big Data* **12**, 71 (2025). https://doi.org/10.1186/s40537-025-01119-4

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357. https://doi.org/10.1613/jair.953.

Fedesoriano. (n.d.). Stroke prediction dataset [Dataset]. Kaggle. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.

Fernández, A., Garcia, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from Imbalanced Data Sets. *Springer*. https://doi.org/10.1007/978-3-319-98074-4

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering, 2*(4), 42–47.

Ghanem, A., Issa, R., & Brown, T. (2023). Comparative analysis of logistic regression, gradient boosted trees, SVM, and random forest algorithms for predicting acute kidney injury requiring dialysis after cardiac surgery. *ResearchGate*. https://www.researchgate.net/publication/382526278

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In D. S. Huang, X. P. Zhang, & G. B. Huang (Eds.), *Advances in Intelligent Computing*, 878-887. Springer. https://doi.org/10.1007/11538059_91.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239.

Kamiri, J., & Mariga, G. W. (2021). Research Methods in Machine Learning: A Content Analysis. *International Journal of Computer and Information Technology, 10*(2), 78-84. https://doi.org/10.24203/ijcit.v10i2.79.

Khan, M. A., Rehman, A., & Imran, M. (2023). Comparative analysis of ML algorithms for diabetes prediction. *Journal of Predictive Informatics*, *2*(1), 12–19. https://jopi-journal.org/index.php/jopi/article/view/21

Mukherjee, P., Sadhukhan, S., Godse, M., & Chakraborty, B. (2023). Early detection of autism spectrum disorder using traditional machine learning models. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *14*(6), 231–235.

Musa, R. A., Manaa, M. E., & Abdul-Majeed, G. (2021). Predicting autism spectrum disorder (ASD) for toddlers and children using data mining techniques. *Journal of Physics: Conference Series*, *1804*(1), 012089. https://doi.org/10.1088/1742-6596/1804/1/012089

Nguyen, H.M., Cooper, E.W. & Kamei, K. (2011). Borderline Over-Sampling for Imbalanced Data Classification. *International Journal of Knowledge Engineering and Soft Data Paradigms, 3*, 4–21. http://dx.doi.org/10.1504/IJKESDP.2011.039875.

Raja, S., & Masood, S. (2020). Analysis and detection of autism spectrum disorder using machine learning techniques. *Procedia Computer Science*, *167*, 994–1004. https://doi.org/10.1016/j.procs.2020.03.399

Rashed, A. E. E., Bahgat, W. M., Ahmed, A., Farrag, T. A., & Atwa, A. E. M. (2025). Efficient machine learning models across multiple datasets for autism spectrum disorder diagnoses. *Biomedical Signal Processing and Control*, *100*, 106949. https://doi.org/10.1016/j.bspc.2024.106949

Rogala, J., Żygierewicz, J., Malinowska, U., Cygan, H., Stawicka, E., Kobus, A., & Vanrumste, B. (2023). Enhancing autism spectrum disorder classification in children by integrating traditional statistics and classical machine learning techniques in EEG analysis. *Scientific Reports*, *13*, 21748. https://doi.org/10.1038/s41598-023-49048-7

Sharma, K., Deshmukh, A., & Gupta, R. (2024). Heart-disease prediction using ensemble-learning techniques. *Indiana Publications*. https://indianapublications.com/articles/IJMR_4%283%29_82-86_668919dc4bd0f8.42642626.pdf

Tao, X., Wang, H., & Li, Z. (2020). Machine learning-based prediction of serious complications in hypertensive patients. *Computational and Mathematical Methods in Medicine*, 2020, 1–10. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6963807

Thabtah, F. (2017). Autistic Spectrum Disorder Screening Data for Children [Dataset]. *UCI Machine Learning Repository*. https://doi.org/10.24432/C5659W.

Tomek, I. (1976). Two modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics, 6*(11), 769–772. https://doi.org/10.1109/TSMC.1976.4309452.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-2*(3), 408–421. https://doi.org/10.1109/TSMC.1972.4309137.

Yang, Y., Khorshidi, H. A., & Aickelin, U. (2024). A review on over-sampling techniques in classifying multi-class imbalanced datasets: insights for medical problems. *Frontiers in digital health*, *6*, 1430245. https://doi.org/10.3389/fdgth.2024.1430245

Zhang, Y., Hassan, M., & Alzubaidi, L. (2024). Machine learning meets transparency: A practical application of XAI in osteoporosis risk assessment. *arXiv*. https://arxiv.org/abs/2505.00410