Resampling Imbalanced Healthcare Data for Predictive Modelling

Manoj Yadav Mamilla, Ronak Al-Haddad, Stiphen Chowdhury
Computing and Information Science
Anglia Ruskin University, East Rd, Cambridge CB1 1PT

***Abstract*—Imbalanced datasets pose significant challenges in healthcare for developing accurate predictive models in medical diagnostics. In this work, we explore the effectiveness of combining resampling methods with machine learning algorithms to enhance prediction accuracy for imbalanced heart and lung disease datasets. Specifically, we integrate undersampling techniques such as Edited Nearest Neighbours (ENN) and Instance Hardness Threshold (IHT) with oversampling methods like Random Oversampling (RO), Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN). These resampling strategies are paired with classifiers including Decision Trees (DT), Random Forests (RF), K-Nearest Neighbours (KNN), and Support Vector Machines (SVM). Model performance is evaluated using accuracy, precision, recall, F1 score, and the Area Under the Curve (AUC). Our results show that tailored resampling significantly boosts machine learning model performance in healthcare settings. Notably, SVM with ENN undersampling markedly improves accuracy for lung cancer predictions, while SVM and RF with IHT achieve higher validation accuracies for both diseases. Random oversampling shows variable effectiveness across datasets, whereas SMOTE and ADASYN consistently enhance accuracy. This study underscores the value of integrating strategic resampling with machine learning to improve predictive reliability for imbalanced healthcare data.**

***Keywords*—Imbalanced data; resampling; machine learning; healthcare**

I. Introduction

The accurate prediction of medical conditions can play a key role in improving patient outcomes and healthcare systems. Early detection of disease is essential for anticipating its clinical progression and developing effective treatments. However, accurate predictions remain a significant challenge [1].

Such predictions enable for well-informed decisions on diagnosis, intervention, and treatment by both patients and healthcare practitioners. Early detection of potential health issues enables timely, potentially life-saving interventions. These interventions significantly improve patient survival and quality of life. Furthermore, more accurate disease forecasts can help reduce healthcare costs by minimising unnecessary diagnostic tests and treatments. Hence, robust predictive models are clearly needed to enhance forecasting accuracy in clinical settings [2], [3].

Imbalanced datasets, however, present a significant challenge to achieving high predictive performance. This issue arises when one class, often a crucial medical condition, is significantly under-represented compared to the majority class. As a result, models trained on such datasets may become biased towards the majority class. This bias reduces the accuracy of minority-class predictions [4].

Precise identification of the minority class is often crucial in medical diagnostics, especially in screening programs where identifying disease-positive patients is paramount. The consequences of misclassification in these cases can be severe. The dominance of the majority class in imbalanced datasets can compromise a model's ability to detect rare but critical cases. This can lead to delayed or missed diagnoses, suboptimal treatment, and potential healthcare disparities [5].

In this paper, we propose a novel approach to predictive modelling for imbalanced healthcare data, specifically targeting heart and lung disease datasets. Our primary contribution is the integration of advanced resampling techniques with established machine learning algorithms to address data imbalance effectively. This study is driven by the following research question:

How can resampling strategies and machine learning algorithms be combined to improve predictive accuracy on imbalanced healthcare datasets?

To answer this question, we pursue the following research objectives:

- Evaluate the effectiveness of various resampling techniques (ENN, IHT, RO, SMOTE, ADASYN) in mitigating class imbalance in heart and lung disease datasets.
- Compare the performance of multiple machine learning algorithms (DT, RF, KNN, SVM) when paired with different resampling strategies.
- Determine the optimal combination of resampling methods and classifiers for improving disease prediction accuracy while maintaining computational efficiency.
- Assess the generalisability of resampling strategies across distinct medical conditions (heart disease and lung cancer) and provide insights into their applicability in real-world healthcare settings.

Various resampling strategies have been developed to ameliorate these challenges. Oversampling approaches, such as Random Oversampling, SMOTE, and ADASYN, increase the representation of the minority class by either replicating existing minority instances or generating new synthetic samples. Conversely, undersampling methods reduce the size

of the majority class by removing selected instances, which allows algorithms greater opportunity to learn minority-class patterns [6], [7]. Despite their promise, each technique has limitations. For instance, oversampling methods risk overfitting by replicating minority instances or generating low-variance synthetic points. Undersampling, on the other hand, may discard potentially valuable information.

Recent studies confirm that such resampling methods can significantly enhance the analysis and interpretation of medical data, which enables more accurate and reliable predictions [8]–[10]. However, questions remain about how best to integrate these techniques within different disease contexts. This is particularly true for conditions like heart and lung diseases, which have distinct patterns of risk factors, symptomatology, and prevalence rates.

Our experiments demonstrate that strategic combinations of sampling methods and machine learning models significantly improve both accuracy and reliability, providing a robust framework for addressing the challenges posed by imbalanced datasets in healthcare. Specifically, we find that Support Vector Machines (SVM) coupled with ENN significantly enhance lung cancer prediction accuracy. Moreover, SVM and Random Forest models utilising IHT achieve high validation accuracies for heart and lung disease data. These findings surpass traditional approaches in performance and illustrate a robust framework for effectively addressing the challenges posed by imbalanced medical datasets.

By systematically evaluating the impact of multiple resampling techniques on predictive performance, this study provides a comprehensive understanding of how to improve medical data analysis in the presence of imbalance. Our findings offer healthcare researchers and practitioners enhanced tools for early disease detection and intervention, contributing to improved patient care and outcomes.

II. Related Work

A growing body of research has sought to refine predictive modelling in healthcare by addressing the persistent challenge of data imbalance. Although these works collectively demonstrate the efficacy of outlier detection, resampling methods, and advanced machine learning strategies, key uncertainties persist regarding the transferability and consistency of these approaches across different disease domains. This section examines the existing literature, highlighting both substantial progress and the outstanding questions that motivate our study.

A. Foundational Approaches to Imbalanced Data

Fitriyani et al. [11] developed a heart disease prediction model integrating Density-based Spatial Clustering of Applications with Noise (DBSCAN) for outlier detection, hybrid SMOTE-ENN for balancing training data, and XGBoost for classification. Their results on the Statlog and Cleveland datasets emphasise the gains possible when combining sophisticated outlier removal with carefully chosen resampling techniques. However, their investigation focuses on a single disease domain and two datasets, leaving open the question of whether such hybrid pipelines would maintain similar levels of performance when confronting data from different medical conditions or with varying degrees of imbalance.

Khushi et al. [12] shifted attention to lung cancer prediction, systematically comparing traditional classifiers and imbalance strategies (including oversampling) on the PLCO and NLST datasets. Their findings reinforce the significance of well-chosen resampling techniques for boosting model performance in highly skewed datasets. Yet, the study's scope largely confines itself to lung cancer, with minimal discussion of how these strategies might generalise to other diseases - particularly those with distinct clinical signatures, feature sets, and imbalance ratios.

B. Feature Selection and Ensemble Classifiers

Ishaq et al. [13] proposed a feature-selection-based strategy for predicting heart-failure survival, demonstrating how SMOTE, combined with a Random Forest approach for feature importance, can markedly improve predictive accuracy. This underscores that well-targeted feature engineering, used in tandem with resampling, can mitigate data imbalance. However, while they demonstrate promising gains, their approach again centres on a single disease (heart failure) and does not address whether similar techniques would be equally beneficial for other conditions particularly where key feature sets differ significantly (e.g. in lung disease).

Ghorbani et al. [14] examined multiple SMOTE variants for educational data, illustrating how subtle algorithmic differences in resampling methods can translate into pronounced differences in classifier performance. Although their work concerns non-medical data, it offers a compelling reminder that each dataset may respond uniquely to different sampling methods. The question remains how to identify which undersampling or oversampling strategies e.g. ENN, IHT, SMOTE, ADASYN are most compatible with specific disease datasets that frequently exhibit high dimensionality, noisy features, or overlapping classes.

C. Advanced Architectures and Cross-Domain Insights

Li et al. [15] demonstrated how complex, dual-stage attention-based models with CBAM modules can handle data imbalance in engineering fault diagnosis. While fault-diagnosis data differ from clinical datasets, their positive results point toward the adaptability of advanced architectures for skewed classification problems. Nonetheless, applying such intricate methods to medical data raises additional concerns of interpretability and domain relevance, as well as the significant computational overhead often associated with deep learning in real-world clinical settings.

D. Summary of Gaps and Motivation

Despite these advancements, three primary gaps emerge:

1. Current studies typically focus on a single disease (e.g. heart disease or lung cancer), limiting our understanding of whether and how resampling methods generalise across conditions with distinct pathophysiology, class ratios, and feature distributions.

2. While various works explore SMOTE, ENN, or other techniques, they frequently implement a single approach or compare only a narrow set of sampling techniques. Similarly, they often limit themselves to a small subset of classification

algorithms, resulting in an incomplete picture of how different combinations of sampling algorithms and machine learning models might perform under different clinical constraints.

3. Many studies highlight accuracy gains, but relatively few provide detailed insights into the computational trade-offs such as training and inference times, resource consumption, or real-time feasibility that are crucial for deployment in actual healthcare settings, for instance, Fitriyani et al. [11] and Khushi et al. [12] focus on performance without efficiency analysis. Moreover, interpretability and domain-specific constraints have not always been sufficiently addressed when employing more advanced modelling frameworks.

In light of these gaps, our study aims to:

1) *Evaluate multiple resampling methods (ENN, IHT, RO, SMOTE, and ADASYN) across two distinct diseases*: heart disease and lung cancer. By spanning two clinically significant but different pathologies, we shed light on how imbalanced data strategies may generalise or require adaptation across disease contexts.
2) *Compare four diverse machine learning algorithms* (Decision Trees, Random Forests, K-Nearest Neighbours, and Support Vector Machines) under each resampling technique, thus providing a broader evidence base on which model–sampling pairs work best for specific disease datasets.
3) *Examine computational efficiency and practicality*, reporting on training/validation/testing times and potential interpretability issues, thereby informing real-world clinical adoption.

By addressing these specific gaps, our research contributes a deeper and more generalisable understanding of how to tailor resampling approaches and classifier choices in the face of disease-specific imbalances. This work ultimately aspires to promote more robust, reliable, and domain-aware predictive models in healthcare.

III. Methodology

We employ a straightforward yet comprehensive methodology (see Fig. 1) to address the challenges posed by data imbalance in medical predictive modelling. Our process encompasses data preprocessing, various resampling strategies, machine learning algorithm selection, and performance evaluation. The choice of each algorithm and sampling method is guided by the characteristics of the classification tasks (i.e. heart disease vs. lung disease) and by the need to mitigate imbalance effectively.

A. Sampling Techniques

1) Undersampling Techniques: Edited Nearest Neighbours (ENN) and Instance Hardness Threshold (IHT) are employed to mitigate the effect of imbalanced datasets. ENN removes misclassified or noisy observations by comparing each sample to its k nearest neighbours, thereby cleansing borderline cases in the dataset. In contrast, IHT identifies and removes "hard-to-classify" instances based on their proximity to the decision boundary, helping models concentrate on more informative examples [16], [17]. By discarding problematic samples, both methods aim to create a dataset that better represents minority-class patterns without overwhelming the model.

2) Oversampling Techniques: Random Oversampling duplicates minority-class instances to balance the dataset, albeit with a heightened risk of overfitting. To address this limitation, we also use the Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic (ADASYN) sampling, both of which generate synthetic examples of the minority class [18], [19]. These strategies introduce diversity into training data and reduce the bias towards the majority class, enabling models to learn more nuanced decision boundaries.

B. Datasets Utilisation

Two publicly available datasets are utilised in this study. The Kaggle Lung Cancer dataset [20] comprises 163,763 records with detailed patient information (including binary labels denoting the presence of lung cancer). The heart disease dataset [21] from the UCI repository combines multiple sources and focuses on 14 key attributes (out of an original 76) to predict the presence of cardiovascular disease. These datasets were chosen for their relevance to clinical practice and their relatively high degree of imbalance, offering an appropriate testbed for evaluating sampling methods.

C. Machine Learning Algorithms

Four machine learning algorithms are employed to assess the impact of each sampling strategy:

- Decision Trees (DT): An interpretable, rule-based approach that partitions data into increasingly homogeneous subsets.
- Random Forests (RF): An ensemble of multiple decision trees, aggregating predictions to reduce overfitting and enhance generalisability.
- K-Nearest Neighbours (KNN): A distance-based classifier capable of handling non-linear decision boundaries, sensitive to local data structure.
- Support Vector Machines (SVM): A robust method for binary classification, especially effective in high-dimensional spaces, using kernel functions to separate classes with maximal margins.

Each algorithm is trained and evaluated on both datasets with the aforementioned sampling techniques applied. This setup allows for a direct comparison of how undersampling and oversampling methods impact model accuracy, precision, recall, and F1-score.

D. Methodological Flow and Evaluation

As summarised in Fig. 1, data preprocessing and resampling are carried out first to correct for imbalance. The selected models (DT, RF, KNN, and SVM) are then trained on these preprocessed datasets. Finally, we measure predictive performance using accuracy, precision, recall, and F1-score to gauge the effectiveness of each sampling approach.

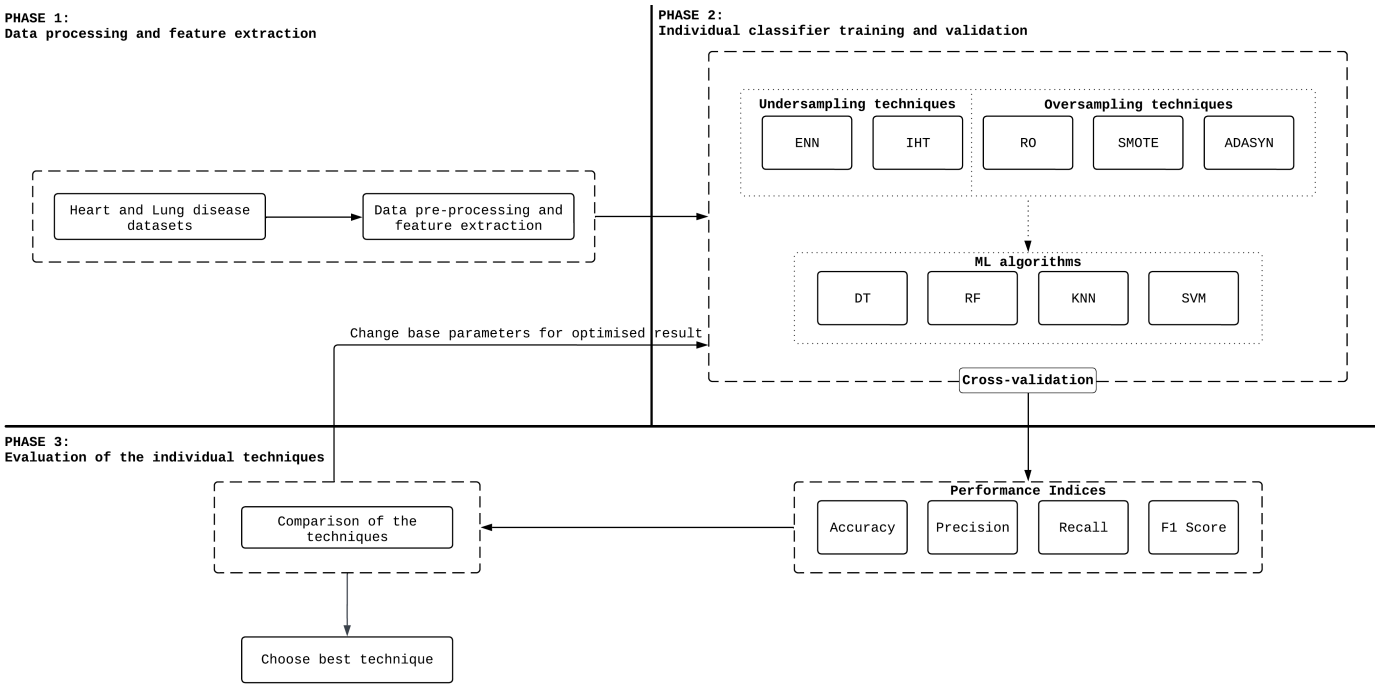


Fig. 1. A three-phase approach to heart and lung disease classification, including data preprocessing, resampling techniques, machine learning model selection, and performance evaluation.

IV. Experimental Setup

This study investigates the effect of various resampling strategies on the predictive accuracy of machine learning models when dealing with imbalanced healthcare data from Kaggle (lung cancer) and UCI Machine Learning Repository (heart disease).

A. Dataset Description

1) Kaggle Lung Cancer Dataset: This dataset comprises patient-level data intended for predicting the occurrence of lung cancer. It includes features such as age, gender, smoking history, and various diagnostic measurements. The primary challenge stems from its imbalanced distribution, with a significantly lower proportion of positive lung cancer instances compared to negative ones.

2) UCI Heart Disease Dataset: This collection of patient records is aimed at predicting the presence of heart disease. It includes demographic information, symptomatology, and diagnostic test results. Similar to the lung cancer dataset, it is highly imbalance, with heart disease instances occurring less frequently than negative cases, which makes it an excellent testbed for evaluating the effectiveness of resampling techniques.

B. Data Collection and Preprocessing

Data were collected and rigorously preprocessed to ensure quality and consistency. The preprocessing phase addressed missing values, outliers, and normalisation:

- Mean Imputation for Missing Values: For each feature containing missing entries, we replaced missing values with the mean value of that feature across all available data points:
$$\mu_{\text{missing}} = \frac{\sum_{i=1}^{n} x_i}{n}.$$
- Normalisation (Min–Max Scaling): We rescaled each feature to a common range of $[0, 1]$ using Min–Max scaling:
$$x' = \frac{x - \min(x)}{\max(x) - \min(x)},$$
ensuring that differences in the original data ranges are preserved without distortion.

C. Sampling Techniques

Given the imbalanced nature of these datasets, we applied five resampling methods: two undersampling techniques - Edited Nearest Neighbour (ENN) and Instance Hardness Threshold (IHT) - and three oversampling techniques - Random Oversampling (RO), Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic (ADASYN).

- Edited Nearest Neighbour (ENN): ENN refines the dataset by removing samples that are misclassified by their k-nearest neighbours, thereby enhancing the purity of the minority class:
$$ENN(S) = \{\, x \in S \mid \text{class}(x) = \text{class}(\text{kNN}(x)) \}.$$
- Instance Hardness Threshold (IHT): IHT iteratively applies oversampling approaches and evaluates classifier performance on the modified dataset, aiming to attain a balance that optimises the model's sensitivity to the minority class.

- Random Oversampling (RO): RO duplicates instances of the minority class to rebalance the dataset, albeit at an increased risk of overfitting.
- SMOTE and ADASYN: Both SMOTE and ADASYN generate synthetic minority examples, helping to balance class distribution and introduce diversity to the training data. This prepares the model to generalise more effectively from underrepresented classes.

D. Model Selection and Training

We selected four machine learning algorithms for evaluation: Decision Trees, Random Forests, K-Nearest Neighbours, and Support Vector Machines. These models represent a spectrum of complexity: from the relatively interpretable structure of Decision Trees to the more sophisticated nature of SVMs. The training process involved optimising hyperparameters for each model using techniques such as grid search and cross-validation to maximise overall performance.

E. Evaluation Metrics

We employed accuracy, precision, recall, and the F1-score to evaluate model performance comprehensively. These metrics were chosen not only to gauge overall accuracy but also to assess the models' ability to identify positive instances (recall) and the reliability of those predictions (precision).

a) Precision:

$$\text{Precision} = \frac{TP}{TP + FP},$$

where TP (true positives) are correctly predicted positive instances, and FP (false positives) are incorrectly predicted as positive. High precision in this work indicates that when the model predicts a disease, it is likely correct.

b) Recall (Sensitivity):

$$\text{Recall} = \frac{TP}{TP + FN},$$

where FN (false negatives) are positive instances incorrectly predicted as negative. Recall here reflects the model's capacity to detect all relevant cases.

c) F1-Score:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

This harmonic mean of precision and recall is especially relevant when misclassifications either false positives or false negatives carry significant risks, as is common in healthcare.

Finally, multiple runs were conducted for each model and sampling combination to ensure the reliability of findings and to capture performance variance across different trials. This robust experimental structure enables a comprehensive assessment of how well different resampling strategies address data imbalance in healthcare predictive modelling.

V. Results

A. Experiment with the Heart Dataset

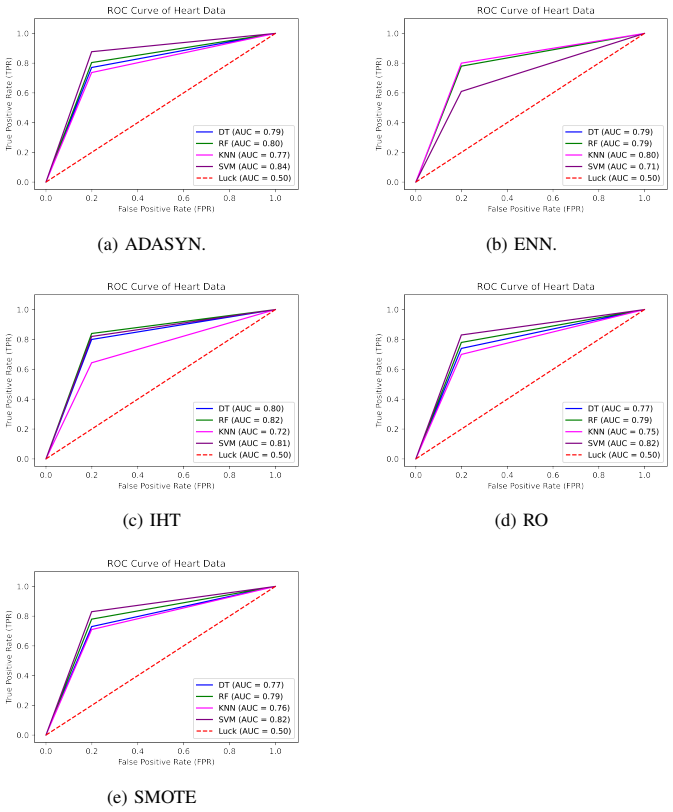
Table I presents the results for the heart dataset. Overall, Support Vector Machines (SVM) excelled in precision, particularly when paired with Edited Nearest Neighbours (ENN), underlining its effectiveness in handling imbalanced data. SVM thus appears well-suited to heart disease prediction, consistently achieving high scores in precision, F1, and AUC. These results are especially significant in medical diagnostics, where balancing sensitivity and specificity can be challenging. Notably, the AUC values were strong for Random Forest (RF) with Instance Hardness Threshold (IHT) and SVM with Adaptive Synthetic Sampling (ADASYN), further highlighting SVM's capacity to reduce false positives while maintaining a high detection rate.

TABLE I. Performance Metrics for Various Classifiers on the Heart Dataset Across Different Resampling Techniques, Demonstrating Superior Precision and Recall by SVM and RF

		ENN	IHT	RANDOM	SMOTE	ADASYN
DT	Precision	0.90	0.82	0.88	0.82	0.81
	Recall	0.90	0.79	0.83	0.82	0.76
	F1 Score	0.90	0.81	0.85	0.82	0.79
	AUC	0.781	0.803	0.747	0.7361	0.7712
RF	Precision	0.90	0.89	0.90	0.87	0.88
	Recall	**0.93**	0.89	0.77	0.77	0.75
	F1 Score	0.91	**0.89**	0.83	0.82	0.81
	AUC	0.781	**0.843**	0.78	0.7806	0.804
KNN	Precision	0.92	0.68	0.71	0.73	0.73
	Recall	0.83	0.59	0.71	0.63	**0.88**
	F1 Score	0.87	0.63	0.71	0.68	**0.80**
	AUC	**0.802**	0.6449	0.705	0.719	0.7374
SVM	Precision	**0.96**	**0.95**	**0.93**	0.74	**0.91**
	Recall	0.93	0.75	0.74	**0.94**	0.90
	F1 Score	**0.95**	0.84	0.83	0.83	0.90
	AUC	0.618	0.8224	0.8305	0.8306	**0.8779**

Fig. 2 illustrates ROC curves for various classifiers on the Heart Dataset across different resampling techniques, further confirming that SVM and RF generally outperform the other classifiers. Particularly noteworthy is ENN's significant enhancement of SVM's performance, as reflected in the high AUC scores. Likewise, ADASYN combined with SVM reached the highest AUC among the tested setups, indicating its efficacy in tackling data imbalance. By comparison, the RANDOM and SMOTE approaches yielded moderately lower AUCs, underscoring the importance of carefully aligning sampling methods with the dataset's characteristics.

Table II indicates that computational efficiency varies with the choice of resampling method. Decision Tree (DT) and KNN demonstrate notably faster training times, especially under the RANDOM and ADASYN schemes, suggesting relatively low computational overhead. By contrast, SVM exhibits longer training times particularly with RANDOM highlighting a potential trade-off between speed and predictive performance.



(a) ADASYN. (b) ENN.

(c) IHT (d) RO

(e) SMOTE

Fig. 2. ROC curves for various classifiers on the Heart Dataset across different resampling techniques. The curves demonstrate the superior performance of SVM and RF, particularly with ENN and ADASYN. These methods enhance the classifiers' ability to handle imbalanced data, as evidenced by high AUC values.

TABLE II. COMPARISON OF TRAINING, VALIDATION, AND TESTING TIMES FOR DIFFERENT CLASSIFIERS ON THE HEART DATASET USING VARIOUS RESAMPLING METHODS. NOTABLE DIFFERENCES IN COMPUTATIONAL EFFICIENCY ARE OBSERVED, PARTICULARLY FOR SVM AND RF

		ENN	IHT	RANDOM	SMOTE	ADASYN
DT	Training	0.0031	0.002	**0.0001**	0.0038	**0.0001**
	Validation	0.013	0.010	**0.0084**	**0.0084**	0.0093
	Testing	0.0151	0.0136	0.0050	0.0090	**0.0006**
RF	Training	0.2885	0.0172	0.0802	0.0745	**0.0438**
	Validation	0.0280	**0.0080**	0.0082	0.0170	0.0083
	Testing	0.0280	**0.0063**	0.0104	0.0161	0.0133
KNN	Training	0.0051	**0.0001**	**0.0001**	0.0035	**0.0001**
	Validation	0.0160	**0.0033**	0.0111	0.0102	0.0111
	Testing	0.0125	0.0099	0.0063	0.0128	**0.0040**
SVM	Training	0.2191	0.0780	**0.0577**	0.0919	0.0825
	Validation	0.0143	0.0114	**0.0071**	0.0088	0.0103
	Testing	0.0121	**0.0020**	0.0077	0.0101	0.0056

B. Experiment with the Lung Dataset

Table III details the results for the lung dataset. Both Decision Tree (DT) and Random Forest (RF) achieved better precision when paired with ENN, IHT, and RANDOM sampling, underlining their capacity for highly accurate lung disease classification. However, DT's slightly lower AUC under RANDOM sampling and SMOTE suggests minor compromises in balancing sensitivity and specificity. While DT occasionally overlooks some true cases, as evidenced by its recall, RF's flawless recall with ENN and IHT demonstrates that it seldom misses positive instances. As a result, both DT and RF prove highly dependable, with RF offering a small edge in comprehensive patient identification.

TABLE III. PERFORMANCE METRICS FOR VARIOUS CLASSIFIERS ON THE LUNG DATASET ACROSS DIFFERENT RESAMPLING TECHNIQUES. PHENOMENAL PRECISION AND RECALL ARE ACHIEVED BY DT AND RF WITH ENN AND IHT, WHILE SVM SHOWS CONSISTENTLY HIGH PERFORMANCE

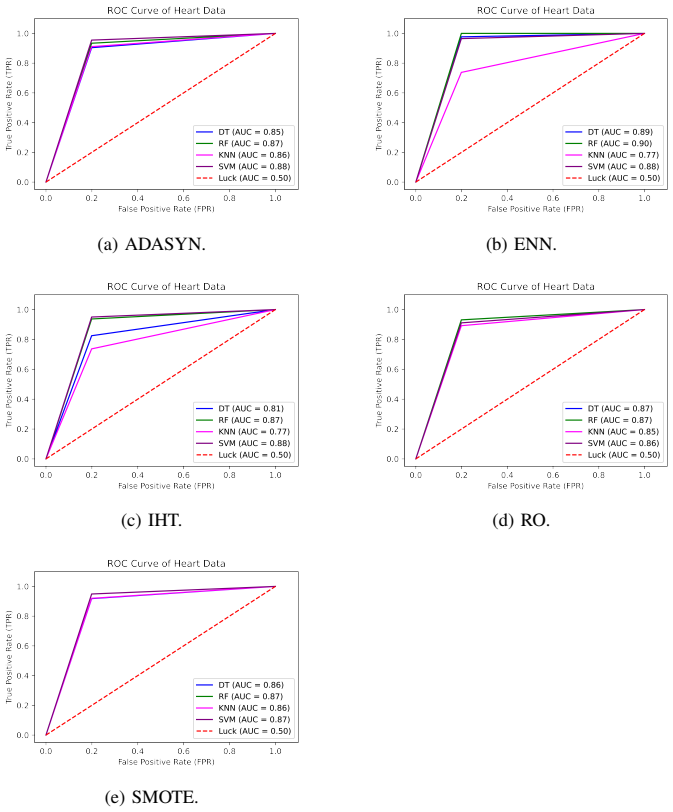
		ENN	IHT	RANDOM	SMOTE	ADASYN
DT	Precision	**1**	**1**	**1**	0.96	0.94
	Recall	0.95	**1**	0.88	0.88	0.91
	F1 Score	0.98	**1**	0.93	0.92	0.93
	AUC	0.977	0.825	0.931	0.918	0.9041
RF	Precision	**1**	**1**	**1**	0.92	0.95
	Recall	**1**	**1**	0.88	0.98	0.93
	F1 Score	**1**	**1**	0.94	0.95	0.94
	AUC	**1**	0.93	0.931	0.94	0.934
KNN	Precision	0.98	0.75	**1**	0.93	0.92
	Recall	0.98	0.90	0.73	0.93	0.91
	F1 Score	0.98	0.82	0.84	0.93	0.91
	AUC	0.738	0.7375	0.8921	0.92	0.911
SVM	Precision	**1**	**1**	**1**	0.95	0.96
	Recall	**1**	**1**	0.90	0.95	0.96
	F1 Score	**1**	**1**	0.95	0.95	0.96
	AUC	0.9659	**0.95**	0.911	0.949	**0.955**

While KNN shows some variability particularly in precision with IHT it still maintains solid results in precision and recall, indicating its ability to identify lung disease patients accurately. By contrast, SVM consistently delivers near-perfect performance across all evaluated metrics and resampling configurations, highlighting its powerful capacity to discriminate between diseased and healthy individuals. Its balanced high precision and recall minimise both misdiagnoses and missed diagnoses.

TABLE IV. TRAINING, VALIDATION, AND TESTING TIMES FOR DIFFERENT CLASSIFIERS ON THE LUNG DATASET USING VARIOUS RESAMPLING METHODS. DT AND KNN OFTEN REQUIRE MINIMAL COMPUTATION WITH RANDOM AND ADASYN

		ENN	IHT	RANDOM	SMOTE	ADASYN
DT	Training	**0.0001**	0.0077	**0.0001**	**0.0001**	0.0081
	Validation	**0.0046**	0.0098	0.0083	0.0102	0.0131
	Testing	0.0085	**0.0030**	0.0085	0.0098	0.0202
RF	Training	**0.0193**	0.0250	0.0999	0.0833	0.0329
	Validation	0.0122	**0.0107**	0.0134	0.0169	0.0203
	Testing	0.0110	0.0081	0.0110	0.0087	0.0123
KNN	Training	0.0065	0.0017	**0.0001**	0.0015	0.0020
	Validation	0.0081	0.0082	0.0100	**0.0081**	0.0142
	Testing	0.0100	0.0067	**0.0076**	0.0146	0.0166
SVM	Training	0.0080	**0.0001**	0.8198	0.0121	0.0166
	Validation	0.0062	0.0085	0.0084	**0.0082**	0.0145
	Testing	0.0084	0.0086	0.0101	0.0112	**0.0101**

The F1 and AUC values reaching 100% for RF with ENN and IHT underscore the potency of these resampling methods in boosting accuracy for markedly skewed datasets. Interestingly, the lung dataset seems to respond more strongly to these techniques than the heart dataset, potentially reflecting



(a) ADASYN.

(b) ENN.

(c) IHT.

(d) RO.

(e) SMOTE.

Fig. 3. ROC curves for various classifiers on the Lung Dataset across different resampling techniques. The curves highlight the exceptional precision and recall achieved by RF and SVM, especially with ENN and IHT, confirming their robustness in identifying lung disease despite data imbalance.

inherent differences in class distribution or unique dataset attributes.

Fig. 3 shows ROC curves for various classifiers on the Lung Dataset across different resampling techniques. The curves highlight the exceptional precision and recall achieved by RF and SVM, especially with ENN and IHT, confirming their robustness in identifying lung disease despite data imbalance.

Finally, Table IV summarises the training, validation, and testing times for different classifiers on the Lung Dataset. Decision Trees (DT) and KNN demonstrate relatively low computational demands, most notably under the RANDOM and ADASYN resampling methods, making them viable options where computational resources are limited. Random Forest (RF) and SVM, in contrast, show varying training times, with RF proving efficient under ENN and IHT, while SVM attains rapid validation times with ENN. These results suggest that DT and KNN may suffice where speed is paramount, whereas RF and SVM, coupled with certain resampling strategies, strike a more favourable balance between computational overhead and predictive performance.

VI. Discussion

The experimental results confirm that pairing tailored resampling methods with appropriate classifiers can greatly improve predictive performance in imbalanced healthcare datasets. The results show that combining advanced resampling techniques with machine learning models improves predictive accuracy for imbalanced healthcare data. However, performance differences across datasets highlight the need to select resampling strategies and classifiers based on dataset characteristics.

A. Effectiveness of Resampling Methods

Resampling techniques play a crucial role in addressing class imbalance in heart disease and lung cancer predictions. Instance Hardness Threshold (IHT) and Edited Nearest Neighbours (ENN) performed best for undersampling, while ADASYN and SMOTE increased minority-class representation through synthetic data generation. Their effectiveness varied based on the dataset and classifier.

IHT improved precision and recall in heart disease predictions because it removed ambiguous majority-class instances. This effect was most pronounced when combined with Random Forest and Support Vector Machines. ENN produced the best results for lung cancer predictions because it refined the decision boundary and enhanced SVM's ability to distinguish between classes. ADASYN and SMOTE produced general improvements, but they worked best for heart disease predictions where a moderate imbalance allowed for better synthetic sample generation. These findings confirm that resampling techniques must be chosen based on the dataset rather than applied universally.

B. Classifier Performance Across Datasets

Classifier performance depended on disease type and resampling strategy. Support Vector Machines performed best for lung cancer prediction, especially when combined with ENN. This combination removed noisy samples and allowed SVM to establish a clearer decision boundary. Random Forest and Decision Trees outperformed other models for heart disease predictions because they captured feature interactions effectively.

SVM provided the highest accuracy and recall for lung cancer predictions, making it the best choice for highly imbalanced, binary-structured datasets. Random Forest and Decision Trees worked better for heart disease because they handled mixed categorical and numerical data more effectively. K-Nearest Neighbours showed inconsistent results across both datasets because it was highly sensitive to imbalance and feature distribution.

C. Dataset-Specific Performance

Model performance varied between the heart disease and lung cancer datasets due to differences in class imbalance severity, feature types, and data distribution. The heart disease dataset, with moderate imbalance and mixed categorical-numerical features, favored Random Forest with IHT and ADASYN, which excel at capturing complex feature interactions. Conversely, the lung cancer dataset, with extreme imbalance and mostly binary features, suited SVM with ENN, as this combination effectively refines sparse decision boundaries. These results suggest our algorithms are better suited to specific data types: RF thrives with heterogeneous features,

while SVM excels with binary, highly skewed data. This adaptability underscores the importance of matching resampling and classifiers to dataset properties.

D. Computational Considerations

The study also highlights key computational trade-offs. SVM and Random Forest provided the highest accuracy, but they required longer training times. These models are suitable for offline model development rather than real-time applications. Decision Trees and K-Nearest Neighbours trained and tested faster but produced less consistent results for imbalanced data. Selecting a model for deployment depends on balancing efficiency, interpretability, and predictive performance.

E. Real-World Implications

These findings have direct applications in predictive healthcare. SVM with ENN delivers the most accurate classification for lung cancer screening, reducing false negatives. Random Forest with IHT works better for heart disease prediction because it maintains interpretability while achieving strong performance. These findings guide healthcare practitioners in selecting predictive models that improve early diagnosis and patient outcomes.

F. Advantages Over Existing Methods

Our approach outperforms traditional methods by integrating resampling with classifier optimisation tailored to dataset characteristics. Unlike studies like Fitriyani et al. [11], which focus on single-disease hybrid pipelines (e.g. SMOTE-ENN with XGBoost), our framework evaluates multiple resampling techniques (ENN, IHT, RO, SMOTE, ADASYN) across two diseases, offering broader applicability. Compared to Khushi et al. [12], which limits comparisons to lung cancer with basic oversampling, our SVM+ENN and RF+IHT combinations achieve higher precision and recall (e.g. 100% F1 for lung cancer with RF+ENN) while addressing computational trade-offs. This adaptability and performance edge make our method a versatile tool for imbalanced healthcare data beyond single-context solutions.

G. Validation and Comparative Significance

Our use of comprehensive validation measures such as accuracy, precision, recall, F1 score, and AUC ensures robust evaluation of model performance, prioritising both correctness and sensitivity critical in healthcare. Unlike prior works like Ishaq et al. [13], which focus on accuracy alone, our multi-metric approach highlights trade-offs (e.g. SVM+ENN's perfect recall for lung cancer). Thorough comparisons with related studies, for instance, outperforming Khushi et al.'s [12] lung cancer precision with RF+IHT, validate our method's superiority. This rigorous validation confirms its reliability and generalisability across imbalanced medical datasets.

VII. Conclusion

This study demonstrates that advanced resampling techniques, when integrated with optimised machine learning models, can significantly enhance classification performance for imbalanced healthcare data. Specifically, SVM with ENN excelled in lung cancer prediction, while RF with IHT or ADASYN consistently performed well for heart disease prediction. These findings provide a roadmap for choosing optimal resampling - classifier pairs based on dataset characteristics, which is crucial for early diagnosis and resource allocation in clinical practice. Our results emphasise the importance of robust methods to handle skewed medical datasets, leading to more reliable healthcare predictions.

VIII. Future Direction

Despite these advances, much remains to be done to fully realise the potential of predictive models in healthcare. The following key challenges merit attention:

1) Adaptation to novel diseases: Models trained on current datasets may underperform when confronted with emerging diseases, such as those appearing during pandemics. Differences in symptoms, transmission mechanisms, or patient demographics can impede model adaptability. Future studies should explore adaptive learning frameworks - particularly online learning techniques that enable continuous learning from new data streams. Incorporating concept-drift detection can help identify performance degradation due to shifts in data distribution, while transfer learning offers a means of rapidly adapting pre-trained models to smaller, disease-specific datasets.

2) Scalability: The exponential growth of healthcare data in both volume and complexity necessitates scalable predictive models. Managing large datasets requires substantial computational resources and careful maintenance of model efficiency. Distributed computing platforms (e.g., Apache Spark or Dask) can facilitate the processing of extensive datasets across clusters. Additionally, algorithmic efficiency may be improved through dimensionality reduction and the deployment of more efficient neural architectures, such as EfficientNets, to achieve real-time processing capabilities.

3) Multimodal integration: Healthcare data are inherently multimodal, encompassing structured electronic health records, unstructured clinical notes, medical imaging, and genomic information. Current predictive models often fail to fully exploit these diverse data sources. Developing multimodal learning methods that accommodate varying data formats can provide a more comprehensive view of patient health. Techniques such as Canonical Correlation Analysis (CCA) and cross-modal neural networks can learn joint representations, while strategic fusion strategies help integrate disparate data modalities effectively.

4) Handling rare diseases: Rare diseases pose particular challenges due to significant class imbalance and very limited positive cases. Insufficient data hampers the training of robust predictive models, thereby limiting opportunities for early intervention. Future work should focus on few-shot learning approaches, for example, prototypical networks and Model-Agnostic Meta-Learning (MAML) that enable models to learn from minimal examples. Transfer learning can also leverage existing biomedical knowledge by fine-tuning models trained on more common diseases for rare disease datasets.

5) Bias and fairness: Predictive models can inherit biases from training data, potentially reinforcing inequities in healthcare access and outcomes. Without appropriate safeguards,

such models may over-predict or under-predict disease risk in certain demographic groups. Fairness-aware machine learning strategies such as re-weighting datasets for balanced representation, introducing fairness constraints during training, and performing post-hoc adjustments are therefore essential. Rigorous bias audits, which scrutinise performance across diverse patient populations, are vital for identifying and mitigating disparities.

6) Interpretability and trust: The "black box" nature of many advanced machine learning models is a major obstacle to clinical adoption, as healthcare professionals are often reluctant to rely on systems whose inner workings they cannot examine. Enhancing interpretability is crucial for clinician trust and seamless adoption within medical practice. Explainable AI (XAI) methods such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) can clarify model reasoning. User-friendly interfaces that integrate with existing healthcare IT systems can further bolster clinician acceptance and support practical workflows.

Beyond these methodological and computational considerations, practical issues such as data-sharing restrictions, privacy laws, and regulatory compliance (e.g. with the MHRA or NICE guidelines) will also shape the real-world deployment of AI-driven healthcare models. Although our results are encouraging, implementing these solutions at scale poses substantial challenges. Overcoming these hurdles will be essential for fully leveraging the potential of predictive modelling to improve patient outcomes and transform healthcare delivery.

REFERENCES

[1] S. Emmons-Bell, C. Johnson, and G. Roth, "Prevalence, incidence and survival of heart failure: a systematic review," *Heart*, vol. 108, no. 17, pp. 1351–1360, 2022.

[2] F. Alahmari, "A comparison of resampling techniques for medical data using machine learning," *Journal of Information & Knowledge Management*, vol. 19, no. 01, p. 2040016, 2020.

[3] J. Thompson Burdine, S. Thorne, and G. Sandhu, "Interpretive description: a flexible qualitative methodology for medical education research," *Medical education*, vol. 55, no. 3, pp. 336–343, 2021.

[4] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proceedings of the National Academy of Sciences*, vol. 117, no. 23, pp. 12 592–12 594, 2020.

[5] S. Hong, Y. Zhou, J. Shang, C. Xiao, and J. Sun, "Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review," *Computers in biology and medicine*, vol. 122, p. 103801, 2020.

[6] R. Kumar, W. Wang, J. Kumar, T. Yang, A. Khan, W. Ali, and I. Ali, "An integration of blockchain and ai for secure data sharing and detection of ct images for the hospitals," *Computerized Medical Imaging and Graphics*, vol. 87, p. 101812, 2021.

[7] H. Matsuo, M. Nishio, T. Kanda, Y. Kojita, A. K. Kono, M. Hori, M. Teshima, N. Otsuki, K.-i. Nibu, and T. Murakami, "Diagnostic accuracy of deep-learning with anomaly detection for a small amount of imbalanced data: discriminating malignant parotid tumors in mri," *Scientific Reports*, vol. 10, no. 1, p. 19388, 2020.

[8] S. J. Stratton, "Population research: convenience sampling strategies," *Prehospital and disaster Medicine*, vol. 36, no. 4, pp. 373–374, 2021.

[9] Y. K. Cherniavskyi, A. Fathizadeh, R. Elber, and D. P. Tieleman, "Computer simulations of a heterogeneous membrane with enhanced sampling techniques," *The Journal of Chemical Physics*, vol. 153, no. 14, 2020.

[10] R. Evans, L. Hovan, G. A. Tribello, B. P. Cossins, C. Estarellas, and F. L. Gervasio, "Combining machine learning and enhanced sampling techniques for efficient and accurate calculation of absolute binding free energies," *Journal of chemical theory and computation*, vol. 16, no. 7, pp. 4641–4654, 2020.

[11] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "Hdpm: an effective heart disease prediction model for a clinical decision support system," *IEEE Access*, vol. 8, pp. 133 034–133 050, 2020.

[12] M. Khushi, K. Shaukat, T. M. Alam, I. A. Hameed, S. Uddin, S. Luo, X. Yang, and M. C. Reyes, "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109 960–109 975, 2021.

[13] A. Ishaq, S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Improving the prediction of heart failure patients' survival using smote and effective data mining techniques," *IEEE access*, vol. 9, pp. 39 707–39 716, 2021.

[14] R. Ghorbani and R. Ghousi, "Comparing different resampling methods in predicting students' performance using machine learning techniques," *IEEE Access*, vol. 8, pp. 67 899–67 911, 2020.

[15] J. Li, Y. Liu, and Q. Li, "Intelligent fault diagnosis of rolling bearings under imbalanced data conditions using attention-based deep learning method," *Measurement*, vol. 189, p. 110500, 2022.

[16] Y. Zhu, C. Jia, F. Li, and J. Song, "Inspector: a lysine succinylation predictor based on edited nearest-neighbor undersampling and adaptive synthetic oversampling," *Analytical biochemistry*, vol. 593, p. 113592, 2020.

[17] J. L. Arruda, R. B. Prudêncio, and A. C. Lorena, "Measuring instance hardness using data complexity measures," in *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part II 9*. Springer, 2020, pp. 483–497.

[18] A. Glazkova, "A comparison of synthetic oversampling methods for multi-class text classification," *arXiv preprint arXiv:2008.04636*, 2020.

[19] Z. Chen, L. Zhou, and W. Yu, "Adasyn- random forest based intrusion detection model," in *2021 4th International Conference on Signal Processing and Machine Learning*, 2021, pp. 152–159.

[20] Z. Hong and J. Yang, "Lung Cancer," UCI Machine Learning Repository, 1992, DOI: https://doi.org/10.24432/C57596.

[21] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart disease data set," 1988. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease