ASIAN JOURNAL OF PHARMACEUTICAL AND CLINICAL RESEARCH

Vol 19, Issue 2, 2026

Online - 2455-3891
Print - 0974-2441

Research Article

REBALANCING DATA FOR CANCER-ASSOCIATED THROMBOSIS: COMPARISON OF DIFFERENT RESAMPLING APPROACH

FAIZA NAIMAT[1]*, KWOK-WEN NG[2]*, MATHUMALAR LOGANATHAN FAHRNI[1], NURUL HANIS AMIRUDDIN JAFRY[3], KHAIRIL ANUAR MD ISA[4], YUSNAINI MD YUSOFF[3]

[1]Department of Pharmacy Practice and Clinical Pharmacy,, Faculty of Pharmacy, Universiti Teknologi MARA, Selangor. [2]Department of Pharmceutical Chemistry, Faculty of Pharmacy, Quest International University, Perak. [3]Pusat Pengajian Citra Universiti, Universiti Kebangsaan Malaysia, Selangor. [4]Department of Basic Sciences, Faculty of Health Science, Universiti Teknologi MARA, Selangor, Malaysia. *Corresponding author: Kwok-Wen NG; Email: kwokwen.ng@gmail.com/Faiza Naimat; Email: naimatfaiza@gmail.com

Received: 12 October 2025, Revised and Accepted: 13 January 2026

ABSTRACT

Objective: Cancer-associated thrombosis (CAT) presents a complex challenge in oncology, exacerbated by data imbalances in related datasets that often lead to suboptimal outcomes in machine learning (ML) classification. Many ML algorithms were originally designed for balanced datasets, prompting this study to evaluate the interaction between logistic regression (LR) and eXtreme Gradient Boost (XGBoost) and data resampling techniques for improving prediction on imbalances in Malaysian data on CAT (MDCAT).

Methods: Random oversampling (ROS), random undersampling (RUS), and a combined oversampling and undersampling approach (BOTH) were applied to MDCAT dataset. Classification tasks were performed using LR and XGBoost in R version 4.3.1. Classifier performance was assessed using accuracy, sensitivity, specificity, and the area under the ROC curve (AUROC) to evaluate the impact of different resampling techniques.

Results: Applying LR and XGBoost to the imbalanced data revealed high specificity but low sensitivity in testing samples. A substantial decline in XGBoost performance was observed, with the AUC decreasing from 0.794 in training to 0.381. Metastasis, surgery, and Indian ethnicity showed statistically significant associated with the CAT event across all resampling techniques. Among XGBoost models, oversampling (XO) exhibited excellent training performance (Accuracy 0.99; AUC 0.98) but showed a large performance drop on the test set (Accuracy 0.82; AUC 0.72). Among LR models, logistic undersampling yielded the highest training accuracy (0.83) and AUC of 0.82. Tuning amplified the differences between resampling strategies and highlighted clear classifier–resampling interactions. XGBoost benefited most, particularly when trained on mixed and oversampled datasets, while LR remained comparatively stable.

Conclusion: This study demonstrated that the effectiveness of prediction models in imbalanced MDCAT dataset is strongly influenced by the interaction between classifier characteristics and resampling strategies. A tuned XGBoost model with mixed resampling outweighed the benefits of LR's simplicity and stability, making it our recommended approach given the primary importance of AUC.

Keywords: Cancer-associated thrombosis, Machine learning classification, Resampling techniques, Classifier-resampling interactions.

© 2026 The Authors. Published by Innovare Academic Sciences Pvt Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/) DOI: http://dx.doi.org/10.22159/ajpcr.2026v19i2.57152. Journal homepage: https://innovareacademics.in/journals/index.php/ajpcr

INTRODUCTION

Cancer-associated thrombosis (CAT) is a significant global health concern, ranking as the second leading cause of death among cancer patients [1]. The complexity of CAT presents major challenges for clinical management, particularly in critically ill patients. Incorporating individual patient factors is essential in personalized medicine, and clinical prediction models have become vital tools for guiding healthcare decision-making.

Conventionally, the risk of venous thromboembolism (VTE) in cancer patients has been assessed using scoring systems derived from multivariable regression analyses. These systems assign scores to various risk factors, allowing clinicians to estimate the likelihood of CAT based on a patient's total score. However, the complexity of cancer and thrombosis calls for more advanced predictive methods. In response, artificial intelligence (AI) and machine learning (ML) have emerged as powerful tools to improve the accuracy of CAT prediction [2-4]. Beyond risk stratification, AI has been shown to significantly enhance the overall quality of clinical workflows, including clinical trials [5]. The growing impact of ML in medicine is further illustrated by its successful application in other scientific domains, such as nanoscience [6,7]. Collectively, these advancements underscore AI's transformative potential in supporting early diagnosis and primary prevention, thereby reducing mortality rates, improving healthcare system efficiency, and lowering overall costs.

Recent advances in CAT prediction models, which employ methods beyond traditional risk assessment scores, have demonstrated superior accuracy [8,9]. However, these advancements face significant challenges, particularly the issue of imbalanced datasets commonly found in CAT research. Imbalanced data, characterized by unequal class distributions, complicates the performance of ML algorithms, often leading to suboptimal outcomes, especially in low-prevalence conditions like CAT [10].

In this regard, effectively addressing data imbalance is crucial for developing robust ML models. Strategies to manage imbalanced datasets fall into three main categories: data-level, algorithm-level, and hybrid approaches. Merging over-sampling and undersampling techniques with ensemble classifiers found to achieve high accuracy for the classification task. Despite the wide array of techniques available, each has its own set of challenges, and no single method has proven universally superior [11].

Undersampling and oversampling techniques remain popular due to their simplicity, yet the challenge of selecting the most suitable method for a specific dataset persists. Despite the wide array of techniques available, each has its own set of challenges, and no single method has proven universally superior.

In the context of Malaysia, research on CAT and its predictive models remains limited. The Malaysian healthcare landscape presents unique challenges, including diverse patient demographics and varying healthcare resources, which necessitate localized research. Existing studies on CAT have predominantly focused on Western populations, highlighting a significant gap in understanding how predictive models perform within the Malaysian context. This gap underscores the need for localized studies to develop and validate prediction models tailored to Malaysian patients, considering their specific risk factors and healthcare settings. This research aims to bridge that gap by to evaluate the interaction between logistic regression (LR) and XGBoost and data resampling techniques for improving prediction on imbalances in Malaysian data on CAT (MDCAT).

METHODS

Experiment design

The methodological pipeline for this study includes dataset description, data resampling, classification, and evaluation of model performance, as illustrated in Fig. 1.

Data preprocessing

Before implementing the sampling methods, essential data preprocessing steps were carried out to ensure the integrity and usability of the MDCAT dataset. Data cleaning procedures addressed missing values, with categorical data imputed using the mode and continuous data using the mean. In addition, variables such as weight, height, and body mass index (BMI) were consolidated into BMI to reduce multicollinearity.

The MDCAT dataset comprises two distinct outcomes: (i) CAT cases, and (ii) cancer patients who did not develop thrombosis (NONCAT). This binary classification dataset includes 146 CAT cases (minority class) and 355 NONCAT cases (majority class), resulting in an imbalance ratio (IR) of 2.43. It encompasses a total of 26 variables, including demographic details such as age group (18–44, 45–54, 55–64, 65–74, >75), gender, and ethnicity; clinical indicators such as ECOG performance status, smoking status, and BMI; and specific medical factors such as central venous catheter use, presence of sepsis, and various comorbidities (hypertension, diabetes, airway disease, thyroid issues, renal insufficiency, ischemic heart disease, heart failure). It also includes types of cancer therapies administered (chemotherapy, radiotherapy, surgery, hormonal therapy), biomarkers (platelet level, white blood cell level, hemoglobin level), and cancer profiles (metastasis, staging, cancer type). Patients with incomplete data for these parameters were imputed and included in the present analysis.

Resampling techniques

To address the class imbalance in the MDCAT dataset, three resampling methods were employed using the ROSE package in R version 4.3.1. These methods included random oversampling (ROS), random undersampling (RUS), and a mixed approach (BOTH). Oversampling involved randomly replicating observations from the minority class CAT to balance class proportions, resulting in a total of 698 samples, comprising 256 CAT and 442 NONCAT instances. Conversely, undersampling involved randomly removing observations from the majority class (NONCAT) to achieve a balanced dataset, yielding 292 samples with an equal distribution of 146 CAT and 146 NONCAT cases. The mixed method (BOTH) combined oversampling of the minority class with undersampling of the majority class to reach a target of 500 samples, effectively balancing the dataset for subsequent analysis and model training.

Classification

The four samples (ROS, RUS, BOTH, and Original) were independently split into training (80%) and testing (20%) sets. LR and eXtreme Gradient Boost (XGBoost) classifiers were applied using the glmnet and xgboost packages in R version 4.3.1. Each training dataset underwent 5-fold cross-validation, where four folds were used for model training and one for validation, repeated across all folds. LR is commonly employed as a benchmark model for assessing the performance of other classifiers in CAT studies [8,9]. It is becoming popular because the LR slope coefficient can be conveniently interpreted as an odds ratio (OR) [12]. XGBoost was chosen as one of the algorithm because it demonstrated the effectiveness in CAT prediction as comprehensively reviewed by Leevy *et al.* [11]. LR and XGBoost classifiers were first trained using default settings and subsequently optimized using grid-based hyperparameter tuning. Model performance for each classifier–resampling combination was evaluated internally through cross-validation and externally on the test set to assess generalizability. Hyperparameters were tuned using grid search to avoid overfitting. We changed two important hyperparameters for LR: lambda to 0.0001 and alpha to 0. We adjusted a number of critical hyperparameters for the XGBoost model: number of boosting iterations set at 100, maximum depth of a tree at 9, and learning rate was 0.1. We also adjusted gamma to 1, fraction of features used per tree to 0.8, minimum sum of instance weights needed in a child to 1, and subsample fraction of samples used for training to 1. Our experiment also found similar model performances for 5-fold and 10-fold cross-validation. Therefore, 5-fold cross-validation process was performed as it is computationally more efficient for each set of hyperparameters, and the test set was used to evaluate the performance of the tuned hyperparameters [12].

Assessment and comparison of classification model performances

The performance of the classification models was assessed using several key metrics: The Area Under the Receiver Operating Characteristic Curve (AUC), which evaluates the model's ability to distinguish between the two classes; Accuracy, which measures the proportion of correctly classified instances among all instances; Sensitivity (Recall), which indicates the model's capability to correctly identify actual positives; and Specificity, which represents the model's accuracy in identifying actual negatives. These metrics were compared to highlight the impact of imbalanced data and resampling methods on classification outcomes, providing valuable insights into the models' performance.

RESULTS

Effect of imbalanced data on classifier's performance

In handling imbalanced data, two classifiers are compared, and performance measures for the original dataset using LR and XGBoost are shown in Table 1. A substantial decline in XGBoost performance was observed, with the AUC decreasing from 0.794 in training to 0.381 in testing, indicating substantial overfitting. This pattern is consistent with the large reductions in sensitivity (94.4–57.1%) and specificity (89.7–78.5%) between training and testing. In contrast, LR showed more stable performance, with a higher test AUC (0.707) and smaller discrepancies across performance metrics. These findings suggest the need for resampling techniques to address data imbalance and improve model performances.

Table 1: Performance of logistic regression and XGBOOST on imbalanced dataset

Performance metrics	XGBoost (%)	LR (%)
Accuracy		
Training	90.8	79.3
Testing	74.0	74.0
Sensitivity		
Training	94.4	67.3
Testing	57.1	55.5
Specificity		
Training	89.7	83.2
Testing	78.5	80.8
AUC		
Training	0.7944	0.8546
Testing	0.3809	0.7067

LR: Logistic regression, AUC: Area under curve, XGBoost: eXtreme gradient boost

Significant variables pattern in different resampling

Table 2 shows the pattern of significant variables across different resampling techniques applied to LR. Metastasis, surgery, and Indian ethnicity showed a consistently elevated and statistically significant association with CAT event across all resampling techniques, depicting both clinical plausibility and methodological stability. Diabetes, sepsis, smoker and radiology demonstrated moderate but meaningful associations, although their significance fluctuated in some resampling techniques, indicating reduced stability. ORs and confidence intervals (CIs) were compared across resampling techniques to assess the stability of variable effect. However, these measures were not used to determine the best resampling method because differences in OR and CI width reflect changes in class distribution rather than true biological effects.

Performance of LR and XGBoost with three resampling techniques

Table 3 represents the performance of classifiers using default parameters, while Table 4 represents the performance after hyperparameter tuning of LR and XGBoost. Across all models, the resampling strategy affecting model's predictive performance, particularly the degree of overfitting. XGBoost with oversampling (XO) exhibited excellent training performance (Accuracy 0.99; AUC 0.98) but showed a large performance drop on the test set (Accuracy 0.82; AUC 0.72). XGBoost with RUS produced lower training accuracy and the lowest test specificity (0.62), while mixed resampling (XB) achieved a more balanced profile, with moderate training accuracy (0.98) and improved test AUC (0.77) compared with XO and XU. Among LR models, (LU) yielded the highest training accuracy (0.83) and AUC of 0.82), while oversampling (LO) improved test specificity (0.84). The mixed resampling logistic model (LB) displayed the lowest test AUC (0.67), indicating limited benefit from the mixed resampling approach.

Hyperparameters tuning were done using the grid search method. Table 4 presents performances of model after hyperparameter tuning. Tuning substantially improved model's performance, particularly for XGBoost, which initially showed severe overfitting on the imbalanced dataset and also on resampling datasets, but after tuning the hyperparameters, models achieved better performance across all resampling techniques. Tuned XGBoost demonstrated the strongest interaction with data balancing. Oversampling increased minority-class representation and improved test AUC to 0.83, but the mixed approach (XB tuned) performed best overall, achieving the highest test AUC (0.87) with balanced sensitivity and specificity. Undersampling remained less effective even after tuning, likely due to the loss of majority-class information. LR, in contrast, showed smaller gains from tuning. Its performance remained relatively stable across resampling techniques, reflecting the model's lower sensitivity to hyperparameter optimization. Among LR combinations, undersampling (LU tuned) yielded the highest test AUC (0.77) and improved sensitivity, though at some cost to specificity.

Overall, tuning amplified the differences between resampling strategies and highlighted clear classifier–resampling interactions. XGBoost benefited most, particularly when trained on mixed or oversampled datasets, while LR remained comparatively stable. These findings showed the importance of understanding how classifiers and resampling techniques interact when developing models for imbalanced clinical data.

DISCUSSION

ML approaches are increasingly utilized in developing clinical prediction models for CAT across various cancer types, including hospitalized cancer patients [3], lung cancer patients [2], and gastric cancer patients [14]. One common challenge in medical data, particularly in cancer datasets, is the issue of imbalanced data [15]. Achieving a balanced dataset for VTE across all cancer subtypes is difficult due to its relatively low incidence rate of 1.67–12.6% [16] and the complex, multifactorial nature of cancer [17]. Although, most

Table 2: The pattern of significant variables for LR

Variables	Original	ROS	RUS	BOTH
Metastasize				
Odds ratio	4.11**	3.19**	7.14**	3.59**
95% CI	(1.8,9.4)	(1.6, 6.3)	(1.9,27.2)	(1.1.6.2)
Surgery				
Odds ratio	0.32**	0.44**	0.11**	0.17**
95% CI	(0.12,0.72)	(0.2, 0.8)	(0.03,0.43)	(0.1, 0.46)
Ethnic [13]				
Odds ratio	0.16**	0.13**	0.18*	0.16**
95% CI	(0.05,0.56)	(0.05, 0.34)	(0.03,0.95)	(0.05,0.52)
Diabetes				
Odds ratio	2.57**	2.59**	1.91	3.26**
95% CI	(1.28,5.15)	(1.5, 4.47)	(0.65, 5.55)	(1.58,6.74)
Sepsis				
Odds ratio	2.92**	1.76	2.06	2.47*
95% CI	(1.36, 6.29)	(0.91, 3.40)	(0.64,6.61)	(1.07, 5.66)
Smoker				
Odds ratio	7.95**	4.49**	1.76	1.83
95% CI	(2.25, 28.07)	(1.98, 10.19)	(0.38,8.20)	(0.66,5.09)
Radiology				
Odds ratio	0.47*	0.60	0.76	0.34*
95% CI	(0.22,0.99)	(0.34, 1.05)	(0.25,2.32)	(0.16,0.72)
Cancertype 2				
Odds ratio	3.98	2.79	22.72*	6.82*
95% CI	(0.94,16.85)	(0.90,8.62)	**(1.66, 311.07)**[#]	**(1.42, 32.95)**
Cancertype 4				
Odds ratio	0.39	0.24*	3.31	0.11*
95% CI	(0.09,1.81)	(0.07,0.82)	(**0.34,32.64**)[#]	(0.19,0.61)
Airway disease				
Odds ratio	0.05*	0.14*	4.93	1.21
95% CI	(0.003,0.76)	(0.02, 0.88)	(**1.10,22.19**)[#]	(0.15, 9.45)

*p<0.05; **p<0.01. ROS: Random oversampling, RUS: Random undersampling, BOTH: Mixed method (oversampling+undersampling), CI: Confidence interval.

[#] Cancer type 2 with significant high odds ratio of cancer-associated thrombosis but showed wide confidence interval likely due to effect of random undersampling dataset having small number of events, within this cancer subgroup.

Table 3: Logistic regression and XGBoost on different resampling techniques (Default model)

Performance metrics	XO	XU	XB	LO	LU	LB
Accuracy						
Training	0.9876	0.7821	0.9751	0.8032	0.8333	0.8454
Testing	0.8156	0.7069	0.8081	0.7660	0.7414	0.7071
Sensitivity						
Training	0.9789	0.8034	0.9694	0.7782	0.8205	0.8214
Testing	0.7183	0.7931	0.7708	0.6901	0.7931	0.6667
Specificity						
Training	0.9964	0.7607	0.9805	0.8286	0.8462	0.8683
Testing	0.9143	0.6207	0.8431	0.8426	0.6897	0.7451
AUC						
Training	0.9789	0.8034	0.9694	0.7782	0.8205	0.8214
Testing	0.7183	0.7931	0.7707	0.6901	0.7931	0.6667

XO: XGBoost+oversampling, XU: XGboost+undersampling, XB: XGBoost+mixed approach, LO: Logistic+oversampling, LU: Logistic+undersampling, LB: Logistic+mixed approach

Table 4: LR and XGBoost on different resampling techniques (tuned model)

Performance metrics	XO tuned	XU tuned	XB tuned	LO tuned	LU tuned	LB tuned
Accuracy						
Training	0.9947	0.8291	0.9825	0.7908	0.7821	0.8279
Testing	0.8723	0.6897	0.7778	0.7801	0.7414	0.6970
Sensitivity						
Training	0.9930	0.8120	0.9847	0.7711	0.8034	0.8112
Testing	0.8310	0.7241	0.7708	0.7042	0.8276	0.7083
Specificity						
Training	0.9964	0.8462	0.9805	0.8107	0.7607	0.8439
Testing	0.9143	0.6552	0.7843	0.8571	0.6552	0.6863
AUC						
Training	0.9923	0.8120	0.9847	0.7711	0.8073	0.8112
Testing	0.8310	0.7241	0.8701	0.7042	0.7679	0.7083

XO: XGBoost+Oversampling, XU: XGBoost+undersampling, XB: XGBoost+Mixed approach, LO: Logistic+oversampling, LU: Logistic+undersampling, LB: Logistic+mixed approach

study in CAT prediction did not mention how the imbalanced data was handled [10,17], it can significantly impact the ability to accurately distinguish between cancer patients with and without VTE risk. Thus, addressing imbalanced data is a critical preprocessing step for developing effective CAT prediction models.

While hybrid algorithmic approaches may offer potential advantages [10], resampling methods remain popular due to their simplicity, ease of implementation, and performed well on imbalanced medical dataset [18]. In our study, the dataset's IR was 2.43, almost the same as IR for the dataset used by Kim *et al*. [19]. This IR value falls within the range of 1.9–9, classifying it as mildly imbalanced according to established guidelines [20]. We employed LR and XGBoost in our experiment because these algorithms are frequently used for predicting CAT [2].

In this study, we aimed to evaluate the interaction between classifiers of choice (LR vs. XGBoost) and data-level resampling techniques (ROS, RUS, BOTH) for improving prediction performance on the imbalanced MDCAT dataset. Imbalanced datasets, by their nature, tend to favor the majority class, which can lead to skewed interpretations of performance metrics like accuracy, especially in cases where the minority class, often the focus of clinical interest, is underrepresented.

The results clearly demonstrate that the performance of each classifier is dependent on the resampling strategy applied, demonstrating an important interaction effect between model architecture and class-imbalance handling. When trained on the original imbalanced dataset (Table 1), XGBoost exhibited a strong tendency toward overfitting, achieving high training performance (AUC 0.79; sensitivity 94.4%) but suffering a dramatic collapse in test AUC to 0.38. XGBoost, albeit often favored for its flexibility and ability to model complex interactions [18], showed a marked decline in performance when applied to the test set, indicating potential overfitting. This severe train–test divergence suggests that the boosting algorithm amplified the majority-class patterns, which in our dataset constituted 71% of the data while undermined its effectiveness in capturing minority class cases.

LR, in contrast, showed more stable behavior on the original dataset, with a moderate training AUC of 0.85 and substantially better preservation of test performance (AUC=0.71), despite of lower sensitivity. This stability is expected given the linear and less flexible nature of LR, which limits its susceptibility to overfitting under imbalance. Despite of stability, interpretability (ORs), and lower computational cost if LR model, it may have a lower ceiling than a well-regularized XGBoost model.

The use of resampling techniques in LR highlights the critical interaction between resampling method and variables behavior. Resampling not only influences overall predictive performance but also affects the interpretability of individual predictors. Variables that retain consistent effect estimates across diverse analytic conditions, such as metastasis, diabetes, and ethnicity, are more likely to reflect true clinical risk factors. Conversely, variables with large ORs but wide CIs, such as cancer type, should be interpreted with caution, as they likely arise from data sparsity or model instability rather than genuine clinical effects. This analysis demonstrates the importance of evaluating predictor stability across modeling strategies, especially in imbalanced clinical datasets. This approach enhances the reliability of variable interpretation and prevents overreliance on statistically unstable predictors that may distort clinical conclusions.

When resampling methods were applied, the magnitude and direction of improvement differed between the classifiers, highlighting their unique interactions with different resampling techniques. XGBoost benefited substantially from resampling, particularly mixed approaches, which reduced overfitting and improved test sensitivity and AUC. Oversampling alone increased sensitivity but tended to expose XGBoost to overfitting, while undersampling improved generalization but introduced volatility by removing informative majority-class instances. LR, on the other hand, demonstrated more predictable and incremental gains from resampling. Oversampling and mixed approaches generally improved sensitivity without excessively compromising specificity, while undersampling sometimes reduced stability due to loss of data, although it occasionally improved minority-class detection.

Taken together, these findings indicate that XGBoost requires active imbalance mitigation, as its intrinsic complexity makes it highly responsive positively, as what was found in Xu's work [14] to resampling strategies. Meanwhile, LR is less sensitive to imbalance but still benefits from resampling, especially when improved minority-class detection is clinically important. The observed interaction emphasizes that the optimal resampling method is not universal but model-dependent: boosting algorithms demand careful control of class distributions, whereas simpler linear models may tolerate moderate imbalance without catastrophic degradation. This underscores the importance of evaluating classifier–resampling combinations rather than treating resampling as a purely preprocessing step.

Resampling methods such as ROS, RUS, and BOTH were effective in improving model accuracy, recall, and specificity as compared to imbalanced data. These findings align with existing literature that recommends resampling as a robust strategy for handling imbalanced datasets [21]. Specifically, applying these techniques led to significant improvements in AUC if coupled with properly regularized LR and XGBoost classifiers.

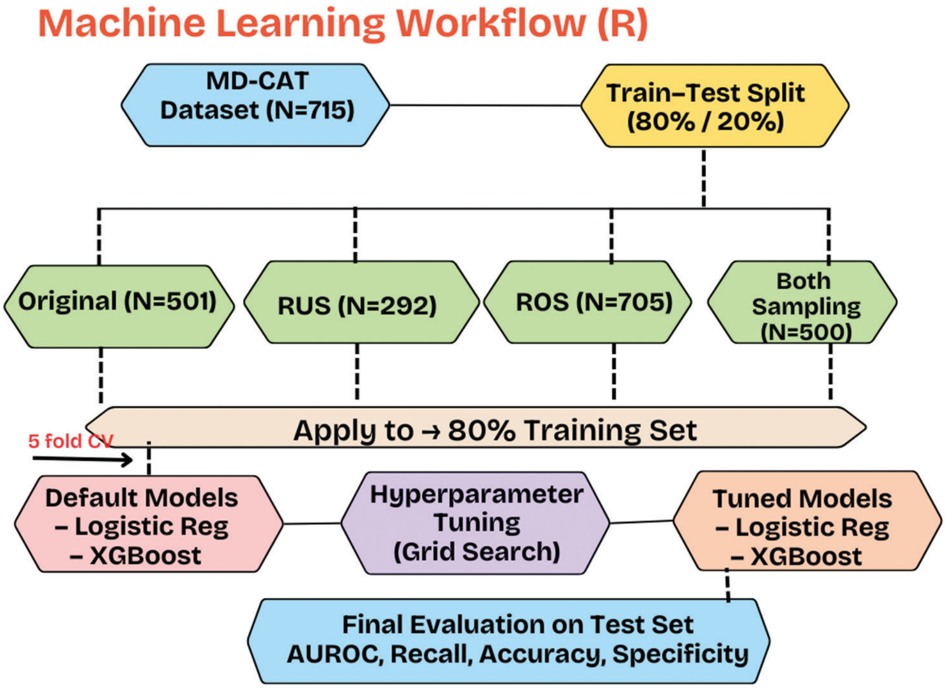
Machine Learning Workflow (R)
MD-CAT Dataset (N=715)
Train–Test Split (80% / 20%)
Original (N=501)
RUS (N=292)
ROS (N=705)
Both Sampling (N=500)
Apply to → 80% Training Set
5 fold CV
Default Models – Logistic Reg – XGBoost
Hyperparameter Tuning (Grid Search)
Tuned Models – Logistic Reg – XGBoost
Final Evaluation on Test Set AUROC, Recall, Accuracy, Specificity

Fig. 1: Methodological flow

Notwithstanding the advantages of resampling, certain limitations were observed. Resampling was performed before splitting the data; some degree of information leakage may have occurred, potentially inflating performance estimates. This limitation should be considered when interpreting the findings. Potential instability of RUS can lead to a loss of valuable data during the down-sampling process, negatively impacting model performance. On the other hand, ROS may result in overfitting due to the duplication of instances.

There are shortcomings in our study. First, only LR and XGBoost were employed in this experiment, even though other algorithms such as Support Vector Machine and Random Forest were commonly used in the prediction of CAT. Second, only the resampling technique for the data level method was applied to handle the imbalanced dataset. Therefore, future research on the CAT dataset should explore more diverse balancing strategies and employ more classifiers to compare the best performance among all the combination strategies in handling imbalanced data.

CONCLUSION

This study demonstrates that the effectiveness of prediction models in the imbalanced MDCAT dataset is strongly influenced by the interaction between classifier characteristics and resampling strategies. XGBoost on the original dataset showed severe overfitting, evidenced by its large train–test performance gap and dramatic drop in test AUC, indicating inadequate learning of minority-class patterns. LR displayed more stable baseline performance but limited sensitivity to minority outcomes. For MDCAT dataset, the performance gains from a tuned XGBoost model with mixed resampling outweighed the benefits of LR's simplicity and stability, making it our recommended approach given the primary importance of AUC. The application of resampling methods, particularly mixed approaches, substantially improved minority-class detection, with each classifier responding differently depending on the resampling technique employed. These findings highlight the necessity of evaluating resampling methods in conjunction with specific classifiers rather than applying them as generic preprocessing steps. For clinical applications where minority class detection is crucial, selecting an appropriate classifier–resampling combination is essential to achieving reliable and generalizable predictive performance.

ACKNOWLEDGMENTS

The authors would like to thank to the dedicated staff of the Radiology and Oncology Medical Record Unit Hospital Kuala Lumpur, for their endless helps and support during the study.

AUTHOR CONTRIBUTIONS

F.N.: Conceptualization, Methodology, Investigation, Formal analysis, Data curation, Writing-original draft. N.K.W.: Validation, Data Curation, Writing – Review and Editing. M.L.F.: Supervision. N.H.A.J.: Software, Resources, Writing – Review and Editing. K.A.M.I.: Supervision. Y.M.Y.: Software, Resources, Writing – Review and Editing.

CONFLICTS OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

ETHICAL STATEMENT

The study was reviewed by the Medical Research Ethics Committee of the Faculty of Pharmacy at Universiti Teknologi Mara, Malaysia UiTM - REC/04/2020 (UG/MR/133). Permission to do research using the Ministry of Health Facilities was obtained from the National Medical Research Register (NMRR-19-4016-46553 (IIR)), and approval for field data collection in Hospital Kuala Lumpur was obtained from CRC HKL.

REFERENCES

1. Falanga A, Marchetti M, Russo L. The mechanisms of cancer-associated thrombosis. Thromb Res. 2015;135 Suppl 1:S8-11. doi: 10.1016/s0049-3848(15)50432-5, PMID 25903541
2. Lei H, Zhang M, Wu Z, Liu C, Li X, Zhou W, *et al*. Development and validation of a risk prediction model for venous thromboembolism in

lung cancer patients using machine learning. Front Cardiovasc Med. 2022;9:845210. doi: 10.3389/fcvm.2022.845210, PMID 35321110

3. Meng L, Wei T, Fan R, Su H, Liu J, Wang L, *et al*. Development and validation of a machine learning model to predict venous thromboembolism among hospitalized cancer patients. Asia Pac J Oncol Nurs. 2022;9(12):100128. doi: 10.1016/j.apjon.2022.100128, PMID 36276886
4. Javaid M, Haleem A, Singh RP, Suman R, Rab S. Significance of machine learning in healthcare: Features, pillars and applications. Int J Intell Netw. 2022;3:58-73. doi: 10.1016/j.ijin.2022.05.002
5. Mahadevappa MK, Krishnan GN, Murthannagari VR, Arun J. Harnessing artificial intelligence: Transforming clinical trials for the future. Int J Appl Pharm. 2025;17:102-10. doi: 10.22159/ijap.2025v17i6.54181
6. Lafi Z, Matalqah S, Asha S, Asha N, Mhaidat H, Asha SY. Advanced fabrication and characterization of silver nanoparticles using AI techniques. Int J Appl Pharm. 2025;17:42-51. doi: 10.22159/ijap.2025v17i5.55011
7. Mohamed MM, Jusril NA, Adenan MI, Wen NG. *In silico* identification of APOBEC3B small molecule inhibitors from DTP-NCI libraries. Int J Appl Pharm. 2021;13:165-70. doi: 10.22159/ijap.2021v13i3.41600
8. Pabinger I, Van Es N, Heinze G, Posch F, Riedl J, Reitter EM, *et al*. A clinical prediction model for cancer-associated venous thromboembolism: A development and validation study in two independent prospective cohorts. Lancet Haematol. 2018;5(7):e289-98. doi: 10.1016/s2352-3026(18)30063-2, PMID 29885940
9. Moik F, Englisch C, Pabinger I, Ay C. Risk assessment models of cancer-associated thrombosis - potentials and perspectives. Thromb Update. 2021;5:100075. doi: 10.1016/j.tru.2021.100075
10. Kaur H, Pannu HS, Malhi AK. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Comput Surv. 2019;52:79. doi: 10.1145/3343440
11. Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N. A survey on addressing high-class imbalance in big data. J Big Data. 2018;5(1):42. doi: 10.1186/s40537-018-0151-6
12. Schober P, Vetter TR. Nonparametric statistical methods in medical research. Anesth Analg. 2020;131(6):1862-3. doi: 10.1213/ane.0000000000005101, PMID 33186171
13. Rilianto B, Kurniawan RG, Prasetyo BT, Windiani PR, Gotama KT, Kusdiansah M, *et al*. Risk factors of cerebral aneurysms rupture in an Indonesian population. Neurol Res. 2024;46(11):989-95. doi: 10.1080/01616412.2024.2376308, PMID 38971160
14. Xu Q, Lei H, Li X, Li F, Shi H, Wang G, *et al*. Machine learning predicts cancer-associated venous thromboembolism using clinically available variables in gastric cancer patients. Heliyon. 2023;9(1):e12681. doi: 10.1016/j.heliyon.2022.e12681, PMID 36632097
15. Tasci E, Zhuge Y, Camphausen K, Krauze AV. Bias and class imbalance in oncologic data - towards inclusive and transferrable AI in large scale oncology data sets. Cancers (Basel). 2022;14(12):2897. doi: 10.3390/cancers14122897, PMID 35740563
16. Angchaisuksiri P. Cancer-associated thrombosis in Asia. Thromb J. 2016;14 Suppl 1:26. doi: 10.1186/s12959-016-0110-4, PMID 27766052
17. Wan ML, Wang Y, Zeng Z, Deng B, Zhu BS, Cao T, *et al*. Colorectal cancer (CRC) as a multifactorial disease and its causal correlations with multiple signaling pathways. Biosci Rep. 2020;40(3):BSR20200265. doi: 10.1042/bsr20200265, PMID 32149326
18. Montomoli J, Romeo L, Moccia S, Bernardini M, Migliorelli L, Berardini D, *et al*. Machine learning using the extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of SOFA score at ICU admission in COVID-19 patients. J Intensive Med. 2021;1(2):110-6. doi: 10.1016/j.jointm.2021.09.002, PMID 36785563
19. Kim JS, Kwon D, Kim K, Lee SH, Lee SB, Kim K, *et al*. Machine learning-based prediction of pulmonary embolism to reduce unnecessary computed tomography scans in gastrointestinal cancer patients: a retrospective multicenter study. Sci Rep. 2024;14(1):25359. doi:10.1038/s41598-024-75977-y
20. Noorhalim N, Ali A, Shamsuddin SM. Handling Imbalanced Ratio for Class Imbalance Problem using SMOTE. In: Proceedings of the Third International Conference on Computing, Mathematics and Statistics; 2019. doi: 10.1007/978-981-13-7279-7_3
21. Rahman HA, Wah YB, Huat OS. Predictive performance of logistic regression for imbalanced data with categorical covariate. Pertanika J Sci Technol. 2020;29:1-10. doi: 10.47836/pjst.29.1.10