Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Boundary-Stabilized Oversampling for Interpretable Clinical Prediction in Imbalanced and Data-Scarce Regimes

Seong-Hyun Yu
Sungkyunkwan University
Minha Kim
Seoul National University
Inhyung Lee
Seoul National University
Won-Gyun Son
Seoul National University
Ji-Hoon Jeong
jihoon.jeong@skku.edu

Sungkyunkwan University

Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11425847/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Boundary-Stabilized Oversampling for Interpretable Clinical Prediction in Imbalanced and Data-Scarce Regimes

Seong-Hyun Yu[1,4,+], **Minha Kim**[2,+], **Inhyung Lee**[2], **Won-Gyun Son**[2,*], **and Ji-Hoon Jeong**[1,3,4,*]

[1]Sungkyunkwan University, Samsung Advanced Institute for Health Sciences & Technology (SAIHST), Dept. of Medical Device Management and Research, Seoul, 06355, Republic of Korea
[2]Seoul National University, Dept. Veterinary Clinical Sciences, College of Veterinary Medicine, Seoul, 08826, Republic of Korea
[3]Samsung Medical Center, Research Institute for Future Medicine, Seoul, 06355, Republic of Korea
[4]Sungkyunkwan University, Medical Imaging Insight Institute, Seoul, 06355, Republic of Korea
[*]corresponding. carpeego@snu.ac.kr, jihoon.jeong@skku.edu
[+]these authors contributed equally to this work

ABSTRACT

Predicting rare adverse events in clinical settings is complicated by extreme class imbalance, severe data scarcity, and the need for physiologically interpretable explanations. We propose Boundary-Stabilized SMOTE (BS-SMOTE), a domain-constrained oversampling method extending standard SMOTE with two constraints: feature-type aware generation applying distinct interpolation rules for continuous, binary, and categorical variables, and physiological feasibility projection clipping synthetic samples within clinically plausible bounds. To prevent data leakage, BS-SMOTE is applied strictly within each training fold of a stratified 5-fold nested cross-validation. The framework was applied to peri-anesthetic hypoxemia prediction in veterinary anesthesia using 138 dogs with 15 confirmed hypoxemic cases (imbalance ratio 1:8.2). A 10-way ablation study demonstrated that BS-SMOTE with fine-tuned GradientBoosting consistently outperformed standard SMOTE, CTGAN, and VAE-based augmentation (F1 $= 0.307$, AUC $= 0.728$, Cohen's $\kappa = 0.256$). Robustness across 20 random seeds yielded stable discriminative performance (AUC $= 0.751 \pm 0.037$). An augmentation ratio sweep from $0\times$ to $8\times$ confirmed monotonically improving class separability. SHAP analyses restricted to held-out original data identified respiratory compliance, age, and anesthesia duration as dominant risk contributors, with compliance exhibiting a nonlinear threshold effect consistent with veterinary respiratory physiology. BS-SMOTE provides a rigorous and interpretable framework for rare clinical event prediction, directly transferable to other data-scarce imbalanced settings.

Introduction

Rare adverse events in clinical medicine present a fundamental analytical challenge because of their very rarity renders them disproportionately difficult to study, yet disproportionately consequential when they occur. Peri-anesthetic respiratory complications exemplify this paradox. In veterinary anesthesia, arterial blood gas-confirmed post-extubation hypoxemia represents a clinically critical outcome that arises from complex and interdependent interactions among anatomical susceptibility, respiratory mechanics, anesthetic drug selection, and perioperative physiological states[1–3]. Unlike pulse oximetry-detected transient desaturation, which is relatively common and often self-limiting, ABG-confirmed hypoxemia under standardized anesthetic protocols represents a rare but clinically meaningful respiratory insult associated with delayed recovery, cardiovascular instability, and adverse neurological outcomes[4–6]. Despite its clinical importance, this rarity places the problem squarely within a data-limited scientific regime in which both total case counts and adverse event frequencies are intrinsically constrained by the nature of the phenomenon itself rather than by limitations of data collection[7]. In our cohort, only 15 of 138 cases exhibited confirmed post-extubation hypoxemia, underscoring the intrinsic scarcity of positive samples and the severe class imbalance that defines this clinical domain.

Machine learning approaches have increasingly been explored to improve peri-anesthetic risk prediction in both human and veterinary medicine[8–13]. These models offer the potential to identify complex nonlinear patterns that elude conventional statistical approaches, particularly in settings characterized by heterogeneous clinical variables and multi-factorial outcome mechanisms[14–16]. However, their application to rare adverse event settings is systematically undermined by two interrelated methodological problems that have received insufficient attention in the clinical machine learning literature. The first is severe

class imbalance between normal and complication cases, which causes standard supervised models to exhibit strong majority-class bias, producing high aggregate accuracy while failing to detect the minority class of primary clinical interest[17–20]. Synthetic oversampling techniques have been widely proposed to mitigate this imbalance, with the Synthetic Minority Oversampling Technique (SMOTE)[21] representing the most commonly applied approach. More recently, conditional generative adversarial networks (CTGAN)[22] and variational autoencoders (VAE)[23] have been applied to tabular clinical data as alternatives capable of modeling complex multivariate distributions. Despite their utility, all three approaches share a fundamental limitation in clinical settings: they do not differentiate between feature types during synthetic sample generation. Generative approaches such as CTGAN and VAE introduce additional model complexity without guaranteeing clinical plausibility, and may produce diffuse synthetic distributions that blur class boundaries in sparsely represented minority-class regions of the feature space[24–26].

The second and more pervasive methodological problem is data leakage introduced by the application of synthetic oversampling prior to train-test splitting. This practice, which is prevalent in the clinical machine learning literature, allows the model to indirectly observe test set distributional characteristics during training through the synthetic samples generated from pooled data[27]. The consequence is systematically inflated performance estimates that do not reflect true generalization to unseen clinical cases[28]. In data-scarce rare-event settings, where minority class samples are few and test set composition is highly sensitive to the specific split, the magnitude of this leakage effect is amplified relative to larger and more balanced datasets. Despite its implications for reproducibility and clinical validity, this issue is rarely acknowledged or formally corrected in published studies, and its interaction with augmentation strategy and sample size has not been systematically characterized.

Beyond prediction accuracy, there is growing recognition in the clinical artificial intelligence community that interpretability is not merely a desirable supplement but a methodological necessity in data-scarce regimes, where model outputs must be reconcilable with established domain knowledge to support hypothesis generation and clinical trust[8]. Shapley Additive Explanations (SHAP) have emerged as a principled and theoretically grounded framework for quantifying individual feature contributions to model predictions across diverse model architectures[29]. However, a critical and frequently overlooked concern arises when SHAP is applied in the context of synthetic augmentation: attribution patterns derived from synthetic samples may not reflect genuine physiological relationships present in real clinical data, potentially introducing optimistic or artifactual explanatory bias. Furthermore, in small-sample settings, the stability of SHAP-based feature rankings across different data partitions is rarely assessed, leaving open the question of whether identified risk patterns are reproducible or partition-specific artifacts. Restricting all interpretability analyses exclusively to held-out original clinical data, and formally quantifying ranking stability across independent partitions, is therefore essential for ensuring that SHAP-derived insights reflect generalizable physiological relationships rather than augmentation artifacts[30–33].

To address these limitations in a unified and principled manner, this study proposes Boundary-Stabilized SMOTE (BS-SMOTE), a domain-constrained oversampling framework that extends standard SMOTE with two mathematically defined constraints: feature-type aware generation that applies distinct interpolation rules for continuous, binary, and categorical variables, and physiological feasibility projection that clips synthetic samples within clinically plausible bounds derived from training data. BS-SMOTE is embedded within a leakage-free stratified nested cross-validation protocol in which augmentation is applied strictly within each training fold, ensuring complete isolation of the test set. The predictive component employs imbalance-optimized gradient boosting with systematic hyperparameter tuning via inner cross-validation[34]. Interpretability analyses are conducted exclusively on held-out original clinical data, with SHAP feature ranking stability formally quantified via Kendall's τ across 20 independent random seeds and 100 cross-validation runs. The complete framework is evaluated through a comprehensive 10-way ablation study on peri-anesthetic hypoxemia prediction in veterinary anesthesia, a domain that exemplifies the structural challenges of rare clinical event prediction and where physiologically coherent risk stratification has direct clinical implications. The contributions of this work are threefold: (i) a novel constrained augmentation method that preserves clinical feature-type integrity and physiological plausibility, (ii) a rigorous leakage-free nested cross-validation design with formal robustness assessment applicable to data-scarce imbalanced settings, and (iii) stable and physiologically coherent risk factor identification through conservative SHAP-based rule discovery restricted to held-out original clinical data.

Result

Dataset Characteristics

A total of 138 dogs met the final inclusion criteria following application of predefined exclusion rules. Among these, 15 dogs exhibited ABG-confirmed post-extubation hypoxemia ($PaO_2 < 80$ mmHg), yielding a class imbalance ratio of approximately 1:8.2 (Normal: $n = 123$; Abnormal: $n = 15$). The class distribution is illustrated in Figure 1a. The hypoxemic group was significantly older than the non-hypoxemic group (11.4 ± 1.9 vs. 8.4 ± 4.4 years; $p < 0.001$) and showed a higher median ASA status ($p = 0.046$), consistent with greater pre-anesthetic comorbidity burden. No significant differences were observed in body condition score ($p = 0.708$), brachycephalic breed prevalence ($p = 1.000$), respiratory sign severity ($p = 0.053$). These findings indicate that while age and anesthetic risk classification differentiated hypoxemic from non-hypoxemic cases at the

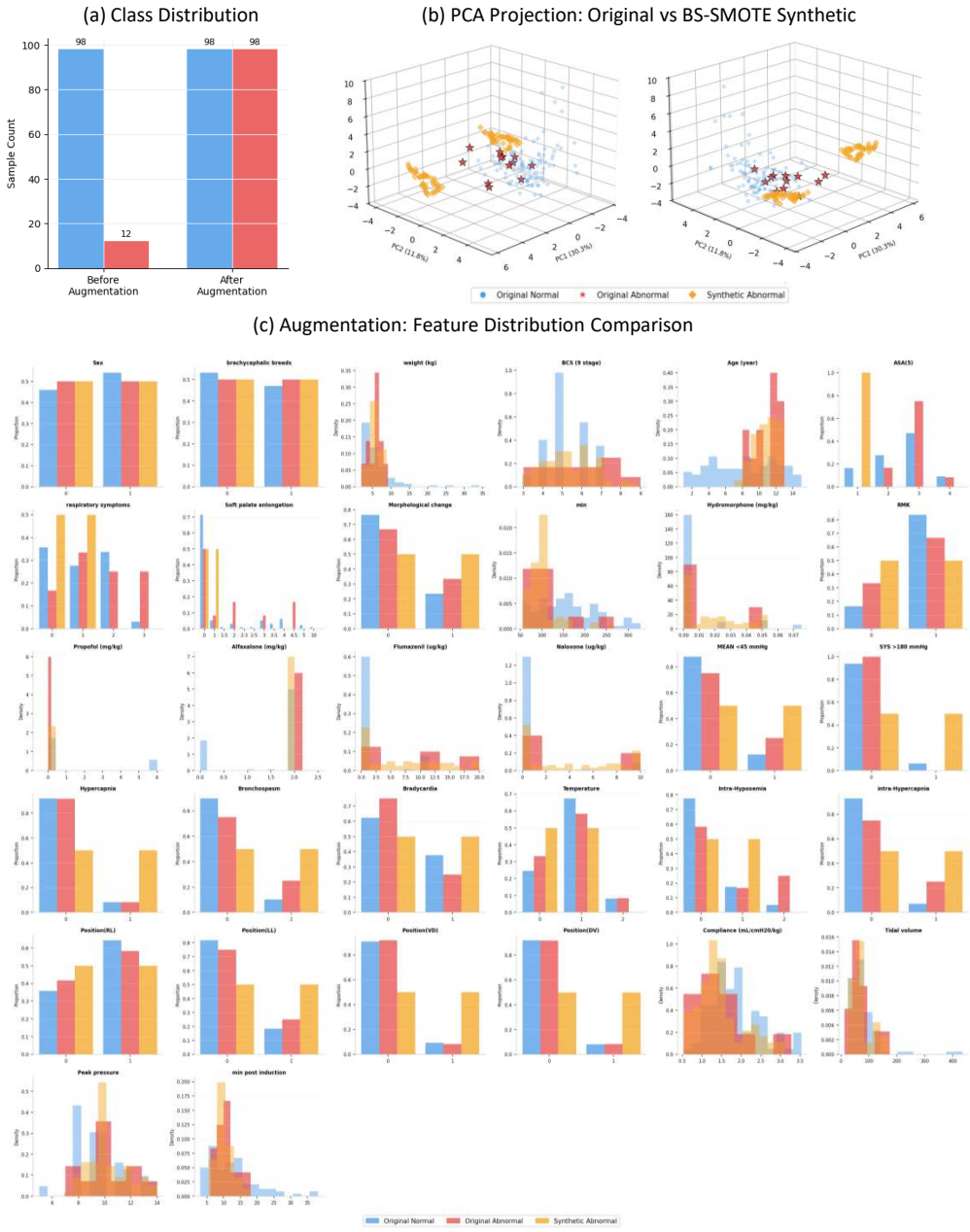


Figure 1. Dataset characteristics and BS-SMOTE augmentation analysis. (a) Class distribution of the final analytic cohort ($n = 138$), illustrating the severe class imbalance between non-hypoxemic (Normal, $n = 123$) and hypoxemic (Abnormal, $n = 15$) cases (imbalance ratio 1:8.2). (b) Principal component analysis (PCA) projection of the BS-SMOTE augmented training set from a representative fold (Fold 1), showing the spatial distribution of original normal cases (blue), original abnormal cases (red stars), and BS-SMOTE synthetic abnormal samples (orange diamonds) in the 3-dimensional PCA space. Improved class separability is evident relative to the unaugmented feature space. (c) Feature-wise distribution comparison between original normal cases, original abnormal cases, and BS-SMOTE synthetic abnormal samples across all 32 input features. Histograms represent density estimates for continuous features; bar charts represent class proportions for binary and categorical features. Synthetic samples preserve the distributional characteristics of original minority cases while expanding the minority-class boundary region.

group level, individual-level prediction remains challenging due to the substantial overlap in baseline characteristics and the extreme scarcity of positive cases.

Augmentation Ratio Sensitivity

To characterize the sensitivity of BS-SMOTE to augmentation intensity and to identify the range within which synthetic oversampling remains beneficial, a ratio sweep experiment was conducted in which the number of generated synthetic minority samples was varied from $0\times$ (no augmentation) to $8\times$ the original minority count. Results are summarized in Figure 2a.

All three performance indicators (F1 score, PR-AUC, and inter-class cluster distance in PCA-projected space) increased monotonically with augmentation ratio across the evaluated range. Relative to the unaugmented baseline ($0\times$), the $8\times$ configuration yielded improvements of $+71.9\%$ in F1 ($0.152 \rightarrow 0.261$), $+26.2\%$ in PR-AUC ($0.326 \rightarrow 0.412$), and $+53.6\%$ in inter-class cluster distance ($2.11 \rightarrow 3.24$). No evidence of performance degradation or increased variance was observed at higher augmentation ratios, suggesting that the physiological feasibility constraints imposed by BS-SMOTE effectively prevent the introduction of noise even as the proportion of synthetic samples increases. A minor non-monotonic fluctuation was observed in PR-AUC at the $2\times$ ratio, which is attributable to the inherent instability of fold composition with only 15 minority cases rather than a systematic effect of augmentation intensity. The $8\times$ augmentation ratio which approximately corresponds to the majority-to-minority class ratio of the dataset was adopted as the default configuration for all subsequent experiments, as it yielded the highest performance across all metrics.

Model Robustness

To assess the stability of the proposed framework across different data partitions, the complete BS-SMOTE with fine-tuned GradientBoosting pipeline was repeated across 20 independent random seeds, each with stratified 5-fold cross-validation, yielding 100 independent evaluation runs in total. The distribution of performance metrics across seeds is illustrated in Figure 2b.

The framework demonstrated consistent discriminative performance across all seeds. AUC exhibited the highest stability among evaluated metrics (AUC $= 0.751 \pm 0.037$, 95% CI: 0.735–0.767), with a narrow interquartile range reflecting robust class separation regardless of data partition. F1 showed moderate variance (F1 $= 0.221 \pm 0.061$, 95% CI: 0.194–0.248), attributable to the inherent sensitivity of fold composition to the exact placement of the 15 minority-class cases rather than systematic model instability. Cohen's κ similarly exhibited moderate variance ($\kappa = 0.147 \pm 0.064$, 95% CI: 0.119–0.176), with the 95% confidence interval excluding zero across all metrics, confirming that the observed performance consistently exceeded chance-level prediction. Sensitivity and specificity remained stable across seeds (Sensitivity $= 0.223 \pm 0.054$; Specificity $= 0.926 \pm 0.016$), with specificity exhibiting the lowest variance of all reported metrics, reflecting the relative abundance of majority-class cases in the test partitions.

Comparative Performance Evaluation

The contribution of each framework component was evaluated through a comprehensive 10-way ablation study comparing five augmentation (no augmentation, standard SMOTE, CTGAN, VAE, and BS-SMOTE) across fixed and fine-tuned GradientBoosting configurations.

Among augmentation strategies evaluated with fine-tuned GradientBoosting, BS-SMOTE achieved the highest performance across all primary metrics (F1 = 0.307, 95% CI: 0.016–0.598; AUC = 0.728, 95% CI: 0.655–0.801; Cohen's $\kappa = 0.256$), as illustrated in Figure 3a. Standard SMOTE yielded substantially lower minority-class detection (F1 = 0.147), consistent with the known limitation of uniform linear interpolation across heterogeneous feature types. CTGAN and VAE-based augmentation demonstrated intermediate performance, with CTGAN achieving higher specificity (0.984) at the cost of reduced sensitivity (0.133), indicating a conservative prediction tendency that limits its utility for rare adverse event detection. VAE produced similarly modest results (F1 = 0.213), suggesting that unconstrained generative approaches may blur class boundaries in sparsely represented minority-class regions.

A classifier comparison using BS-SMOTE augmentation with fine-tuned hyperparameters revealed that GradientBoosting consistently outperformed RandomForest[35] and XGBoost[31] across F1 and Cohen's κ. While RandomForest achieved the highest AUC (0.755), its F1 (0.167) and sensitivity (0.133) were substantially lower than those of GradientBoosting (F1 = 0.280, sensitivity = 0.267), indicating that AUC alone may overestimate minority-class detection capability under severe class imbalance. XGBoost demonstrated competitive F1 (0.250) but lower specificity (0.870), reflecting a higher false-positive rate that may limit clinical applicability. GradientBoosting provided the most balanced trade-off between minority-class detection and specificity preservation, supporting its selection as the primary classifier. The multi-metric radar chart in Figure 3b provides a comprehensive visual comparison of the proposed framework against representative baseline methods across five performance dimensions, confirming the balanced superiority of BS-SMOTE with fine-tuned GradientBoosting.

Figure 3c further illustrates the effect of hyperparameter fine-tuning across augmentation strategies. Fine-tuning yielded negligible or marginally negative gains in F1 for no augmentation and standard SMOTE, suggesting that decision boundary

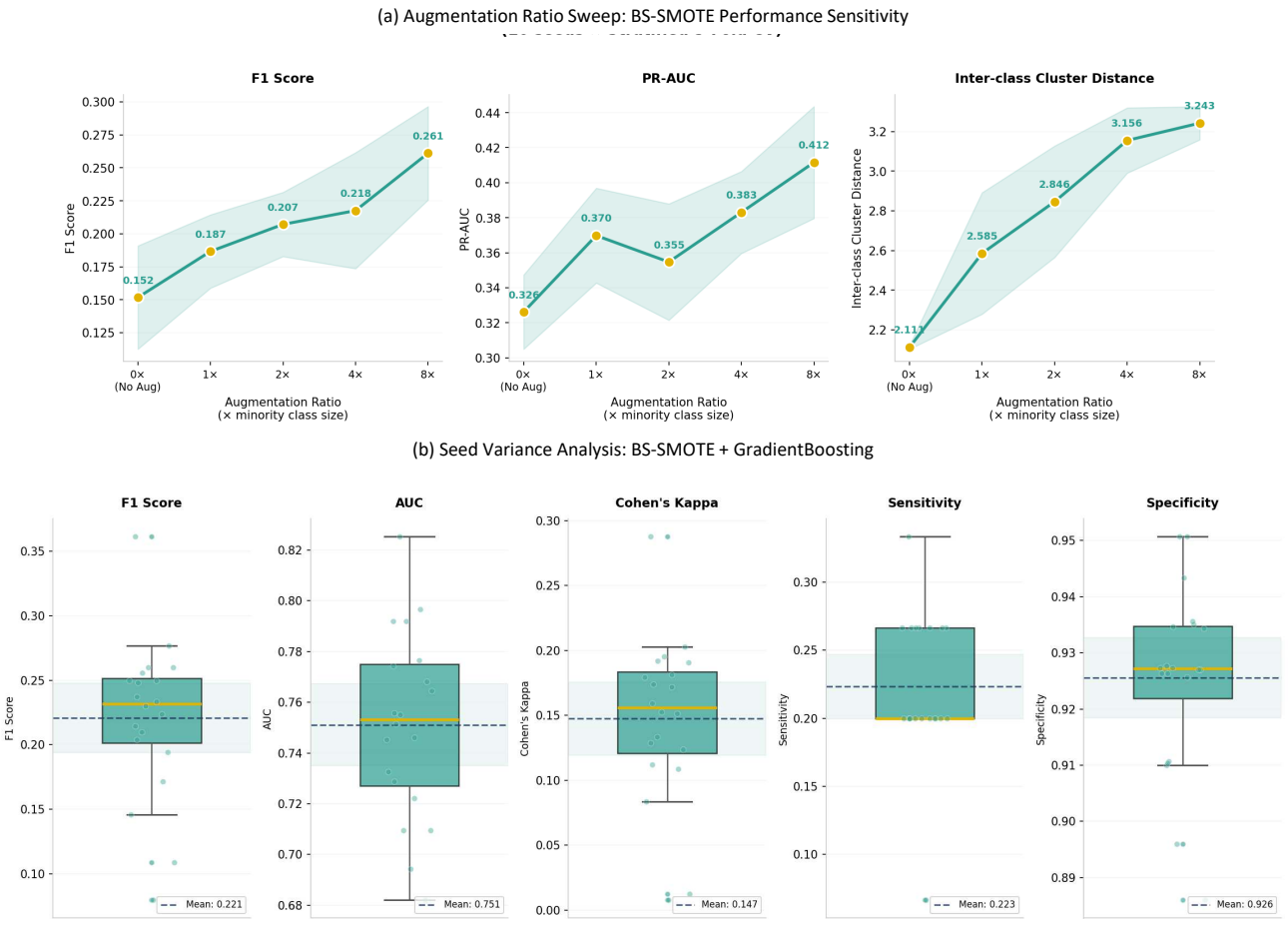


Figure 2. Augmentation ratio sweep and model robustness analysis. (a) Sensitivity of BS-SMOTE to augmentation intensity, evaluated across ratios from 0× (no augmentation) to 8× the original minority class size. Three performance indicators are reported as mean values with 95% confidence intervals across 10 independent random seeds and stratified 5-fold cross-validation. All metrics increased monotonically with augmentation ratio, with no evidence of performance degradation at higher ratios. (b) Distribution of performance metrics across 20 independent random seeds, each evaluated under stratified 5-fold cross-validation (100 runs total). Each boxplot summarizes the mean performance across 5 folds for a given seed. The dashed line indicates the mean across seeds, and the shaded region represents the 95% confidence interval. Scatter points represent individual seed means.

instability introduced by inadequate augmentation constrains the effectiveness of parameter optimization. In contrast, BS-SMOTE with fine-tuned GradientBoosting consistently outperformed its fixed-parameter counterpart across all three metrics, demonstrating that physiologically constrained augmentation and systematic hyperparameter optimization act synergistically.

SHAP-Based Physiological Rule Discovery

ggregated across 100 independent runs (20 seeds × 5 folds), identified respiratory compliance, age, and anesthesia duration as the three dominant risk contributors, as illustrated in Figure 4a. These three features exhibited substantially higher mean absolute SHAP values (0.894, 0.881, and 0.972, respectively) and narrower 95% confidence intervals compared to all remaining features, indicating that their prominence is robust to data partition variability rather than an artifact of a specific cross-validation split. The SHAP summary plot in Figure 4b further illustrates the directional contributions of individual features across all held-out test observations, confirming consistent attribution patterns for the top-ranked predictors.

SHAP dependence plots for the three dominant features, constructed using original-scale feature values aggregated across all 100 runs, are presented in Figure 4c. Respiratory compliance exhibited a nonlinear, threshold-dependent relationship with hypoxemia risk. SHAP values were positive indicating risk-increasing contributions at low compliance values, and transitioned

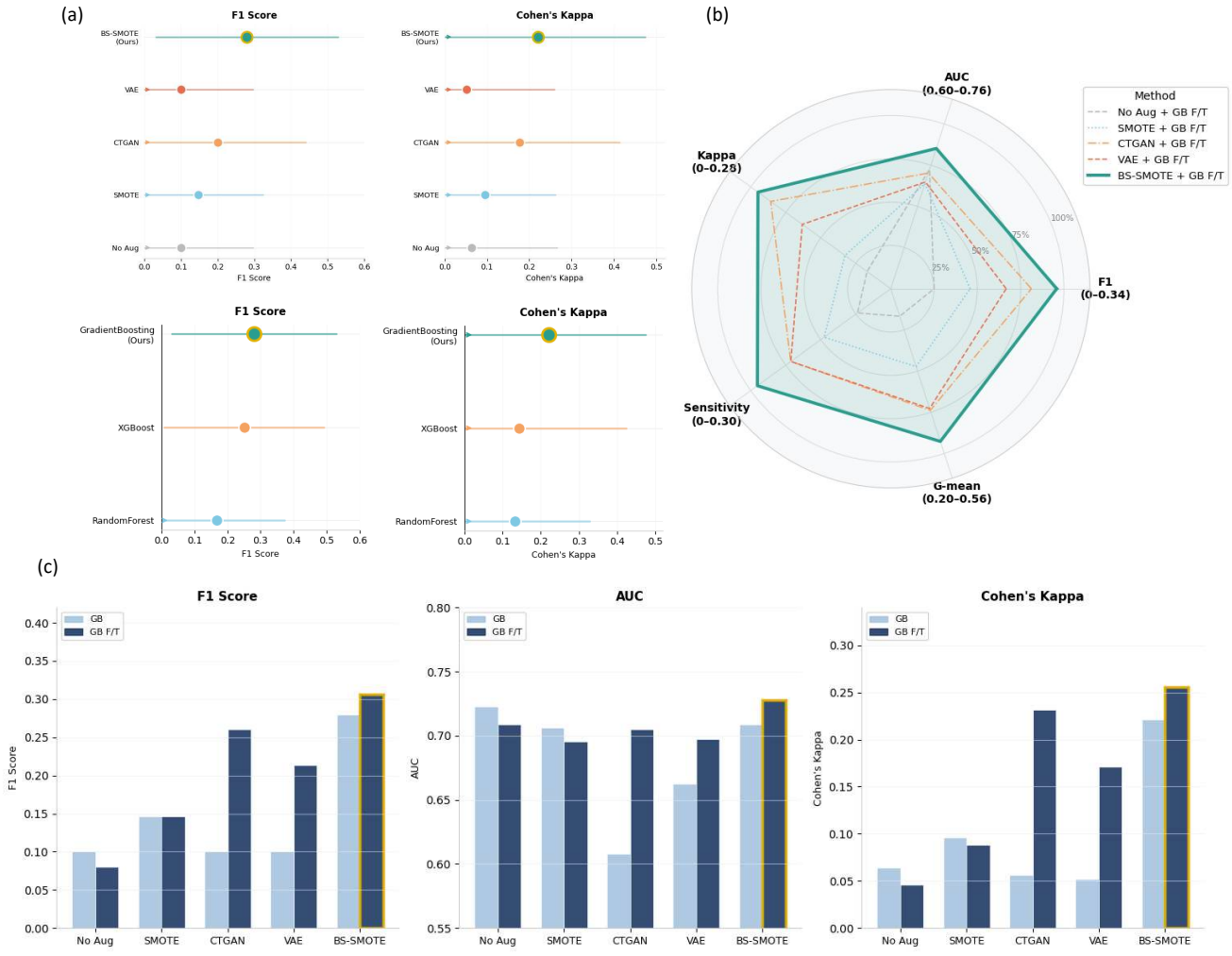


Figure 3. Comparative performance evaluation across augmentation strategies and classifiers. (a) Cleveland dot plot comparing five augmentation strategies under fine-tuned GradientBoosting, reporting F1, AUC, and Cohen's κ with 95% confidence intervals. Error bars represent 95% confidence intervals across stratified 5-fold cross-validation. CI lower bounds for F1 and Kappa are clipped at 0 for display purposes. (b) Multi-metric radar chart comparing BS-SMOTE with fine-tuned GradientBoosting against representative baseline methods across five performance dimensions (F1, AUC, Kappa, Sensitivity, G-mean). Values are normalized within each metric range for display. (c) Bar chart comparing five augmentation strategies under fixed (GB) and fine-tuned (GB F/T) GradientBoosting configurations, reporting F1, AUC, and Cohen's κ. The left bars represent fixed hyperparameter configurations and the right bars represent fine-tuned configurations. Gold border indicates the best-performing combination.

sharply to negative values above approximately 1.5 mL/cmH_2O/kg, suggesting a physiological threshold below which reduced lung distensibility disproportionately elevates respiratory compromise risk under anesthesia. Below this threshold, impaired dynamic compliance is likely to restrict effective tidal volume delivery, promote alveolar collapse, and predispose patients to ventilation-perfusion mismatch, thereby reducing arterial oxygenation in the post-extubation period. This nonlinear pattern is consistent with established respiratory physiology, in which reduced dynamic compliance impairs effective alveolar ventilation and predisposes patients to ventilation-perfusion mismatch and hypoxemia[36–38]. The identification of a clinically interpretable compliance threshold further supports the utility of intra-anesthetic spirometry monitoring as an early warning signal for post-extubation hypoxemia risk stratification. Age exhibited a nonlinear risk profile, with positive SHAP contributions emerging above approximately 10 years, consistent with the age-related decline in respiratory reserve capacity and anesthetic tolerance observed in geriatric veterinary patients[39,40]. In older dogs, reduced chest wall compliance, diminished respiratory muscle

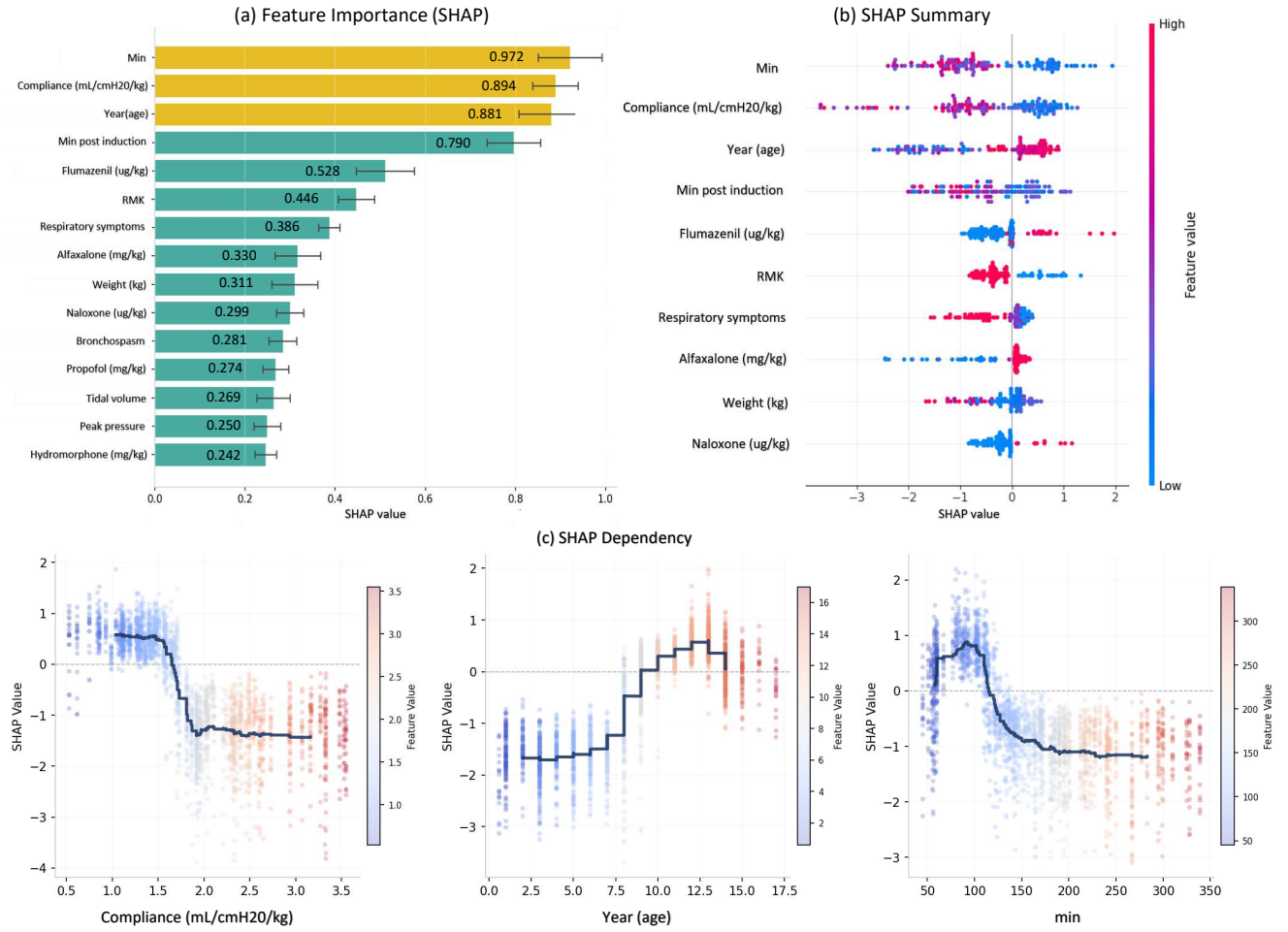


Figure 4. SHAP-based physiological rule discovery. (a) Mean absolute SHAP values for the top 15 features, aggregated across 100 independent runs (20 seeds × 5-fold cross-validation). Error bars represent 95% confidence intervals. All analyses were restricted to held-out original clinical data. (b) SHAP summary plot illustrating the direction and magnitude of individual feature contributions across all held-out test observations. Each point represents a single observation, with color indicating feature value (red = high, blue = low). Features are ranked by mean absolute SHAP value. (c) SHAP dependence plots for the three dominant risk contributors aggregated across 100 independent runs. X-axis values are shown on the original clinical scale. The navy trend line represents a moving average across pooled observations. A nonlinear threshold effect is evident for respiratory compliance at approximately 1.5 mL/cmH_2O/kg, above which the risk-increasing contribution transitions sharply to a protective effect.

strength, and increased anatomic dead space collectively impair the ability to maintain adequate gas exchange under anesthetic-induced respiratory depression, rendering this population disproportionately susceptible to post-extubation hypoxemia[41]. The threshold effect observed above 10 years aligns with established clinical observations that geriatric dogs face substantially elevated anesthetic mortality risk relative to younger counterparts, underscoring the importance of age-stratified monitoring protocols in veterinary anesthesia practice. Anesthesia duration demonstrated a monotonically decreasing SHAP profile with increasing duration, with positive contributions concentrated at shorter durations and negative contributions at longer durations. This finding appears counterintuitive, as prolonged anesthesia is generally associated with progressive atelectasis, impaired oxygenation, and other anesthetic complications. A more plausible explanation is confounding by indication rather than a protective effect of longer anesthesia[42]. In clinical practice, dogs with compromised respiratory function or poor general condition are more likely to undergo brief diagnostic or abbreviated procedures, whereas lengthy elective surgeries are more commonly performed in patients with relatively preserved cardiopulmonary status. Notably, this pattern persisted despite the

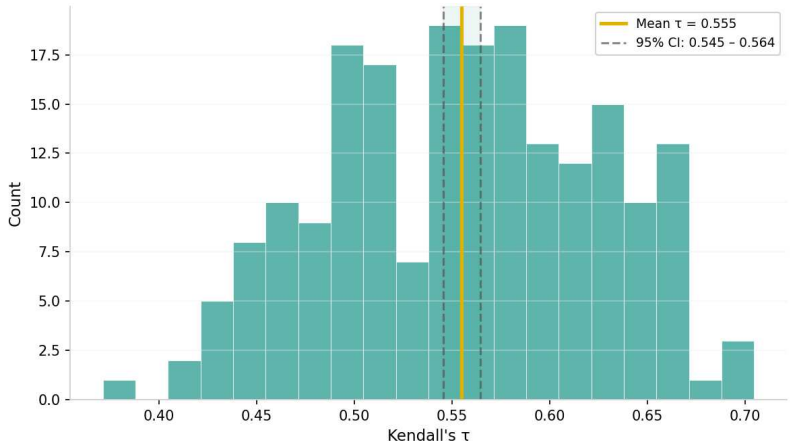


Figure 5. Distribution of pairwise Kendall's τ rank correlation coefficients across 200 fold comparisons (20 seeds $\times \binom{5}{2} = 10$ pairwise fold combinations per seed), quantifying the consistency of SHAP-based feature importance rankings across data partitions. The gold vertical line indicates the mean $\tau = 0.555$, and dashed lines denote the 95% confidence interval (0.545–0.564). The distribution indicates moderate ranking consistency, reflecting the inherent variability of feature attribution estimates derived from small held-out test partitions with as few as three minority-class cases per fold.

inclusion of ASA physical status and procedure type as model inputs, suggesting that anesthesia duration captures aspects of clinical severity not fully represented by these categorical variables. Because SHAP values reflect model-learned associations rather than causal effects[43], anesthesia duration is best regarded as a surrogate marker of underlying patient condition rather than a modifiable risk factor.

The stability of SHAP-based feature rankings across data partitions was quantified using Kendall's τ, computed across all pairwise fold comparisons within each seed (200 comparisons in total). The distribution of τ values is illustrated in Figure 5. The mean Kendall's τ was 0.555 (95% CI: 0.545–0.564), indicating moderate ranking consistency across partitions. While this value falls below the threshold conventionally associated with high consistency ($\tau > 0.7$), it is consistent with the expected instability of feature rankings derived from fold compositions containing as few as three minority-class test cases per fold. The top three features (respiratory compliance, age, and anesthesia duration) were consistently ranked among the highest contributors across the majority of folds and seeds, as evidenced by their narrow confidence intervals in Figure 4a, supporting the reliability of the identified physiological patterns as generalizable risk signals rather than partition-specific artifacts.

The SHAP analyses suggest that peri-anesthetic hypoxemia in dogs is not driven by isolated predictors but emerges from the interplay of respiratory mechanics, patient age, and anesthetic duration. These are consistent with the hypothesis that dogs with reduced pulmonary compliance whether attributable to brachycephalic upper airway anatomy, obesity, or subclinical respiratory disease are disproportionately vulnerable to hypoxemia under conditions of prolonged anesthesia, particularly in older individuals with diminished physiological reserve. The identified patterns constitute clinically interpretable hypotheses that may inform prospective risk stratification protocols and targeted monitoring strategies in veterinary anesthesia practice.

Methods

Study Design and Data Collection

An overview of the proposed four-stage framework is provided in Figure 6. This retrospective single-cohort observational study reviewed anesthetic records of dogs that underwent inhalation anesthesia at the Seoul National University Veterinary Medical Teaching Hospital between January and June 2022. The medical records were originally generated during routine clinical practice and were subsequently reviewed for research purposes. The study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines[44]. Ethical approval for the retrospective review and analysis of these clinical records was obtained from the Institutional Animal Care and Use Committee of Seoul National University (Approval number: SNU-250322-1; Approval date: 10 April 2025) prior to commencement of the retrospective record review. Written informed consent was obtained from the owners for the use of the animals' clinical data during their admission for anesthesia.

Patient characteristics and ABG data obtained during the intra-anesthetic and post-extubation periods were recorded in the electronic medical record system (e-Friends version 5.1; Woorien Co., Ltd., Korea). Pre-anesthetic assessments, intra-anesthetic spirometry, vital sign data, body position, anesthesia duration, and administration of reversal agents were collected from an

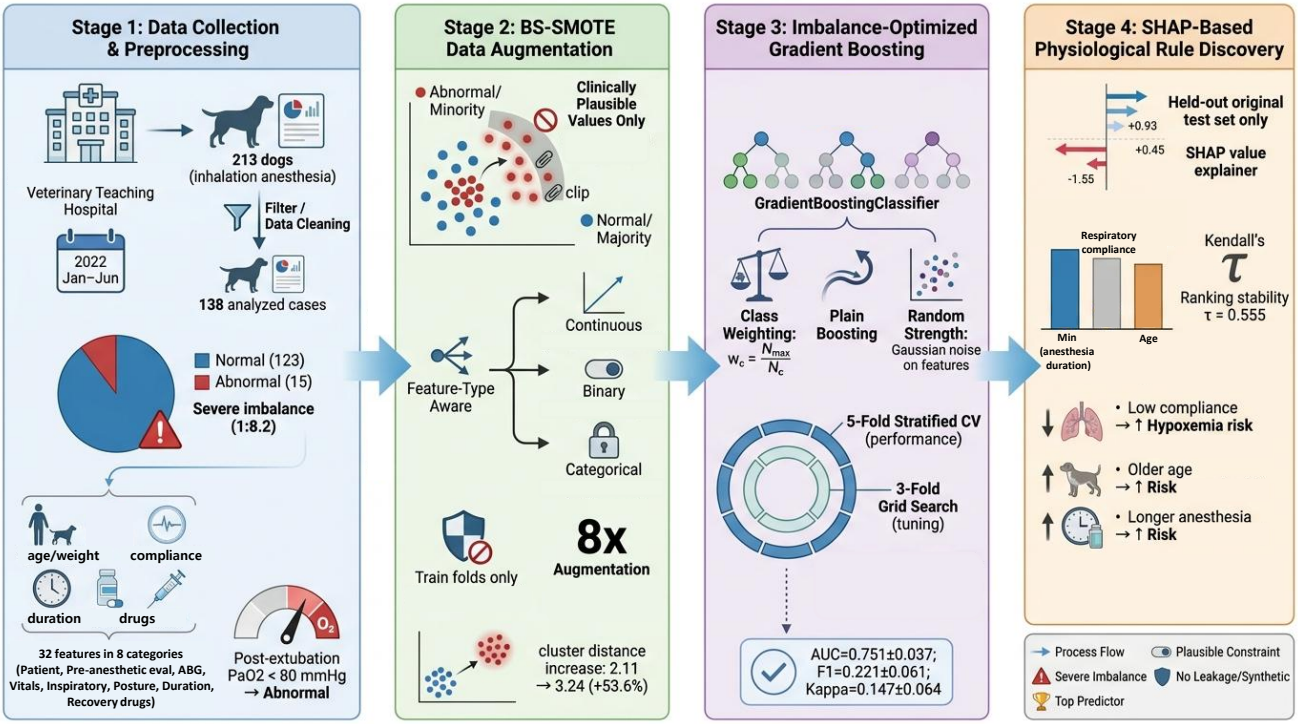


Figure 6. Overview of the BS-SMOTE framework for interpretable peri-anesthetic hypoxemia prediction. The pipeline comprises four stages: (1) data collection and preprocessing of 138 dogs with severe class imbalance. (2) BS-SMOTE augmentation with feature-type aware generation and physiological feasibility constraints applied strictly within training folds. (3) imbalance-optimized GradientBoosting with class weighting, plain boosting, and random strength injection under leakage-free nested cross-validation. (4) SHAP-based physiological rule discovery restricted to held-out original data, identifying respiratory compliance, age, and anesthesia duration as dominant risk contributors

electronic spreadsheet (Excel; Microsoft Corp, v16.41, WA, USA).

ABG analyses were performed using a single analyzer (ABL 90 FLEX PLUS; Radiometer, Denmark), and all values were corrected for body temperature. For intra-anesthetic analysis, arterial blood samples were collected after intubation, once hemodynamic stability had been achieved. For post-extubation analysis, samples were collected after the patient had regained full consciousness and at least 5 minutes after discontinuation of oxygen supplementation to ensure accurate measurement in room air. Arterial blood sampling was carried out by the anesthesiologist responsible for each case, while a veterinary student previously trained in ABG interpretation performed the analysis under the direct supervision of a faculty member from the Division of Veterinary Anesthesia and Analgesia.

Intra-anesthetic vital signs and spirometry data were continuously displayed on a multi-parameter patient monitor (Carescape B650; GE Healthcare, MA, USA) and recorded at 5-minute and 15-minute intervals, respectively. Spirometry data were obtained from a spirometry module connected to the breathing circuit via a pitot-based flow sensor (D-lite sensor; GE Healthcare, MA, USA). Pulmonary compliance was measured as dynamic compliance calculated from expired tidal volume and the pressure differential (PIP − PEEP). The C_{dyn} value closest to the time of post-extubation PaO_2 measurement was normalized to body weight and used for analysis.

Data Preprocessing

A total of 32 peri-anesthetic variables were retained as input features following a structured preprocessing procedure, as summarized in Table 1. Post-extubation PaO_2 and $PaCO_2$ were excluded from the feature set to prevent label leakage, as these values were used to define the primary outcome. The remaining features spanned five clinical domains: patient characteristics, pre-anesthetic assessment, intra-anesthetic drug administration, intra-anesthetic monitoring (spirometry, vital signs, and body position), and reversal agent administration during the recovery period. RMK denotes a combination of Remifentanil, Midazolam, and Ketamine administered intra-anesthetically for analgesia and sedation.

All continuous and ordinal features were converted to numeric format. Cases with missing values in any input feature were

Table 1. List of variables used as input features and primary outcome for model development. Variables were extracted from either the electronic medical record system (e-Friends version 5.1; Woorien Co., Ltd., Korea) or an electronic spreadsheet (Excel; Microsoft Corp, v16.41, WA, USA).

Variable	Type	Unit	Definition / Notes
Patient characteristics			
Breed (brachycephalic)	Binary	0/1	–
Body weight	Numerical	kg	–
Body condition score	Ordinal	1–9	–
Age	Numerical	years	–
Pre-anesthetic assessment			
ASA status	Ordinal	1–5	–
Respiratory signs	Ordinal	0–3	0=none; 1=mild; 2=moderate; 3=severe
Soft palate elongation	Binary	0/1	–
Morphological change	Binary	0/1	Upper airway anatomical abnormality
Intra-anesthetic drug administration			
Hydromorphone	Numerical	mg/kg	Pre-medication opioid
Propofol	Numerical	mg/kg	Induction agent
Alfaxalone	Numerical	mg/kg	Induction agent
Flumazenil	Numerical	μg/kg	Benzodiazepine antagonist
Naloxone	Numerical	μg/kg	Opioid antagonist
RMK	Binary	0/1	Intra-anesthetic analgesic combination
Intra-anesthetic spirometry			
Compliance (C_{dyn})	Numerical	mL/cmH_2O/kg	Dynamic compliance, normalized to body weight
Tidal volume	Numerical	mL	–
Peak inspiratory pressure	Numerical	cmH_2O	–
Intra-anesthetic vital signs			
Hypotension	Binary	0/1	MAP $<$ 45 mmHg
Hypertension	Binary	0/1	SAP $>$ 180 mmHg
Hypercapnia	Binary	0/1	$PaCO_2 > 45$ mmHg
Bronchospasm	Binary	0/1	Clinician-documented bronchospasm
Bradycardia	Binary	0/1	HR $<$ 60 bpm
Hypothermia	Binary	0/1	BT $<$ 36.5°C
Intra-anesthetic hypoxemia	Binary	0/1	$PaO_2 < 80$ mmHg during anesthesia
Intra-anesthetic hypercapnia	Binary	0/1	$PaCO_2 > 45$ mmHg during anesthesia
Intra-anesthetic position			
Right lateral	Binary	0/1	–
Left lateral	Binary	0/1	–
Ventro-dorsal	Binary	0/1	–
Dorso-ventral	Binary	0/1	–
Intra-anesthetic time			
Duration of anesthesia	Numerical	min	Total duration from induction to extubation
Time post induction	Numerical	min	Time from induction to first ABG sampling
Primary outcome			
Post-extubation PaO_2	Binary	0/1	Hypoxemia = $PaO_2 < 80$ mmHg

removed via complete-case analysis, as the proportion of excluded cases was minimal ($n = 1$, $<1\%$), rendering imputation unnecessary and avoiding the introduction of artificial variance into the feature distributions[45]. Cases with duplicate chart numbers were identified and removed by retaining only the first recorded anesthetic event per patient. Following these exclusion steps, the final analytic cohort comprised 138 cases, as described in Section 2.1.

Boundary-Stabilized SMOTE (BS-SMOTE)

Continuous features of the minority-class training samples were standardized using z-score normalization to mitigate distance bias toward high-variance features. The standardized feature space was used exclusively for neighbor determination, while synthetic interpolation and all subsequent model training were conducted in the original feature scale. Standard SMOTE generates synthetic minority-class samples by interpolating between existing minority instances and their k-nearest neighbors:

$$\tilde{x} = x_i + \lambda \cdot (x_{nn} - x_i), \quad \lambda \sim \mathscr{U}(0,1) \tag{1}$$

where x_i is a randomly selected minority-class seed sample, x_{nn} is one of its k-nearest neighbors, and λ is drawn uniformly from $[0,1]$. While effective at increasing minority-class representation, this formulation applies identical linear interpolation across all feature types, irrespective of whether features are continuous, binary, or categorical. This uniformity can produce physiologically implausible values, for instance, intermediate values such as 0.3 or 0.7 for binary clinical indicators such as brachycephalic breed status or reversal agent administration.

To overcome these limitations, we propose Boundary-Stabilized SMOTE (BS-SMOTE), which extends standard SMOTE with two mathematically defined constraints. The first constraint is feature-type aware generation, which applies distinct

interpolation rules according to the measurement scale of each feature:

$$\tilde{x}^{(j)} = \begin{cases} x_i^{(j)} + \lambda \cdot (x_{nn}^{(j)} - x_i^{(j)}) & \text{if } j \in \mathscr{F}_{\text{cont}} \\ \mathbb{1}[\lambda > 0.5] & \text{if } j \in \mathscr{F}_{\text{bin}} \\ x_i^{(j)} & \text{if } j \in \mathscr{F}_{\text{cat}} \end{cases} \tag{2}$$

where $\mathscr{F}_{\text{cont}}$, $\mathscr{F}_{\text{bin}}$, and $\mathscr{F}_{\text{cat}}$ denote the index sets of continuous, binary, and categorical features, respectively. Continuous features are interpolated linearly as in standard SMOTE; binary features are assigned via Bernoulli sampling with the interpolation weight λ serving as the sampling probability; and categorical features retain the value of the seed sample to avoid the introduction of spurious intermediate category codes.

The second constraint is physiological feasibility projection, which clips each generated feature value within a clinically plausible range derived from the training data:

$$\tilde{x}^{(j)} \leftarrow \text{clip}\left(\tilde{x}^{(j)},\, \ell_j,\, u_j\right) \tag{3}$$

where ℓ_j and u_j denote the first and 99th percentiles of feature j computed from the training set. This projection prevents the generation of samples with extreme or physiologically impossible values that may arise from interpolation near the boundary of the minority-class distribution.

To prevent data leakage, BS-SMOTE was applied strictly within the training partition of each cross-validation fold. The held-out test set in every fold consisted exclusively of original clinical cases, ensuring that synthetic samples did not influence performance evaluation or interpretability analyses. Figure 1b illustrates the PCA[46] projection of the augmented training data from a representative fold, demonstrating improved class separability relative to the unaugmented feature space. Feature-wise distribution comparisons between original normal cases, original abnormal cases, and BS-SMOTE synthetic samples across all 32 input features are shown in Figure 1c, confirming that synthetic samples preserve the distributional characteristics of original minority cases while expanding the minority-class boundary region.

To assess the sensitivity of model performance to augmentation intensity, a ratio sweep experiment was conducted in which the number of synthetic minority samples was varied from 0× to 8× the original minority count, using 10 independent random seeds and stratified 5-fold cross-validation.

Imbalance-Optimized Gradient Boosting

The predictive component of the framework employs a gradient boosting classifier configured for rare event learning under severe class imbalance. Gradient boosting was selected for its established capacity to model complex nonlinear interactions in heterogeneous tabular clinical data, its native compatibility with SHAP-based attribution analysis, and its ability to incorporate sample-level weighting without architectural modification[47,48]. Three complementary imbalance-aware mechanisms were integrated into the training procedure.

The first mechanism is class weighting[49] in which minority-class samples are assigned elevated loss weights to counteract the dominance of the majority class during gradient computation:

$$w_c = \frac{N_{\max}}{N_c} \tag{4}$$

where $N_{\max}$ denotes the size of the largest class and N_c denotes the number of samples in class c. This formulation penalizes misclassification of minority-class instances proportionally to their underrepresentation, reducing majority-class bias without requiring additional data.

The second mechanism is plain boosting, in which weak learners are trained sequentially on the residuals of the current ensemble[50–52], with the class weight incorporated into the update rule:

$$F_m(x) = F_{m-1}(x) + \eta \cdot w_c \cdot h_m(x) \tag{5}$$

where η denotes the learning rate and $h_m(x)$ denotes the weak learner fitted at iteration m. Plain boosting was adopted over more aggressive variants to promote convergence stability under small-sample conditions.

The third mechanism is random strength injection, in which Gaussian noise is added to feature importance scores prior to split selection at each boosting iteration:

$$\tilde{I}_j = I_j + \varepsilon, \quad \varepsilon \sim \mathscr{N}(0, \sigma^2) \tag{6}$$

where I_j denotes the impurity-based importance of feature j. This perturbation reduces repeated selection of a narrow subset of dominant predictors and promotes diversity across weak learners in the ensemble, which is particularly beneficial in small-sample settings where a few features may otherwise monopolize the tree construction process.

Hyperparameter optimization was conducted using a nested cross-validation design to prevent overfitting to the validation set. The outer loop consisted of a stratified 5-fold cross-validation for unbiased performance estimation. Within each outer training fold, an inner 3-fold grid search was applied to identify the optimal hyperparameter configuration from a predefined search space comprising $n_{\text{estimators}} \in \{100, 200, 300\}$, max_depth $\in \{2, 3, 4\}$, learning_rate $\in \{0.01, 0.05, 0.1\}$, and subsample $\in \{0.7, 0.8, 1.0\}$, yielding 81 candidate configurations per fold. The inner search was optimized with respect to the F1 score, which is more informative than accuracy under severe class imbalance. The most frequently selected configuration across outer folds (learning rate = 0.05, max depth = 2, $n_{\text{estimators}}$ = 200, subsample = 0.7) was reported as the representative hyperparameter setting.

SHAP-Based Interpretability Analysis

Interpretability analysis was conducted using Shapley Additive Explanations (SHAP), a game-theoretic framework that decomposes individual model predictions into additive feature contributions with theoretical guarantees of consistency and local accuracy. SHAP was employed not as a post hoc visualization tool, but as a hypothesis-generating instrument for physiological rule discovery, consistent with its established use in data-scarce clinical settings where controlled experimentation is infeasible[8].

To ensure that identified attribution patterns reflect genuine physiological relationships rather than artifacts introduced by synthetic augmentation, all SHAP analyses were restricted exclusively to held-out original clinical data. Synthetic samples generated by BS-SMOTE were used solely during model training and were entirely excluded from the interpretability analysis. This separation prevents optimistic bias in feature attribution estimates and ensures that the identified risk patterns are grounded in real, previously unseen clinical observations.

Global feature relevance was assessed using mean absolute SHAP values aggregated across all held-out test observations:

$$\bar{\phi}_j = \frac{1}{n} \sum_{i=1}^{n} \left| \phi_j^{(i)} \right| \tag{7}$$

where $\phi_j^{(i)}$ denotes the SHAP value of feature j for observation i and n is the total number of held-out test cases. To obtain stable and reliable importance estimates, SHAP values were aggregated across all folds and all 20 independent random seeds, yielding 100 independent runs in total. Mean absolute SHAP values are reported with 95% confidence intervals computed across these 100 runs.

SHAP dependence plots were constructed for the top three features identified by mean absolute SHAP value, using original-scale feature values recovered via inverse transformation of the fold-wise standardization scaler. These plots visualize the nonlinear relationship between feature values and their corresponding SHAP contributions, enabling identification of threshold effects and monotonic risk patterns. Trend lines were computed as moving averages across pooled observations from all 100 runs to characterize the aggregate directional effect of each feature.

The stability of SHAP-based feature rankings across data partitions was formally quantified using Kendall's rank correlation coefficient τ[53]. For each pair of folds within each random seed, the Kendall's τ between the corresponding SHAP importance rankings was computed, yielding a total of $20 \times \binom{5}{2} = 200$ pairwise comparisons. The mean and 95% confidence interval of τ across these comparisons are reported as a measure of ranking consistency. Values of $\tau > 0.7$ are generally interpreted as indicating high consistency, while values in the range $0.5 < \tau \leq 0.7$ indicate moderate consistency[54].

Limitations

Several limitations of this study warrant consideration. The most fundamental constraint is the small analytic cohort of 138 dogs with only 15 confirmed hypoxemic cases, which limits statistical power and generalizability. Although BS-SMOTE was specifically designed for this data-scarce regime, the resulting performance estimates exhibit wide confidence intervals that reflect genuine uncertainty, and external validation on an independent multi-institutional cohort is required before the identified patterns can inform prospective clinical protocols. The single-center design further limits representativeness, as referral bias inherent to a tertiary teaching hospital may have produced a patient population with higher anesthetic risk than would be observed in general veterinary practice.

Second, while SHAP analyses were restricted to held-out original data and ranking stability was formally quantified via Kendall's τ, the moderate consistency observed ($\tau = 0.555$) reflects the inherent variability of attribution estimates derived from small test partitions. The identified physiological rules, particularly the nonlinear compliance threshold and the duration-dependent risk pattern, should therefore be interpreted as hypothesis-generating findings rather than established clinical guidelines, and prospective studies with larger and more diverse cohorts are needed to confirm their directionality and clinical relevance.

Third, the current BS-SMOTE implementation derives physiological feasibility bounds from training data percentiles, which may not fully capture all clinically meaningful constraints. Incorporation of domain-expert-defined absolute bounds based on established veterinary pharmacological and physiological guidelines represents a natural extension of the framework,

as does its application to multi-class outcomes, temporally structured monitoring data, and other rare adverse event domains beyond veterinary anesthesia.

Conclusion

This study presented BS-SMOTE, a domain-constrained synthetic oversampling framework designed for rare adverse event prediction under severe class imbalance and data scarcity. By extending standard SMOTE with feature-type aware generation and physiological feasibility projection, BS-SMOTE preserves clinical feature integrity during synthetic sample generation while expanding the minority-class decision boundary in a physiologically plausible manner. Embedded within a leakage-free stratified nested cross-validation protocol, the framework provides unbiased performance estimates that correct a systematic methodological flaw prevalent in augmentation-based clinical machine learning studies.

In a comprehensive 10-way ablation study applied to peri-anesthetic hypoxemia prediction in veterinary anesthesia, BS-SMOTE with fine-tuned GradientBoosting consistently outperformed standard SMOTE, CTGAN, and VAE-based augmentation across all primary metrics, achieving an F1 of 0.307 (95% CI: 0.016–0.598), AUC of 0.728 (95% CI: 0.655–0.801), and Cohen's κ of 0.256. Robustness assessment across 20 independent random seeds confirmed stable discriminative performance ($\text{AUC} = 0.751 \pm 0.037$), and an augmentation ratio sweep demonstrated monotonically improving class separability without evidence of performance degradation within the evaluated augmentation range. SHAP-based interpretability analyses restricted to held-out original clinical data identified respiratory compliance, anesthesia duration, and age as dominant and stable risk contributors, with compliance exhibiting a nonlinear threshold effect consistent with established veterinary respiratory physiology. The moderate Kendall's τ of 0.555 across 100 independent runs reflects the inherent instability of small-sample partitioning rather than systematic attribution inconsistency, and the top three features were consistently identified across the majority of evaluation runs.

The proposed framework addresses a class of problems that recurs across diverse clinical domains including neonatal medicine, intensive care, and surgical complications. The core principles of BS-SMOTE, namely feature-type integrity, physiological plausibility, and leakage-free evaluation, are domain-agnostic and directly transferable to other rare-event clinical settings. Future work will focus on external multi-institutional validation, incorporation of temporally structured monitoring data, and extension to multi-class outcome settings to further establish the generalizability and clinical utility of the proposed framework.

Acknowledgments

This work was partly supported by the New Faculty Startup Fund from Seoul National University, and the BK21 FOUR program and Research Institute of Veterinary Science, College of Veterinary Medicine, Seoul National University and was partly supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00456480, RS-2026-25479483), a grant of Korean ARPA-H Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (RS-2026-25613012) and Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (IPET) through Agriculture and Food Convergence Technologies Program for Research Manpower Development Program funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (RS-2024-00398561). The authors thank professor Klaus-Robert Müller for insightful discussions and constructive feedback that helped refine the conceptual aspects, S.-N. Jang, H.-Y. Park S.-W. Kim, and J.-Y. Oh for the useful discussion of the data analysis.

Author contributions: S.-H. Yu and M. Kim contributed to the writing, editing, and reviewing of the research and manuscript preparation. S.-H. Yu contributed to machine learning modeling and data interpretation. M. Kim proposed the study and was responsible for data acquisition, statistical analysis, and data interpretation. I. Lee contributed to data interpretation, and to the writing, editing, and reviewing of the research. W.-G. Son and J.-H. Jeong contributed to the study design, writing, editing, and reviewing of the research, and provided supervision. All authors reviewed and approved the final version of the manuscript. The corresponding authors had full access to all the data and had final responsibility for the decision to submit for publication.

Competing interests: There are no competing interests to declare.

Data availability statement: The clinical data, including patient demographics and anesthetic monitoring vital signs, are not publicly available due to privacy and ethical restrictions. However, these data may be made available from the corresponding author upon reasonable request for academic research purposes. Requests for relevant data can be made to Dr. Son (carpeego@snu.ac.kr). Responses can be expected within 3 weeks. Processed and summarized datasets that support the findings of this study, including those used for additional analyses, are provided in the supplementary materials. Source data are provided with this paper.

References

1. Kim, M., Choi, J., Lee, I. & Son, W.-G. Intubation-related reflex atelectasis of lung lobes during computed tomography in two dogs. *Vet. Anaesth. Analg.* **51** (2024).
2. Hoareau, G., Jourdan, G., Mellema, M. & Verwaerde, P. Evaluation of arterial blood gases and arterial blood pressures in brachycephalic dogs. *J. Vet. Intern. Med.* **26**, 897–904 (2012).
3. Rieser, T. M. Arterial and venous blood gas analyses. *Top. Companion Anim. Med.* **28**, 86–90 (2013).
4. Kim, M., Sung, T., Nam, C., Lee, I. & gyun Son, W. Evaluating the ratio of arterial partial pressure of oxygen to inspired oxygen fraction and dynamic compliance following anesthetic induction for predicting postanesthetic hypoxemia in dogs. *Vet. Anaesth. Analg.* **52**, 794–802 (2025).
5. Berry, C. R., Hawkins, E. C., Hurley, K. J. & Monce, K. Frequency of pulmonary mineralization and hypoxemia in 21 dogs with pituitary-dependent hyperadrenocorticism. *J. Vet. Intern. Med.* **14**, 151–156 (2000).
6. Piemontese, C. *et al.* The incidence of hypoxemia in dogs recovering from general anesthesia detected with pulse-oximetry and related risk factors. *Vet. J.* **305**, 106135 (2024).
7. Zhu, M., Liu, Y., Luo, Z. & Zhu, T. Bridging data gaps of rare conditions in icu: A multi-disease adaptation approach for clinical prediction. *npj. Digit. Med.* **9** (2026).
8. Lundberg, S. *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2** (2018).
9. Bellini, V. *et al.* Machine learning in perioperative medicine: a systematic review. *J. Anesth. Analg. Crit. Care* **2** (2022).
10. Kim, Y. *et al.* Machine learning-based risk prediction model for canine myxomatous mitral valve disease using electronic health record data. *Front. Vet. Sci.* **10** (2023).
11. Shelley, B. & Shaw, M. Machine learning and preoperative risk prediction: the machines are coming. *Br. J. Anaesth.* **133**, 925–930 (2024).
12. Jeong, J.-H. *et al.* EEG-based cross-subject prediction for consciousness state transitions under sedation using a deep learning framework. *IEEE J. Biomed. Heal. Inform.* 1–14 (2025).
13. Miyaguchi, N., Takeuchi, K., Kashima, H., Morita, M. & Morimatsu, H. Predicting anesthetic infusion events using machine learning. *Sci. Rep.* **11**, 23648 (2021).
14. Ayuso-Muñoz, A., Prieto-Santamaría, L., Ugarte-Carro, E., Serrano, E. & Rodríguez-González, A. Uncovering hidden therapeutic indications through drug repurposing with graph neural networks and heterogeneous data. *Artif. Intell. Med.* **145**, 102687 (2023).
15. Huang, L. *et al.* Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *J. Biomed. Inform.* **99**, 103291 (2019).
16. Duffee, L., Stewart, S., Banerjee, T. & Khanna, C. Retrospective evaluation of the impact of hypoxemia during thoracic surgery in dogs on outcome. *Vet. Anaesth. Analg.* **50**, 341–348 (2023).
17. He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21**, 1263–1284 (2009).
18. Vairetti, C., Assadi, J. L. & Maldonado, S. Efficient hybrid oversampling and intelligent undersampling for imbalanced big data classification. *Expert. Syst. Appl.* **246**, 123149 (2024).
19. Wang, S., Dai, Y., Shen, J. & Xuan, J. Research on expansion and classification of imbalanced data based on smote algorithm. *Sci. Rep.* **11** (2021).
20. Douzas, G. & Bacao, F. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert. Syst. Appl.* **91**, 464–471 (2018).
21. Dablain, D., Krawczyk, B. & Chawla, N. V. Deepsmote: Fusing deep learning and smote for imbalanced data. *IEEE Trans. Neural Netw. Learn. Syst.* **34**, 6390–6404 (2023).
22. Xu, L., Skoularidou, M., Cuesta-Infante, A. & Veeramachaneni, K. Modeling tabular data using conditional gan. In *Adv. Neural Inf. Process. Syst.*, vol. 32 (2019).
23. Pu, Y. *et al.* Adversarial symmetric variational autoencoder. In *Adv. Neural Inf. Process. Syst.*, vol. 30 (2017).
24. Mukherjee, M. & Khushi, M. Smote-enc: A novel smote-based method to generate synthetic data for nominal and continuous features. *Appl. Syst. Innov.* **4** (2021).

25. Kuo, N. I.-H. *et al.* Generating synthetic clinical data that capture class imbalanced distributions with generative adversarial networks. *J. Biomed. Inform.* **144**, 104436 (2023).

26. Lee, Y., Chae, Y. & Jung, K. Leveraging vq-vae tokenization for autoregressive modeling of medical time series. *Artif. Intell. Med.* **154**, 102925 (2024).

27. Demircioğlu, A. Applying oversampling before cross-validation will lead to high bias in radiomics. *Sci. Rep.* **14** (2024).

28. Freitas, N. R. *et al.* Multiple mask and boundary scoring r-cnn with cgan data augmentation for bladder tumor segmentation in wlc videos. *Artif. Intell. Med.* **147**, 102723 (2024).

29. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Adv. Neural Inf. Process. Syst.*, 4768–4777 (2017).

30. Hur, S. *et al.* Comparison of shap and clinician friendly explanations reveals effects on clinical decision behaviour. *npj. Digit. Med.* **8**, 578 (2025).

31. Yi, F. *et al.* Xgboost-shap-based interpretable diagnostic framework for alzheimer's disease. *BMC Med. Inform. Decis. Mak.* **23** (2023).

32. Jin, Y. *et al.* Shap-based interpretable machine learning for parkinson's disease severity prediction. *Front. Neurol.* **16** (2025).

33. Dorogush, A., Ershov, V. & Gulin, A. Catboost: gradient boosting with categorical features support. *arXiv* (2018).

34. Zhang, Z., Zhao, Y., Canes, A., Steinberg, D. & Lyashevska, O. Predictive analytics with gradient boosting in clinical medicine. *Ann. Transl. Med.* **7** (2019).

35. Li, X., Wang, Y., Basu, S., Kumbier, K. & Yu, B. A debiased mdi feature importance measure for random forests. In *Adv. Neural Inf. Process. Syst.*, 8047–8057 (2019).

36. Raillard, M. *et al.* Characterization of dynamic compliance of the respiratory system in healthy anesthetized dogs. *Front. Vet. Sci.* **11** (2024).

37. Neto, A. S. *et al.* Association between driving pressure and development of postoperative pulmonary complications in patients undergoing mechanical ventilation for general anaesthesia. *Lancet Respir. Med.* **4**, 272–280 (2016).

38. Rodrigues, R. *et al.* Intraoperative protective mechanical ventilation in dogs: A randomized clinical trial. *Front. Vet. Sci.* **9**, 842613 (2022).

39. Brodbelt, D. C. *et al.* The risk of death: the confidential enquiry into perioperative small animal fatalities. *Vet. Anaesth. Analg.* **35**, 365–373 (2008).

40. Brodbelt, D. C., Pfeiffer, D. U., Young, L. E. & Wood, J. L. N. Results of the confidential enquiry into perioperative small animal fatalities regarding risk factors for anesthetic-related death in dogs. *J. Am. Vet. Med. Assoc.* **233**, 1096–1104 (2008).

41. Hughes, L. Anaesthesia for the geriatric dog and cat. *Ir. Vet. J.* **61**, 380–387 (2008).

42. Kyriacou, D. N. & Lewis, R. J. Confounding by indication in clinical research. *JAMA* **316**, 1818–1819 (2016).

43. Molnar, C. *et al.* General pitfalls of model-agnostic interpretation methods for machine learning models. In *ICML*, 39–68 (2020).

44. von Elm, E. *et al.* The strengthening the reporting of observational studies in epidemiology (strobe) statement: guidelines for reporting observational studies. *Lancet* **370**, 1453–1457 (2007).

45. Sterne, J. A. C. *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* **338** (2009).

46. Daffertshofer, A., Lamoth, C. J., Meijer, O. G. & Beek, P. J. Pca in studying coordination and variability: a tutorial. *Clin. Biomech.* **19**, 415–428 (2004).

47. Friedman, J. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29** (2001).

48. Lundberg, S. *et al.* From local explanations to global understanding with explainable ai for trees. *Nat. Mach. Intell.* (2019).

49. Lee, B.-H., Jeong, J.-H. & Lee, S.-W. SessionNet: Feature similarity-based weighted ensemble learning for motor imagery classification. *IEEE Access* **8**, 134524–134535 (2020).

50. Mienye, I. D. & Sun, Y. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access* **10**, 99129–99149 (2022).

51. Cao, Y., Geddes, T., Yang, J. & Yang, P. Ensemble deep learning in bioinformatics. *Nat. Mach. Intell.* **2**, 1–9 (2020).

52. Ganie, S., Dutta Pramanik, P. & Zhao, Z. Ensemble learning with explainable ai for improved heart disease prediction based on multiple datasets. *Sci. Rep.* **15** (2025).

53. Brown, G. W. & Hayden, G. F. Nonparametric methods: Clinical applications. *Clin. Pediatr.* **24**, 490–498 (1985).

54. Schober, P., Boer, C. & Schwarte, L. Correlation coefficients: Appropriate use and interpretation. *Anesth. Analg.* **126** (2018).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Sci.Rep.Supplementaryfile.pdf