Post-PCI Cardiac Death Prediction via Synthetic Minority Augmentation and Stress-Tested Probability Quality

Daniil Burakov [1] **Dmitrii Khelimskii** [2] **Mikhail Lazarev** [1]

[1]*HSE University, Moscow, Russian Federation* [2]*E.N. Meshalkin National Medical Research Center, Ministry of Health of the Russian Federation, Novosibirsk, Russian Federation*. Correspondence to: Mikhail Lazarev mvlazarev@hse.ru.

1. Motivation

Coronary artery disease remains a leading cause of morbidity and mortality worldwide despite advances in cardiology and interventional technologies. Risk stratification after percutaneous coronary intervention (PCI) supports follow-up planning, secondary prevention, and allocation of clinical resources. In real-world registries, however, clinically relevant endpoints such as long-term cardiac death are rare, producing severe class imbalance that can mask model brittleness: headline metrics (e.g., accuracy) may look strong even when the model fails to identify any high-risk patients. We study 3-year cardiac mortality prediction for PCI bifurcation lesions using tabular clinical, angiographic, and procedural variables. In preliminary experiments, we observed a mismatch between test-set performance and clinical plausibility: models that look good under conventional cross-validation may assign unrealistically low risk to severe yet plausible patient profiles. We therefore focus on robustness and probability quality under stress tests. Our hypothesis is that augmenting training with (i) in-distribution synthetic minority samples and (ii) expert-validated extreme "edge-case" profiles can improve sensitivity to rare fatal outcomes and yield more clinically reasonable risk estimates without sacrificing overall discrimination.

2. Data

We analyzed a multicenter registry of 2044 PCI patients with bifurcation lesions. The primary outcome is cardiac death within 3 years. We use a patient-level split into training (N=1,635, positives n=127) and test (N=409, positives n=31) sets. Features include patient status (e.g., age, comorbidities), angiographic characteristics, and procedure-related variables. Missing values were imputed (categorical: mode; continuous: iterative imputation) and categorical variables were one-hot encoded. To probe generalization, we additionally evaluate on an external cohort from a separate center (N=158, positives n=29) consisting of acute myocardial infarction patients treated with PCI in 2011–2013.

3. Methods

Predictive models. We trained six tabular classifiers: logistic regression, random forest, CatBoost, XGBoost, TabPFN, and Kolmogorov–Arnold Networks (KAN). Hyperparameters were selected using cross-validation on the training set.

Synthetic augmentation and stress tests. To mitigate imbalance and probe robustness, we compared three training/evaluation settings (see Fig. 1):

- **No augmentation** (baseline training on the imbalanced cohort).
- **ARF oversampling**: +500 synthetic minority-class samples generated by Adversarial Random Forest (ARF) [1], a two-forest generator/discriminator scheme that approximates the minority distribution. (We also explored other generators such as CTGAN/TVAE and Gaussian copula in the full paper.)
- **Edge-case stress test**: a separate evaluation on 500 hand-crafted "severe" profiles representing clinically plausible high-risk constellations, used to audit whether predicted risks increase appropriately under deterioration.

Evaluation. Besides discrimination (AUC-ROC, F1, precision/recall, accuracy), we report probability quality via Brier score, expected calibration error (ECE), average confidence, and predictive entropy. This combination is essential in imbalanced clinical prediction, where a model can appear "well calibrated" simply by predicting the majority class with high confidence.

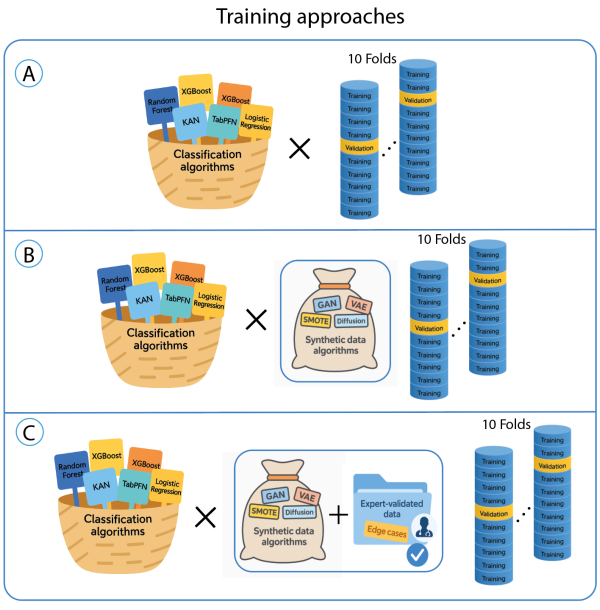

Fig. 1: Three training regimes: a) Standard 10 folds cross validation. b) Synthetic in distribution data used to enrich the dataset. c) Synthetic data and expert-validated edge cases. Taken from [2].

4. Results

Baseline models are accurate but miss the minority class. Without augmentation, all models achieve high accuracy (≈ 0.92–0.93) and moderate AUC-ROC (0.76–0.82), yet they almost completely ignore the minority class at the default decision threshold: F1 and recall are essentially zero for most models (Table 1). In this regime, calibration metrics can be misleading: average confidence is ≈ 0.92 with low ECE (< 0.04), reflecting majority-class certainty rather than clinically useful risk estimation.

ARF augmentation improves minority-class detection, but shifts probability behavior. Adding 500 ARF-generated minority samples Fig. 2 consistently increases recall and F1 for several models. For example, random forest improves from (F1 = 0, Recall = 0) to (F1 = 0.29, Recall = 0.29) at a similar AUC-ROC (0.78), while CatBoost reaches F1 = 0.14 (Table 1). At the same time, ARF tends to reduce overconfident majority predictions (lower average confidence and higher entropy), which is desirable, but it may increase calibration error for some architectures (e.g., KAN achieves high recall yet poor ECE), highlighting a trade-off between detection and probability quality.

Stress tests reveal clinically relevant differences not captured by headline metrics. On constructed severe profiles, well-behaved models should assign substantially higher predicted risk. We observe that augmentation strategies can materially change the *distribution* of predicted probabilities on edge cases, making stress testing a practical complement to standard test-set metrics.

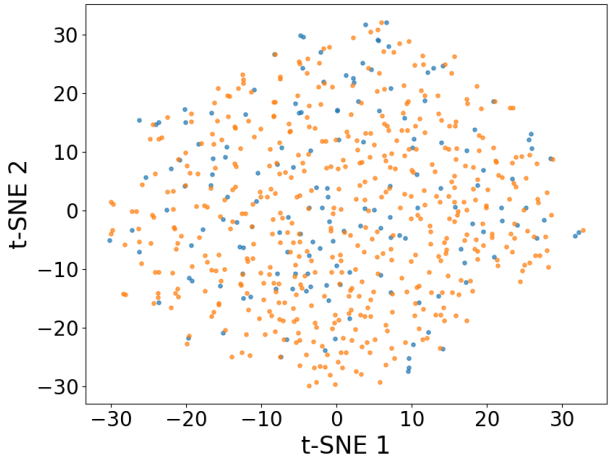

Fig. 2: t-SNE projections of real and synthetic minority-class samples generated by ARF. Real (N = 158) and synthetic (N = 500) samples are shown in blue and orange, respectively.

Feature importance highlights clinically plausible drivers. Permutation feature importance identifies Age, Ejection Fraction, Peripheral Artery Disease, and Cerebrovascular Disease as the most influential features. A follow-up experiment that removes non-informative features improves generalization: with ARF augmentation plus feature selection, all models exceed AUC–ROC > 0.8 on the test set, and external validation reaches AUC–ROC ≈ 0.69–0.74 with materially better F1/recall for tree ensembles (full paper [2]).

Table 1: Compact summary on the test set: baseline vs. ARF and Edge-cases augmentation. ECE is from probabilistic evaluation; higher recall/F1 indicates improved minority detection. (Values reproduced from the full paper's Tables 1 and 4.) RF - Random Forest, CB - CatBoost

Setting	Model	AUC-ROC	F1	Recall	ECE
Baseline	RF	0.78	0.00	0.00	0.03
	CB	0.76	0.00	0.00	0.02
	KAN	0.79	0.06	0.03	0.04
ARF (500)	RF	0.78	0.29	0.29	0.11
	CB	0.77	0.14	0.10	0.06
	KAN	0.75	0.27	0.55	0.31
Edge Cases (500)	RF	0.78	0	0	0.04
	CB	0.77	0.12	0.06	0.03
	KAN	0.76	0.14	0.10	0.05
ARF (500) + Edge Cases (500)	RF	0.78	0.27	0.29	0.21
	CB	0.76	0.14	0.10	0.13
	KAN	0.76	0.30	0.65	0.31

5. Conclusion

Straightforward synthetic augmentation with realistic and extreme cases can expose, quantify, and partially reduce brittleness in imbalanced clinical prediction using only tabular records. In a highly imbalanced post-PCI registry, strong AUROC/accuracy can mask clinically unusable models that never predict the minority outcome. Training with in-distribution synthetic minority samples and expert-validated edge-case stress tests makes models more sensitive to rare cardiac death and produces more plausible risk estimates [2].

References

[1] David S Watson, Kristin Blesch, Jan Kapar, and Marvin N Wright. Adversarial random forests for density estimation and generative modeling. In *International Conference on Artificial Intelligence and Statistics*, pages 5357–5375. PMLR, 2023.

[2] Daniil Burakov, Ivan Petrov, Dmitrii Khelimskii, Ivan Bessonov, and Mikhail Lazarev. Cardiac mortality prediction in patients undergoing pci based on real and synthetic data. `https://arxiv.org/abs/2512.22259`, 2025.