Supplementary

Appendix 1

Methods

Composite Score Calculation

The composite score was calculated following the method described by Dong J, Jin Z, Li C, *et al.* Machine Learning Models With Prognostic Implications for Predicting Gastrointestinal Bleeding After Coronary Artery Bypass Grafting and Guiding Personalized Medicine: Multicenter Cohort Study. J Med Internet Res. 2025;27:e68509. doi:10.2196/68509.

For each of the 40 candidate models, we computed four metrics from 5-fold, 10-repeat cross-validation:

1. Mean AUC
2. Mean Brier score
3. Standard deviation of AUC (SD_AUC)
4. Standard deviation of Brier score (SD_Brier)

These metrics were then normalized using min–max scaling across all candidate models to a [0,1] range:

- Normalized AUC = (AUC - min_AUC) / (max_AUC - min_AUC)
- Normalized Brier = 1 - [(Brier - min_Brier) / (max_Brier - min_Brier)]
- Normalized SD_AUC = 1 - [(SD_AUC - min_SD_AUC) / (max_SD_AUC - min_SD_AUC)]
- Normalized SD_Brier = 1 - [(SD_Brier - min_SD_Brier) / (max_SD_Brier - min_SD_Brier)]

The Brier score and both SD terms were inverted (subtracted from 1) because lower values indicate better performance. The composite score was computed as the arithmetic mean of these four normalized values:

Composite Score = (Normalized_AUC + Normalized_Brier + Normalized_SD_AUC + Normalized_SD_Brier) / 4

Higher composite scores indicate better overall performance combining discrimination, calibration, and stability.

https://dx.doi.org/10.21037/tp-2026-1-0007

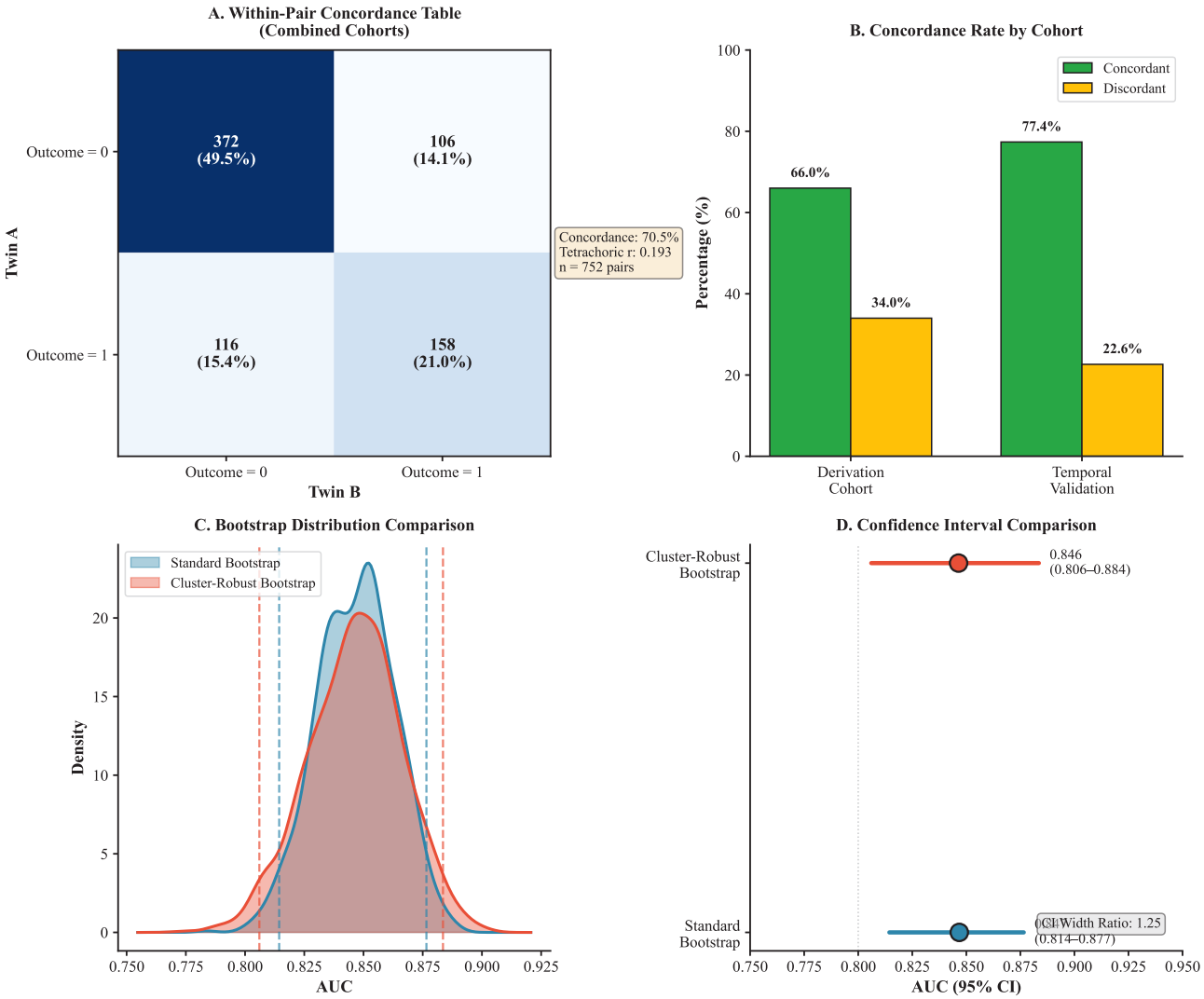

Figure S1 Within-pair correlation and cluster bootstrap sensitivity analysis. (A) Two-by-two concordance table showing outcome agreement between twin pairs across combined cohorts (n=456 pairs). Numbers represent pair counts with percentages in parentheses. Concordance rate was 70.5% with tetrachoric correlation of 0.193. (B) Concordance rates comparing derivation and temporal validation cohorts, showing the proportion of twin pairs with concordant (green) versus discordant (yellow) outcomes. (C) Bootstrap distribution comparison between standard bootstrap (blue) and cluster-robust bootstrap (red) for AUC estimation in the temporal validation cohort (1,000 iterations). Dashed lines indicate 95% confidence interval boundaries. (D) Forest plot comparing confidence intervals from standard bootstrap and cluster-robust bootstrap methods. The CI width ratio of 1.25 indicates modest widening of the confidence interval under cluster resampling, while point discrimination remained stable.

https://dx.doi.org/10.21037/tp-2026-1-0007

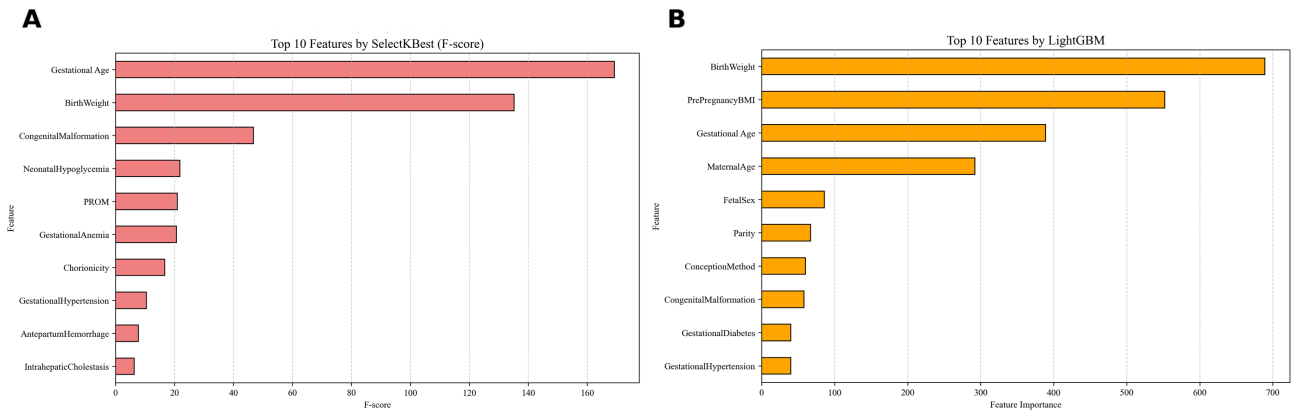

Figure S2 Feature importance rankings for alternative feature selection methods. (A) K-Best (ANOVA F-value) scores for candidate predictors. Features with higher scores contribute more strongly to model discrimination. LGBM, Light Gradient Boosting Machine. (B) LightGBM-based feature importance scores for candidate predictors.

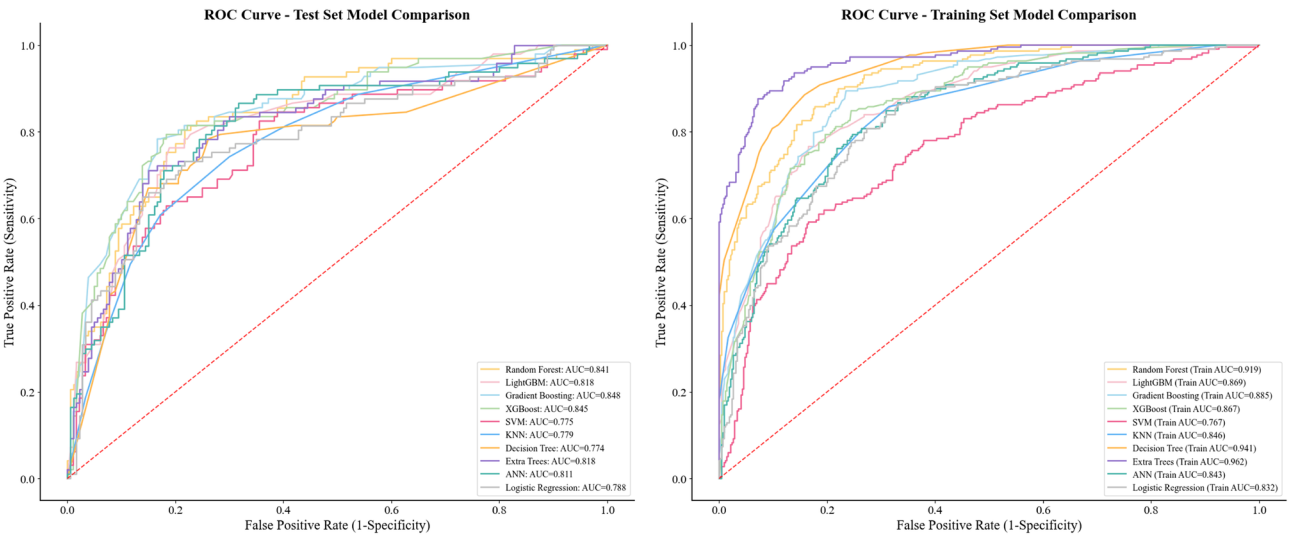

Figure S3 Receiver operating characteristic (ROC) curves for all ten machine learning algorithms in the derivation cohort. (A) ROC curves in the internal testing set. (B) ROC curves in the training set. Algorithms include logistic regression (LR), artificial neural network (ANN), decision tree (DT), extremely randomized trees (ET), gradient boosting machine (GB), k-nearest neighbors (KNN), Light Gradient Boosting Machine (LGBM), random forest (RF), support vector machine (SVM) and Extreme Gradient Boosting (XGB). AUC, area under the ROC curve.

https://dx.doi.org/10.21037/tp-2026-1-0007

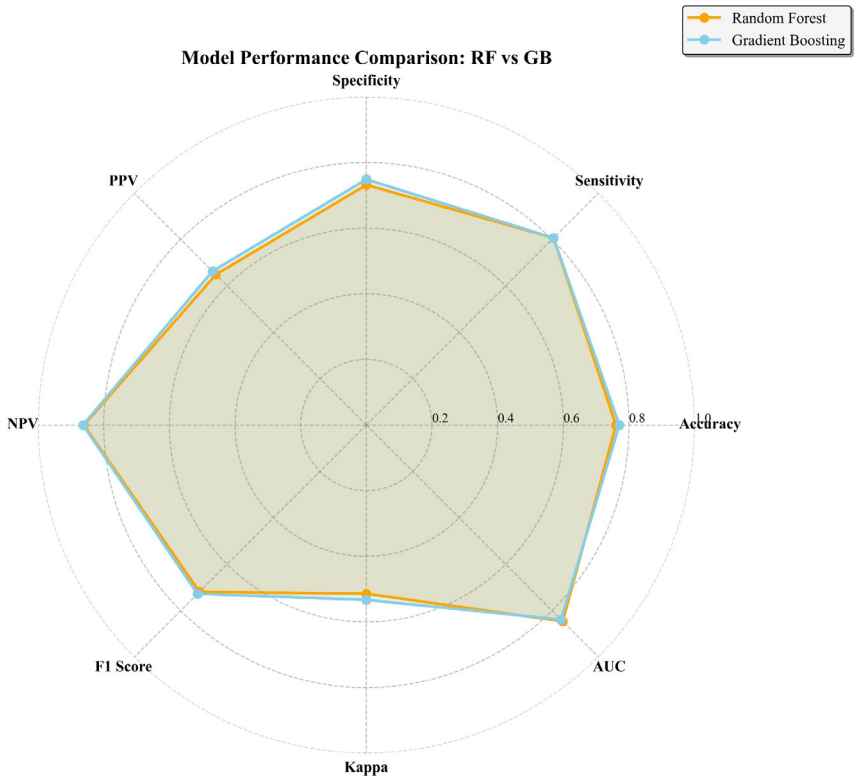

Figure S4 Radar plot summarizing classification metrics for Random Forest (RF) and Gradient Boosting (GB) in the temporal validation cohort. The plot displays accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score, Cohen's kappa and area under the ROC curve (AUC) on a normalized 0–1 scale.

https://dx.doi.org/10.21037/tp-2026-1-0007