An Optimization Algorithm for Medical Tabular Data Analysis Using Machine Learning and Deep Learning*

So-Hee Lim[1], Yu-Jin Ha[2], Jong-Chan Park[2], and Gun-Woo Kim[1]†

[1] Department of Computer Science and Engineering,
[2] Department of AI Convergence Engineering,
Gyeongsang National University, Jinju, Republic of Korea
limsohee@gnu.ac.kr, nrt5077@gmail.com, pakw2015@naver.com, gunwoo.kim@gnu.ac.kr

Abstract

This study presents an optimization framework that integrates machine learning (ML) and deep learning (DL) to analyze medical tabular data with coexisting missingness and class imbalance coexist. Prior work has noted that electronic health records (EHR) and in-hospital clinical datasets with mixed variables and imbalance make it difficult for a single algorithm to achieve adequate predictive performance [1, 2]. Tabular DL models (TabNet, SAINT, FT-Transformer) learn higher order relations but remain unstable and hyperparameter-sensitive. We propose staged preprocessing, dual feature selection, and stacking. On two datasets, the framework lifts macro F1 by 5.3 and 3.9 percentae points over the best single model.

1 Introduction and Proposed Method

Medical tabular data presents unique challenges including heterogeneous features with systematic missingness and severe class imbalance. Electronic health records combine categorical and numerical variables with missingness and imbalance that make single models brittle. Tree-based models offer stability but struggle with complex interactions while DL models capture patterns but suffer from overfitting [1]. We propose a compact pipeline that keeps rules explicit and trains diverse learners under one protocol to stack only when diversity and accuracy warrant it.

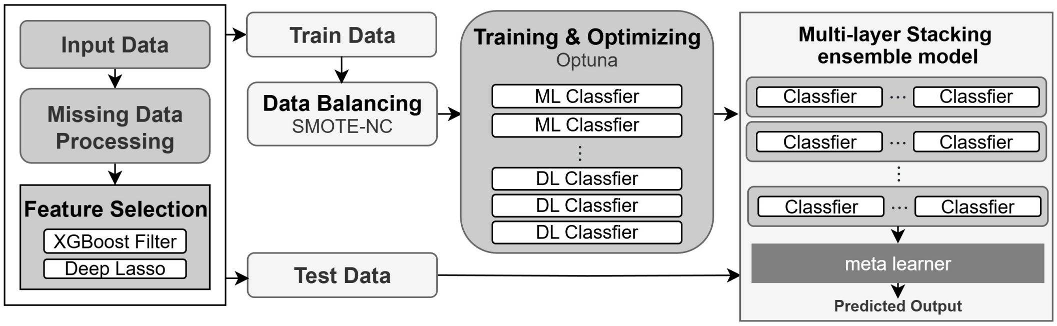

Figure 1. ML/DL Optimization Pipeline for Medical Tabular Data.

*Proceedings of the 9th International Conference on Mobile Internet Security (MobiSec'25), Article No. P-21, December 16-18, 2025, Sapporo, Japan. © The copyright of this paper remains with the author(s).

†Corresponding author. This research was supported by the Regional Innovation System & Education (RISE) program through the RISE Center, Gyeongsangnam-do, funded by the Ministry of Education (MOE) and the Gyeongsangnam-do Provincial Government, Republic of Korea(Grant No. 2025-RISE-16-001).

Pipeline. (1) **Preprocessing:** iFor missing rates below 10% use mean or mode imputation. For 10–5% use K-nearest neighbors (KNN) imputation with K chosen by Kolmogorov–Smirnov (KS) distance. Drop variables exceeding 50% missingness. Handle imbalance with SMOTE for numeric features and SMOTE-NC for mixed features. When the ratio exceeds 1:10 combine minority oversampling with majority undersampling. (2) **Dual selection:** Use the intersection of XGBoost top 60% and Deep Lasso candidates (gradient sparsity with group regularization). (3) **Training:** Train 4 ML models (Random Forest, ExtraTrees, XGBoost, LightGBM) and 7 DL models (MLP, TabNet, SAINT, FT-Transformer, DNF-Net, 1D-CNN, NODE) with Optuna 5-fold search and early stopping for DL. Report macro and weighted F1. (4) **Stacking:** Retain models above F1 threshold. Remove correlated pairs (Pearson $r \geq 0.7$) keeping the better performer. Use a meta-learner from logistic regression, LightGBM, or MLP until validation F1 plateaus.

2 Experimental Results and Conclusion

Datasets and protocol. D1 cardiac readmission: 101,766 samples, 50 features, strong imbalance. D2 disease classification: 10,000 samples, 9 features, approximately 23% missing values. Both share the same pipeline, a single search budget, 5-fold cross-validation, and final training with the best fold.

Table 1. Comparative Performance (Macro F1 score)

Dataset	Best ML	Best DL	Ensemble	Improvement	Selected components
D1	LightGBM (0.4302)	TabNet (0.3653)	**0.4833**	+5.3 pp	LightGBM, XGBoost, TabNet
D2	ExtraTrees (0.3141)	SAINT (0.3133)	**0.3532**	+3.9 pp	ExtraTrees, SAINT, MLP

Results. D1 achieves **0.4833** macro F1 with the stacked ensemble of LightGBM, XGBoost, and TabNet using logistic regression as meta-learner (**+5.3** percentage points over LightGBM alone). D2 reaches **0.3532** with ExtraTrees, SAINT, and MLP (**+3.9** percentage points over ExtraTrees alone). Gains increase with imbalance and feature heterogeneity, confirming the value of correlation pruning and dual selection.

Takeaways. Staged preprocessing, intersection based selection, and a small diverse stack consistently outperform the best single model. The systematic integration of complementary model families enhances both accuracy and stability. Future work includes multi layer stacking, automatic diversity targets, cost sensitive tuning, and temporal validation.

References

[1] Ravid Shwartz-Ziv and Amit Armon. Tabular data: Deep learning is not all you need. *arXiv preprint arXiv:2106.03253*, 2021.

[2] Valeriia Cherepanova, Roee Levin, Gowthami Somepalli, Jonas Geiping, C. Bayan Bruss, Andrew Gordon Wilson, Tom Goldstein, and Michael Goldblum. A performance-driven benchmark for feature selection in tabular deep learning. *arXiv preprint arXiv:2311.05877*, 2023.