



RaX-CHF: A Resource-Efficient and Explainable Small-Model Pipeline for Congestive Heart Failure Prediction from ICU-Scale Tabular Data

Ziwei Wang¹, Haoyun Zhang², Di Zhu³, Chen Xie⁴

¹Carnegie Mellon University, Pittsburgh, United States

²University of Pennsylvania, Philadelphia, United States

³Santa Clara University, Santa Clara, United States

⁴University of Massachusetts Amherst, Amherst, United States

ARTICLE INFO	ABSTRACT
Published Online: 29 January 2026 Corresponding Author: Chen Xie	Congestive Heart Failure (CHF) is a common and high-impact cardiovascular condition, and early risk identification can support proactive clinical decision-making. We present RaX-CHF, a resource-efficient and explainable small-model pipeline that predicts the presence of CHF using structured tabular features derived from MIMIC-III ICU data. RaX-CHF integrates demographics and aggregated laboratory measurements at the patient level and compares classical machine learning baselines (logistic regression, Random Forest, Gradient Boosting, and XGBoost) against a locally deployed small language model (Gemma) prompted with serialized patient features. On a dataset of roughly 44,000 adult patients with 115 attributes, tree-based models achieve the best performance, with XGBoost reaching 0.858 accuracy and 0.861 sensitivity (recall for CHF). The prompted Gemma baseline achieves 0.734 accuracy, near the majority-class baseline, highlighting the continued strength of compact tabular predictors for disease presence classification. We also summarize interpretability signals from linear coefficients and tree-based feature importance, and discuss practical risks related to missingness assumptions, de-identification policy, and deployment in safety-critical settings.
KEYWORDS: congestive heart failure, disease prediction, MIMIC-III, tabular machine learning, XGBoost, logistic regression, small language models, resource-efficient pipeline, interpretability	

I. INTRODUCTION

Congestive Heart Failure (CHF) is a major contributor to morbidity and mortality and remains a leading cause of hospitalization worldwide [6]. Early detection and risk stratification can help clinicians identify patients who may benefit from closer monitoring, follow-up testing, or earlier intervention. At the same time, real-world clinical deployment favors solutions that are (i) low cost, (ii) easy to operate on common infrastructure, and (iii) interpretable for clinicians and stakeholders.

In this work, our goal is to build a predictive model that can identify whether a patient has CHF from routinely available structured data. We focus on a

resource-efficient small-model approach: compact classical models trained on engineered tabular features, complemented by an optional small language model (SLM) that can produce human-readable reasoning. The motivation is to provide a practical workflow that can be used by healthcare professionals without requiring specialized GPU infrastructure or opaque large-model deployments.

We present RaX-CHF, a pipeline that extracts a patient-level dataset from MIMIC-III, cleans and imputes missing values under explicit assumptions, trains multiple ML classifiers, and compares them to a prompted Gemma SLM baseline. This work follows the broader RaX pipeline philosophy of emphasizing efficiency and interpretability in

safety-critical prediction tasks [2]. We emphasize sensitivity (recall for CHF) as a primary objective, since false negatives (missed CHF) are typically more concerning than false positives in screening-oriented settings.

- **RaX-CHF pipeline:** a compact, reproducible workflow for CHF presence prediction using demographics and aggregated lab statistics from ICU-scale EHR data.
- **ML vs. SLM comparison:** an empirical comparison between classical tabular predictors and a locally deployed Gemma SLM prompted with serialized features.
- **Interpretability-oriented analysis:** coefficient-based interpretation for logistic regression and feature-importance-based insights for XGBoost, plus a discussion of practical deployment risks.

II. RELATED WORK

A. Disease Prediction with Tabular EHR Data

Clinical risk prediction using structured EHR features is a long-standing area where logistic regression and tree-based ensembles remain strong baselines due to their robustness, efficiency, and interpretability. Prior studies have shown that logistic regression and tree-based ensembles are competitive and interpretable baselines for cardiovascular disease and heart failure prediction from EHR data [7]. In many hospital settings, the key value is not only accuracy but also calibration, sensitivity, and transparent feature effects. Similar principles have been validated in other safety-critical domains; for example, RaX-Crash shows that compact, interpretable tabular predictors can outperform prompt-based language models at city scale while remaining operationally practical [2].

B. Language Models on Structured Data

Recent work has explored using language models for tabular classification by serializing records into prompts. While this can simplify feature engineering and produce narrative explanations, prompt-only classification often struggles to surpass tuned classical baselines on medium-to-large structured datasets. In safety-critical medical domains, the ability to control errors and interpretability is especially important. From a systems perspective, resource-aware design and efficiency constraints are also central to real-world deployment on shared infrastructure [3].

III. DATA AND FEATURE SCHEMA

A. Data Source

We use MIMIC-III, a large, de-identified ICU database hosted on PhysioNet, containing data for over forty

thousand patients admitted to Beth Israel Deaconess Medical Center between 2001 and 2012 [1]. The database includes demographics, admissions, diagnoses (ICD-9), laboratory results, vitals, medications, and more. Access requires completing training aligned with privacy and human-subject guidelines.

B. Cohort Construction and Label

We construct a patient-level dataset with approximately 44,000 adult patients as observations. The binary label CHF is derived from ICD-9 diagnosis codes indicating congestive heart failure. The prevalence of CHF in our cohort is 9,829 patients (25.5%), implying a majority-class baseline accuracy of 74.5% if always predicting “no CHF.”

C. Features

Each patient has 115 attributes: 15 demographic features (e.g., age, gender, ethnicity, insurance, language, marital status) and 100 laboratory-derived features computed from 25 common blood tests, each aggregated as min, max, mean, and std over the patient stay (25 tests $\times$ 4 statistics = 100 features). These labs include electrolytes, renal function markers, hematology indices, coagulation tests, blood gas measurements, and glucose.

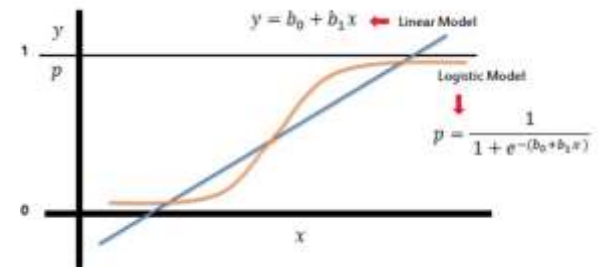


Fig. 1: Comparison between linear regression and logistic regression. Logistic regression maps a linear predictor into a bounded probability via the sigmoid function, enabling binary disease prediction.

D. Age De-identification and Cleaning Notes

A practical issue is that de-identification policies obscure precise ages for certain groups. We compute age from shifted dates where possible; patients above the de-identification threshold are imputed to a capped age (e.g., 92) to preserve a meaningful signal while respecting privacy constraints. We also remove features that directly leak outcomes (e.g., date of death, expiration flag) to avoid label leakage.

E. Missingness and Imputation Strategy

Laboratory values contain missing entries. Some missing values reflect boundary notation (e.g., “<10”, “>1”) and are mapped to boundary values. For models that require complete inputs (e.g., logistic regression, SGD), we impute missing values using medically “normal” reference values under the assumption that missing labs may indicate that clinicians did not suspect abnormality. For tree-based models, we keep missingness when supported by the implementation or apply simpler imputations, since tree models are generally more robust.

IV. METHODS

A. Problem Formulation

Let each patient be represented by a feature vector $\mathbf{x}_i \in \mathbb{R}^d$ and a label $y_i \in \{0, 1\}$ indicating CHF presence. The goal is to learn a classifier $f(\mathbf{x})$ that predicts CHF while prioritizing

sensitivity ($\text{Recall}_{\text{CHF}}$).

B. Logistic Regression

Logistic regression maps a linear combination of features to a probability using the sigmoid function:

$$p(y = 1 | \mathbf{x}) = \frac{1}{1 + e^{-(b_0 + \mathbf{b}^T \mathbf{x})}}.$$

This bounded probabilistic form supports threshold tuning to trade off false negatives and false positives.

We use L1 (lasso) regularization for feature selection and interpretability, tune the penalty parameter C via cross-validation, and adjust the classification threshold to improve sensitivity under class imbalance.

C. Stochastic Gradient Descent (SGD)

We also train a linear classifier via SGD with logistic loss. SGD can scale efficiently and shares similar preprocessing steps with logistic regression (e.g., scaling, imputation). Hyperparameters are tuned via cross-validation and we optionally adjust decision thresholds to prioritize sensitivity.

D. Tree-Based Models

We train Random Forest, Gradient Boosting (GBoost), and XGBoost classifiers. Tree-based models capture nonlinear interactions among labs and demographics and often outperform linear baselines on heterogeneous clinical features. They also provide feature importance measures that can support clinical interpretation.

E. Optimization Objective: Sensitivity

In disease screening contexts, sensitivity ($\text{Recall}_{\text{CHF}}$) is often prioritized. Clinically, it can be preferable to flag a

patient for additional testing (false positive) rather than miss a patient with CHF (false negative). Therefore, we report both accuracy and sensitivity and tune thresholds accordingly for linear models.

F. Prompted Small Language Model Baseline (Gemma)

To mirror our prior “small model” approach, we also evaluate a locally deployed Gemma model (gemma3:4b) using in-context learning (ICL). We serialize each patient record into a text prompt of the form: “Patient data: feature1: value1, feature2: value2, ... Does this patient have CHF? Reply with 1 or 0.” We first feed labeled examples (supervised feeding), then run a mock test with feedback, and finally evaluate on a held-out test split without feedback. This yields an accuracy of 73.4% in our evaluation, which is close to the majority baseline.

V. RESULTS

A. Logistic Regression Evaluation

The baseline logistic regression achieves 0.801 accuracy but relatively low sensitivity (0.423) at the default threshold of

0.5. After tuning (L1 regularization and threshold adjustment), sensitivity improves substantially (up to 0.750) with a modest change in accuracy.

B. Interpretability: Logistic Coefficients

With L1 regularization, many coefficients shrink to zero, and the remaining coefficients provide a human-readable ranking of influential factors. In our model, red blood cell related features and MCH-related measures appear among the strongest signals, aligning with clinical intuition that oxygen-carrying capacity and anemia-related patterns can be associated with CHF risk.

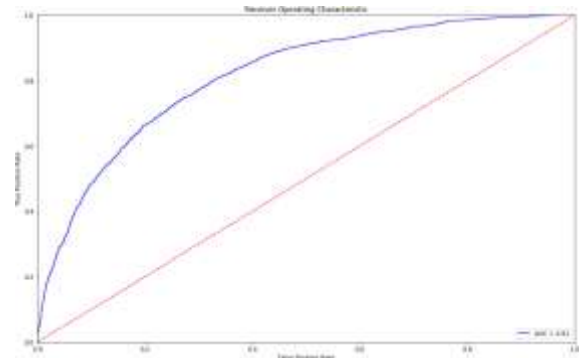


Fig. 2: Receiver Operating Characteristic (ROC) curve for the optimized logistic regression model. The AUC of 0.81 indicates strong discriminative ability despite class imbalance.

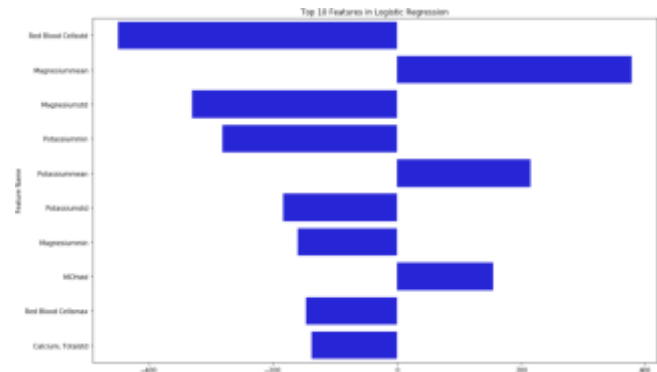


Fig. 3: Top ten features ranked by coefficient magnitude in the L1-regularized logistic regression model. Positive coefficients increase CHF likelihood, while negative coefficients reduce it.

C. Tree-Based Model Performance and Feature Importance

Tree ensembles provide the best overall results. Random Forest achieves 0.855 accuracy and 0.854 sensitivity, while XGBoost achieves 0.858 accuracy and 0.861 sensitivity (best overall). Feature importance from XGBoost highlights that extreme laboratory values—including glucose, coagulation markers (PTT), renal indicators (urea nitrogen), platelet counts, and oxygenation measures (pO2)—dominate CHF prediction.

D. ML vs. Gemma Comparison

Table I summarizes performance across our models.

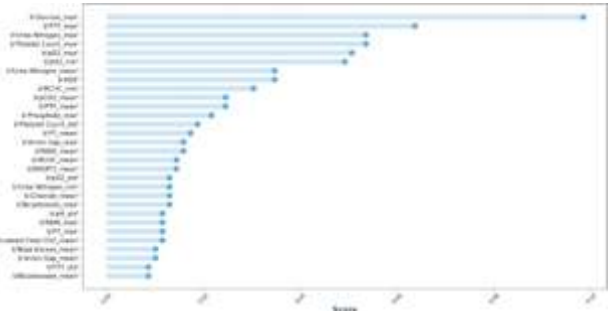


Fig. 4: Feature importance scores from the XGBoost classifier. Laboratory extreme values, including glucose, coagulation markers, renal indicators, and oxygenation measures, dominate CHF prediction.

VI. DISCUSSION AND RISKS

RaX-CHF reinforces a practical lesson: for medium-to-large structured clinical datasets, compact tabular models remain strong baselines, often outperforming prompt-only language models in predictive reliability. At the same time, SLMs can still be useful as a *communication layer*—for example, summarizing model outputs, drafting

Tree-based models substantially outperform the prompt-only Gemma baseline. Gemma’s accuracy (0.734) is close to the majority baseline (0.745), suggesting that without fine-tuning or stronger numeric reasoning constraints, the SLM struggles to extract robust decision boundaries from serialized tabular inputs. This observation is consistent with prior findings that prompt-only language models often underperform tuned tabular classifiers on medium-to-large structured datasets [8].

TABLE I: Performance Comparison: Machine Learning Models vs. Gemma Small Language Model

Model	Accuracy	Sensitivity (Recall _{CHF})	Notes
Logistic Regression (base)	0.801	0.423	threshold=0.50
Logistic Regression (tuned)	0.796	0.750	threshold=0.263
Logistic Regression + NearMiss	–	0.797	under-sampling
SGD (tuned)	–	0.729	log loss
Random Forest (best)	0.855	0.854	tree ensemble
Gradient Boosting (best)	0.859	0.858	boosted trees
XGBoost (best)	0.858	0.861	best overall
Gemma (SLM, ICL)	0.734	–	prompt-only baseline

clinician-facing explanations, or supporting patient education—as long as the primary decision is produced by a calibrated tabular predictor.

We also highlight key risks and assumptions:

- **Imputation assumptions:** imputing medically “normal” values may introduce bias if missingness is not consistent with the assumed mechanism.
- **Edge populations:** extreme physiological states (e.g., pregnancy, anemia) may require subgroup-specific modeling or additional features.
- **Comorbidities:** CHF often co-occurs with other conditions (diabetes, stroke), which may confound signals without explicit multi-disease modeling.
- **Legal and operational pressures:** clinical liability and over-testing incentives can shape data generation and missingness patterns.
- **De-identification effects:** age-capping and anonymization policies can reduce signal in vulnerable age groups and impact fairness.

VII. CONCLUSION AND FUTURE WORK

We presented RaX-CHF, a resource-efficient and explainable pipeline for CHF presence prediction using MIMIC-

III demographics and aggregated laboratory measurements. XGBoost achieves the best overall performance (0.858 accuracy, 0.861 sensitivity), and interpretability signals from logistic coefficients and XGBoost feature importance align with clinically meaningful lab patterns. A prompted Gemma small language model baseline reaches 0.734 accuracy, near the majority baseline, suggesting that prompt-only SLM classification is not yet competitive with tuned tabular models in this setting.

Future work includes: (i) building age-specific models (pediatric and elderly cohorts), (ii) multi-modal extensions using imaging or ECG signals, (iii) cross-hospital validation to test generalization, (iv) multi-disease prediction engines, and (v) fine-tuning small language models to better handle numeric cues and uncertainty while keeping the tabular model as the primary decision component.

REFERENCES

1. E. Johnson, T. J. Pollard, L. Shen *et al.*, “MIMIC-III, a freely accessible critical care database,” *Scientific Data*, vol. 3, 2016.
2. Zhu, C. Xie, Z. Wang, and H. Zhang, “RaX-Crash: A Resource Efficient and Explainable Small Model Pipeline with an Application to City Scale Injury Severity Prediction,” *arXiv preprint arXiv:2512.07848*, 2025. [Online]. Available: <https://arxiv.org/abs/2512.07848>
3. Z. Bao, D. Zhu, L. Jiang, S. Sheng, Z. Wang, and H. Zhang, “SLOFetch: Compressed-Hierarchical Instruction Prefetching for Cloud Microservices,” *arXiv preprint arXiv:2511.04774*, 2025. [Online]. Available: <https://arxiv.org/abs/2511.04774>
4. Zhang, S. Zhao, Z. Zhou, W. Zhang, and Y. Meng, “Domain-Specific RAG with Semantic Normalization and Contrastive Feedback for Document Question Answering,” in *Proc. 2025 7th Int. Conf. Internet of Things, Automation and Artificial Intelligence (IoTAAI)*, 2025, pp. 750–753, doi: <https://doi.org/10.1109/IoTAAI66837.2025.11213473>.
5. Z. Zhou, C. Zhao, X. Li, H. Zhang and R. Chang, “Diverse Stacking Ensemble for Attributing LLM Outputs via Relational Reasoning,” in *Proc. 2025 8th Int. Conf. on Computer Information Science and Application Technology (CISAT)*, 2025, pp. 1089–1092, doi: <https://doi.org/10.1109/CISAT66811.2025.11181963>
6. P. Ponikowski, A. A. Voors, S. D. Anker *et al.*, “2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure,” *European Heart Journal*, vol. 37, no. 27, pp. 2129–2200, 2016, doi: <https://doi.org/10.1093/eurheartj/ehw128>.
7. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, “Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1589–1604, 2018, doi: <https://doi.org/10.1109/JBHI.2017.2767063>.
8. S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang, and D. Sontag, “TabLLM: Few-shot classification of tabular data with large language models,” in *Proc. 26th Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2023, pp. 5549–5581.