ISSN: 3079-6342

Interpretable and Lightweight Predictive Modeling for Congestive Heart Failure Using ICU Electronic Health Records

Ziwei Wang[1], Haoyun Zhang[2] ,Di Zhu[3], Chen Xie[4]

[1]Carnegie Mellon University, Pittsburgh, United States

[2]University of Pennsylvania, Philadelphia, United States

[3]Santa Clara University, Santa Clara, United States

[4]University of Massachusetts Amherst, Amherst, United States

Abstract

Congestive Heart Failure (CHF) is a prevalent cardiovascular condition and a major contributor to hospitalizations and mortality worldwide. Early identification of CHF risk from electronic health records (EHR) can support proactive clinical monitoring and intervention. This paper presents an interpretable and lightweight predictive modeling workflow for CHF prediction using structured ICU data from the MIMIC-III database. A patient-level dataset of approximately 44,000 adult ICU patients with 115 demographic and laboratory-derived features is used to evaluate classical tabular machine learning models, including logistic regression, stochastic gradient descent classifiers, Random Forest, Gradient Boosting, and XGBoost. Tree-based ensemble models achieve the strongest performance, with XGBoost reaching an accuracy of 0.858 and sensitivity of 0.861 for CHF detection. The study also examines interpretability through logistic regression coefficients and feature-importance analysis, and compares these models with a prompted small language model baseline. The findings suggest that compact and interpretable machine learning models provide an effective and deployable approach for disease risk prediction using structured EHR data, especially in resource-constrained clinical environments.

Keywords

congestive heart failure; disease prediction; MIMIC-III; tabular machine learning; XGBoost; logistic regression; small language models; interpretability

1. Introduction

Congestive Heart Failure (CHF) is a major contributor to morbidity and mortality worldwide and remains a leading cause of hospitalization among cardiovascular diseases. Early detection and risk stratification can help clinicians identify patients who may benefit from closer monitoring, follow-up testing, or earlier intervention. With the increasing availability of electronic health record (EHR) data, machine learning approaches have become an important tool for identifying disease risk patterns using structured clinical features such as demographics, laboratory measurements, and physiological indicators. A large body of research has explored predictive modeling using structured EHR data. Classical machine learning models, including logistic regression and tree-based ensemble methods, remain widely used due to their robustness, computational efficiency, and interpretability. In many hospital environments, model interpretability and reliability are critical requirements because clinical stakeholders must understand and trust model outputs before integrating them into decision-making workflows. Recent work has also explored the use of language models for

ISSN: 3079-6342

structured prediction tasks by converting tabular records into textual prompts. Such approaches aim to simplify feature engineering and generate human-readable reasoning. However, prompt-based language model classification often struggles to outperform well-tuned classical models on medium-to-large structured datasets. From a systems perspective, resource efficiency and deployability are also important considerations when integrating predictive models into real-world infrastructure. This work studies an interpretable and lightweight machine learning framework for predicting congestive heart failure using structured ICU data derived from the MIMIC-III database. The goal is to develop a practical and interpretable workflow that can support clinical decision-making while remaining computationally lightweight and deployable on standard infrastructure without specialized hardware requirements. The main contributions are: (1) a practical machine learning workflow for CHF prediction using structured ICU EHR data, (2) a systematic comparison of multiple classical machine learning methods on a cohort of approximately 44,000 patients with 115 clinical features, (3) interpretability analysis through logistic regression coefficients and feature-importance measures, and (4) a comparison with a prompted small language model baseline.

2 Related Work

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 these approaches are competitive and interpretable baselines for cardiovascular disease and heart failure prediction from EHR data. Language models have also been explored for structured reasoning and tabular prediction tasks by converting structured records into textual prompts. Approaches such as TabLLM demonstrate that large language models can perform few-shot classification on tabular datasets by serializing structured features into natural language prompts. Other work explores structured reasoning, retrieval-augmented approaches, and attribution techniques for analyzing language model outputs. In safety-critical medical domains, however, predictive reliability, interpretability, and deployment efficiency remain particularly important.

3 Data and Feature Schema

The study uses 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. The database includes demographics, admissions, diagnoses (ICD-9), laboratory results, vitals, medications, and more. A patient-level dataset with approximately 44,000 adult patients is constructed. The binary label CHF is derived from ICD-9 diagnosis codes indicating congestive heart failure. The prevalence of CHF in the cohort is 9,829 patients (25.5%), implying a majority-class baseline accuracy of 74.5% if always predicting “no CHF.” Each patient has 115 attributes: 15 demographic features (for example, age, gender, ethnicity, insurance, language, and marital status) and 100 laboratory-derived features computed from 25 common blood tests, each aggregated as min, max, mean, and standard deviation over the patient stay. These labs include electrolytes, renal function markers, hematology indices, coagulation tests, blood gas measurements, and glucose. A practical issue is that de-identification policies obscure precise ages for certain groups. Age is computed from shifted dates where possible, and patients above the de-identification threshold are imputed to a capped age (for example, 92) to preserve a meaningful signal while respecting privacy constraints. Features that directly leak outcomes are removed to avoid label leakage. Laboratory values contain missing entries. Some missing values reflect boundary notation and are mapped to boundary values. For models that require complete inputs, missing values are imputed using medically “normal” reference values under the assumption that missing labs

ISSN: 3079-6342

may indicate that clinicians did not suspect abnormality. For tree-based models, missingness is retained when supported by the implementation or otherwise handled with simpler imputations.

4 Methods

Each patient is represented by a feature vector x and a binary label y indicating CHF presence. The goal is to learn a classifier that predicts CHF while prioritizing sensitivity (recall for CHF).

Logistic regression maps a linear combination of features to a probability through the sigmoid function and supports threshold tuning to trade off false negatives and false positives. The study uses L1 regularization for feature selection and interpretability, tunes the penalty parameter C via cross-validation, and adjusts the classification threshold to improve sensitivity under class imbalance. A linear classifier trained via stochastic gradient descent with logistic loss is also evaluated. SGD shares similar preprocessing steps with logistic regression, such as scaling and imputation, and is tuned via cross-validation. Tree-based models include Random Forest, Gradient Boosting, and XGBoost. These models capture nonlinear interactions among laboratory and demographic features and provide feature-importance measures that support clinical interpretation. Because disease screening applications prioritize catching positive cases, sensitivity is treated as the primary optimization objective. A prompted small language model baseline using a locally deployed Gemma model is also evaluated by serializing each patient record into a textual prompt and asking the model to predict whether the patient has CHF.

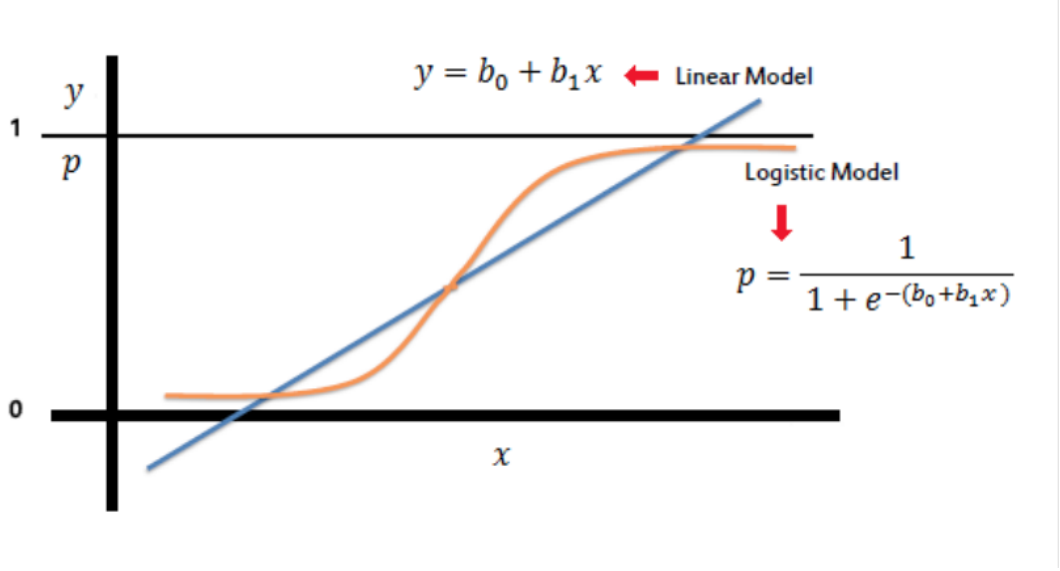

Figure 1. Comparison between linear regression and logistic regression. Logistic regression maps a linear predictor into a bounded probability via the sigmoid function, enabling binary disease prediction.

5 Results

The baseline logistic regression achieves 0.801 accuracy but relatively low sensitivity (0.423) at the default threshold of 0.5. After tuning with L1 regularization and threshold adjustment, sensitivity improves substantially (up to 0.750) with a modest change in accuracy. The optimized logistic regression model also achieves an AUC of 0.81, indicating strong discriminative capability despite the class imbalance in the dataset. With L1 regularization, many logistic regression coefficients shrink to zero, while the remaining coefficients provide a human-readable ranking of influential factors. 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. Tree ensembles provide the best overall performance. Random Forest achieves 0.855 accuracy and 0.854 sensitivity, Gradient Boosting achieves the highest accuracy of 0.859 with 0.858 sensitivity, and

ISSN: 3079-6342

XGBoost achieves 0.858 accuracy with the highest sensitivity of 0.861, making it the most balanced model for CHF detection in this evaluation. Tree-based models substantially outperform the prompt-only Gemma baseline. Gemma achieves 0.734 accuracy, close to the majority baseline (0.745), suggesting that without additional fine-tuning or stronger numeric reasoning constraints, the small language model struggles to extract robust decision boundaries from serialized tabular inputs.

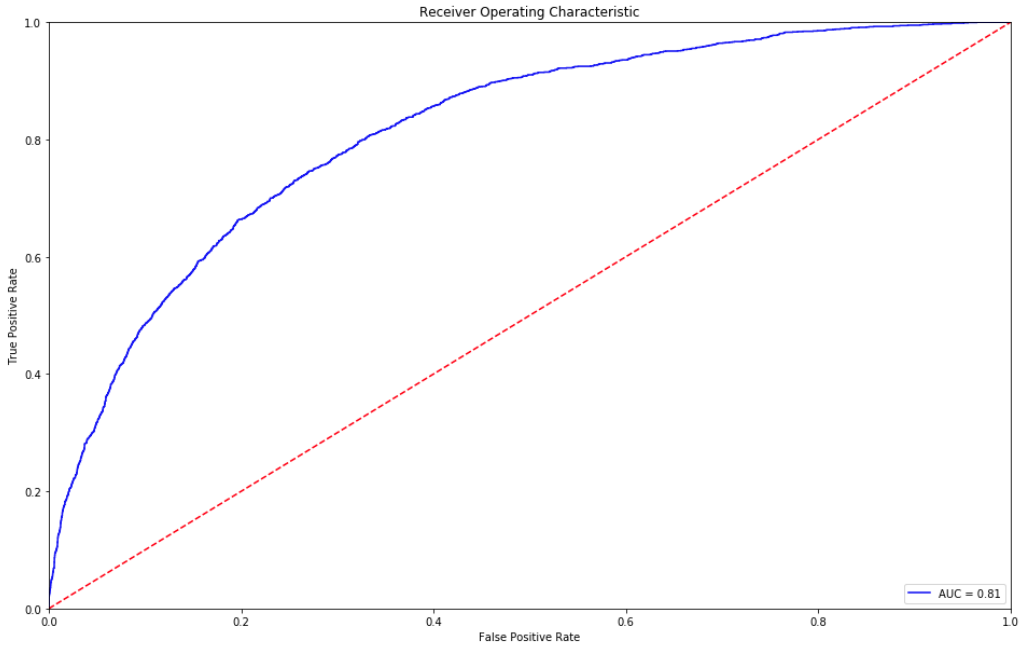

Figure 2 : Receiver Operating Characteristic (ROC) curve for the optimized logistic regression model. The AUC of 0.81 indicates strong discriminative ability despite class imbalance.

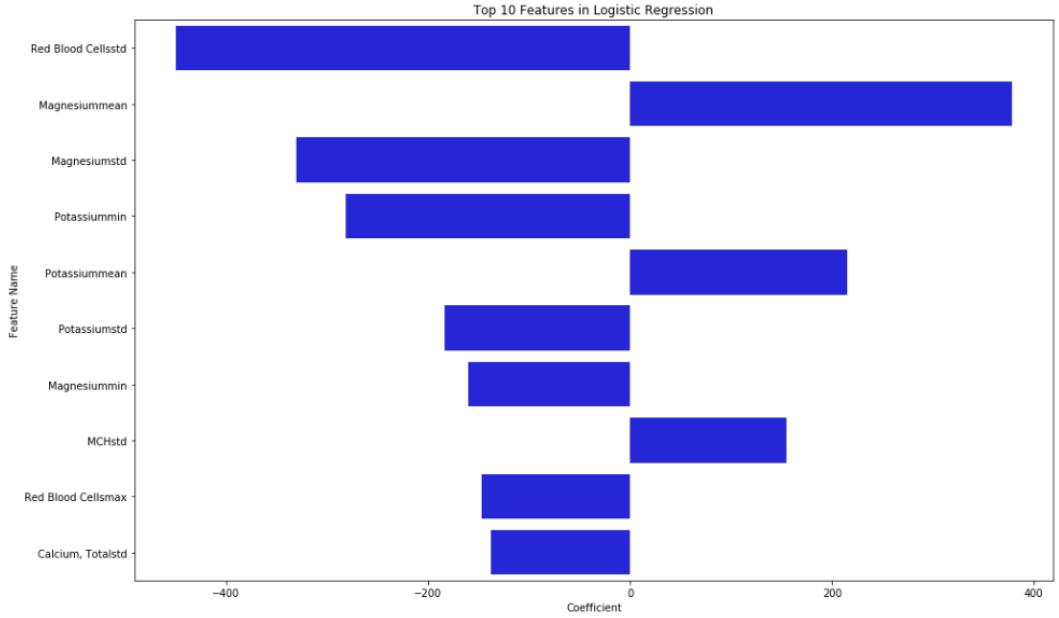

Figure 3: Top ten features ranked by coefficient magnitude in the L1-regularized logistic regression model. Positive coefficients increase CHF likelihood, while negative coefficients reduce it.

ISSN: 3079-6342

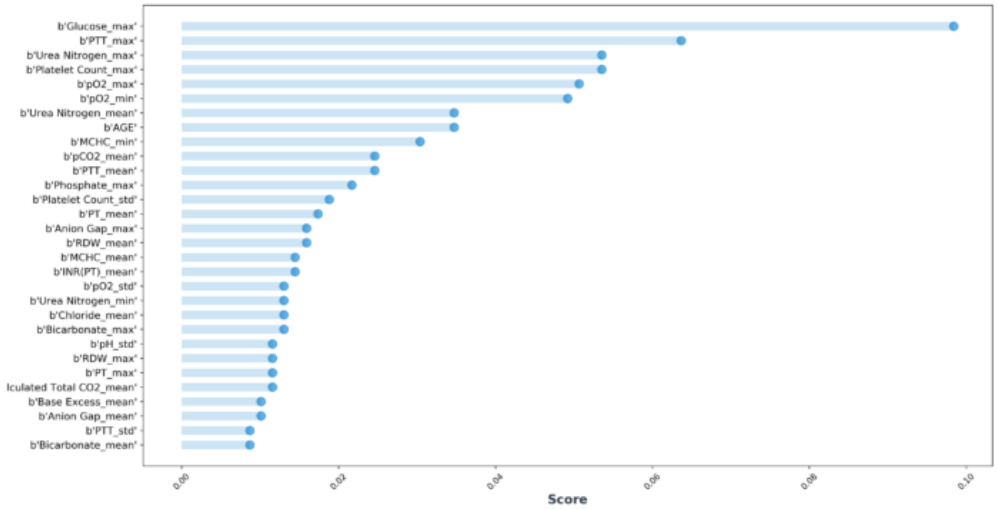

Figure 4: Feature importance scores from the XGBoost classifier. Laboratory extreme values, including glucose, coagulation markers, renal indicators, and oxygenation measures, dominate CHF prediction.

Table 1: Performance Comparison: Machine Learning Models vs. Gemma Small Language Model

Model	Accuracy	Sensitivity	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

6 Discussion

For medium-to-large structured clinical datasets, compact tabular models remain strong baselines and often outperform prompt-based language models in predictive reliability. At the same time, small language models can still be useful as a communication layer, for example by summarizing model outputs, drafting clinician-facing explanations, or supporting patient education, as long as the primary decision is produced by a calibrated tabular predictor. An important advantage of the proposed approach is computational efficiency. Unlike larger deep learning systems that may require GPU-based training or deployment, the models studied here can be trained and executed efficiently on standard CPU-based environments, making them more practical for integration into hospital IT workflows with limited computational resources.

ISSN: 3079-6342

Important limitations and risks include imputation assumptions, edge populations with unusual physiology, confounding from comorbidities, legal and operational pressures that can shape data generation, and de-identification effects that may reduce signal in vulnerable age groups. In addition, the study is based on a single publicly available dataset (MIMIC-III), so future work should evaluate the proposed workflow on additional hospital systems to assess generalizability across institutions.

7 Conclusion and Future Work

This paper presents an interpretable and lightweight predictive modeling framework 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 laboratory patterns. The prompted Gemma baseline reaches 0.734 accuracy, near the majority baseline, suggesting that prompt-only small language model classification is not yet competitive with tuned tabular models in this setting.

Future work includes building age-specific models for pediatric and elderly cohorts, extending the workflow to multimodal data such as imaging or ECG signals, testing generalization across institutions, developing multi-disease prediction engines, and exploring better ways to integrate language models for explanation and decision support while keeping the tabular model as the primary predictor.

References

[1] Johnson, A. E., Pollard, T. J., Shen, L., et al. “MIMIC-III, a freely accessible critical care database.” Scientific Data, vol. 3, 2016.

[2] Zhu, D., Xie, C., Wang, Z., and Zhang, H. “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.

[3] Bao, Z., Zhu, D., Jiang, L., Sheng, S., Wang, Z., and Zhang, H. “SLOFetch: Compressed-Hierarchical Instruction Prefetching for Cloud Microservices.” arXiv preprint arXiv:2511.04774, 2025.

[4] Zhang, H., Zhao, S., Zhou, Z., Zhang, W., and Meng, Y. “Domain-Specific RAG with Semantic Normalization and Contrastive Feedback for Document Question Answering.” In Proceedings of the 2025 7th International Conference on Internet of Things, Automation and Artificial Intelligence, 2025, pp. 750-753.

[5] Zhou, Z., Zhao, C., Li, X., Zhang, H., and Chang, R. “Diverse Stacking Ensemble for Attributing LLM Outputs via Relational Reasoning.” In Proceedings of the 2025 8th International Conference on Computer Information Science and Application Technology, 2025, pp. 1089-1092.

[6] Ponikowski, P., Voors, A. A., Anker, S. D., 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.

[7] Shickel, B., Tighe, P. J., Bihorac, A., and Rashidi, P. “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.

[8] Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., and Sontag, D. “TabLLM: Few-shot classification of tabular data with large language models.” In Proceedings of the 26th International Conference on Artificial Intelligence and Statistics, 2023, pp. 5549-5581.