Diagnosing the Silent Killer: AI-Powered Cardiovascular Disease Prediction

Savali Sandip Deshmukh
sdeshmukh@ucdavis.edu

Shivani Suryawanshi
ssuryawanshi@ucdavis.edu

Tanvi Mehta
tanmehta@ucdavis.edu

Introduction and Goals

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, accounting for approximately 17.9 million deaths annually, or 31% of all global deaths, according to the World Health Organization. These conditions often progress silently and unpredictably, making early and accurate diagnosis vital for improving outcomes. Coronary artery disease, a prominent form of CVD, poses a substantial risk even to asymptomatic individuals, highlighting the clinical urgency of noninvasive, data-driven screening techniques. Conventional diagnostic methods often rely on subjective assessment of symptoms or static threshold-based risk scores, which may not fully capture the nuanced interplay between demographic, behavioral, and clinical factors.

Our overarching goal is to investigate whether recent advances in neural architectures for tabular data—specifically, NODE (Neural Oblivious Decision Ensembles) and TabNet—as well as ensemble learning approaches, can yield meaningful improvements in predictive performance and model interpretability over traditional methods. By bridging methods from classical machine learning (e.g., Random Forest, Support Vector Machines, k-Nearest Neighbors, Gaussian Processes) and deep learning (e.g., RNNs, transformer-inspired attention mechanisms), we aim to assess both accuracy and practical utility of these models in clinical decision-making. We will also explore the integration of models across frameworks such as PyHealth, PyTorch, TensorFlow, and scikit-learn to construct ensemble learners that are optimized for healthcare applications.

Research Questions

To address the challenges in heart disease prediction, we have formulated the following research questions:

RQ1: Do deep learning architectures specifically designed for tabular clinical data (e.g., NODE, TabNet) outperform classical models in predictive accuracy for heart disease detection?

RQ2: What patterns and correlations among clinical parameters (e.g., age, cholesterol, blood pressure, ECG results) are most strongly associated with the severity of cardiovascular disease, and how can these insights inform clinical prioritization or resource allocation?

RQ3: Based on the most accurate predictive model, how would early intervention strategies shift if high-risk individuals were identified using machine learning-based screening rather than traditional clinical scoring tools?

The proposed research questions collectively address critical aspects of cardiovascular disease prediction. **RQ1** validates the technical performance of modern deep learning models against established diagnostic tools, ensuring their relevance in clinical practice. **RQ2** enhances interpretability by uncovering key feature correlations, allowing the model to function not just as a predictor but as a clinical insight generator—similar to how multivariate analyses have advanced stress detection. **RQ3** evaluates the

real-world impact of such models, examining how ML-based screening could reshape clinical decision-making compared to traditional risk-scoring methods.

Approach: Data, Methods, and Tasks

We will use the publicly available UCI Heart Disease dataset, accessible at https://archive.ics.uci.edu/dataset/45/heart+disease [1]. This dataset includes 76 total attributes collected from four different sources (Cleveland, Hungarian, Switzerland, and Long Beach), though the majority of analyses focus on the 13 most clinically relevant features such as age, sex, resting blood pressure, serum cholesterol, fasting blood sugar, ECG results, and exercise-induced angina [2]. We will begin with the Cleveland subset for consistency with previous studies and later explore cross-domain generalization using the full dataset.

Our methodological approach will unfold in two stages. The first stage involves implementing and benchmarking classical models including Random Forest, SVM, KNN, Gaussian Processes, and RNNs [3]. These models will be trained and validated using stratified k-fold cross-validation, and we will perform extensive hyperparameter tuning using grid search and Bayesian optimization. In the second stage, we will implement and train NODE and TabNet using PyTorch and TensorFlow, respectively. These models will be evaluated for both predictive accuracy and interpretability using built-in attribution mechanisms (such as TabNet's attentive masks) and post-hoc methods like SHAP and LIME. Additionally, we will construct model ensembles across different libraries, e.g., combining the predictive distributions of models built with PyHealth and sklearn to evaluate whether hybrid ensembles can exploit complementary strengths.

All stages of modeling will be accompanied by rigorous preprocessing, including normalization, imputation, outlier detection, and potential feature engineering. We will also assess the effects of data augmentation and synthetic oversampling on minority classes to account for potential label imbalance. Final model evaluation will be conducted using a hold-out test set and metrics including accuracy, F1-score, AUC-ROC, calibration error, and decision curve analysis to simulate real-world diagnostic value.

Timeline and Milestones

The project will follow an eight-week timeline. In the first two weeks, we will finalize literature review, complete exploratory data analysis, and preprocess the dataset. Weeks three and four will be dedicated to implementing and validating classical models, followed by training of advanced deep learning models (NODE and TabNet) in weeks five and six. Week seven will focus on ensemble learning, model comparison, and interpretability evaluation. Finally, in week eight, we will compile our findings into a comprehensive project report, accompanied by code, visualizations, and methodological documentation.

References

[1] Janosi, Andras, et al. "Heart Disease." UCI Machine Learning Repository, 1989, https://doi.org/10.24432/C52P4X.

[2] Detrano, R et al. "International application of a new probability algorithm for the diagnosis of coronary artery disease.", 1989, https://pubmed.ncbi.nlm.nih.gov/2756873/.

[3] Mohammad Alfadli, Khadijah, et al. "Feature-Limited Prediction on the UCI Heart Disease Dataset", 2024, https://www.sciencedirect.com/org/science/article/pii/S1546221822002892#ref-5.