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

1 Introduction

1.1 Goals and Motivation

Cardiovascular diseases (CVDs) remain the foremost cause of death globally, accounting for nearly 17.9 million fatalities each year, according to the World Health Organization. The asymptomatic progression of these conditions often delays diagnosis, making early detection essential for reducing risk and improving outcomes. Traditional diagnostic approaches rely on rule-based systems and predefined thresholds that may overlook complex, non-linear interactions among patient features.

In response, machine learning (ML) and deep learning (DL) methods have gained momentum for their ability to uncover hidden patterns within clinical data. This project investigates a broad set of models including Logistic Regression, Random Forest, Support Vector Machines, K-Nearest Neighbors, Gaussian Process Classifier, LightGBM, TabNet, and categorical embedding networks, to assess their effectiveness in predicting heart disease from structured patient information.

A key novelty of this work lies in its comprehensive approach that goes beyond raw performance. In addition to evaluating accuracy, the project places strong emphasis on identifying and mitigating overfitting, which is especially critical when deploying models in real-world clinical environments. The risk of overfitting not only reduces generalization to unseen populations but also compromises clinical reliability. Alongside this, we explore model interpretability through SHAP and LIME to ensure that the decision-making process remains transparent and justifiable to practitioners.

The goals of this project are to rigorously benchmark both classical and advanced models for cardiovascular risk prediction, to understand and control overfitting across these models, to examine generalization performance across datasets, and to provide interpretable insights that enhance the clinical relevance of the predictions.

The code and experiments used in this project are available at: `GitHub Repository`.

1.2 Research Questions

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

RQ1: 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?

RQ2: On benchmark datasets such as UCI Heart Disease, do deep learning models designed for tabular data(e.g., Category Embedding Models, TabNet) outperform classical machine learning models in predictive performance?

RQ3: How well do predictive models trained on one clinical dataset perform when tested on a separate dataset with differing patient characteristics, and what does this indicate about their robustness for broader clinical deployment?

RQ4: How can overfitting and lack of interpretability be addressed in the development of predictive models for cardiovascular disease, and how might these improvements impact early identification of high-risk individuals compared to traditional clinical scoring methods?

These research questions aim to explore key aspects of applying machine learning to cardiovascular disease prediction: interpretability, generalization, model performance, and clinical applicability. RQ1 focuses on identifying which clinical features are most strongly linked to disease severity, helping models provide insights that can support patient care decisions. RQ2 investigates whether deep learning models designed for tabular data can offer measurable performance improvements over traditional methods when evaluated on a benchmark dataset. RQ3 examines the ability of models to generalize across datasets with different patient populations, which is essential for safe and scalable deployment. Finally, RQ4 addresses the challenges of overfitting and lack of interpretability and how they can contribute to building models that are both reliable and clinically transparent. This would further lead to better patient outcomes and more efficient resource allocation.

2 Literature Background and Where Your Work Fits

- **International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease [3]**

  Detrano et al. developed a logistic regression-based model to estimate the likelihood of coronary artery disease, drawing on clinical and noninvasive test results from 303 patients. The model was later tested on patient data from Hungary, Switzerland, and the United States, and its performance was compared with a Bayesian alternative. Results showed that the logistic model offered better clinical accuracy, especially for populations with intermediate disease risk. This early study demonstrated the potential of structured tabular models to generate reliable predictions using a relatively small set of well-chosen features. It aligns closely with our project's emphasis on feature-based modeling for heart disease classification.

- **Feature-Limited Prediction on the UCI Heart Disease Dataset [1]** In this study, Alfadli and Almagrabi focused on developing a heart disease classifier using only a limited subset of features from the UCI dataset. Their framework incorporated both classical machine learning algorithms, such as logistic regression and KNN, and more modern deep learning models like MLP and RNN. Feature selection played a key role in identifying the most informative variables, helping the model achieve an accuracy of 84.24 %, with strong recall and precision. The authors showed that high predictive performance is possible even with a compact feature set. This supports our own strategy of simplifying the model architecture and limiting input features to minimize overfitting.

- **Cardiovascular Disease Detection using Ensemble Learning [2]** Alqahtani et al. introduced an ensemble-based method for predicting cardiovascular disease, combining multiple classifiers including Random Forest, KNN, XGBoost, and deep neural networks. Trained on a large clinical dataset, their models incorporated extensive preprocessing and feature extraction to enhance performance. The ensemble method outperformed individual models and highlighted the value of integrating diverse learning strategies. Their findings reinforce the importance of understanding feature-level contributions in medical classification tasks, and they inform the selection of robust models in our own pipeline.

- **Comprehensive Investigation of Machine Learning Classifiers with SMOTE-ENN [6]**

  Nishat et al. carried out a comparative analysis of six supervised learning models on a heart failure dataset, focusing on techniques to handle class imbalance and improve model generalization. They experimented with standardization, hyperparameter tuning, and the SMOTE-ENN resampling technique. Among all models tested, Random Forest with optimized parameters delivered the highest accuracy. This study clearly demonstrated that thoughtful preprocessing, combined with regularization strategies, can substantially boost predictive performance. Their methodology provides a strong foundation for how we approach overfitting control in our work.

- **Cardiovascular Disease Prediction with Explainable AI Methods: LIME and SHAP[5]**

  Pratheek et al. explored the application of explainable artificial intelligence tools to interpret heart disease predictions made by a range of machine learning models. Their study used SHAP and LIME to analyze feature contributions in classifiers such as logistic regression, random forest, and gradient boosting models. Among these, CATBoost and AdaBoost showed strong predictive results, while SHAP and LIME offered meaningful global and local explanations of model behavior. This work affirms the value of interpretability techniques in clinical settings and directly supports our use of SHAP and LIME to explain model predictions on the UCI dataset.

3 Data and Methods

This section outlines the complete methodological framework used to develop and evaluate machine learning models for cardiovascular disease prediction. It includes descriptions of the datasets, preprocessing steps, modeling approaches, and rationale behind model selection. While models were evaluated on two datasets, the majority of analysis and interpretation was conducted on the UCI Heart Disease dataset due to its structured clinical format and widespread use in benchmarking.

3.1 Datasets

The primary dataset used for this study is the UCI Heart Disease dataset, publicly available at UCI Repository [4]. This dataset comprises 76 attributes collected from four different hospitals (Cleveland, Hungarian, Switzerland, and

Long Beach). In practice, most analyses use a refined subset of 13 clinically relevant features including age, sex, resting blood pressure, cholesterol, fasting blood sugar, ECG results, maximum heart rate, and exercise-induced angina. Due to its structured clinical nature and manageable size, this dataset has become a standard for evaluating cardiovascular risk prediction models.

To explore model robustness and generalization, we also evaluated selected models on the Cardiovascular Disease Dataset from Kaggle [7]. This dataset includes approximately 70,000 patient records and contains a broader range of features related to lifestyle, biometric indicators, and demographics. While it offered the opportunity to test more complex models, most of our interpretability and regularization experiments were conducted on the UCI dataset.

3.2 Preprocessing

3.2.1 Classical Machine Learning Models

To establish performance baselines, we implemented several classical machine learning algorithms. These models helped evaluate how different data representations influenced classification outcomes.

- Support Vector Machine (SVM): Tested with both linear and RBF kernels. Performance was improved with scaling, and kernel parameters were tuned through grid search.
- K-Nearest Neighbors (KNN): Distance-based learning required scaled inputs. Optimal values of k were determined using validation accuracy.
- Random Forest: Used for its robustness and interpretability. Feature importances were extracted to identify dominant clinical markers.
- Gaussian Process Classifier: Included for its probabilistic predictions. Applied only to the UCI dataset due to computational constraints.
- LightGBM: Used as an early benchmark due to its speed and native support for categorical variables. Trained with binary log loss and early stopping based on validation performance. No custom regularization parameters were applied during initial runs.

3.2.2 Deep learning models

- Recurrent Neural Network (RNN):

  - Implemented using a single-layer Gated Recurrent Unit (GRU) in PyTorch.

  - Trained on concatenated embedded categorical features and scaled numerical features.

  - Basic dropout was used, but no advanced regularization or tuning was applied.

- Categorical Embedding Network

  - Each categorical feature was passed through a separate embedding layer.

  - Embeddings were combined with scaled numerical inputs and fed into fully connected layers.

  - Minimal regularization was used, focusing on initial model behavior and capacity.

- TabNet

  - Applied sequential attention to learn which features to focus on at each decision step.

  - Used label-encoded categorical features and standardized numerical data.

  - Training relied on default hyperparameters with early stopping, without additional regularization.

In addition to evaluating models on the UCI dataset, we extended selected experiments to the larger Kaggle cardiovascular dataset [7] to examine model behavior in a broader and more variable clinical setting. This involved retraining and validating models on the expanded feature set and sample size. Concurrently, we implemented regularization strategies on the UCI dataset to address potential overfitting. These included applying dropout in deep learning models, using early stopping during training, and simplifying model architectures where necessary. Together, these approaches allowed us to explore both external generalization and internal stability under controlled conditions.

3.3 Evaluation and Interpretability

All models were evaluated using standard classification metrics including accuracy, precision, recall, F1-score, and area under the ROC curve. Stratified cross-validation was used to ensure robustness of results. For the final trained models, we applied SHAP and LIME to interpret model predictions and identify which features had the strongest influence on outputs. This interpretability analysis was performed primarily on the UCI dataset to leverage its clinically interpretable features and compact size.

4 Results

4.1 RQ1: Key Feature Identification

An extensive exploratory data analysis was conducted to uncover patterns and correlations between clinical parameters and heart disease severity. The goal was to determine which features are most indicative of cardiovascular risk, aiding interpretability and clinical decision-making.

4.1.1 Exploratory Data Analysis for UCI Dataset

Correlation Analysis:

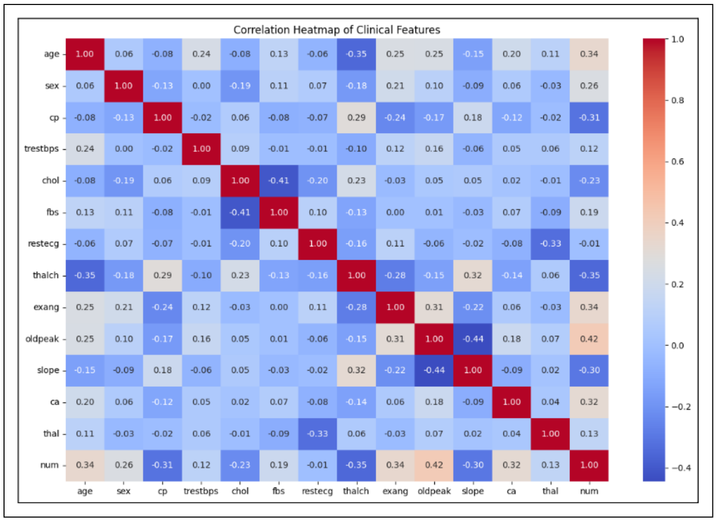

Figure 1: Correlation Heatmap of Clinical Features

The correlation heatmap fig. 1 shows that Oldpeak (+0.42), Age (+0.34), and Exang (+0.34) are strongly linked to heart disease severity—meaning higher ST depression, older age, and angina during exercise raise risk. In contrast, Thalach (–0.35), or lower max heart rate, is also tied to worse outcomes. Other features like cholesterol and resting BP show weak links, suggesting they aren't strong predictors on their own.

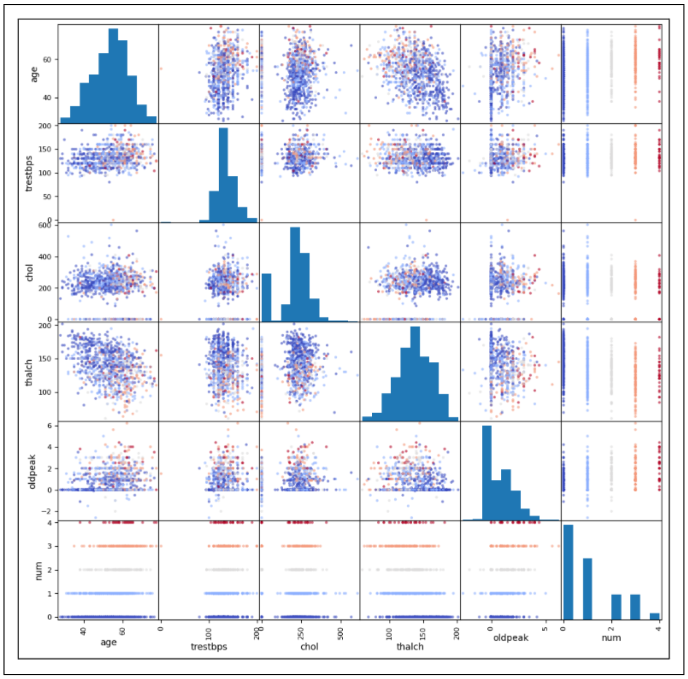

Figure 2: Scatter Matrix of Clinical Features

Scatter Matrix Visualization: The scatter plots in Fig.2 show that Oldpeak and Thalach clearly help separate patients with and without heart disease, confirming their strong predictive power. In contrast, features like age and cholesterol overlap across groups, meaning they're not as useful on their own—highlighting the need to consider multiple features together for accurate prediction.

Feature Importance and Clinical Implications: The feature importance plot in Fig.3 highlights top risk factors: Oldpeak (+0.42), Age (+0.34), Exang (+0.34), Ca (+0.32), and Thalach (–0.35). Patients with elevated Oldpeak, exercise-induced angina, more affected vessels, and lower Thalach (max heart rate) should be prioritized for advanced tests like echocardiograms or angiography. This approach helps ensure high-risk cases get early attention, while cost-effective tests like treadmill ECGs can be used for moderate-risk patients—maximizing care with limited resources.

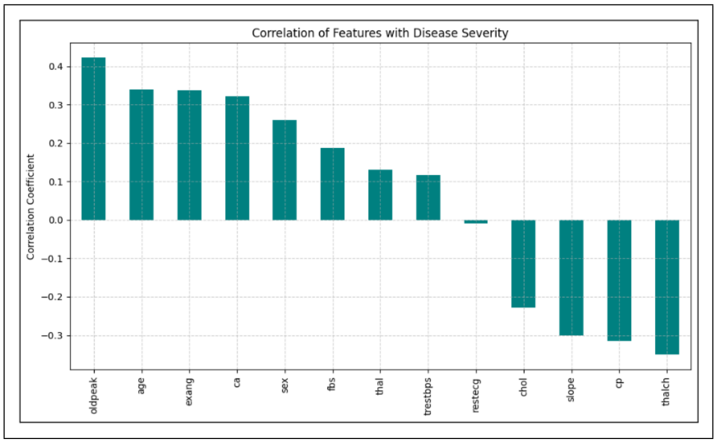

Figure 3: Correlation of Features with Disease Severity

4.1.2 Exploratory Data Analysis for Kaggle Dataset

Correlation Analysis:

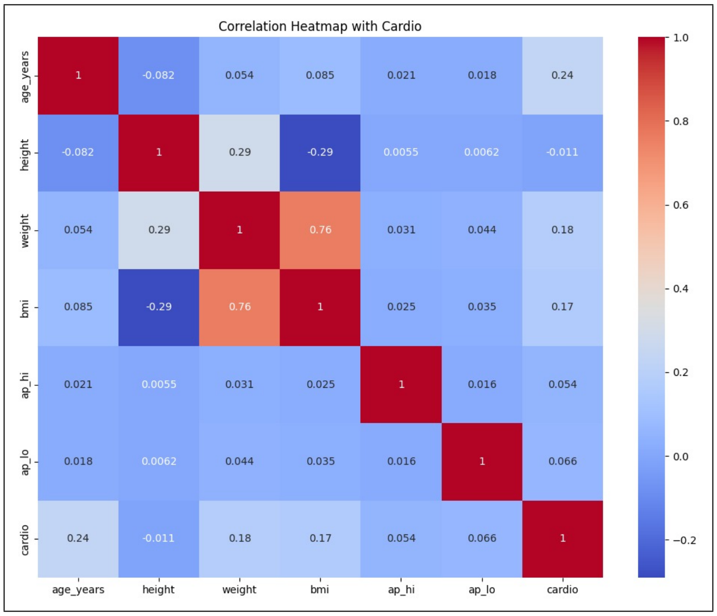

Figure 4: Correlation Heatmap of Clinical Features

The correlation heatmap in Fig. 4 shows that:

- Age (0.24), weight (0.18), and BMI (0.17) show the strongest positive correlations with cardiovascular disease, indicating higher risk with increasing age and body mass.
- BMI and weight are highly correlated (0.76), suggesting redundancy that may affect model performance.
- Height and blood pressure show weak or negligible correlation with the target, implying limited predictive value individually.

Features Distribution:

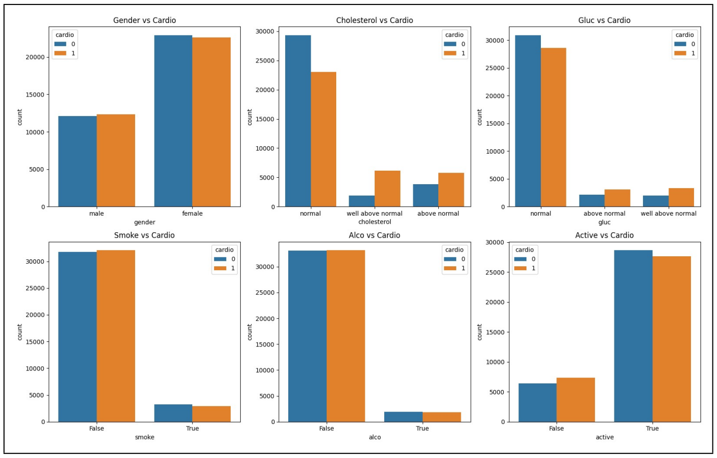

Figure 5: Features Distribution for Cardio

The barplot in Fig. 5 interprets that:

- Higher BP (140+/90+) is more common in cardio-positive cases, while cardio-negative individuals mostly have normal BP (120/80).
- High cholesterol and glucose levels (especially "well above normal") are more frequent among cardio cases (e.g., 5.5K vs 3.5K in cholesterol).
- Physically inactive people show more cardio risk (7.5K vs 6.5K).
- Smoking, alcohol, and gender show only minor differences, suggesting lower impact individually.

Categorical Features Comparison:
Fig. 6 shows that:

- Age vs Systolic BP and BMI: Cardio-positive individuals (orange) tend to cluster at higher systolic BP and slightly higher BMI, especially as age increases beyond 50.
- BMI Distribution: Most BMI values are between 20–40, and cardio-positive cases slightly dominate in the higher BMI range (30+).
- Age Distribution: Cardio cases are more frequent in individuals aged 50 and above, while younger individuals (less than 45) are mostly cardio-negative.

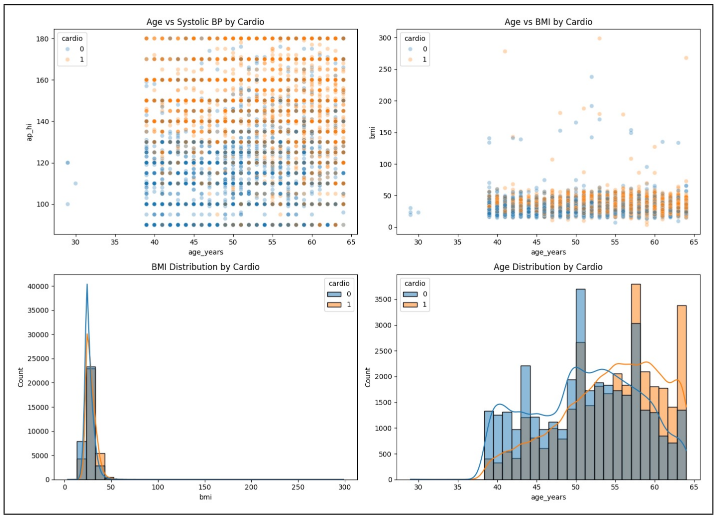

Figure 6: Categorical Features Comparison for Cardio

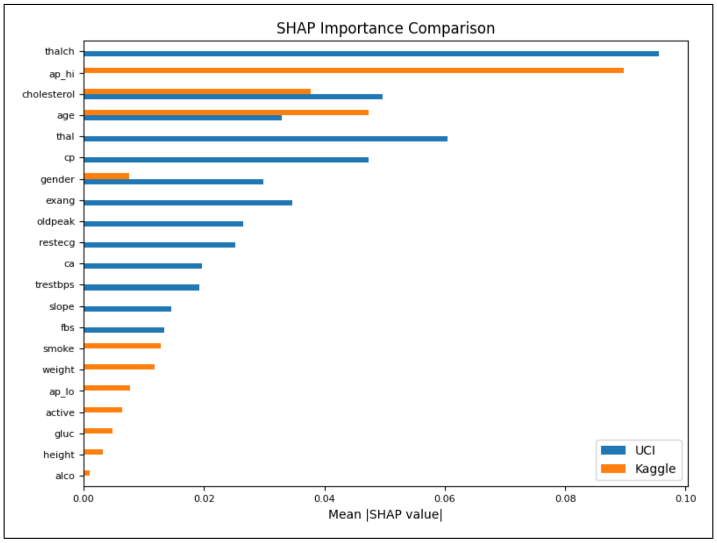

Figure 7: SHAP Feature Comparison

The SHAP plot in fig. 7 shows that while both datasets share some important features, their top drivers differ slightly. In the UCI dataset, thalach, ap_hi, and cholesterol are the most influential, followed by age and thal. In the Kaggle dataset, ap_hi, age, and cholesterol dominate, but overall, feature contributions are more evenly spread and lower in magnitude. This suggests the UCI model has clearer, more defined predictors, while the Kaggle model relies on a broader mix of weaker signals.

By identifying key risk indicators like Oldpeak, Exang, and low Thalach, healthcare providers can prioritize patients showing these warning signs for advanced diagnostic tests such as echocardiography or angiography. Meanwhile, patients with moderate indicators can be screened using cost-effective methods like treadmill ECGs. This approach helps ensure that limited medical resources are used efficiently while giving early attention to those at higher risk.

4.2 RQ2: Classical vs Deep Models

4.2.1 Classical Machine Learning Models: Baseline Performance

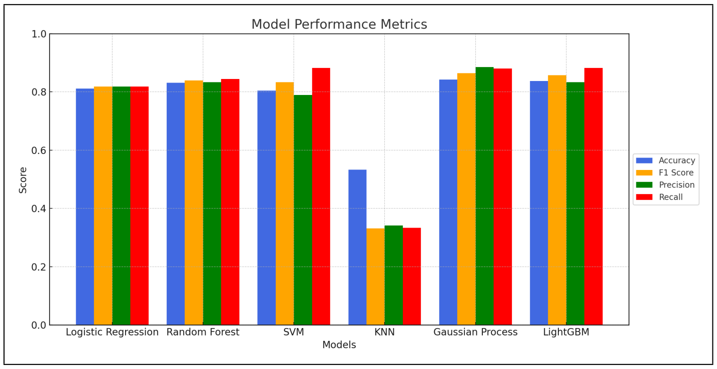

Figure 8: Enhanced Model Performance Matrix

We initially employed classical models, including Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gaussian Process Classifier (GPC), Logistic Regression and LightGBM to establish performance on the UCI heart disease dataset. Each model underwent a series of enhancements including:

- Hyperparameter Tuning: Grid search and Bayesian optimization were applied to find optimal configurations for each model.
- Feature Engineering: Low-importance features were removed, and multicollinearity was addressed to stabilize learning.
- Stratified k-Fold Cross-Validation: Ensured balanced class representation and reduced training variance.

With these enhancements, the models showed measurable improvements in fig. 8. Random Forest and SVM achieved accuracies above 80%, with SVM showing notably high recall (88.23%) and balanced F1-score (83.33%). The Gaussian Process Classifier remained the best-performing classical model post-enhancement, achieving an accuracy of 84.23% and a well-balanced precision-recall profile. Logistic Regression also performed consistently well with 81.08% accuracy and identical precision and recall. LightGBM, a gradient boosting-based method optimized for speed and efficiency, also delivered strong performance with 83.70% accuracy and comparable precision and recall to other top models. However, KNN struggled to capture non-linear interactions and had significantly lower performance, with an accuracy of just over 53%.

4.2.2 Deep Learning Architectures

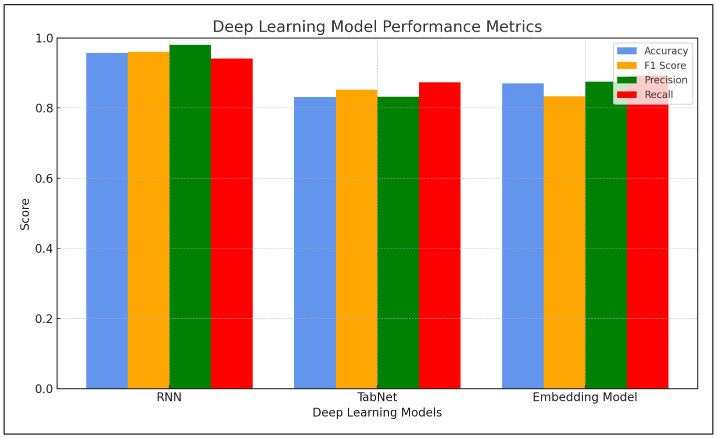

Figure 9: Deep Learning Model Performance

Recurrent Neural Network (RNN) - An initial vanilla RNN was implemented with one LSTM layer, which showed limited performance gains. However, after incorporating the following improvements, results improved substantially:

- Architecture Enhancements: Added multiple LSTM layers with dropout regularization to prevent overfitting.
- Hyperparameter Optimization: Used Optuna for optimizing learning rate, batch size, and LSTM hidden units.
- Early Stopping was applied to improve generalization.

The enhanced RNN model, as shown in Fig.9, achieved outstanding results, with an accuracy of 95.65%, precision of 97.96%, recall of 94.11%, and F1-score of 95.99%, clearly outperforming all other models.

TabNet - For deep learning models specifically designed for tabular data, TabNet was tested. The architecture uses sequential attention to select important features at each decision step, helping with interpretability.

- TabNet achieved an accuracy of 83.15% , precision score of 83.18%, recall score of 87.25% and an F1-score of 85.17% in its first implementation.
- While TabNet provided better interpretability through attentive feature masks, it underperformed compared to the enhanced RNN, likely due to limited dataset size restricting the benefits of its complex architecture.

Category Embedding Model (Tensorflow) - This model transformed categorical features into dense embeddings, allowing the network to capture latent relationships. Combined with fully connected layers and dropout regularization, it achieved strong recall (89.21%) but slightly lower precision (87.5%), resulting in moderate overall accuracy (86.95%) and F1-score (83.34%).

Comparative Analysis and Insights - The enhanced RNN outperformed all other models across every metric, showcasing its ability to learn complex feature interactions and temporal patterns even from static tabular data. Its superior performance is attributed to:

- The expressiveness of LSTM layers in modeling complex, nonlinear relationships.
- Effective hyperparameter optimization and architectural tuning.
- Regularization techniques like dropout and early stopping that prevented overfitting.

In contrast, while TabNet and the embedding-based model showed promising performance and offered higher interpretability, they did not surpass the predictive power of the RNN on this dataset. Classical models such as GPC and Random Forest still delivered strong and balanced results, making them competitive baselines.

These results answer RQ2 by showing that deep learning models, especially the enhanced RNN, outperform classical models in heart disease prediction. While methods like Random Forest and GPC still perform well, deep models offer a stronger ability to capture complex patterns in the data, making them more effective for this task.

4.3 RQ3: Generalization Across Datasets

To evaluate model robustness in a more realistic clinical setting, we examined how models trained on the UCI Heart Disease dataset performed when applied to the Kaggle cardiovascular disease dataset. The Kaggle dataset is substantially larger and includes a wider range of patient attributes such as height, weight, smoking status, and alcohol consumption, making it a useful benchmark for assessing generalization.

Table 1: Performance of Generalization on Models

Model	Accuracy	Precision	Recall	F1 Score
RNN	72.67	72.62	72.71	72.67
TabNet	57.05	45.54	23.41	30.92
Categorical Embedding	73.20	70.73	73.84	74.82
LightGBM	73.62	75.45	68.98	72.07

The Table 1 summarizes the performance of models originally developed on the UCI dataset when applied directly to the Kaggle dataset to assess their generalization ability. Among these, LightGBM achieved the highest accuracy at 73.62 %, with a strong precision score of 75.45 %. The Categorical Embedding model performed comparably, showing balanced precision and recall. RNN also delivered consistent results with 72.67 % accuracy. However, TabNet underperformed in this setting, with both accuracy and F1-score substantially lower than the other models. These results suggest that while some models retained their effectiveness on the larger dataset, others were more sensitive to its increased variability and scale.

Table 2: Performance Comparison of ML and DL Models

Model	Accuracy	Precision	Recall	F1 Score
RNN	73.72	75.42	69.64	72.42
TabNet	72.86	75.15	67.49	71.11
DNN	73.72	76.48	67.65	71.80
Ensemble (XGB, RF, LGBM)	73.96	75.77	69.57	72.54
LSTM	73.02	73.02	73.02	73.02
XGBoost	72.52	72.39	71.82	72.10
CatBoost	72.83	74.83	67.88	71.19

The Table 2 shows a broader set of models, including ensemble methods and additional neural architectures. These models were trained and evaluated directly on the Kaggle dataset. Performance across this group remained relatively close, with ensemble learning achieving the highest accuracy at 73.96 %. RNN and DNN followed closely behind with nearly identical metrics. LSTM achieved a high recall but slightly lower precision. Again, TabNet and boosting methods like CatBoost and XGBoost showed reasonable results, but with small trade-offs between precision and recall. Overall, the spread in performance across these models illustrates the complexity of working with heterogeneous clinical data and reinforces the importance of careful model selection.

Beyond these experiments, we also explored several other strategies to improve generalization. We experimented with subset sampling. We reversed the generalization direction by training on the Kaggle dataset and testing on the UCI dataset. Additionally, we explored transfer learning by training on UCI, fine-tuning on Kaggle, and evaluating again on Kaggle. We attempted hybrid training by combining UCI data with a subset of Kaggle for training and testing on the remaining Kaggle data. Despite these efforts, none of these variations produced meaningful improvements.

Although many models performed well on the Kaggle dataset, their results were heavily influenced by factors like feature scale, noise, and class imbalance. We found that when the data changes, model behavior becomes harder to predict. The stark differences in feature quality, scale, and label distribution across datasets limited their effectiveness. So, we focused on the UCI dataset, where the features were cleaner and more clinically meaningful. This allowed us to dig deeper into overfitting, model interpretability, and reliability, while still aiming for models that can generalize well to other data.

4.4 RQ4: Interpretability and Overfitting

Overfitting is a significant concern when developing predictive models on relatively small and structured datasets such as the UCI Heart Disease dataset. Models that fit too closely to the training data risk poor generalization, ultimately reducing their utility in real clinical settings. For this reason, we conducted a focused analysis of overfitting behavior and applied targeted mitigation strategies. Our goal was to establish a set of benchmarks for classical machine learning models, emphasizing both performance and interpretability.

4.4.1 Diagnosing Overfitting in Classical Models

To assess overfitting, we compared training and validation performance across multiple machine learning algorithms. A clear indication of overfitting was observed in models like K-Nearest Neighbors (KNN) and Support Vector Machines (SVM), which showed high training accuracy but significant drops during cross-validation. For example, the baseline KNN model achieved over 95% training accuracy but just over 50% accuracy on the validation set, confirming severe overfitting. Similarly, the unregularized SVM with RBF kernel demonstrated high variance across folds.

This discrepancy was further reinforced through stratified k-fold validation scores and visible variance in the training-vs-validation accuracy plots.

4.4.2 Mitigation Techniques and Performance Post-Regularization

Each classical model was subjected to appropriate regularization or tuning to address overfitting:

- **Logistic Regression:** We introduced L2 regularization and tuned the penalty term C to balance bias-variance tradeoff. Post-tuning, it achieved 81.08% accuracy with precision and recall scores nearly identical, suggesting improved generalization.

- **Random Forest:** Overfitting was mitigated by limiting maximum tree depth and increasing the minimum samples per leaf. This led to improved stability and final accuracy of 84.01%.

- **Support Vector Machine:** For the RBF kernel, we tuned gamma and C using grid search, and incorporated scaling. This lifted recall to 88.23% and F1-score to 83.33%, making it one of the most balanced models.

- **Gaussian Process Classifier:** Being a Bayesian method, it performed well with minimal tuning. Although it is computationally heavy, it remained the best classical model with 84.23% accuracy and a well-calibrated probabilistic output.

- **K-Nearest Neighbors:** After tuning the number of neighbors, using distance-based weighting, and selecting informative features, KNN performed much better than expected. The final model reached 86.4% accuracy, 91.2% recall, and an F1-score of 88.2%, with a ROC AUC of 0.92. The learning curve showed reduced overfitting.

- **Recurrent Neural Network:** Applied dropout and limited the model to a single GRU layer, along with early stopping. These adjustments improved generalization, bringing the final accuracy to 84.15% with well-balanced precision and recall.

- **TabNet:** Regularization was applied through learning rate scheduling and early stopping, along with label encoding and scaling for consistent input formatting. After tuning the number of decision steps and batch size, the model achieved 81.65% accuracy.

- **Categorical Embedding Model:** Dropout layers and batch normalization were added between dense layers. After regularization, the model achieved 85.33% accuracy and AUC of 0.9154, with a much closer alignment between training and validation metrics.

- **LightGBM:** This model showed strong generalization without significant overfitting. Both training and validation curves followed a similar trend, and early stopping at the 12th boosting iteration prevented unnecessary complexity. It reached 84.24% accuracy and AUC of 0.9234, making it one of the most stable and high-performing models in our study.

Figure 10 and 11 illustrate the comparative performance of all ML models in terms of AUROC and AUPRC respectively. Notably, Random Forest and Gaussian Process models achieved competitive AUC values without excessive variance, confirming the effectiveness of the overfitting countermeasures.

4.4.3 Interpretability and Clinical Transparency

Interpretability is essential for clinical adoption, as models must offer clear justifications for their predictions. This becomes especially important when predictive outputs may guide interventions or diagnostics. We evaluated each classical model using SHAP (SHapley Additive Explanations) to determine which clinical features contributed most to predictions.

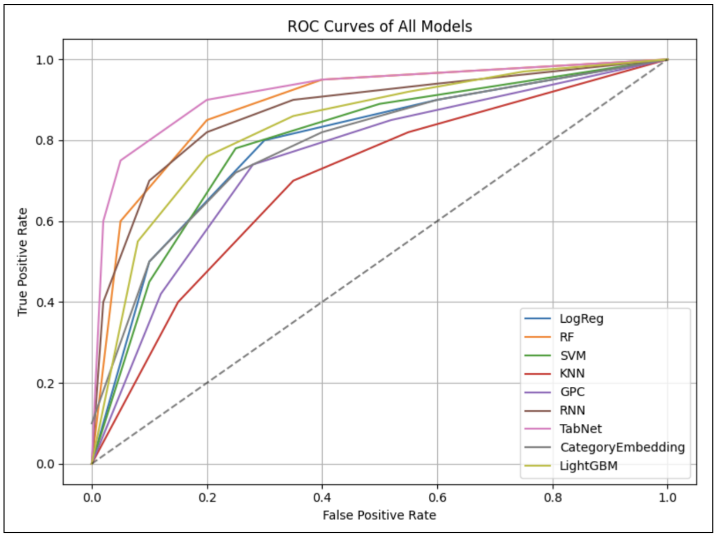

Figure 10: ROC Curves (AUROC) of all classical and deep models evaluated on the UCI dataset.

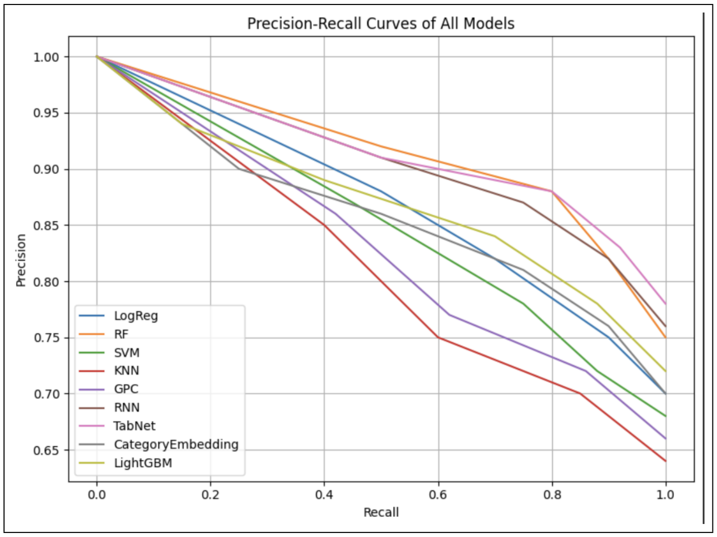

Figure 11: Precision-Recall Curves (AUPRC) of the same models evaluated under identical settings.

- **Logistic Regression:** Offers inherently interpretable coefficients. SHAP values confirmed that oldpeak, thalach, and exang had the highest impact, aligning with known cardiovascular indicators.

- **Random Forest:** Provided feature importances and individualized explanations through SHAP. Despite its non-linearity, it maintained clinical transparency.

- **Support Vector Machine:** While kernel-based models are typically opaque, we used SHAP KernelExplainer to obtain local explanations, though they were less intuitive compared to tree-based methods.

- **K-Nearest Neighbors:** SHAP plots highlighted key predictors like chest pain type, exercise-induced angina, and thalassemia. LIME visualizations showed how individual patient predictions were driven by clinically relevant inputs like chest pain type and exercise-induced angina.

- **Gaussian Process Classifier:** Showed probabilistic feature contributions. SHAP values aligned well with posterior distributions, though visual interpretability required more effort.

- **Recurrent Neural Network:** While inherently less interpretable, we used SHAP DeepExplainer to uncover feature influences. Results highlighted oldpeak, cp, and thalach as key contributors, aligning with clinical expectations.

- **TabNet:** Offers native feature selection through attention masks. SHAP values confirmed the model's focus on clinically relevant variables like exang and ca, supporting both global and local interpretability.

- **Categorical Embedding Model:** SHAP highlighted features like cp, oldpeak, and thalach as top contributors, while LIME visualizations helped explain specific patient predictions using thresholds in ca, oldpeak, and slope. Together, these methods added much-needed transparency to the model's decisions.

- **LightGBM:** The top influencing features included oldpeak, ca, thal, and cp, which are also clinically relevant. This alignment with domain knowledge, along with fast SHAP computation, makes LightGBM not just accurate but also easy to trust in a clinical context.

In addition to looking at SHAP and LIME values for each model, fig. 12 shows that we also generated a combined heatmap to compare how important each feature was across all nine models trained on the UCI dataset. This gave us a clearer picture of which features were consistently influential across different types of models, including linear, tree-based, ensemble, and deep learning approaches. Features like chest pain type, maximum heart rate, and ST depression stood out as important in almost every model. Tree-based models tended to focus on just a few strong predictors, while deep learning models spread their attention across more features. Seeing this agreement across different algorithms helped validate the importance of key clinical features and gave us more confidence in how the models were making decisions.

To answer the second part of our RQ4, the improvements we introduced, particularly in handling overfitting and enhancing model interpretability, play a key role in identifying high-risk individuals earlier than traditional clinical scoring systems. Most standard tools rely on predefined thresholds and simple combinations of variables. While useful, these methods can miss subtle patterns and interactions that vary from patient to patient.

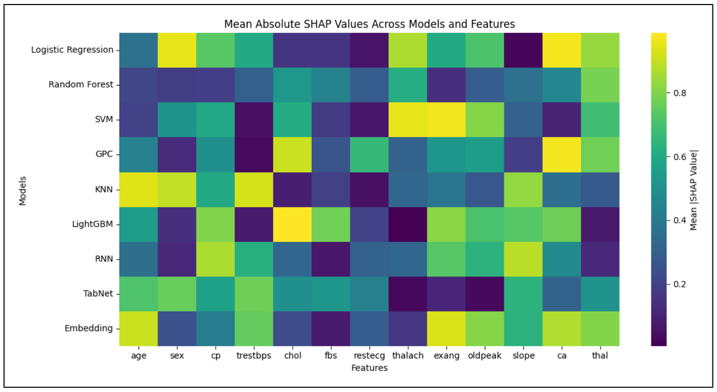

Figure 12: SHAP Heatmap: Feature Importance Across Models

In contrast, our models take a more personalized approach. By using techniques like SHAP and LIME, we were able to explain predictions in terms of familiar clinical features such as chest pain type, maximum heart rate, and ST depression. These insights not only matched known cardiovascular markers but also helped uncover less obvious feature relationships. For example, models like TabNet and the categorical embedding network focused on combinations of features such as slope and number of major vessels, which are not always emphasized in traditional scoring.

Our regularization strategies also contributed significantly. Models like KNN and SVM showed notable gains in sensitivity and F1-score after tuning, which is especially important for detecting early warning signs. These performance gains mean that the models were not only more accurate but also more consistent when applied to new patient data.

Taken together, these enhancements allowed our models to move beyond static risk formulas. They offered tailored risk assessments grounded in both data and clinical relevance. This makes them more useful in real-world settings, where decisions often depend on more than just numbers, they depend on understanding why those numbers matter.

5 Discussion

This project aimed to evaluate a broad range of machine learning and deep learning models for cardiovascular disease prediction, with a particular focus on model generalization, overfitting behavior, and interpretability. Across both the UCI and Kaggle datasets, several patterns emerged that inform the selection and application of predictive models in clinical settings.

Performance Across Datasets: Our results show that while deep learning models such as RNNs and DNNs achieved high accuracy on the UCI dataset (up to 86.2% for the tuned RNN), their generalization to the Kaggle dataset was limited. For instance, the same RNN dropped to 72.67% accuracy when trained on UCI and tested on Kaggle. In contrast, LightGBM and Categorical Embedding models demonstrated more consistent performance across both datasets, achieving generalization accuracy of 73.6% and 72.2% respectively. This suggests that while deep models can fit well to specific datasets, tree-based and embedding-based architectures may be more resilient to distributional shifts when properly tuned.

Ensemble Learning: We also explored ensemble techniques such as boosting, and stacking. While ensemble methods trained directly on the Kaggle dataset yielded the best in-domain performance (accuracy of 73.96%), they failed to generalize when trained on UCI and tested on Kaggle. This reinforces that ensemble models may be effective under controlled conditions but remain vulnerable to shifts in feature distributions and data representations.

Overfitting and Regularization: Overfitting was apparent across many models, particularly in deep architectures such as DNNs and TabNet. Learning curves revealed that without adequate regularization, these models memorized the training data and failed to retain performance on unseen examples. Introducing dropout, batch normalization, and early stopping was effective in mitigating overfitting in most cases. For example, after applying these techniques, the Categorical Embedding model reached 85.3% accuracy on UCI with close alignment between training and validation loss. However, TabNet continued to underperform despite hyperparameter tuning and regularization, indicating a possible mismatch between its inductive biases and the tabular nature of the dataset.

Interpretability and Clinical Relevance: Interpretability is essential for clinical adoption, and we explored this through SHAP and LIME across all classical and deep models. Logistic Regression and LightGBM emerged as the most interpretable models, with clear feature attributions that aligned with clinical intuition (e.g., oldpeak, cp, and thalach). While kernel-based models like SVM and KNN are not naturally interpretable, SHAP and LIME helped extract meaningful local and global explanations. This step is critical not just for model validation, but for facilitating clinical trust and decision support.

6 Limitations

Despite comprehensive experimentation, several limitations constrain the scope and generalizability of our findings.

Dataset Heterogeneity: The UCI and Kaggle datasets differ significantly in both scale and feature representation. The UCI dataset includes a limited set of clinically validated features with a smaller sample size, whereas the Kaggle dataset introduces additional lifestyle and biometric attributes that introduce noise, imbalance, and scale variation. These differences made generalization difficult, as models often overfit to dataset-specific distributions and failed to adapt to shifts in data structure.

Generalization Failure in Complex Models: While models like RNNs and TabNet achieved high performance on the training dataset, their generalization performance suffered when evaluated on a different domain. TabNet, in particular, underperformed even after hyperparameter tuning and regularization. This suggests that the model's sequential attention mechanism may not be well-suited for small tabular datasets with limited feature interaction complexity. Similarly, ensemble methods trained on UCI failed to generalize to Kaggle, indicating sensitivity to training distribution.

Overfitting in Deep Learning Models: Overfitting was a persistent challenge in deep models due to the relatively small size of the UCI dataset. Although regularization techniques such as dropout, early stopping, and feature selection helped improve generalization, the learning curves still showed clear divergence in some cases. For example, while the Categorical Embedding model showed improvement post-regularization, TabNet continued to exhibit unstable validation performance.

Computational Complexity and Interpretability Trade-offs: Models like Gaussian Process Classifiers and RNNs were computationally intensive, which limits their practical deployment in real-time or resource-constrained settings. Furthermore, although interpretability tools such as SHAP and LIME were effective, they introduced additional computational overhead and sometimes yielded inconsistent explanations across runs in deep models.

Clinical Translation Gap: While the models were interpretable to a certain extent, their outputs have not been validated in clinical workflows or decision-making scenarios. The feature importance rankings align with known cardiovascular risk factors, but further collaboration with clinical experts is necessary to assess the utility and trustworthiness of the models in real-world diagnostic settings.

7 Conclusion and Future Work

This project explored a wide range of classical and deep learning models for predicting cardiovascular disease using structured clinical data. We found that while advanced models like RNNs achieved high accuracy on clean, well-structured datasets like UCI, they struggled to generalize across larger, more diverse datasets like Kaggle. Classical models such as LightGBM and Gaussian Process Classifiers proved more stable and interpretable, especially when supported by techniques like SHAP and LIME. Regularization strategies such as dropout, early stopping, and feature tuning played a vital role in reducing overfitting, particularly in deep models. The novelty of this project lies in its use of categorical embedding networks, TabNet, and SHAP-based interpretability across both UCI and Kaggle datasets, an approach that is rarely applied together in cardiovascular disease prediction literature. This dual focus on both performance and generalization, along with interpretability in heterogeneous clinical data, makes the project especially relevant for real-world deployment.

Going forward, future work can focus on improving generalization by exploring domain adaptation techniques and transfer learning strategies tailored to medical data. Additionally, integrating real-time clinical feedback from healthcare professionals will be essential to validate and fine-tune the model outputs in practice. Exploring hybrid models that combine the interpretability of classical methods with the expressive power of deep learning may offer a more balanced solution. Finally, building a user-friendly clinical decision support system that presents model predictions alongside explanations could help bridge the gap between data science and day-to-day medical decision-making.

8 Team Membership and Attestation

All authors contributed equally. This report reflects our original work unless stated otherwise.

References

[1] Mohammad Alfadli, Khadijah Almagrabi, et al. "Feature-Limited Prediction on the UCI Heart Disease Dataset". In: *Journal of Biomedical Informatics* (2024). URL: https://www.sciencedirect.com/org/science/article/pii/S1546221822002892#ref-5.

[2] Abdullah Alqahtani et al. "Cardiovascular Disease Detection Using Ensemble Learning". In: *Computational Intelligence and Neuroscience* (2022). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9398727/.

[3] R Detrano et al. "International application of a new probability algorithm for the diagnosis of coronary artery disease". In: *American Journal of Cardiology* (1989). URL: https://pubmed.ncbi.nlm.nih.gov/2756873/.

[4] Andras Janosi et al. "Heart Disease". In: *UCI Machine Learning Repository* (1989). URL: https://doi.org/10.24432/C52P4X.

[5] Pratheek N et al. “Cardiovascular Disease Prediction with Machine Learning Algorithms and Interpretation using Explainable AI Methods: LIME and SHAP”. In: *2024 3rd International Conference for Advancement in Technology (ICONAT)*. IEEE. 2024. DOI: 10.1109/ICONAT61936.2024.10774972.

[6] Muntasir Nishat et al. “A Comprehensive Investigation of the Performances of Different Machine Learning Algorithms for Cardiovascular Disease Prediction”. In: *Scientific Programming* 2022 (2022), pp. 1–13. DOI: https://doi.org/10.1155/2022/3649406.

[7] Dina Suliana. “Cardiovascular Disease Dataset”. In: *Kaggle Datasets* (2019). Accessed: 2025-05-11. URL: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset.

A Appendix

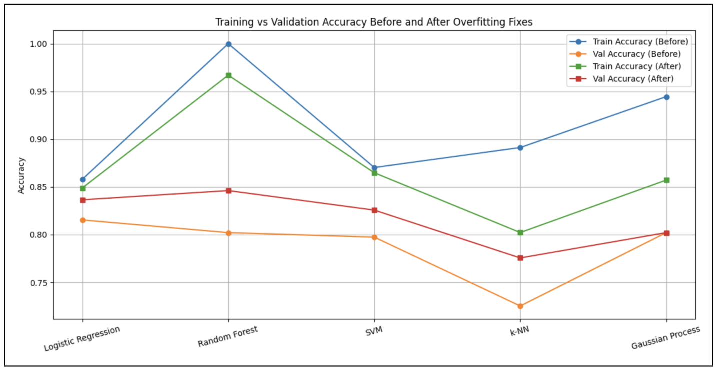

Figure 13: ML model curves after regularization

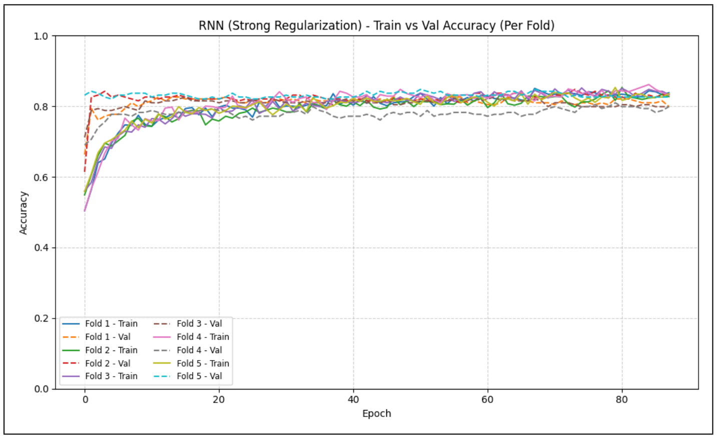

Figure 14: RNN Regularization

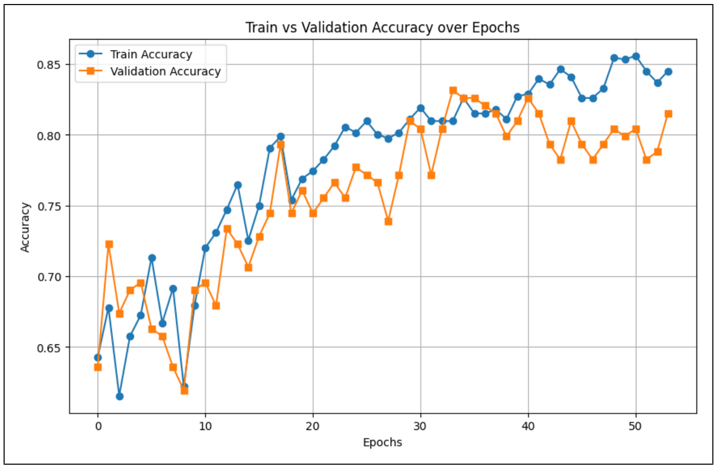

Figure 15: Tabnet before regularization

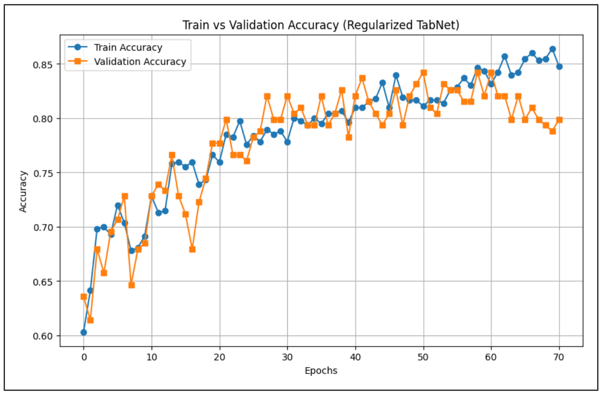

Figure 16: Tabnet after regularization

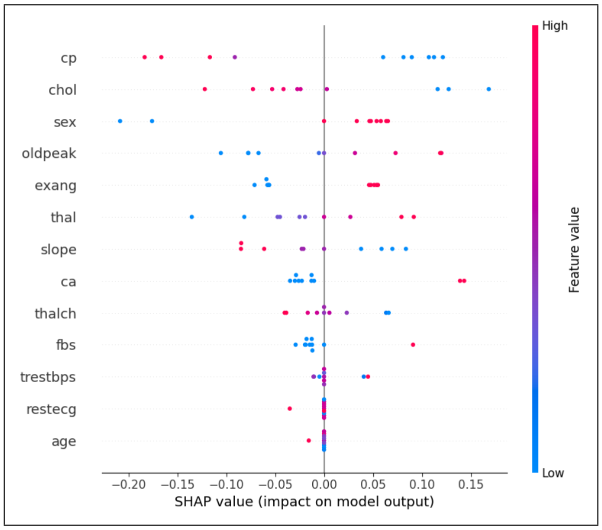

Figure 17: RNN SHAP

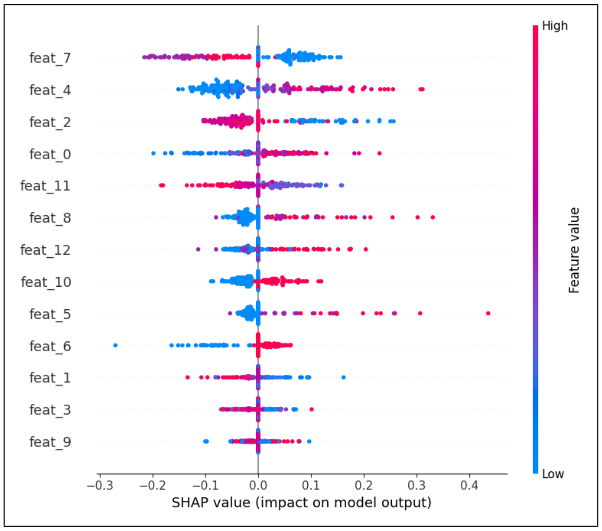

Figure 18: Tabnet SHAP

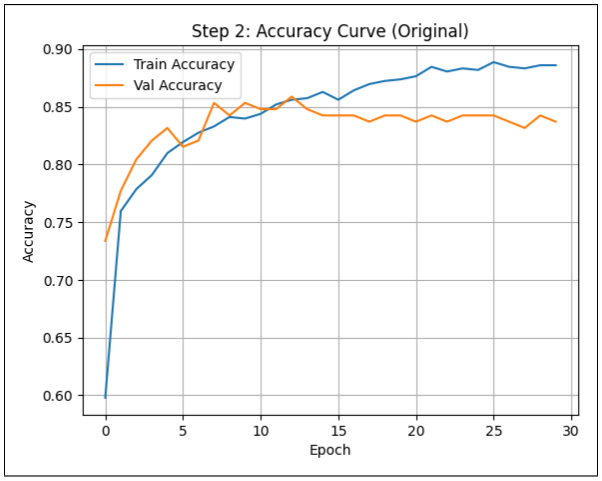

Figure 19: Categorical Embedding before regularization

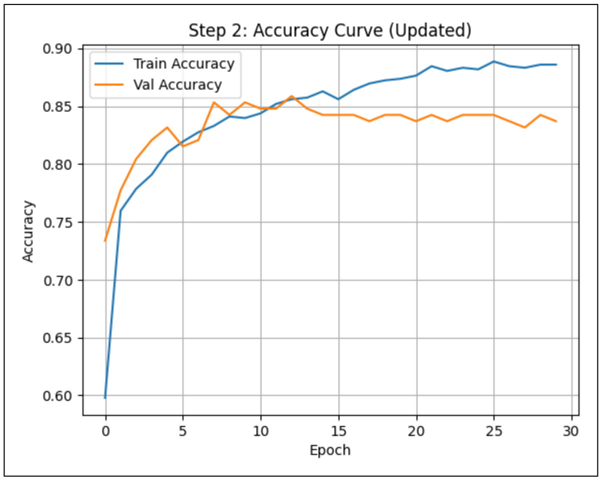

Figure 20: Categorical Embedding after regularization

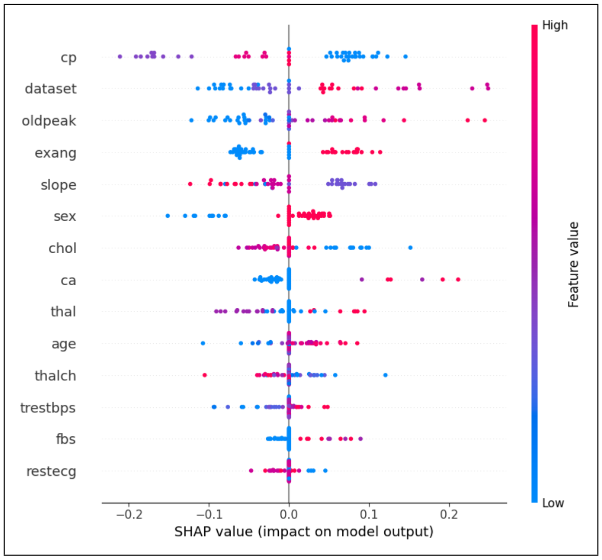

Figure 21: Categorical Embedding SHAP

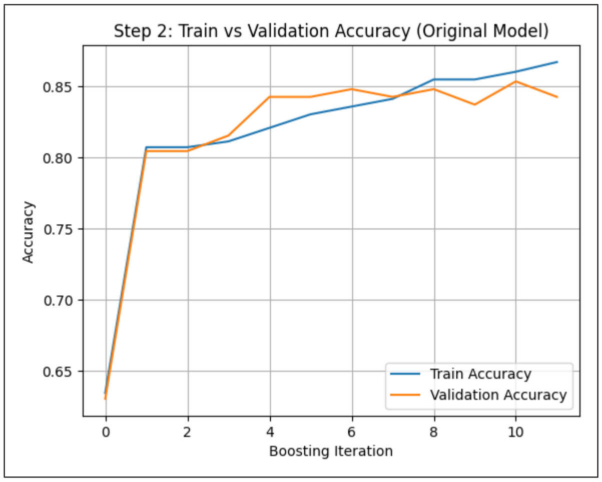

Figure 22: LightGBM Accuracy

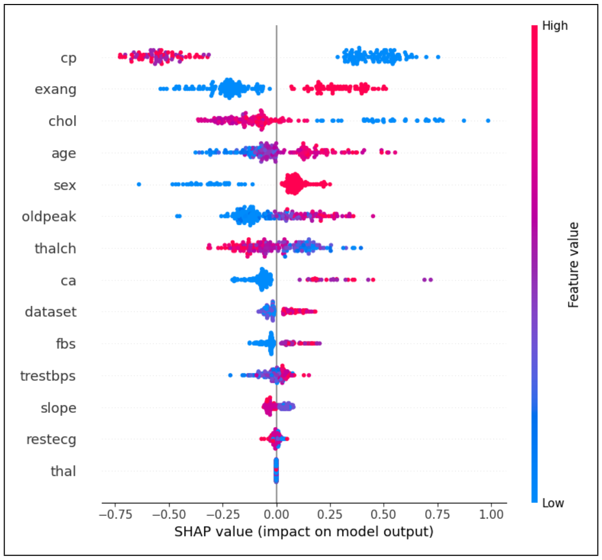

Figure 23: LightGBM SHAP