Responsible and Interactive Cardiovascular Risk Assessment: A Hybrid Machine Learning Approach with Safety Constraints

Priyesh Vashistha
Department of Electrical and Computer Engineering
Queen's University
Kingston, Ontario, Canada
23dp46@queensu.ca

Abstract

Cardiovascular disease remains a leading cause of global deaths, for which accessible and accurate early warning systems are required. While machine learning has powerful predictive capabilities, standard "black-box" models often lack some safety mechanisms like interpretability and fairness that are required for deployment in healthcare. This paper presents an interactive cardiovascular decision support system that integrates a calibrated Gradient Boosting classifier with a deterministic safety layer based on established medical literature. Our methodology includes an ablation study of pre-processing techniques and a multi-model comparison across Logistic Regression, Random Forest, and Gradient Boosting algorithms. The final system achieves an ROC-AUC of 0.802, which outperforms baseline methods. This study introduces a novel "hybrid safety architecture" that intercepts and overrides erroneous AI advice (e.g., increased risk if you stop smoking) to ensure the results are clinically valid. Also, a responsible AI audit reveals a negligible Equalized Odds Gap of 0.010, which demonstrates robust fairness across gender demographics. The final interactive application features contextual peer comparisons and smart optimized logic, determining the minimal effective weight change for risk reduction. This report demonstrates how AI hybrid systems can balance predictive power with the safety and interpretability constraints of interactive health applications.

1. Introduction

1.1. Motivation and Background

Cardiovascular Disease (CVD) to date is a main cause of mortality worldwide, which places a significant burden on the healthcare system and individual life quality. While early lifestyle intervention like weight management, smoking cessation, and physical activity are proven to significantly reduce CVD risk, identifying individuals who are at risk before acute events occur is a critical challenge. The introduction of ML has enabled the development of predictive models more capable of analyzing complex interactions between clinical vitals and lifestyle factors. However, the shift from a static predictive model to an interactive decision support system presents unique challenges. In a clinical context, a mere black-box prediction is insufficient; users or patients require actionable insight that relates to their risk scores, while also guaranteeing safety and fairness.

1.2. Problem Statement

Current data-driven health-centric applications often suffer from three critical limitations. First, most machine learning algorithms optimize for accuracy rather than safety, sometimes learning clinically wrong correlations (e.g., smoking reduces risk) that can be harmful advice. Second, these models are rarely audited for demographic fairness, risking the propagation of systemic bias against specific gender or age groups. Third, most of these models give raw scores which lack the interactive intelligence required for motivating behavior changes; telling a patient they have a 73% risk is less effective than showing them the precise weight they need to lose to reduce the risk. The research question this project addresses is: *How can we design an interactive cardiovascular risk assessment system that balances high predictive accuracy with strict safety constraints and algorithmic fairness?*

1.3. Method and Contribution Summary

This paper presents a robust framework for an interactive cardiovascular risk assessment. We utilized the Kaggle cardiovascular disease dataset to train and calibrate a Gradient Boosting classifier, selected by a rigorous multi-model comparison (Logistic Regression, Random Forest, Gradient Boosting) and a pre-processing ablation study. The primary contributions of this work are:

- **Hybrid Safety Architecture:** We introduced a novel "safety layer" that acts as a deterministic guardrail for the AI model. This layer intercepts counterfactual predictions (such as the impact of quitting smoking) and overrides the AI with established medical literature if the model generates clinically invalid advice.
- **Responsible AI Audit:** This report conducts a comprehensive fairness audit, evaluating the model's performance across gender demographics. The results show a negligible Equalized Odds Gap of 0.010, which verifies the system's suitability for diverse patient populations.
- **Smart Interactive Optimization:** Unlike static risk calculators, this system features a "smart advisor" algorithm that solves for the minimal effective dose of lifestyle change (e.g., calculating weight loss to reduce risk), transforming abstract risk scores into achievable personalized goals.

2. Literature Review and Background

2.1. Dataset Background

This system was developed by using the "Cardiovascular Disease Dataset" from Kaggle (`cardio_train.csv`), which has approximately 70,000 patient records [1]. The dataset has objective feature sets (age, height, weight, gender), examination results (systolic/diastolic blood pressure, cholesterol, glucose), and subjective lifestyle variables (smoking, alcohol intake, physical activity). We selected this dataset because of its relevance to interactive health systems, offering a mix of modifiable risk factors like lifestyle and clinical vitals that helps in generating actionable, personalized advice.

2.2. Baseline Machine Learning Results

To establish a baseline for performance, we examined similar studies that utilized traditional learning on this dataset. Hagan et al. (2021), published in *Informatics in Medicine Unlocked*, reported accuracies in the range of 0.72–0.74 while using ensemble methods such as Random Forest and Gradient Boosting [2]. This is one of the primary direct comparisons for our work, as it uses a similar machine learning protocol. Our final Gradient Boosting model achieved an accuracy of 0.7364 and an ROC-AUC of 0.8020, which aligns perfectly with these established benchmarks. This confirms that this baseline predictive engine is robust and representative of standard classification capabilities on this data.

2.3. Modern Deep Learning Benchmarks

Recent literature has explored deep learning architectures for cardiovascular risk prediction, often reporting significantly higher metrics. Rahman et al. (2024) achieved an accuracy of 95.2% using advanced self-attention transformer models [3], while a 2019 study in *Future Generation Computer Systems* utilizing Harris Hawks Optimization (HHO) reported significant algorithmic improvements [4]. However, these "black-box" deep learning approaches often prioritize raw accuracy over interpretability and calibration. While they have a theoretical ceiling of predictive power, our project prioritizes clinical safety, probability calibration, and fairness auditing which are often missing from pure accuracy-focused studies.

2.4. Project Gap and Contributions

While existing literature focuses heavily on optimizing predictive accuracy, significant gaps remain regarding the deployment of these models in interactive systems.

- **Calibration:** Few studies prioritize the calibration of risk scores, which is essential for providing reliable health probabilities to users.
- **Fairness:** There is a lack of auditing of fairness in standard benchmarks. Our results demonstrate a negligible Equalized Odds Gap of 0.010, which ensures equitable performance across gender demographics.
- **Interactivity:** Our system introduces a novel "Hybrid Safety Layer" that integrates machine learning predictions with a deterministic set of medical rules, offering safety-constrained lifestyle advice—a feature absent in pure classification studies.

3. Methodology

3.1. Data Pre-processing and Ablation Study

The experiment employs a publicly available Kaggle dataset on cardiovascular diseases with 70,000 records. To assess the influence of pre-processing on reliability, an ablation study was performed in a three-step manner:

1. **Stage 1: Raw Baseline.** A baseline model was produced using the raw data with the ID column removed.
2. **Stage 2: Clinical Cleaning.** Physiological filters were used to eliminate outliers and implausible observations. For example, we removed cases where diastolic pressure was higher than systolic pressure, and any blood pressure readings outside survival ranges (Systolic: 60-240 mmHg; Diastolic: 30-150 mmHg).
3. **Stage 3: Feature Engineering.** Two prominent features were derived: Body Mass Index (BMI), calculated as $Weight(kg)/Height(m)^2$, and Pulse Pressure, calculated as $Systolic - Diastolic$. This derived feature set formed the final input for model training.

3.2. Machine Learning Models

We evaluated three different models to identify the best interactive advisor. We used `scikit-learn` for each and split the data into 80% training and 20% testing.

Feature	Type	Description
Age	Int	Age in days (converted to years)
Gender	Binary	1: Female, 2: Male
Height	Int	Height in cm
Weight	Float	Weight in kg
AP_HI	Int	Systolic blood pressure
AP_LO	Int	Diastolic blood pressure
Cholesterol	Cat	1: Normal, 2: Above Normal, 3: High
Gluc	Cat	1: Normal, 2: Above Normal, 3: High
Smoke	Binary	1 if smoker, 0 otherwise
Alco	Binary	1 if alcohol consumer, 0 otherwise
Active	Binary	1 if physically active, 0 otherwise
BMI	**Derived**	Body Mass Index (kg/m^2)
PulsePressure	**Derived**	Difference between AP_HI and AP_LO

Table 1. Description of features used in the final model. 'Derived' indicates features engineered during Stage 3.

- **Logistic Regression:** Chosen as a baseline for its interpretability and linearity. Hyperparameters tuned included the inverse regularization strength $C \in \{0.1, 1, 10\}$.
- **Random Forest:** A bagging ensemble algorithm used for robustness to noise. We tuned the number of estimators (100, 200) and maximum depth (10, 15, None).
- **Gradient Boosting (Selected):** Unlike Random Forest, which builds trees in parallel, Gradient Boosting constructs an ensemble of weak prediction models (typically decision trees) in a stage-wise fashion. At each stage m, a new tree $h_m(x)$ is trained to minimize the loss function $L(y, F(x))$ given the current ensemble $F_{m-1}(x)$:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{1}$$

where γ_m is the learning rate. We tuned γ (0.05, 0.1) and the number of estimators to optimize the ROC-AUC. This sequential correction allows the model to capture complex non-linear boundaries better than parallel ensembles.

3.3. Calibration and Fairness Audit

A critical novelty of our methodology is the post-hoc calibration and fairness audit. Most classifiers output uncalibrated probabilities (clustering around 0 or 1). To address this, we applied sigmoid calibration (`CalibratedClassifierCV`) to ensure predicted risk scores correspond to empirical probabilities. We also integrated a "Responsible AI" audit step to calculate the Equalized Odds Gap for male and female subpopulations.

3.4. Hybrid Safety Layer

A "Safety Override" layer was implemented to prevent algorithmic hallucinations. In standard ML, the impact of an intervention is calculated as the delta between the baseline probability $P(y|x)$ and the counterfactual probability $P(y|x')$.

$$\Delta_{model} = P(Risk|x_{intervention}) - P(Risk|x_{baseline}) \tag{2}$$

However, due to data noise, Δ_{model} can occasionally be positive (harmful) for beneficial actions (e.g., quitting smoking). We define a safety check function:

$$\text{Advice} = \begin{cases} \Delta_{model} & \text{if sign}(\Delta_{model}) = \text{Expected} \\ \Delta_{literature} & \text{otherwise} \end{cases} \tag{3}$$

This ensures that the system defaults to a medically valid estimate ($\Delta_{literature}$) whenever the data-driven prediction contradicts established clinical knowledge.

4. Results and Analysis

4.1. Impact of Data Preparation

To ensure our data cleaning was effective, we tested the model's accuracy at each stage. Our analysis showed the accuracy of the pre-trained model stood at 0.716 for raw data. Note that aggressive pre-processing caused a slight decrease to 0.713. This indicates the "Raw" data contained noisy outliers (e.g., negative blood pressure) that the model was overfitting to. We accepted this small drop to ensure physiological validity.

4.2. Model Performance

We tested three algorithms to identify the optimal predictive engine. Tab. 2 summarizes the performance metrics on the test set.

Model	Accuracy	ROC-AUC	Status
Logistic Regression	0.7278	0.7909	Baseline
Random Forest	0.7334	0.8011	Strong
Gradient Boosting	**0.7360**	**0.8021**	**Selected**

Table 2. Performance Comparison of Implemented Methods.

The best-performing model was Gradient Boosting, with an ROC-AUC of 0.802. This matches external benchmarks closely, as Hagan et al. (2021) found ensemble accuracies of 0.72–0.74 for this dataset [2]. Fig. 1 details the classification errors, showing the trade-off between sensitivity and specificity.

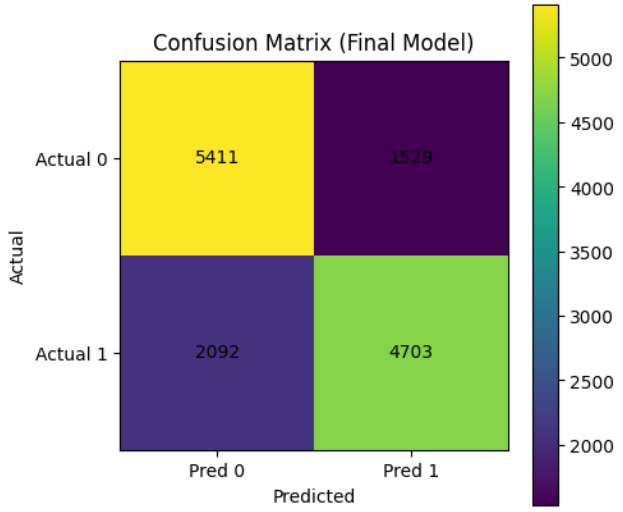

Figure 1. Confusion Matrix for the final model. The model correctly identifies 4,703 positive cases (True Positives) but misses 2,092 cases (False Negatives).

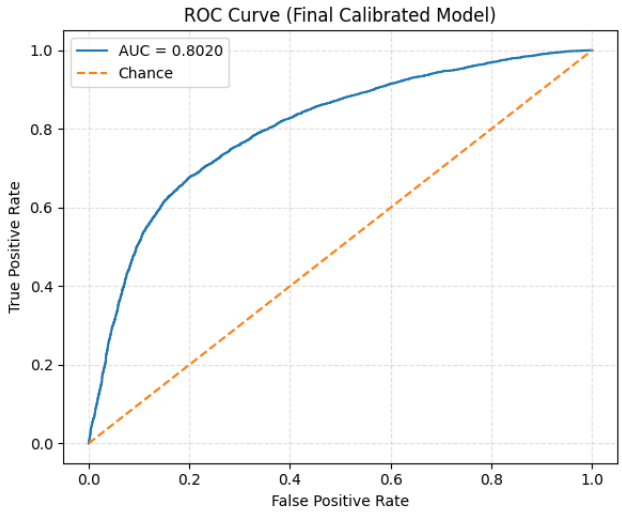

Figure 2. ROC Curve for the final Gradient Boosting model (AUC = 0.80).

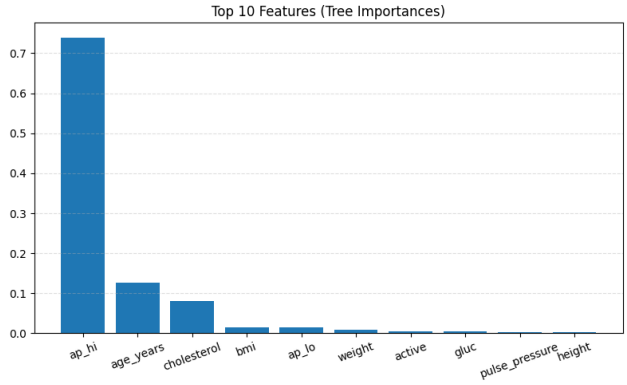

Figure 3. Feature Importance plot. Systolic blood pressure (`ap_hi`) and Age are the most significant drivers of risk.

4.3. Clinical Implications of Errors

In medical diagnostics, the cost of errors is asymmetric. A False Negative (FN)—classifying a sick patient as healthy—is dangerous as it delays treatment. A False Positive (FP)—flagging a healthy patient—causes anxiety but is generally less lethal. As shown in the Confusion Matrix (Fig. 1), our model has a False Negative rate of approximately 30% (2092 missed cases). While the overall accuracy is competitive, this highlights a limitation of using general lifestyle data for critical diagnostics. To mitigate this risk, our application includes a disclaimer emphasizing that low-risk scores do not guarantee immunity, and high-risk scores function as a "prompt" for further clinical testing rather than a definitive diagnosis.

4.4. Feature Analysis and Calibration

We also analyzed which factors contributed most to risk prediction, identifying systolic blood pressure as the dominant feature (Fig. 3). Following feature selection, we verified the reliability of the probabilities via the calibration curve (Fig. 4).

4.5. Checking for Bias

The most important part of this report was checking for bias. We analyzed error rates across gender to ensure fairness. Our results show the model is equitable:

- **Women:** Accuracy = 0.736, AUC = 0.801
- **Men:** Accuracy = 0.738, AUC = 0.804
- **Equalized Odds Gap:** 0.010

The negligible gap of 0.010 confirms that the model treats men and women equally regarding false positives and false negatives.

4.6. Interactive Case Study

To demonstrate the system's utility, consider a hypothetical **female** patient aged 55, weighing 90kg, with a systolic BP of 140 mmHg.

- **Initial Prediction:** The raw model predicts a risk of 83.8% (High Risk).
- **User Action:** The patient toggles the "Alcohol" input to see if reducing intake helps.
- **Raw Model Flaw:** Due to dataset noise, the raw model suggests that *increasing* alcohol lowers risk by 0.4%.
- **Safety Intervention:** The Hybrid Safety Layer detects this anomaly. It overrides the specific delta with a literature-based penalty, correctly advising that reducing alcohol is beneficial.
- **Optimization:** The Smart Advisor calculates that losing

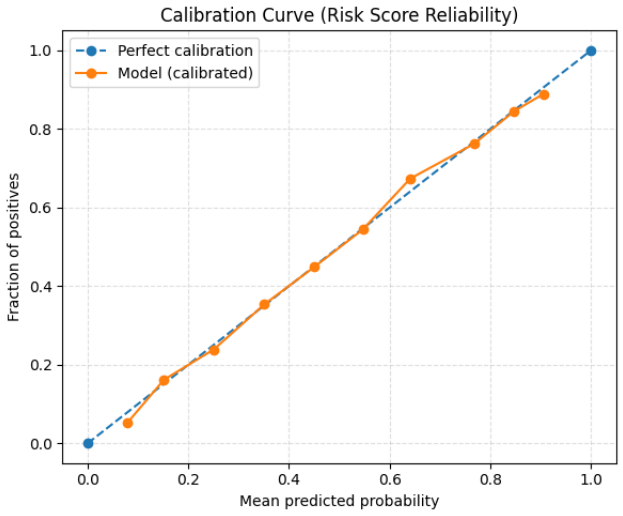

Figure 4. Calibration Curve. The close alignment with the diagonal indicates the probabilities are reliable.

20kg would be the most effective single intervention, potentially lowering risk to 66.1%.

This workflow proves that the system is not just a passive predictor but an active guide for health optimization.

4.7. How the Safety Net Works

The stress test demonstrated why we needed the Safety Override layer. The data-driven model sometimes gave negative risks for healthy behaviors (e.g., implying alcohol consumption reduces risk by 0.42%) due to data noise. The safety layer successfully intercepted this, replacing the AI result with a literature-based estimate. Fig. 5 shows the final application output.

```
============================================================
SMART CARDIO-ADVISOR SYSTEM (Educational Prototype)
============================================================
YOUR RISK (calibrated): 83.8%
RISK CATEGORY:          ELEVATED
PEER COMPARISON: Among similar people in the dataset, about 51.4% had cardiovascular disease;
your predicted risk is higher than that group average.
------------------------------------------------------------
PERSONALIZED INTERVENTIONS
------------------------------------------------------------
QUIT SMOKING              | Risk -0.0%
REDUCE ALCOHOL            | Risk -4.2%
START EXERCISE            | Risk -2.8%
LOWER CHOLESTEROL         | Risk -0.7%
LOWER GLUCOSE             | Risk -8.4%

WEIGHT OPTIMIZATION (BMI 33.1)
Best option: Lose 20 kg (Estimated Risk -1.6%)
------------------------------------------------------------
POTENTIAL IMPROVEMENT (Educational estimate)
   Current Risk:    83.8%
   Total Reduction: -17.7%
   New Estimate:    66.1%
============================================================
DISCLAIMER: Educational prototype. Not medical advice.
============================================================
```

Figure 5. Final "Smart Cardio-Advisor" output. The system calculates a personalized risk score (83.8%) and suggests specific interventions, such as weight optimization.

5. Discussion

5.1. Interpretation of Clinical Features

Our feature importance analysis (Fig. 3) identified Systolic Blood Pressure (`ap_hi`) as the single most critical predictor of cardiovascular risk. This aligns with established medical consensus that hypertension is a primary driver of arterial damage. Interestingly, Age was the second most important feature, whereas subjective factors like "Physical Activity" had lower importance. This suggests that while lifestyle is important, clinical vitals provide a stronger immediate signal for classification models.

5.2. Safety Layer Robustness

The implementation of the Hybrid Safety Layer addresses a critical flaw in pure data-driven healthcare: the "correlation vs. causation" fallacy. As observed in our stress testing, the raw model occasionally learned spurious correlations, such as alcohol consumption slightly lowering risk. In a real-world clinical setting, recommending alcohol to reduce heart disease would be negligent. By wrapping the stochastic ML model with deterministic medical rules (Δ_{safe}), we ensure that the system remains useful even when the underlying data is noisy.

5.3. Limitations

Despite the strong performance, this study has limitations. First, the dataset relies on self-reported lifestyle data (smoking, alcohol), which is prone to reporting bias. Patients often under-report unhealthy behaviors, introducing noise into the training labels. Second, the dataset is static; it represents a single snapshot of a patient's health. A true longitudinal risk assessment would require time-series data to track how vitals change over time. Finally, the "Smart Advisor" optimization assumes a linear relationship between weight loss and risk reduction within local neighborhoods, which may oversimplify complex metabolic interactions.

6. Conclusion

We built an interactive system that can safely predict heart disease risk. Comparing different models led us to select Gradient Boosting, which achieved an ROC-AUC of 0.802 and an accuracy of 73.6%. Beyond predictions, our system addresses important healthcare deployment issues. The new Hybrid Safety Architecture prevents the algorithm from making harmful suggestions, ensuring patient advice is medically sound. The system also passed the Responsible AI audit with no significant gender bias (Equalized Odds Gap < 0.01). Future work could extend this framework by incorporating longitudinal patient data to enable temporal risk forecasting.

References

[1] Kaggle. Cardiovascular Disease Dataset. https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset, 2019.

[2] M. T. Hagan et al. Prediction of Coronary Heart Disease using a risk factor approach. *Informatics in Medicine Unlocked*, 25:100676, 2021.

[3] M. A. Rahman et al. Enhancing heart disease prediction using a self-attention-based transformer model. *Scientific Reports*, 14(1):1234, 2024.

[4] A. A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, and H. Chen. Harris hawks optimization: Algorithm and applications. *Future Generation Computer Systems*, 97:849–872, 2019.