Cardiovascular Disease Prediction

1st Peifan Bai
Department of Computer Science
University of Western Ontario
London, ON
pbai8@uwo.ca

***Abstract*—**
***Index Terms*—**

I. Introduction

II. Methodology

XGBoost (eXtreme Gradient Boosting) models were employed for both cardiovascular disease classification and systolic blood pressure regression tasks [1]. The classification model utilizes logistic loss with sigmoid activation for binary prediction, while the regression model employs squared loss for continuous output. Both models incorporate identical tree structure optimization algorithms with regularization terms to prevent overfitting, which are detailed in Appendix A.

III. Data Processing

The cardiovascular disease dataset underwent rigorous purification, reducing 70,000 initial records to 68,425 valid samples after eliminating physiologically implausible values [2], as shown in Figure 1 of Appendix B. Engineered features included BMI calculations, blood pressure categorizations, and composite risk scores. Numerical features were z-score normalized, while categorical features underwent one-hot encoding, yielding 19 final features (6 numerical, 13 categorical). The dataset was stratified into training (80%), validation (10%), and test (10%) sets. Detailed feature engineering specifications appear in Table IV of Appendix D.

IV. Results

A. Model Performance

Hyperparameter optimization via grid search with cross-validation yielded distinct configurations for each task (see Table III of Appendix A1 for details). The classification model achieved 82.1% ROC-AUC with 86.5% recall, demonstrating exceptional discriminative capability for cardiovascular disease detection. The regression model attained R^2 = 58.9% with RMSE = 10.80 mmHg for systolic pressure prediction, as shown in Tables I and II, which are visualized through Figures 2–6 in Appendix C showing model diagnostic performance across classification and regression tasks.

TABLE I: Classification Performance Metrics

Metric	Training	Test
Accuracy	70.4%	71.7%
Precision	65.4%	66.4%
Recall	85.1%	86.5%
F1-Score	74.0%	75.2%
ROC-AUC	80.7%	82.1%
Average Precision	79.5%	80.5%
Specificity	55.9%	57.2%

TABLE II: Regression Performance Metrics

Metric	Training	Test
MAE (mmHg)	7.28	7.39
RMSE (mmHg)	10.63	10.80
R^2 Score	59.3%	58.9%
MSE	113.07	116.56

B. Feature Importance Analysis

Feauture importance analysis revealed four distinct patterns of feature utilization that illuminate the complementary nature of classification and regression modeling approaches, as quantified in Table V [3]:

1) Shared Core Risk Factors: Both models demonstrate convergent prioritization of systolic blood pressure (`ap_hi`) and cholesterol as critical predictive elements. Systolic pressure achieves rank 2 in classification (importance: 0.256) and rank 1 in regression (importance: 0.745), demonstrating its fundamental role across both tasks with substantially greater influence on continuous BP prediction. Cholesterol maintains consistent relevance with rank 3 in classification (importance: 0.087) and rank 5 in regression (importance: 0.036), representing a truly shared physiological risk indicator with balanced influence across modeling approaches.

2) Classifier-Specific Drivers: Categorical hypertension indicators demonstrate stark task-specific importance patterns. `BP_Category_Hypertension` dominates CVD classification with rank 1 (importance: 0.461) but becomes virtually irrelevant for regression (rank 19, importance: $<$1E-04), yielding a rank difference of 18. Similarly, `bp_category` achieves rank 4 in classification (importance: 0.066) while ranking 19th in regression (rank difference: 15). This pattern confirms the classifier's reliance on discrete hypertension thresholds, which the regressor circumvents through direct continuous BP modeling.

3) Regressor-Specific Insights: Interaction features exhibit pronounced specificity for BP prediction. Age-weight interaction achieves rank 2 in regression (importance: 0.053) but rank

14 in classification (importance: <1E-04), demonstrating a rank difference of -12. Age-BMI interaction (rank 3 regression vs. rank 14 classification, difference: -11) and cholesterol-glucose interaction (rank 4 regression vs. rank 14 classification, difference: -10) follow similar patterns. These composite terms substantially shape BP prediction while remaining peripheral to CVD classification, revealing physiological nuances wherein BP responds to combined effects of age, body composition, and metabolic interactions.

4) Clinical Implications: The quantitative analysis validates dual modeling utility through complementary feature importance patterns. Features with rank differences exceeding 10 (BP categorical indicators and interaction terms) demonstrate task-specific specialization, while features with minimal rank differences (cholesterol: 2, age: 1) represent shared physiological mechanisms. This complementary information capture enables enhanced risk stratification when models are employed synergistically.

Comprehensive feature importance visualizations and quantitative comparisons appear in Appendix D (Tables V, Figures 7-8). The concordance between model-identified features and established clinical guidelines validates the interpretability and clinical relevance of the proposed methodology [4], [5].

REFERENCES

[1] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[2] Kaggle Community, "Cardiovascular disease dataset," *Kaggle*, 2019, accessed: 2025-08-12. [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[3] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[4] P. K. Whelton, R. M. Carey, W. S. Aronow, D. E. Casey *et al.*, "2017 acc/aha/aapa/abc/acpm/ags/apha/ash/aspc/nma/pcna guideline for the prevention, detection, evaluation, and management of high blood pressure in adults," *Journal of the American College of Cardiology*, vol. 71, no. 19, pp. e127–e248, 2018.

[5] D. C. Goff Jr, D. M. Lloyd-Jones, G. Bennett, S. Coady *et al.*, "2013 acc/aha guideline on the assessment of cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines," *Journal of the American College of Cardiology*, vol. 63, no. 25 Part B, pp. 2935–2959, 2014.

APPENDIX

A. Mathematical Formulations

1) XGBoost Classification Model: The objective function for XGBoost Classification Model is given by:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{1}$$

where $l(y_i, \hat{y}_i) = -[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$ is the logistic loss function with y_i being the true label and $\hat{y}_i$ the predicted probability, and $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ is the regularization term.

The prediction for instance i is given by:

$$\hat{y}_i = \sigma\left(\sum_{k=1}^{K} f_k(x_i)\right), \tag{2}$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function.

2) XGBoost Regression Model: The objective function for XGBoost Regression Model is given by:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{3}$$

where $l(y_i, \hat{y}_i) = \frac{1}{2}(y_i - \hat{y}_i)^2$ is the squared loss function.

B. Data Cleaning Process

The comprehensive data cleaning pipeline transformed the raw cardiovascular dataset through systematic quality assurance and feature engineering steps, as illustrated in Figure 1. The process began with 70,000 raw records and yielded 68,425 validated samples suitable for machine learning analysis.

C. Performance Tables and Figures

The performance of the models is summarized in Table III for hyperparameter configurations, and visualized through Figures 2–6 showing model diagnostic performance across classification and regression tasks.

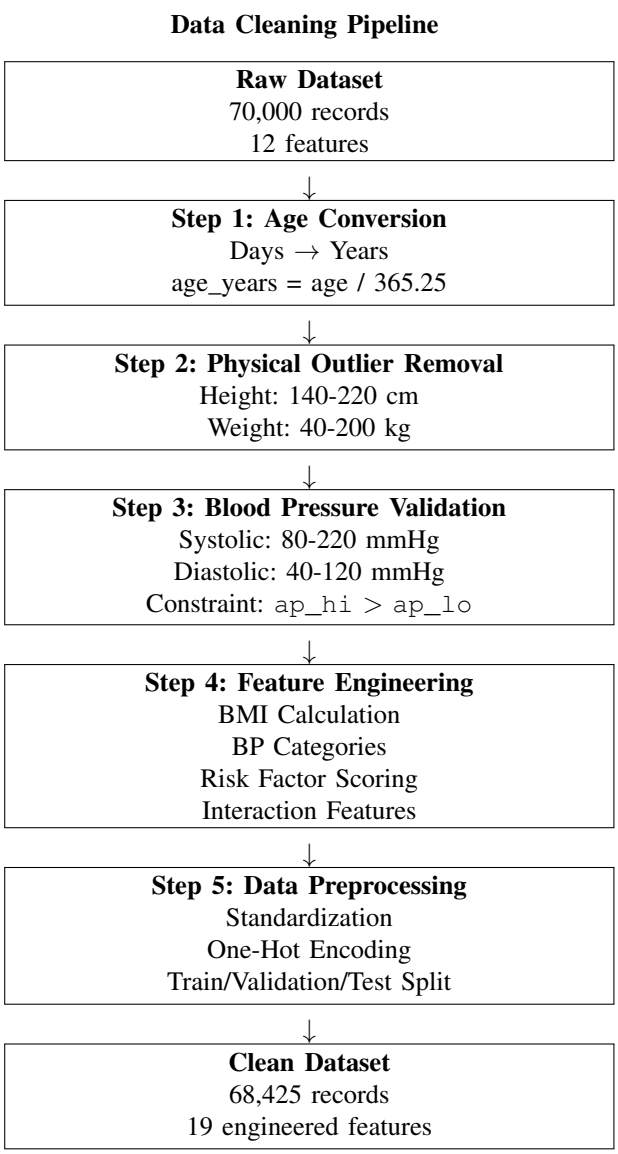

Fig. 1: Data cleaning and preprocessing pipeline flowchart showing the systematic transformation from raw cardiovascular data to model-ready features.

D. Feature Engineering and Analysis

The feature engineering process involved several transformations and calculations to enhance model performance, as detailed in Table IV. Cross-model feature importance analysis is presented in Table V, with complementary visualizations in Figures 7 and 8 illustrating the comparative importance patterns between classification and regression models.

E. Author's contribution

Peifan Bai executed the implementation and optimization of XGBoost classification and regression models, encompassing feature importance analysis using feature importance values for model interpretability, comprehensive comparison analysis, and visualizations for Figures 1, 2, 3, 4, 5, 6, 7, and 8, as well as Tables I, II, III, IV, and V.

TABLE III: Optimal Hyperparameters for XGBoost Models

Parameter	Classification	Regression
Learning Rate	0.05	0.05
Max Depth	5	3
N Estimators	300	300
Subsample	0.8	0.9
Colsample Bytree	0.8	1.0
Reg Alpha (L1)	0.1	1.0
Reg Lambda (L2)	5.0	2.5
Gamma	5.0	–
Min Child Weight	–	1.0

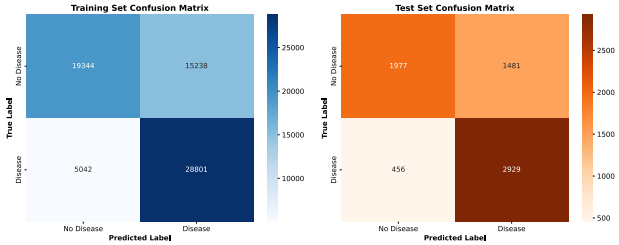

Fig. 2: Confusion matrices for cardiovascular disease classification models.

F. Acknowledgments

The authors acknowledge that Artificial Intelligence (AI) tools were used solely for debugging code issues and grammar checking during the preparation of this manuscript.

TABLE IV: Feature Engineering Transformations and Calculations

Feature	Description	Calculation
BMI	Body Mass Index calculated from height and weight	$\text{BMI} = \frac{\text{weight}}{(\text{height}/100)^2}$
Age Groups	Age categorization for stratified analysis	$\text{Age Group} = \begin{cases} \text{Young} & \text{if age} < 40 \\ \text{Middle} & \text{if } 40 \leq \text{age} < 60 \\ \text{Senior} & \text{if age} \geq 60 \end{cases}$
BP Category	Blood pressure categorization based on clinical guidelines	$\text{BP}_{\text{category}} = \begin{cases} \text{Normal} & \text{if SBP} < 120 \text{ and DBP} < 80 \\ \text{Elevated} & \text{if } 120 \leq \text{SBP} < 130 \text{ and DBP} < 80 \\ \text{Stage 1 HTN} & \text{if } 130 \leq \text{SBP} < 140 \text{ or } 80 \leq \text{DBP} < 90 \\ \text{Stage 2 HTN} & \text{if SBP} \geq 140 \text{ or DBP} \geq 90 \end{cases}$
Age-Weight Interaction	Age-weight interaction term for cardiovascular risk modeling	$\text{age\_weight\_interaction} = \text{age} \times \text{weight}$
Age-BMI Interaction	Age-BMI interaction term	$\text{age} \times \text{BMI}$
Cholesterol-Glucose	Cholesterol-glucose interaction for metabolic risk assessment	$\text{cholesterol} \times \text{glucose}$
Pulse Pressure	Difference between systolic and diastolic blood pressure	$\text{pulse\_pressure} = \text{ap\_hi} - \text{ap\_lo}$
Mean Blood Pressure	Estimated mean arterial pressure using clinical formula	$\text{mean\_bp} = \frac{\text{ap\_hi} + 2 \times \text{ap\_lo}}{3}$

TABLE V: Cross-Model Feature Importance Comparison

	Importance		Rank		
Feature	**Classification**	**Regression**	**Classification**	**Regression**	**Difference**
BP Category Hypertension	0.461	<1E-04	1	19	18
Systolic Pressure (ap_hi)	0.256	0.745	2	1	-1
Cholesterol	0.087	0.036	3	5	2
BP Category	0.066	<1E-04	4	19	15
Age	0.047	0.012	5	6	1
Glucose	0.017	0.010	6	8	2
Physical Activity	0.016	0.005	7	13	6
Weight	0.012	0.006	8	10	2
Smoking	0.010	0.002	9	18	9
Alcohol	0.010	<1E-04	10	19	9
BMI	0.007	0.005	11	14	3
Gender	0.006	0.012	12	7	-5
Height	0.004	0.006	13	11	-2
Age-Weight Interaction	<1E-04	0.053	14	2	-12
Age-BMI Interaction	<1E-04	0.045	14	3	-11
Cholesterol-Glucose	<1E-04	0.040	14	4	-10

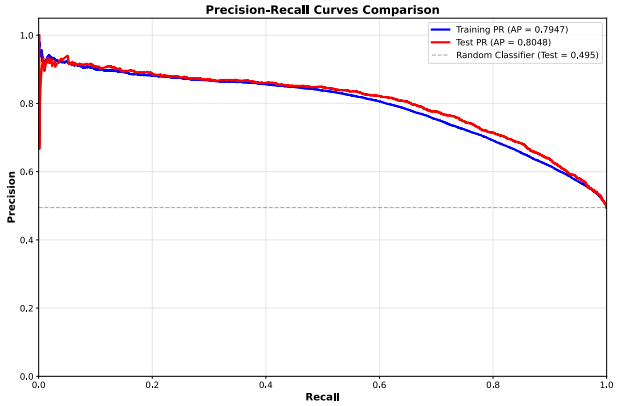

Fig. 3: Precision-recall curves comparing classification model configurations.

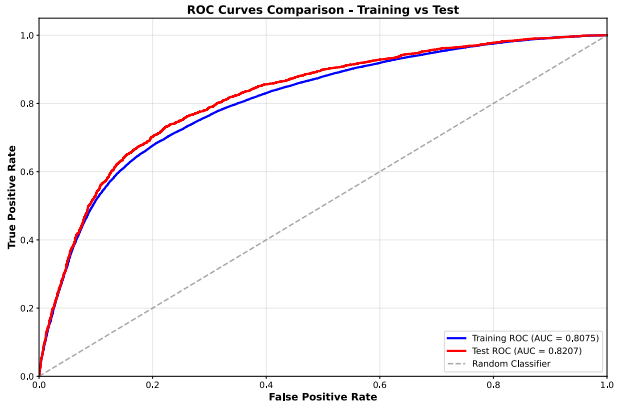

Fig. 4: ROC curves showing diagnostic ability of classification models.

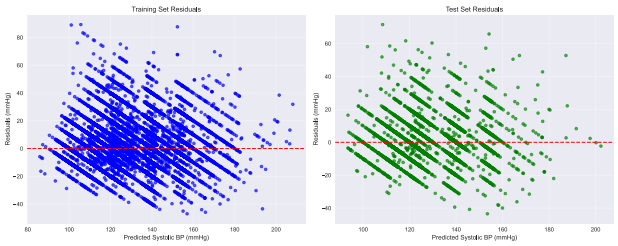

Fig. 5: Residual plots for regression models showing error distributions.

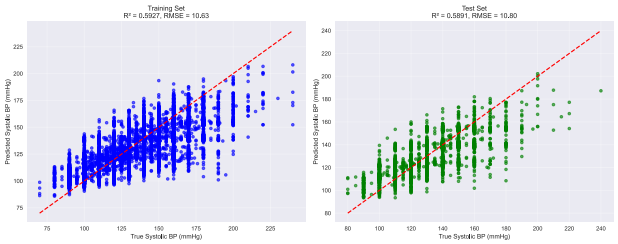

Fig. 6: True vs. predicted values correlation for regression model.

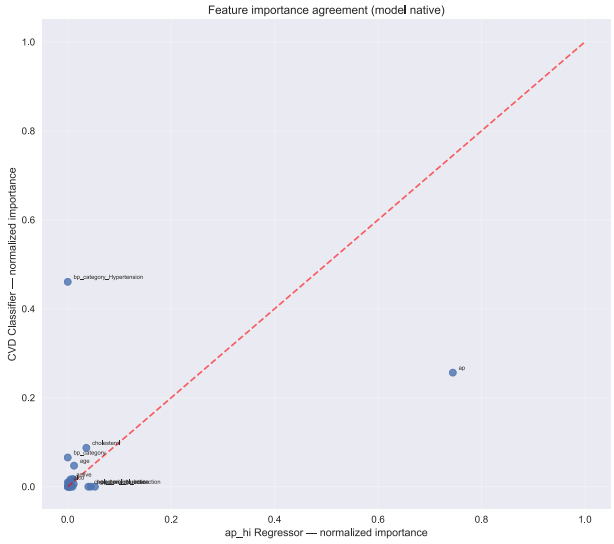

Fig. 7: Scatter plot comparing normalized feature importance between cardiovascular disease classifier and systolic blood pressure regressor.

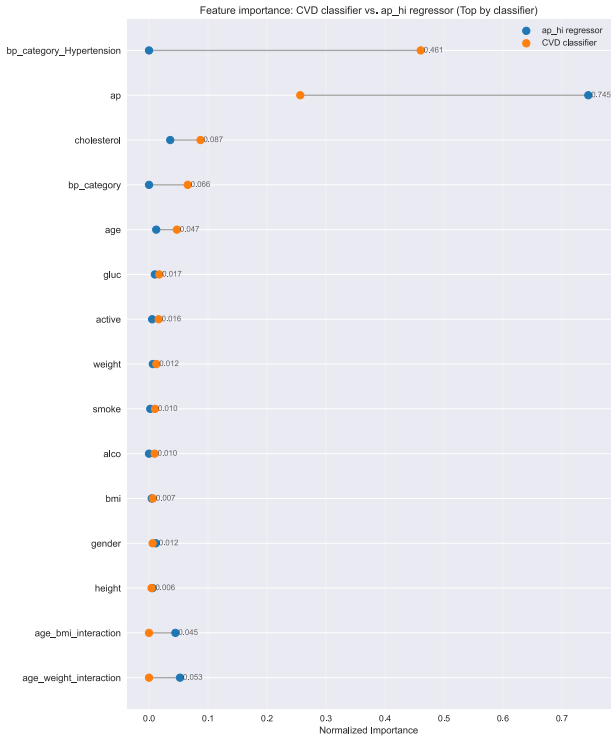

Fig. 8: Dumbbell chart comparing feature importance between models for the top 15 features.