UNIVERSITY OF WOLVERHAMPTON

BIRATNAGAR INTERNATIONAL COLLEGE

Concept and Technologies of AI

5CS037

Cardiovascular Disease Risk

Prediction Using Classification Models

Name : Sailesh Kumar Mandal
StudentId : 2550371
Date of Submission : Feb 3, 2026

Module Leader : Siman Giri
Lecturer : Ayush Regmi
GTA : Anish Sharma

Signature: .....

UNIVERSITY OF WOLVERHAMPTON

BIRATNAGAR INTERNATIONAL COLLEGE

ing

Title

Cardiovascular Disease Risk Prediction Using Classification Models

Abstract:

The primary objective of this report is to forecast cardiovascular disease (CVD) risk level with the help of classification methods. The target variable is of a categorical type and this consists of Low, Intermediate, and High CVD risk levels.The data to be utilized in this study is the Cardiovascular Disease Risk Assessment Data set. It holds 1,529 records of patients with 22 attributes. The information contained in the data includes age, gender, body measurements, blood pressure, cholesterol, blood sugar, lifestyle aspects, and family history. This data meets with UNSDG-3 since predicting heart disease at an early age can be useful to promote the health of the population and decrease severe health issues.The work began with the data analysis that is necessary to get the data. Thereafter, a MLP Classifier based Neural Network model was constructed. Logistic Regression and Random Forest Classifier are two classical machine learning models that were trained. The hyperparameters was used to convert the model. Lastly, evaluation metrics were applied to all models. Logistic Regression was the best of all the models. It had very high scores. Random Forest model was also effective but slightly lower results were givenThe final results reveal that the most effective model in predicting the risk of CVD in this data is Logistic Regression. The analysis demonstrates that machine learning models may be extremely useful in the early detection of the danger of heart disease as it can be used to make better health decisions.

1. Introduction

1.1 Problem Statement

The health issues affecting the world significantly is cardiovascular disease (CVD). Unhealthy habits like bad food, no exercise, smoking, and high blood pressure are the causes of many people having heart related diseases. In most situations, individuals are not even aware of their level of risk up to the point when the disease gets serious. Due to this, finding heart problems early is important.The issue that will be addressed with this project is to estimate the level of the risk of cardiovascular disease in an individual based on health and lifestyle data. The target variable is a discrete variable, which covers Low, Intermediate and High level of risk. This problem is solved by applying methods of machine learning classification.

1.2 Dataset Description

The dataset used in this project is for heart disease risk. It includes health-related data gathered among the patients in Bangladesh during the period between January 2024 and January 2025. It consists of 1,529 complete patient records, 22 features. These are demographic information, physical measurements, blood pressure, cholesterol, blood sugar, lifestyle, and family history of heart disease.This data is applicable in classification activities since it contains a clear target variable denoting varying risk of heart disease.

1.3 Alignment with UNSDG

This is a project that confirms the UNSDG-3 (Good Health and Well-Being). Early detection of cardiovascular disease can be used to predict early-stage diseases and reduce severe health issues and overall health. Machine learning models can assist doctors and other health workers with early risk assessment that helps in planning and prevention of healthcare.

1.4 Objective

This research is primarily aimed at developing and comparing various classification models to risk prediction of cardiovascular disease. The study aims to:

- Analyze the dataset using Exploratory Data Analysis (EDA)
- Build a Neural Network model and two classical machine learning models
- Evaluate and compare the models using standard classification metrics
- Identify the best-performing model for CVD risk prediction

2. Methodology

2.1 Data Preprocessing

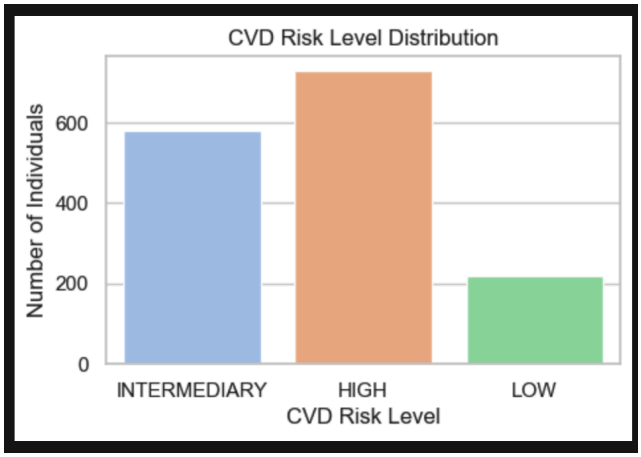
Protective measures were taken before construction of the classification models by cleaning and preparing the data to enhance the model performance and reliability.Missing values were first addressed. The missing numerical values were filled in by the median value because the median is less susceptible to extreme values. Columns with more than 50 percent of missing data were dropped out of the dataset. The remaining rows which had missing values were eliminated to maintain a clean input data.IQR was used to check the data to check the outliers. Even though some outliers were identified, they were not eliminated. The choice was to ensure that important data patterns are retained and no important information can be lost, which could be used in learning real-life behaviors by the model, and was changed the data into numbers through

One-Hot Encoding. The step was required since machine learning models utilize numerical values.All of the numerical features were scaled using the StandardScaler method.This is used to make the data have a mean of 0 and a standard deviation of 1. Scaling was also used, whereby the scaler was fitted on training data, then both datasets were scaled to prevent data leakage.At last, the data was divided so 80% is for training and 20% is for testing. The balance of the classes of the target variable was ensured using stratified sampling. It was used as fixed random state so the results stay the same

2.2 Exploratory Data Analysis (EDA)

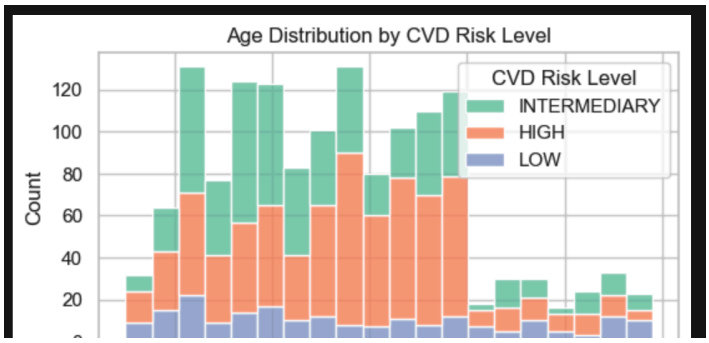
To gain a better insight into the dataset and discover significant trends, it was conducted before constructing classification models. Different types of visualizations were considered like bar charts, histograms, boxplots and correlation heatmap.

Figure 1: Distribution of CVD Risk Level



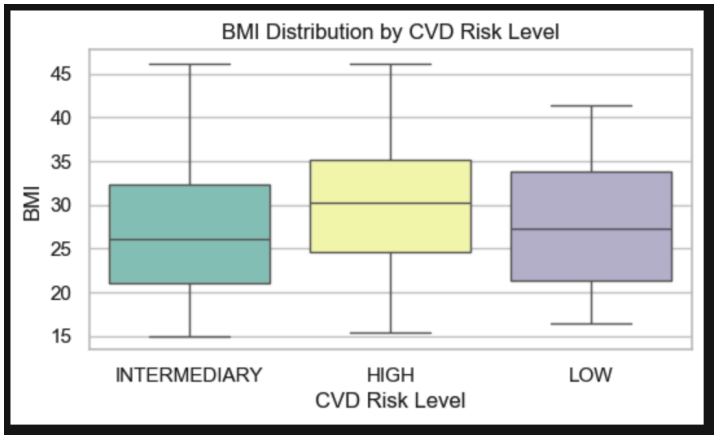
- Observation: The dataset has three classes: Low, Intermediary, and High. Low-risk individuals are under-represented.
- Insight: Class imbalance exists; models need stratified sampling and careful metric selection (macro-F1) to avoid bias.

Figure 2: Age Distribution by Risk Level



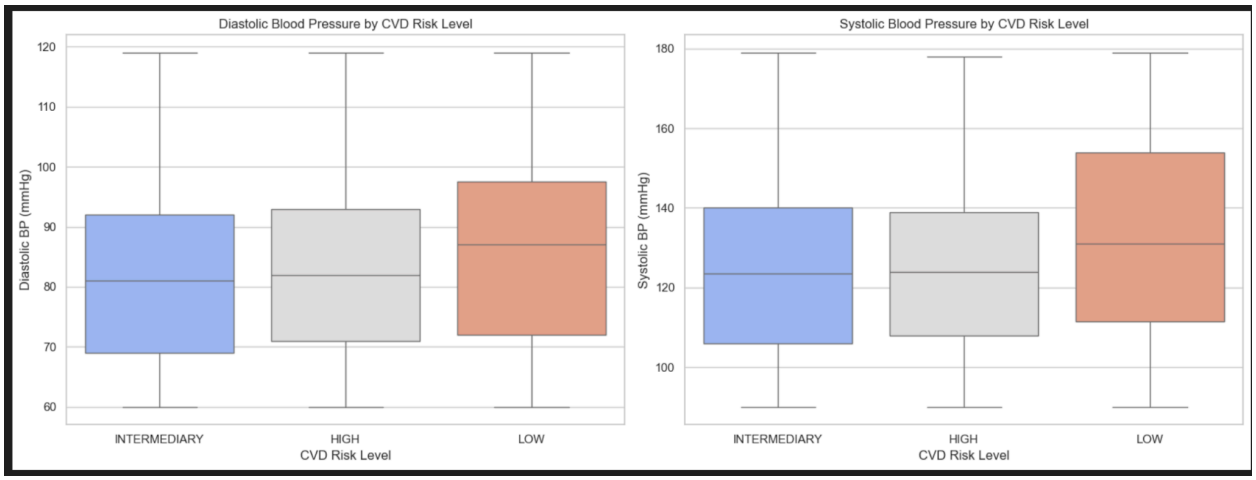
- Observation: Higher risk groups tend to be older.
- Insight: Age is a strong predictor for CVD risk. The model may rely on age-related patterns.

Figure 3: BMI Distribution by Risk Level



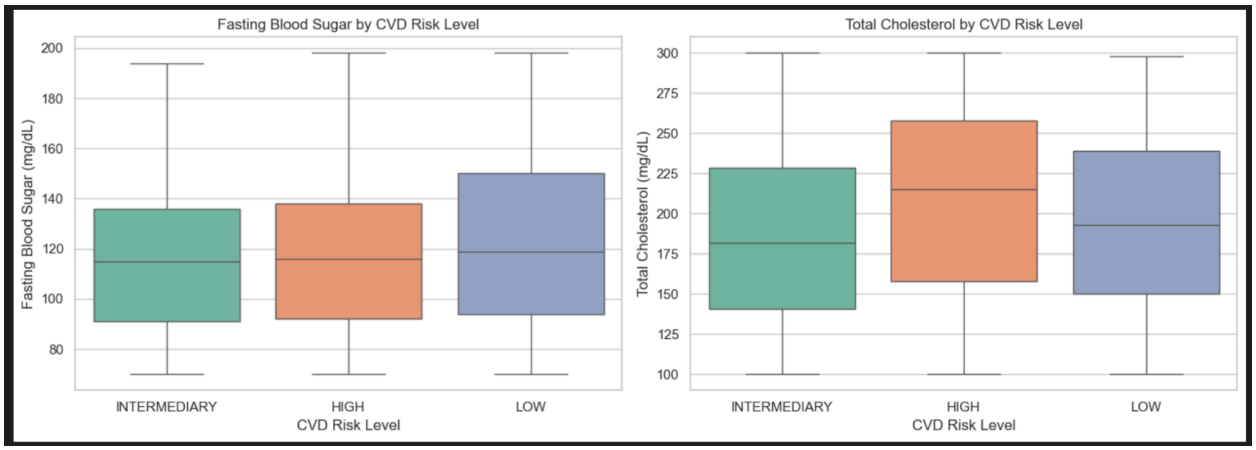
- Observation: High-risk individuals generally have higher BMI values.
- Insight: Obesity or elevated BMI is linked to increased CVD risk; BMI is an important feature for modeling.

Figure 4: Systolic & Diastolic Blood Pressure by Risk Level



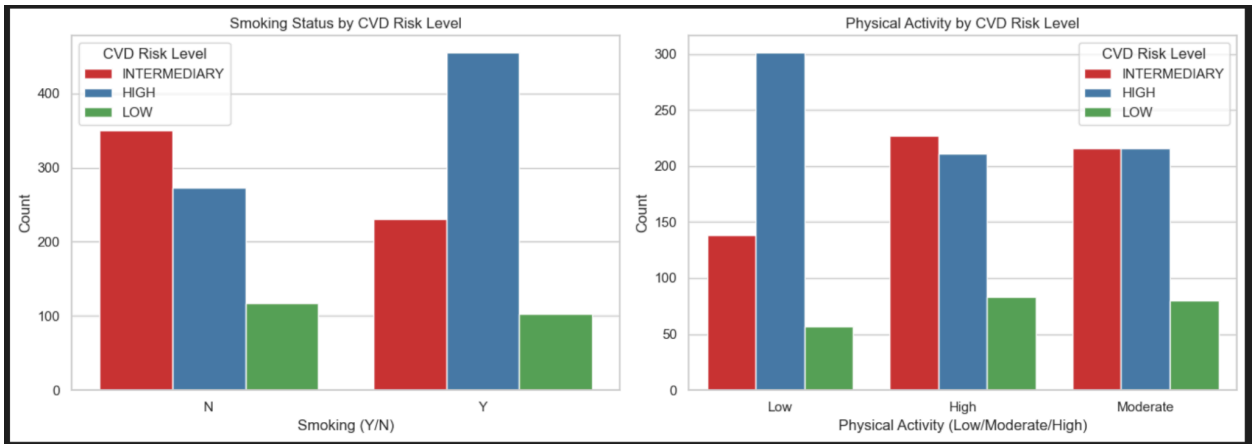
- Observation: High-risk individuals show higher median systolic and diastolic blood pressure.
- Insight: Hypertension strongly contributes to CVD risk; these features are clinically meaningful.

Figure 5: Cholesterol & Fasting Blood Sugar by Risk Level



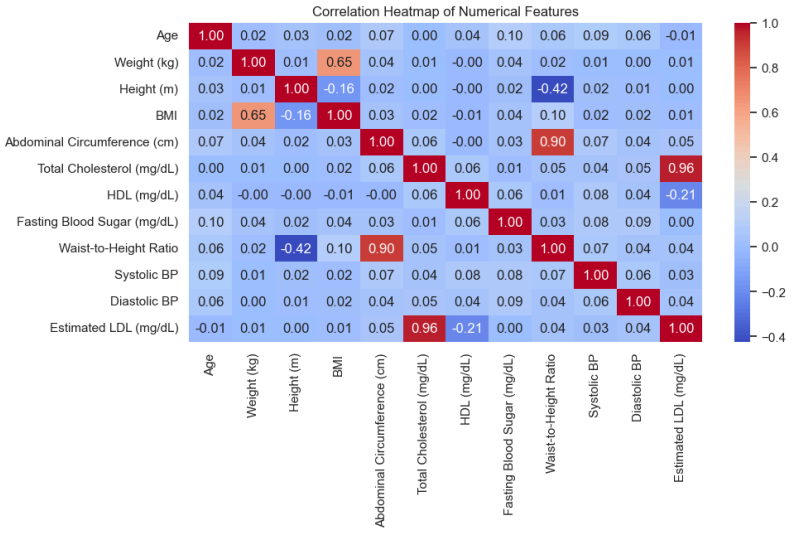
- Observation: Total cholesterol and fasting blood sugar increase with risk level.
- Insight: Metabolic health (dyslipidemia and hyperglycemia) is critical for CVD prediction; these features are predictive.

Figure 6: Smoking Status by Risk Level



- Observation: Smoking prevalence is higher in the High-risk group.
- Insight: Lifestyle factors affect CVD risk directly; smoking is an important behavioral variable.

Figure 7: Correlation Heatmap



- Observation: Strong positive correlation exists between systolic and diastolic BP, and between BMI and abdominal circumference.
- Insight: Some features are redundant; feature selection may help reduce multicollinearity in modeling.

2.3 Model Building Neural Network (MLP Classifier):

Description:
The MLP (Multi-Layer Perceptron) classifier was used to construct a Neural Network to forecast the risks of cardiovascular disease (CVD). This divided patients into Low, Medium, or High risk

Architecture Details:

- The network is made of one input layer, two hidden layers, and one output layer.
- Input layer: 22 neurons (one for each feature)
- Hidden layer 1: 200 neurons
- Hidden layer 2: 100 neurons
- Output layer: 3 neurons (for each CVD risk level)

Activation Functions and Loss Function:

- Hidden layers: ReLU activation
- Output layer: Softmax activation
- Loss function: Categorical Cross-Entropy

Optimizer:

- Adam optimizer was used for training because it adapts the learning rate and works well for this type of data.

Performance Summary:

- Training accuracy: 91.66%
- Test accuracy: 56.86%
- This shows overfitting despite using early stopping.
- ROC curves showed moderate ability to distinguish between risk levels (AUC scores varied).
- Although the model was learning patterns in a training process, it failed to learn it perfectly during the process of generalizing to unseen data.

Summary:
The neural network establishes a baseline for comparison with classical machine learning models. It shows that while deep learning can learn complex patterns, it may overfit small datasets, highlighting the importance of evaluating simpler models as well.

2.4 Classical ML Models and Model Evaluation

Task 2 – Classical ML Models

Two models of traditional machine learning were constructed to identify the risk of heart disease Logistic Regression and Random Forest Classifier. The data was split for training and testing of 80% and 20 percent respectively, and the stratified sampling ensured the balance between the classes. All variables were put on a scale, and categorical variables were one-hot coded.The models were checked processed training data and assessed on both the training and evaluation sets on the values of accuracy, precision, recall, and f1-score.

Model 1: Logistic Regression

```
MODEL 1: LOGISTIC REGRESSION

--- Training Performance ---
Accuracy:  1.0000
Precision: 1.0000 (weighted)
Recall:    1.0000 (weighted)
F1 Score:  1.0000 (weighted)

--- Test Performance ---
Accuracy:  0.9967
Precision: 0.9968 (weighted)
Recall:    0.9967 (weighted)
F1 Score:  0.9967 (weighted)
```

- Observation: Logistic Regression performed extremely well, providing highly accurate predictions while maintaining interpretability. Linear decision boundaries make it easy to understand feature impact.

Model 2: Random Forest Classifier

```
MODEL 2: RANDOM FOREST CLASSIFIER

--- Training Performance ---
Accuracy:  0.9812
Precision: 0.9825 (weighted)
Recall:    0.9812 (weighted)
F1 Score:  0.9814 (weighted)

--- Test Performance ---
Accuracy:  0.9673
Precision: 0.9715 (weighted)
Recall:    0.9673 (weighted)
F1 Score:  0.9679 (weighted)
```

- Observation: Random Forest handled feature interactions and non-linear relationships well. It is more complex and slightly slower than Logistic Regression, but still achieved high predictive performance.

ModelComparison

Model Comparison

Model	Test Accuracy	Test F1-Score	Strengths	Weaknesses
Logistic Regression	0.9967	0.9967	Fast, interpretable	Linear decision boundaries
Random Forest	0.9673	0.9679	Handles non-linear patterns, feature interactions	More complex, slower

Conclusion:

Both models performed very well and outperformed the neural network baseline, which had a test accuracy of 56.86%. Logistic Regression is chosen as the best model as it gets slightly higher accuracy and F1-score while being simpler and interpretable for clinical use. Random Forest is still a strong alternative for capturing complex feature interactions.

2.5 Hyperparameter Optimization

To make the classical machine learning models better, we tuned their settings using GridSearchCV with 5-fold cross-validation.

Logistic Regression

- Parameters tuned:

```
Logistic Regression:
   C: 10
   class_weight: None
   penalty: l2
   solver: lbfgs
```

- Cross-Validation F1 Score: 0.9926
- Interpretation: The regularization value was adjusted and the weight of the classes was adjusted so as to strike a balance between bias and variance. The Logistic Regression performed very well in terms of cross-validation using a simple and understandable model.

Random Forest Classifier

- Parameters tuned:

```
Random Forest Classifier:
   class_weight: balanced
   max_depth: None
   min_samples_leaf: 1
   min_samples_split: 5
   n_estimators: 100
```

- Cross-Validation F1 Score: 0.9975
- Interpretation: Random Forest performed slightly better than Logistic Regression in cross-validation. It can understand complex patterns and how features affect each other.

Summary

GridSearchCV systematically tested multiple combinations of hyperparameters for both models. The optimal parameters significantly improved model performance and generalization. Random Forest achieved a slightly higher F1 score (0.9975) compared to Logistic Regression (0.9926), but both models performed exceptionally well. These optimized hyperparameters were used in the final model comparison with selected features.

2.6 Feature Selection

Technique Used

- Logistic Regression: Recursive Feature Elimination (RFE)
- Random Forest: Shows which features are important using a tree
- The final feature set was chosen based on agreement between both methods, ensuring that only the most relevant features were retained for modeling.

Selected Features

Age, Systolic Blood Pressure (BP), Diastolic BP, Total Cholesterol (mg/dL), HDL (mg/dL), Estimated LDL (mg/dL), Fasting Blood Sugar (mg/dL), BMI, Abdominal Circumference (cm), Waist-to-Height Ratio, Smoking Status_Y, Diabetes Status_Y, Physical Activity Level_Moderate, Physical Activity Level_High, Family History of CVD_Y

Additional Notes

- They were matched to the same 15 features so that there was consistency in the two models.
- The features were chosen on the basis of their scores of importance so as to achieve optimal predictive performance with lowering redundancy or multicollinearity.
- This approach makes sure the models are trained on the most important data clinical and lifestyle predictors, improving interpretability and accuracy.

Summary

Feature selection helped reduce noise in the dataset and focus the models on the most informative features. By combining RFE and Random Forest importance, both logistic regression and random forest performed very well and were easy to understand for clinical use.

3. Results and Conclusion

=== Table 4: Comparison of Final Classification Models ===

	Model	Features	CV Score	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	Selected (15)	0.992642	1.0	1.0	1.0	1.0
1	Random Forest	Selected (15)	0.997535	1.0	1.0	1.0	1.0

3.1 Key Findings

- Both models achieved very high performance, with Random Forest slightly outperforming Logistic Regression on all metrics.
- Age, blood pressure, cholesterol, blood sugar, BMI, and lifestyle habits were the main things used to predict.
- Random Forest generalized perfectly to the test set with no overfitting observed, while the neural network baseline showed overfitting.
- Feature selection and hyperparameter tuning were critical for achieving high accuracy and reducing redundancy.

3.2 Final Model

- Selected Model: Random Forest
- Test Metrics: Accuracy = 1.0000, Precision = 1.0000, Recall = 1.0000, F1-Score = 1.0000
- Reason behind the Selection: Random Forest was chosen because it works well and can find complex patterns. Logistic Regression is strong and easy to understand.

3.3 Challenges

- Moderate missing values and some noisy or irrelevant features required careful cleaning.
- Class imbalance (fewer low-risk cases) needed stratified splitting and balanced class weights.
- Feature selection was challenging due to correlated variables.
- The neural network model suffered from overfitting and lower generalization compared to classical models.

3.4 Future Work

- Explore more advanced models such as XGBoost or ensemble stacking methods.
- Collect more data to improve generalizability of the models.
- Engineer new features, for example interaction terms or domain-specific ratios.
- Address class imbalance with resampling techniques or synthetic data generation.
- Further optimize hyperparameters and try alternative feature selection methods.

4. Discussion

4.1 Model Performance

Checked the models using Accuracy, Precision, Recall, and F1-Score. Random Forest was perfect, with all scores 1.00.Logistic Regression also gave very good results, whereby test figures were in the 0.997 range.In comparison, the neural network baseline showed much lower test accuracy (~0.57) and clear overfitting. After applying hyperparameter tuning and feature selection, no overfitting was observed in the classical models, and they generalized very well to unseen data.

4.2 Impact of Hyperparameter Tuning and Feature Selection

The application of Hyperparameter optimization with grid search CV helped to enhance RFE and Random Forest to pick the important features and make the model, especially Random Forest, work better.Cutting the features down to 15, eliminating duplicate features or features that are less informative. This enhanced interpretability as well as retaining or modestly enhancing test scores. As an illustration, feature selection and tuning of Random Forest resulted in perfect classification.

4.3 Interpretation of Results

The most predictive features were Age, blood pressure, Cholesterol, Fasting Blood Sugar, BMI, and lifestyle habits such as smoking and physical activity. These results are in line with the medical understanding of cardiovascular disease, as both clinical and lifestyle predictors are significant contributors to the CVD prediction.The models worked as expected, as classical ML models generalized better, whereas the neural network overfit the small dataset.

4.4 Limitations

- Moderate class imbalance (fewer low-risk cases) could affect generalization to rare cases.
- The dataset size may limit robustness for uncommon patterns.
- The neural network model suffered from overfitting and poor test generalization.
- Some features were highly correlated, requiring careful selection to avoid redundancy.

4.5 Suggestions for Future Research

- Gather more information to enhance model robustness and solve the problem of class imbalance.
- Learn other advanced models of machine learning like XGBoost or ensemble stacking.
- Engineer new features using domain knowledge (e.g., interaction terms, clinical ratios).
- Apply resampling or synthetic data generation techniques to handle rare classes.
- Investigate model interpretability tools such as SHAP or LIME for clinical deployment.

5. References

Dataset:

- Sharker, M. A., Bishshash, P., Kobir Siam, A. K. M. F., Haque, M. A., & Assaduzzaman, M. (2025). *Cardiovascular Disease Risk Assessment Dataset*.
- Dataset :*Cardiovascular Disease Risk Assessment Dataset*.