

Cardiovascular Disease Risk Classification Using Machine Learning with Weighted Feature Fusion and Explainable AI on Bangladeshi Clinical and Lifestyle Data

*Tasnimul Intazam Asif¹, Bishwaprotap Ray¹, Md. Alomgir Hossain¹, Faisal Imran², Prime Biswajit Barua¹, Nishat Salsabil Anisha¹, Ariful Haque Minhaj¹, Amit Roy¹

¹Department of Computer Science and Engineering, International University of Business and Agricultural Technology, Dhaka, Bangladesh

²Department of Computer Science and Engineering (CSE) & CSIT, Shanto-Mariam University of Creative Technology, Dhaka, Bangladesh.

Article Info

Article history:

Received Nov 1, 2025

Revised Jan 21, 2026

Accepted Feb 8, 2026

Keywords:

Cardiovascular Disease,
Weighted Feature Fusion,
Explainable AI,
SHAP,
LIME,
Random Forest,
Risk Prediction

ABSTRACT

Cardiovascular disease (CVD) is one of the top causes of death across the world, and there is a need to develop early risk prediction models that can be accurate and interpreted. This study introduces a weighted feature fusion (WFF) model of machine learning to integrate clinical, lifestyle, and engineered features into an integrated machine learning model to improve the classification of CVD risk and the interpretability of the model. Several classifiers, such as the Random Forest (RF), K-Nearest Neighbors (KNN), XGBoost (XGB), Bagging, Decision Tree were trained and tested based on fusion-based methods. The experimental findings indicated that the highest classification accuracy of the model at 91% obtained by the Random Forest model. Moreover, the model was better in terms of discrimination as ROC-AUC scores were over 0.980447 in all categories of CVD risk. Explainable AI algorithms, such as SHAP and LIME were used to provide transparency, when combined with feature fusion, leads to a significant improvement in accuracy, reliability, and interpretability of CVD risk prediction models that can lead to the development of data-driven healthcare decision support systems of trust.

Copyright © 2026 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Tasnimul Intazam Asif
Department of Computer Science and Engineering,
International University of Business and Agricultural Technology,
4 Embankment Drive Road, Sector-10, Uttara Model Town, Dhaka-1230,
Email: intazam2896@gmail.com

1. INTRODUCTION

Cardiovascular disease (CVD) has been the number one cause of death in the world and is a major health problem in the world because it is a major problem in developing nations such as Bangladesh where the medical facilities are inadequate, and the preventive care mechanisms are yet to be established. Having been the leading cause of death in Bangladesh with close to 30 percent contributing to all fatalities in the country, CVD has impacted both city and rural areas with ailments like cardiac arrest, high blood pressure, and stroke. Some of the risk factors cited to have largely contributed to the rise in prevalence of CVD include aging of population, sedentary living, unhealthy diets, obesity, diabetes, and tobacco consumption. Despite the advances in the country in terms of diagnostic services like ECG, echocardiography, and stress testing, the lack of cardiologists, the lack of specialized centers, the high cost of treatment, and the fact that the population has not been sensitized on the topic especially in the rural areas, still weighs heavily on the healthcare system. Such a state of affairs highlights the urgent need to develop data-driven, cost-efficient and interpretable predictive models to help in early risk identification and better management of CVD in resource-strained areas.

Although machine learning (ML)-based CVD prediction studies have greatly advanced in the world, there are still a few limitations that do not allow them to be translated to the real world. Most of the available

research is based on publicly available data, e.g., Cleveland, Statlog and Framingham, which represents the West or urban population, and does not reflect the clinical and lifestyle realities of the South Asian or rural population. These models tend to refer to a small range of clinical or demographic characteristics that solely ignore very important behavioral and socioeconomic factors like diet, physical activity, smoking habits, and stress which are very important concerning cardiovascular health. In addition, most current studies reduce the prediction problem to binary classification (CVD or not), ignoring the fact that the gradient of risk levels low, intermediary, and high are more clinically relevant. The other important problem is the absence of interpretability whereby most ML models are black boxes and hence, clinicians cannot know how the prediction was made and thus, restricting trust and practicability in the health sector. Lastly, a lot of the past works have been built on small or homogeneous datasets resulting in overfitting and low generalizability to different populations.

Analysis of the literature in the field reveals that several models have demonstrated high predictive levels and have weaknesses in terms of contextual and methodological constraints. Ozcan et al. and Yahya et al. used decision trees and random forest models on benchmark data, which showed high accuracy but did not consider lifestyle or socioeconomic factors [1,2]. On the same note, Hossain et al. implemented a dataset that only included 391 samples of Bangladesh specificity but only binary classification and a limited number of risk factors [3]. Ensemble-based algorithms, including the ones suggested by Chandrasekhar et al., used a mix of Random Forest, KNN, and Gradient Boosting with 95 percent of accuracy, but their research did not include a strong feature selection and seemed to have used small datasets [4]. Baghdadi et al. combined several global datasets and achieved a high level of performance, but lack of contextual specificity [5]. The model proposed by other studies, such as Pal et al. and Mahmud et al., had a high accuracy but did not use rigorous cross-validation and interpretability [6,7]. Very little research like Al Gharib et al. used Explainable AI (XAI) methods like SHAP and LIME, but no data has used such explainable and optimized ML systems with localized Bangladeshi datasets that combine both clinical and lifestyle data [8].

Most of the studies done on ML-based CVD are based on publicly available datasets (like Cleveland, Statlog, or Framingham) that include mostly Western or urban data. Although their predictive power is quite high, such works do not always work in situations of South Asian or rural population, whose lifestyle, eating habits, and access to health care vary significantly. As an example, the decision trees and random forests used in [1] and [2] on benchmark datasets achieved accuracy of over 90 percent, but the models did not include lifestyle or socioeconomic factors that are important in non-Western settings. Likewise, Baghdadi et al. [5] summarized several international datasets, but they were not regionalized. Our paper bridges this gap by presenting an original dataset, CAIR-CVD-2025, comprising of 1,529 patient records of Jamalpur Medical College Hospital located in the rural Bangladesh. This data set combines clinical, biochemical, anthropometric and lifestyle data, which forms a basis of risk modeling at the regional level.

The risk of CVD is simplified to a binary classification problem (CVD vs. non-CVD) by many of the existing frameworks and the gradual risk severity that is clinically significant is not taken into consideration. As an example, Hossain et al. used ML on a sample of Bangladeshi participants but analyzed their data in terms of binary outcomes [3]. The simplification of this kind neglects the fact that stratified interventions are needed to work with low-, intermediate-, and high-risk individuals. Conversely, our model uses a three-class risk stratification (low, intermediate, high) as it is compatible with clinical guidelines and allows us to assess the risk of individuals at risk and provide more subtle preventive care.

ML interpretability is still a major impediment to clinical adoption. Majority of the studies focus on accuracy rather than explainability hence producing black-box systems which do not provide much information about the contribution of features. Although some recent studies, including Al Gharib et al., have implemented the use of Explainable AI (XAI) methods, including SHAP and LIME, such attempts are limited and tend to be done over local data and not a mixture of clinical and lifestyle data. We explicitly integrated XAI into our framework where SHAP is used as the global interpretability and LIME as the local explanations. Such a two-sided approach will guarantee that clinicians become familiar with general model behavior, as well as the individual forecasts, which will result in the development of trust and the ability to act.

The representation of features is usually made on simple clinical variables, and very less attention is paid to derived or interactive features that exhibit complex physiological relationships. Whereas there are studies that utilize feature selection techniques like Chi-square or Correlation-Based Feature Selection (CFS), there are rare studies that consider multimodal fusion of heterogeneous data sources [9]. Furthermore, the accuracy of ensemble methods such as Chandrasekhar et al. is high; however, they use small datasets without strong feature combination. To address these constraints, we proposed Weighted Feature Fusion (WFF) mechanism which is dynamically sensitive to combine clinical, lifestyle, and engineered features and uses optimized weights assigned to every feature subset, in contrast to simple concatenation, based on Optuna-based Bayesian hyperparameter tuning, to make sure that the model does not overvalue contributions by one data modality over another. It is supplemented with a strict feature engineering pipeline which produces new

indices, i.e. the Cardiovascular Index, BP-Cholesterol Interaction and Metabolic Index, all aimed at capturing multifactorial risk interactions.

Though machine learning (ML) has demonstrated significant potential in CVD prediction, available models are frequently characterized by a variety of severe limitations: (i) they use western-centric datasets (e.g., Cleveland, Framingham) that do not have to do with South Asian or rural areas; (ii) they overly simplify the notion of risk, making it binary (CVD vs. non-CVD); (iii) they are opaque, and thus can be used as black boxes in clinical practice; and (iv) they do not combine heterogeneous types of data (e.g., clinical measures). Recent progress in multimodal machine learning showed that combining heterogeneous data sources can significantly enhance the model robustness, its accuracy and its interpretability. In recognising emotions, transformer-based and adversarial fusion networks are used to concurrently process audio, visual and textual features to improve performance in noisy, natural dialogues [10,11]. Likewise, Wang et al. takes advantage of contextual interactions to optimize semantic alignment between modalities [12], whereas Zhu et.al presents a visualization framework on the predictive on the side of the end user, which renders multimodal outcomes interpretable. These methods demonstrate the significance of adaptive fusion algorithms and interpretability which are also immediately applicable to medical prediction where inputs are commonly multimodal (e.g., clinical, behavioral, imaging) and outputs are expected to be transparent [13].

Minimization Medical imaging and physiological signal analysis medical imaging and physiological signal analysis have used ideas like Multi-Branch Weighted Feature Fusion [14] and Generative Random Modality to address missing data and enhance generalization [15]. These approaches place various streams of features during training to encourage robustness, a technique we would adjust in our framework of Weighted Feature Fusion (WFF) to trade-off clinical, lifestyle, and engineered features.

Our conceptual framework of WFF agreement with the recent multimodal fusion paradigms. An example is that in the case of DNLN, deformable attention and weighted branches are used to orchestrate the multi-scale image features, similarly, we are using optimized weights to orchestrate clinical, lifestyle, and engineered feature streams. On the same note, RAFT and Generative Random Modality Dropout robustness techniques suggest that approach breaks missing data and class imbalance through SMOTE-ENN, stratified cross-validation. In addition, interpretability emphasis in CIME and EMVAS is the reason that prompts us to use SHAP and LIME to illustrate the feature contributions and the decision-making pathways.

Overall, the study provides a context-sensitive, comprehensible, and extensible predictive model that may aid in early diagnosis and preventative interventions of CVD in resource-inadequate rural areas, providing important perspectives on subsequent use in the community health and clinical decision support systems.

Table 1. Methodological Advancements

Category	Common Practice in Literature	Our Work's Novelty
Dataset	Mostly UCI, small or aggregated, non-local	CAIR-CVD-2025: Rural Bangladeshi clinical + lifestyle data [16]
Risk Stratification	Binary classification (CVD vs. non-CVD)	Three-class risk levels (low, intermediate, high)
Imbalance Handling	Oversampling applied outside cross-validation	SMOTE-ENN inside stratified CV pipeline
Feature Selection	Chi-square, CFS, or manual pruning	RFE, ANOVA-F, Mutual Info, PCA integrated in preprocessing
Feature Fusion	None or simple concatenation	Weighted feature fusion (clinical vs lifestyle) with optimized modality weights
Optimization	GridSearchCV or heuristic methods (PSO/SGO)	Optuna-based Bayesian tuning for multi-model hyperparameter optimization
Evaluation Metrics	Accuracy, F1 score	ROC-AUC, PR-AUC, Recall, K-Fold Cross Validation
Explainability	Rarely addressed; occasional SHAP/LIME use	Integrated SHAP & LIME for global and local interpretability

2. RESEARCH METHOD

This section endeavors to the framework of the study's plan including data collection, Data descriptions, data preprocessing, data cleaning, feature engineering, optimization (optuna), model training-testing along with evaluation metrics and cross validation, contribution. Throughout this process we will classify the CVD state classification as severity level (High, Intermediary, Low) considering the clinical and lifestyle data.

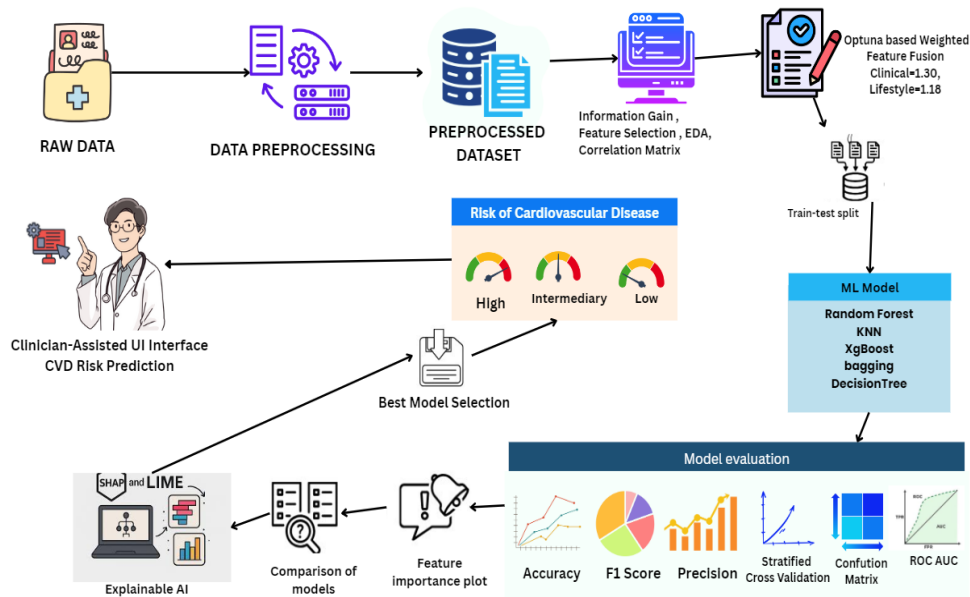


Figure 1. Proposed Methodology

2.1. Data collection

The dataset, obtained in Mendeley [16], contains 1,529 records of patients of different ages and has 22 attributes reflecting clinical, demographic, and lifestyle characteristics. The levels of CVD risk distribution in the dataset involve 717 patients (47.57%) high risk, 565 patients (37.75%) intermediary risk and 229 patients (15.03%) low risk. All features covered in the study have been described in Table 2.

Table 2. Characteristics of the dataset

No.	Column Name	Non-Null Count	Data Type	Feature group
1	Sex	1,529	object	Demographic
2	Age	1,451	float64	Demographic
3	Weight (kg)	1,448	float64	Anthropometric
4	Height (m)	1,462	float64	Anthropometric
5	BMI	1,465	float64	Anthropometric
6	Abdominal Circumference (cm)	1,462	float64	Anthropometric
7	Blood Pressure (mmHg)	1,529	object	Clinical
8	Total Cholesterol (mg/dL)	1,456	float64	Clinical
9	HDL (mg/dL)	1,449	[	Clinical
10	Fasting Blood Sugar (mg/dL)	1,462	float64	Clinical
11	Smoking Status	1,529	object	Lifestyle Factors
12	Diabetes Status	1,529	object	Lifestyle Factors
13	Physical Activity Level	1,529	object	Lifestyle Factors
14	Family History of CVD	1,529	object	Lifestyle Factors
15	CVD Risk Level	1,529	object	Anthropometric
16	Height (cm)	1,455	float64	Anthropometric
17	Waist-to-Height Ratio	1,450	float64	Anthropometric
18	Systolic BP	1,458	float64	Clinical
19	Diastolic BP	1,447	float64	Clinicals
20	Blood Pressure Category	1,529	object	Engineered Features
21	Estimated LDL (mg/dL)	1,460	float64	Clinical
22	CVD Risk Score	1,459	float64	Anthropometric

2.2. Data Preprocessing

Preprocessing provides data quality and consistency prior to model training. This raw data was organized and cleansed into an analysis form that could be modeled. Also, duplicate records were eliminated, and feature names were made standard to enhance ease of reading and interpreting data. Principal component analysis (PCA), Recursive Feature Elimination (RFE), ANOVA-F, and Mutual Information (MI), Simple

Imputer and RobustScaler were some of the major components integrated in the preprocessing pipeline. The attribute of Cardiovascular Index had missing values that were replaced with median value.

2.3. Handling Missing Values

Nominal features e.g. Sex, Smoking status, Diabetes status, and Family History of CVD were one-hot coded. Binary values (0 and 1) were assigned to binary attributes (e.g., Smoking, Alcohol Consumption, Diabetes). Ordinal items, such as Physical Activity Level, CVD Risk Level and Blood Pressure Category, were encoded according to intrinsic order, as follows: Level of physical activity: Low - 0, Intermediary -1 and High-2. Level of CVD Risk: Low - 0, Intermediary - 1, High - 2. Blood Pressure Type: normal - 0, elevated - 1, hypertension stage 1 -2, hypertension stage 2 - 3. To indicate the level of risk which were [0, 1, 2], the target variable (CVD Risk Level) was labelled-encoded.

Normalization Min-max scaling was performed on the continuous variables, Age, Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Cholesterol and BMI. Class Balancing The dataset had a slight imbalance between the categories of CVD risks. To counter this, the Synthetic Minority Oversampling Technique (SMOTE) was implemented to create synthetic samples of the classes that are underrepresented.

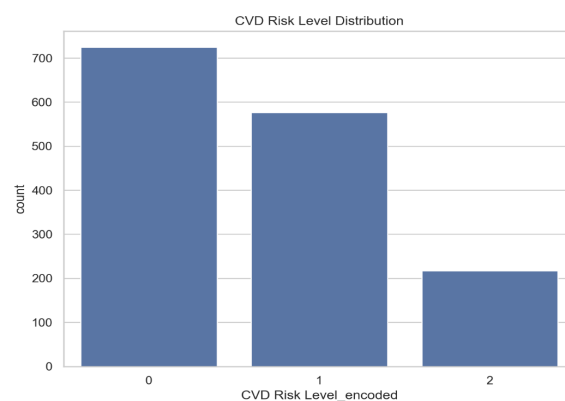


Figure 2. Distribution of Count plot of patients with CVD Risk Levels

Figure 2 is a bar chart indicating the distribution of CVD Risk Levels (encoded) in the dataset. The x-axis will be the three categories of the CVD risk, which will be 0, 1, and 2, and the y-axis will indicate the number of people in each category. As can be seen on the chart, risk level 0 (low or no cardiovascular risk), contains the highest number of instances of 700+, which is why it is the overwhelming class in the dataset. Risk level 1 (Intermediary risk) comes next with almost 550-600 cases, and risk level 2 (high risk) is the least represented with about 200-250 cases.

2.4. Feature Engineering

In Table 3, the F-test determines features that have a high linear relationship with the target variable. The best F-scores were observed in such features as CVD Risk Score, Cholesterol Ratio and Smoking Status which implies strong statistical significance and predictability.

Table.3 F-test (ANOVA)

Rank	Feature	F-Score
1	CVD Risk Score	62.709010
2	Cholesterol Ratio	48.551889
3	Smoking Status	36.314851
4	Estimated LDL (mg/dL)	35.642280
5	Diabetes Status	33.233532
6	Family History of CVD	30.854917
7	BMI	28.364318
8	Age-BMI Interaction	26.000121
9	HDL (mg/dL)	23.809444
10	Total Cholesterol (mg/dL)	22.171215

Selection based on Mutual Information (MI) analyzes the nonlinear relationships among predictors and the target in table 4. Features such as Blood Pressure, BMI and Estimated LDL had the highest MI scores indicating that there are complex associations between them that could not be established by linear tests. By incorporating both F-test and MI based feature selection, this approach creates a balanced strategy which is very significant in improving model robustness, interpretability and predictive accuracy.

Table 4. Mutual Information

Rank	Feature	MI-Score
1	Height (cm)	0.067329
2	BMI	0.063234
3	Height (m)	0.059070
4	CVD Risk Score	0.054815
5	Cholesterol Ratio	0.048691
6	Family History of CVD	0.041738
7	Blood Pressure (mmHg)	0.041323
8	Cardiovascular Index	0.036363
9	Estimated LDL (mg/dL)	0.035412
10	Diabetes Status	0.035188

Table 5. New Feature after preprocessing

Statistic	BP_Ratio	Cholesterol_Ratio	Metabolic_Index	Cardiovascular_Index	Age_BMI_Interaction	BP_Cholesterol_Interaction
Count	1518	1518	1518	1518	1518	1518
Mean	1.556704	3.827790	33.467044	98.050013	1337.300101	24.956434
Std	0.368640	1.591374	11.960525	30.117514	482.523417	8.532099
Min	0.794872	1.136364	10.78	40.5	475.6	9.36
25% (Q1)	1.287918	2.643327	24.157415	75.2775	969.675	18.125
50% (Median)	1.513047	3.517857	31.917125	93.18	1253.874	24.034500
75% (Q3)	1.779412	4.637897	41.499750	113.752	1646.45	30.82825
Max	2.854839	9.8	76.298260	221.99	3065.99	50.75

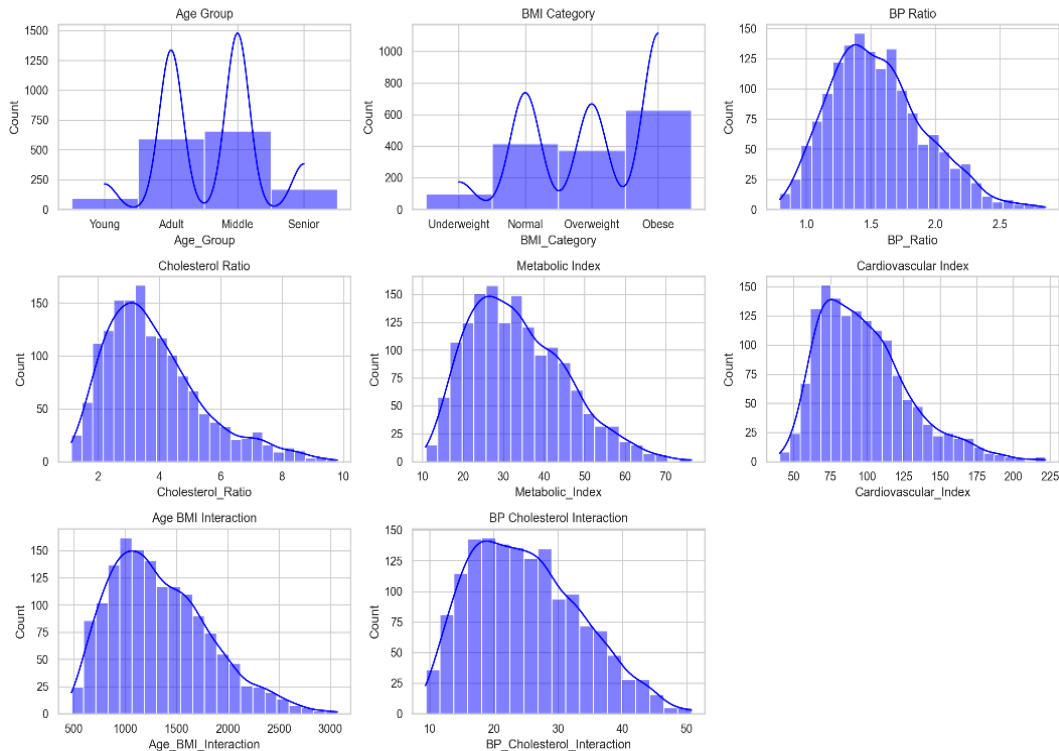


Figure 3. Feature Engineering with new features

In Table 5, the new features are calculated through feature encodings. We distributed the age groups and BMI category into four categories as Young=0 to 30, Adult=30 to 45, Middle=45 to 60, Senior= 60 to 100 and Underweight= 0 to 18.5, Normal= 18.5 to 25, Overweight=25 to 30, Obese=30 to 100. BPRatio was computed as a ratio of Systolic Blood Pressure and Diastolic Blood Pressure, which provides a normalized value of blood pressure dynamics. Cholesterol_Ratio corresponded to the ratio of Total Cholesterol to HDL Cholesterol which assesses the balance between the harmful and protective lipid components. Body Mass Index (BMI) multiplied by Fasting Blood Sugar (mg/dL) was used to estimate the metabolic load through construction of a Metabolic_Index which is calculated as the result of these two factors divided by 100. The dataset contained 1,518 non-missing entries for each feature. Cardiovascular_Index has been calculated as the product of Systolic and Diastolic Blood Pressure Times Age divided by 100 giving an age-adjusted hemodynamic risk score. Also, to control the synergistic effect between hypertension and dyslipidemia, two

interaction terms were added: Age_BMI_Interaction, the product of Age and BMI, to represent the compounded effect of aging and body composition on CVD risk; and BP_Cholesterol_Interaction, the product of Systolic Blood Pressure and Total Cholesterol divided by 1000.

Figure 3 displays a train of histograms that are curve based on a kernel density estimation (KDE) that show the distributions of the innovated features introduced in the dataset when doing the feature engineering process. The continuous variables are BP-Ratio, Cholesterol Ratio, Metabolic Index, Cardiovascular Index, Age-BMI-Interaction, and BP-Cholesterol-Interaction, and each of them has a right-skewed distribution, with long right tails in their KDE curves.

2.5. Baseline Clinical Risk Model Comparison

To put the performance of our machine learning models into perspective, we used the WHO/ISH CVD Risk Prediction Charts of the South-East Asia Region D as a clinical baseline. WHO/ISH model is a 10-year estimate of risk of cardiovascular event, both fatal and non-fatal, because of age, sex, systolic blood pressure, smoking, status diabetes, and total cholesterol. To match our three-class outcome (low, intermediate, high risk), we adjusted the categories of risks in the chart based on the recommended thresholds:

- Low risk: <10%
- Intermediate risk: 10-20%
- High risk: $\geq 20\%$

The WHO/ISH lookup tables were used to calculate risk scores of individual patients in our dataset. The WHO/ISH model was assessed using the identical stratified 5-fold cross-validation system and measures (accuracy, macro F1-score, ROC-AUC) of our ML models.

3. RESULTS AND DISCUSSION

To summarize the key features of the impact of CVD Risks and its consequences we need descriptive analysis.

3.1. Descriptive analysis

Figure 4 shows the Pearson correlation coefficients of numerous clinical, demographic, and lifestyle characteristics with the CVD Risk Level. HDL (mg/dL), Systolic Blood Pressure (SBP), Blood Pressure (mmHg) (encoded), and Diastolic BP are positively correlated with mild correlations. According to these associations, an elevated blood pressure level and change in the level of HDL is correlated with cardiovascular risk. Also, there are Intermediary positive correlations between Fasting Blood Sugar, Abdominal Circumference and Waist-to-Height Ratio, which is consistent with established clinical data that high levels of glucose and central obesity are major risk factors of CVD. The features are negatively correlated. On the other hand, the features to the left of the plot (negative correlation) have negative relationships with the CVD risk. The least positively related characteristics are Family History of CVD (encoded), CVD Risk Score, Diabetes Status (encoded), Estimated LDL (mg/dL) and Smoking Status (encoded).

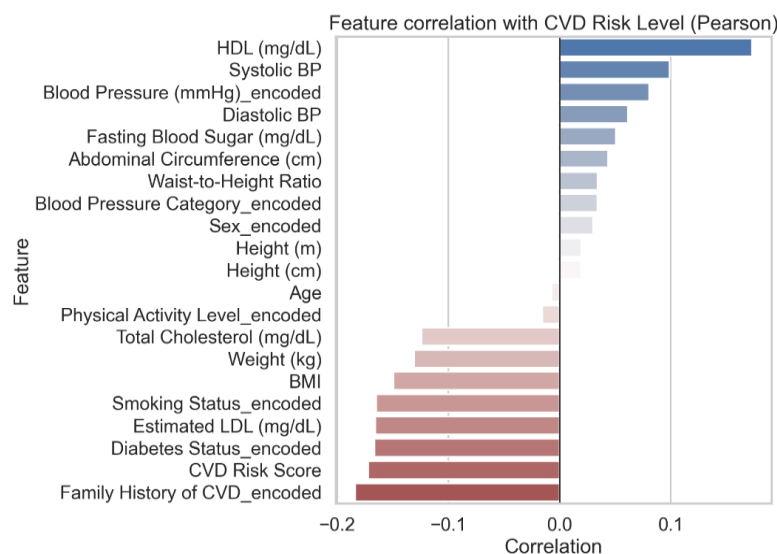


Figure 4. Pearson correlation coefficients

In Figure 5, a scatter diagram is used to demonstrate the correlation between the estimated LDL (Low-Density Lipoprotein) levels, in milligrams per deciliters (mg/dL), and the Risk Level of Cardiovascular Disease (CVD) as well as the fitted regression line. The x-axis is associated with the estimation of LDL levels of 0-200mg/dL and the y-axis correspond with the levels of CVD risks, which are discrete and assume values: 0, 1, and 2. There is a red regression line drawn across the data indicating the overall trend and encircled by a shaded pink band of confidence indicating the 95% confidence interval. The regression line is steeped in a downward direction and therefore there is a negative correlation between the predicted LDL and the predicted CVD risk level. This implies that when the LDL levels are higher, the CVD risk level is lower based on this linear model.

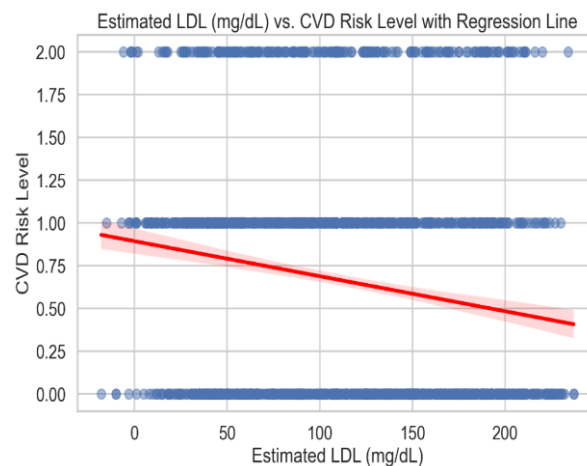


Figure 5. Distribution of Estimated LDL (mg/dL) versus CVD Risk Level

Figure 6 is in the form of a scatter plot that shows the correlation between the Body Mass Index (BMI) and the cardiovascular disease (CVD) Risk Level with a fitted regression line. The horizontal axis portrays the values of BMI which lie within an approximate 15 to 45 and the vertical axis portrays the CVD risk level which is discrete and is a variable that takes values of 0, 1 and 2. The blue circles are an individual observation and indicate the varying risk of CVD at varying levels of BMI. An over-plot red regression line shows the overall trend and a pink band shaded around the regression line represents the 95% confidence interval which shows the uncertainty in the regression estimate. The line is also slightly negative, which means BMI and CVD risk level are weakly related to each other, therefore, the higher the BMI, the lower the predicted risk level of CVD is expected to be.

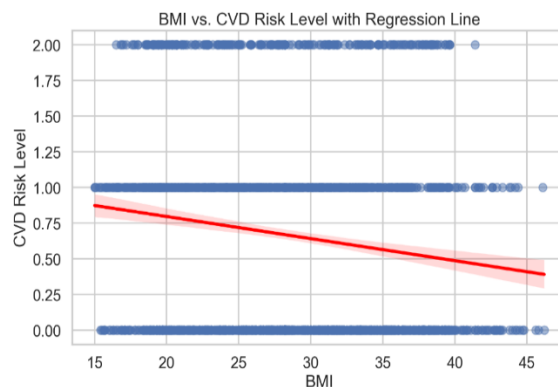


Figure 6. Distribution of BMI vs CVD Risk Level

Figure 7 presents a violin plot that shows the distribution of the cardiovascular disease (CVD) Risk Level on the various Blood Pressure Categories. The Category are distinguished into Normal (0), Elevated (1), Hypertension Stage 1(2), Hypertension Stage (3). The x-axis is the blood pressure worsening, which is given a number between 0 and 3, whereas the y-axis is the CVD risk level, with an approximation of -0.5 to 2.5. The interquartile range (IQR) is denoted by a central black bar in each of the violins, and the median CVD risk level in that category of blood pressure is denoted by a white dot. The density is high around CVD risk levels of 0 at levels of 1 in each category, which shows that the majority of people fall within lower and intermediary levels of risk irrespective of the blood pressure group.

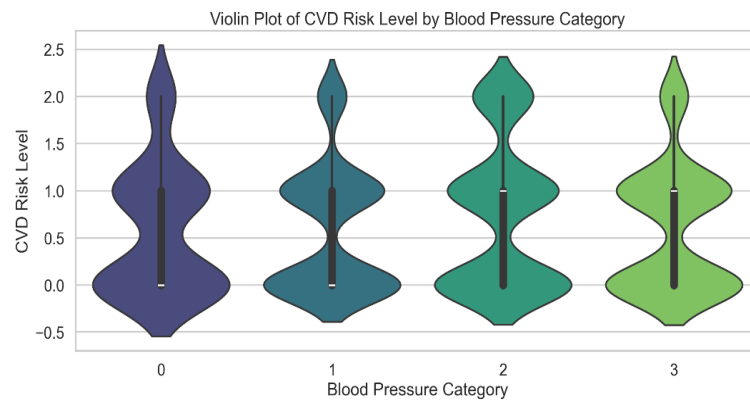


Figure 7. Blood Pressure category vs CVD Risk Level

3.2. Train-Test Model

Figure 8 shows the distribution of classes in training and test dataset after train-test split with stratification. A train-test split of 80 percent-20 percent has been made, that is, 80 percent of the entire dataset was used to train and 20 percent was used to test.

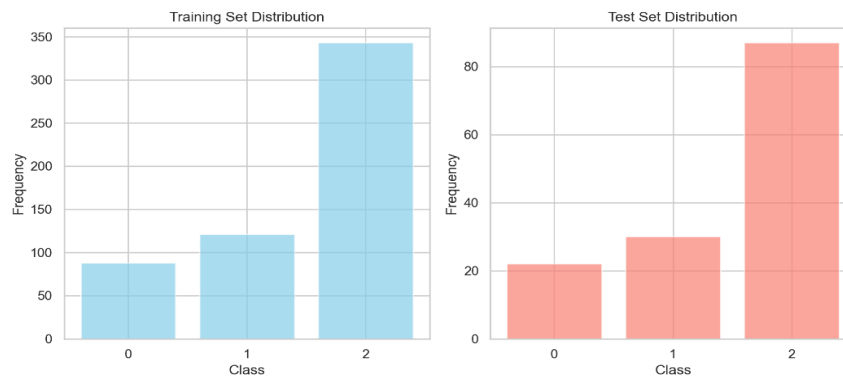


Figure 8. Distribution of Train-Test

3.3. Evaluation with Weighted Feature fusion

This part will illustrate the evaluation process of our models with Weighted feature fusion (Optuna) with classification report, ROC-AUC, Validation methods. The Weighted Feature Fusion methodology used in this study involved combination of three different categories of features namely clinical, lifestyle/behavioral, and engineered clinical features to increase the overall predictive power of cardiovascular disease (CVD) risk classification models. They were the variables of Age, BMI, Abdominal Circumference (cm), Total Cholesterol (mg/dl), HDL, Estimated LDL, Fasting Blood Sugar (mg/dl), Systolic Bp, Diastolic Bp, Waist to height ratio, Height (cm), Height (m), and weight (kg).

Moreover, a group of constructed clinical characteristics was obtained to describe the interactions and ratios between the most important physiological parameters, which gives greater information about metabolic and cardiovascular associations. Such artificial attributes were BP_Ratio, Cholesterol_Ratio, Metabolic_Index, Cardiovascular_Index, Age_BMI_Interaction and BP_Cholesterol_Interaction.

Random Forest Classifier

Table.6 Classification report RF

Class	Precision	Recall	F1-Score	Support
0(Low)	0.95	0.86	0.90	43
1(Intermediary)	0.95	0.78	0.85	49
2(High)	0.88	0.97	0.92	120
Accuracy			0.91	212
Macro Avg	0.93	0.87	0.89	212
Weighted Avg	0.91	0.91	0.90	212

Table 6 illustrates the classification report on Random Forest where we performed on 212 test cases and distributed into three categories where 0= Low risk, 1= Intermediary risk, 2 = High risk. The confusion matrix in figure 9 of the Random Forest classifier shows the success of the model to predict three categories of risk of cardiovascular disease (CVD) 0 (low risk), 1 (Intermediary risk), and 2 (high risk). The real labels are the ones shown on the y-axis and the predicted labels on the x-axis. In the case of Class 0, Random Forest was able to identify 37/43 of the instances and only 6 of them was an incorrect classification of a Class 1 and Class 3 as Class 2. Equally, with Class 1, the model successfully predicted 38 of 49 cases with 11 samples being mispredicted as Class 2. The accuracy of Class 2 (high-risk) group was the highest: 117 correct predictions were made out of 120 cases, and only 3 cases were wrongly classified as Class 1.

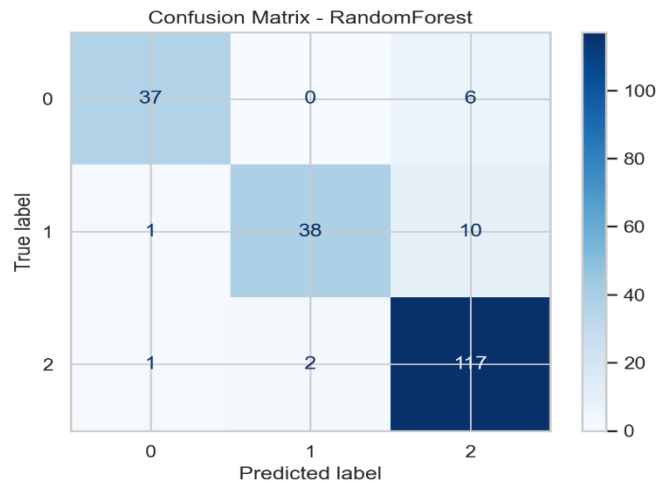


Figure 9. Confusion Matrix Random Forest Classifier

3.4. ROC-AUC

It was observed that the discriminative performance of the Random Forest classifier under the configuration of the WFF was assessed through Receiver Operating Characteristic (ROC) curves of each risk class and their combination measures. The model as shown in Figure 10 was highly separable in all three categories of CVD risks. The values of per-class Area Under the Curve (AUC) are 0.84 (low risk), 0.964 (intermediate risk), and 0.984 (high risk), which means that there is the excellent ability of the model to differentiate between each of the risks. The macro-average AUC was 0.9799 and the micro-average AUC was 0.978, which is far above the random classifier baseline (AUC = 0.5).

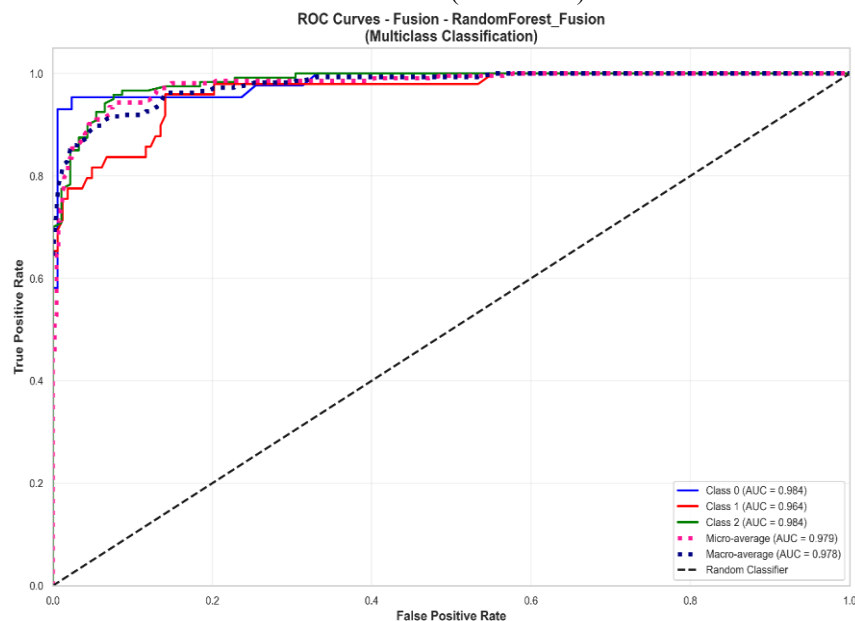


Figure 10. ROC-AUC for models

The findings validate that a combination of the clinical, lifestyle, and engineered properties weighted through the proposed WFF mechanism improves the ability of the classifier to identify subtle risk patterns along the multi-class risk spectrum of CVD.

3.5 Model Performance

Table 7 shows the comparative analysis of machine learning classifiers based on the framework of the Weighted Feature Fusion (WFF) showed that there were significant differences in predictive accuracy and generalization ability. Out of all the five tested models, ensemble methods performed significantly better than their simpler counterparts with the highest test set accuracy (91%) and the best stratified 5-fold cross-validation score (0.923) showing good discrimination as well as good generalization. XGBoost was next with 90% accuracy and a cross-validation score of 0.920, which once again supported the usefulness of XGBoost in identifying complicated interactions among features. Accuracy of the Bagging classifier was 89% with a cross-validation value of 0.890, which indicates a stable ensemble performance and K-Nearest Neighbors (83% accuracy, 0.876 cross-validation value) showed moderate and consistent performance. Conversely, the Decision Tree model was the least accurate (82) and the lowest cross-validation (0.805) with respect to overfitting and generalizability in this context of multimodal clinical features.

Table 7. Model performance report

Model	Accuracy (%)	Stratified K-Fold Score
Random Forest	91	0.922642
Decision Tree	82	0.804849
KNN	83	0.876415
XGBoost	90	0.919811
Bagging	89	0.889623

These results highlight the benefits of ensemble learning methods, especially the Random Forest and XGBoost, in cardiovascular disease risk stratification using a combination of clinical, lifestyle, and engineered features.

3.6 Feature Importance Description

The weights on each feature subset (clinical = 1.30, lifestyle = 1.18) were not predetermined but were selected together with classifier hyperparameters with a Bayesian optimization framework (Optuna). The clinical and Lifestyle subsets, weights, were sampled as continuous hyperparameters, and they were searched on 100 trials in the range [0.5, 3.0], with the goal to maximize the macro averaged F1 score of the validation set. The fact that obtained optimal weights were 1.30 on clinical and 1.18 on lifestyle features implies that engineered composite variables (e.g., Cardiovascular_Index, BP_Cholesterol_Interaction) added about twice the amount as that of lifestyle features, with clinical measurements adding 1.30 times more. This weighting scheme represents the predictive power of each of the modalities in the rural Bangladeshi cohort and played a role in the performance improvements.

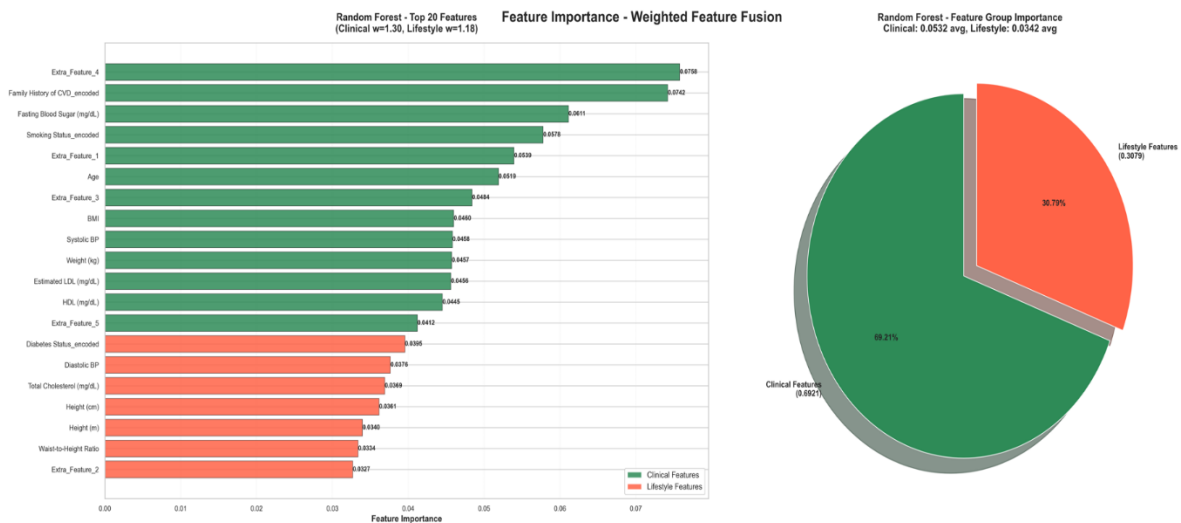


Figure 11. Feature importance top 20 Clinical vs Lifestyle Data (Random Forest)

Figure 11 shows a Random Forest feature importance analysis with the Weighted Feature Fusion (WFF) model, weighted clinical features are 1.30 and weighted lifestyle features are 1.18. The 5 most contributing features are Family history of CVD (encoded), Fasting blood sugar (mg/dl), Smoking status (encoded), and two other engineered features which are named Extra_feature_4 and Extra_feature_1 (Cardiovascular_Index, BP_Cholesterol_Interaction) with an important score of 0.0476 each. As indicated by the bar graph, clinical and engineered features take the lead in decision making of the model with the importance of the group having an average importance of 0.0532 on clinical features versus 0.0342 on lifestyle features. The findings confirm the capability of the WFF mechanism to focus on clinically relevant and interaction-based variables without neglecting a significant, albeit less significant, role of behavioral variables including smoking and family history.

3.7. Explainable AI



Figure 12. SHAP summary plot for the Random Forest model. Features are ranked by importance; red points indicate high feature values, while blue points indicate low values.

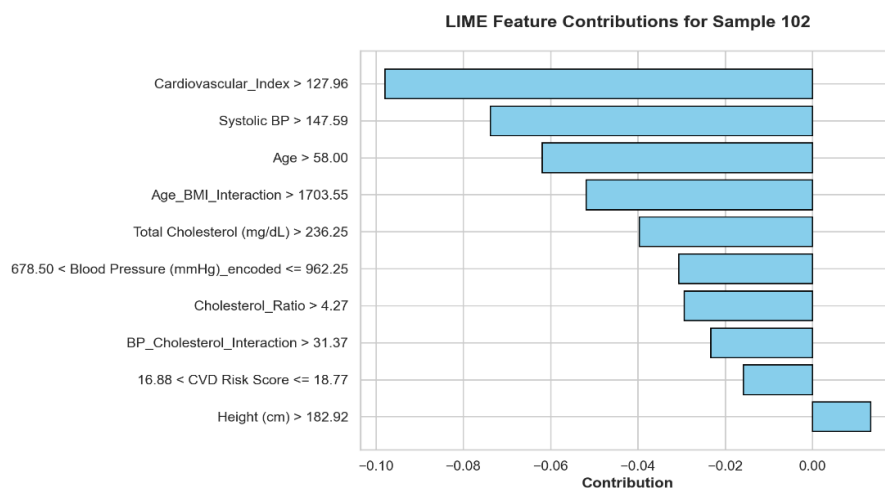


Figure 13. LIME explanations for selected test cases. Bars represent the contribution of features to the prediction for that specific individual.

According to the directions given by the SHAP analysis of the random forest model, the figure 12 depicts the directional effect of the most important features on the prediction of cardiovascular diseases. Positive values of SHAP are those characteristics that enhance the amount of predicted risk, whereas negative

figures denote reduction of the risk. Systolic BP, Blood Pressure (mmHg), and CardiovascularIndex have mixed directional effects (SHAP = -0.45 to +0.60), and there is also evidence of context-dependent effects of these variables on patient subgroups. On the contrary, the contribution of MetabolicIndex, Height (m), and BPRatio to increased risk is always high, and SHAP values range between 0.55 and 0.70.

The lifestyle variables (Smoking Statusencoded, Diabetes Statusencoded) also show positive risk contributions (0.25-0.55), which adds to their applicability in multimodal risks assessment. It is interesting to observe that the traditional biomarkers such as HDL and Estimated LDL have little effect in this model with the SHAP magnitudes of the interaction terms being less than 0.10 meaning that the engineered composite indices and interaction terms could be more representative of predictive signal within this population specific dataset. The results of this SHAP summary confirm the interpretability of the Random Forest model in the WFF framework and emphasize the complicated, usually non-linear, contribution of clinical, behavioral, and derived features to the risk stratification of CVD.

The LIME analysis of the Random Forest model in figure 13, will allow giving a local prediction of a particular prediction, showing how the risk is affected by the thresholds of features. It is worth noting that Systolic BP > 147.59 mmHg and Height > 182.92 cm had the strongest negative contribution to the risk prediction (contributions of -0.07 and -0.08 respectively), indicating that in this case, high systolic pressure and increased height were related to a lower risk classification. In the same vein, Age > 58 years, Age_BMI_Interaction > 1703.55, and Total Cholesterol > 236.25 mg/dL, respectively, had a minor negative effect on the risk predicted (-0.04 to -0.05). Cardiovascular_Index > 127.96 and Cholesterol_Ratio > 4.27 were the least relevant features because they had negative impact of -0.01 which means that these features are not critical in this prediction. These results that can be locally interpreted close to the nuanced, context-dependent clinical and physiological variable contributions and support the importance of LIME in complementing global model explanations with case-specific results in clinical decision-making.

3.8. Model Performance with weighted feature fusion

Weighted Feature Fusion was introduced; the model was effective in its ability to improve its knowledge of risk factors of CVD and inter-feature association by assigning optimized weights to clinical, behavioral and engineered features. The outcomes proved significant improvement of all the classifiers. In table 8, it is a summary of the Stratified K-Fold Cross-Validation (5 splits) of the performance of different machine learning models with the use of Weights features fusion. The highest overall performance occurred with the Random Forest with the mean cross-validation accuracy of 0.922642 and ROC-AUC of 0.980447 which is evidence of good discriminative ability and inter-fold generalization. Close behind it was the XGBoost model that was accurate at 0.919811 and the AUC score and 0.977865 respectively - indicating strong and consistent predictive reliability.

Table 8. Model Cross validation report

Model	CV Acc Mean	CV Acc Std	CV AUC (ovo) Mean	CV AUC (ovo) Std
RandomForest	0.922642	0.015954	0.980447	0.004953
XGB	0.919811	0.009894	0.977865	0.004964
bagging	0.889623	0.013540	0.966706	0.007052
KNN	0.876415	0.010505	0.957639	0.008821
DecisionTree	0.808491	0.034611	0.834911	0.033493

These findings in Table 8, are a clear indication that the process of feature fusion enhanced the capability of the classifier to distinguish various levels of low, moderate and high risks of CVD. The fusion model minimized the bias in classes and it also reflected the multifactorial nature of cardiovascular disease, because the classical clinical predictors (e.g., blood pressure, cholesterol, glucose levels) were combined with the metrics that depend on lifestyle and interaction (e.g., BMI-Age, BP-Cholesterol interactions). Therefore, the Weighted Feature Fusion methodology led to the improvement of the reliability of risk classification by using both physiological measurements and lifestyle behaviors, which is consistent with the medical complexity in practice.

3.9. Contribution and Reflection on CVD Risk Level Classification

The framework enhanced the ability to classify CVD risks and minimize the misclassification between risk categories by incorporating multi-domain features with the help of the Weighted Feature Fusion algorithm. The SHAP and LIME combination offered global and local interpretability, which enabled clinicians to know how the model justifies its risk level of CVDs. The fusion methodology compromised the clinical and

behavioral characteristics and allowed a reduction of the mistakes in Intermediary-risk individuals, which is a crucial category of people to intervene.

- In the case of the models, especially in the Random Forest classifier, the performance was superior (AUC = 0.980447, Accuracy = 91%), which proves the consistency of the method in a practical risk classification setting.
- Essentially, the suggested framework does not only increase predictive accuracy, but it also closes the gap between data-driven and clinical insight modeling. The model captures the interaction between biological, behavioral and engineered risk factors and could therefore be useful in providing personalized and understandable CVD risk assessment to help in the early detection and prevention of the disease.

3.9. Equations

To build analytic models which are highly predictive, different ML classifiers are used including Random Forest (RF), and XGBoost. The main aim of these classifiers is to come up with high effective models of disease detection.

Random Forest

Random Forest is known as ensemble learning technique which works for classification and tasks related to regression that leverages the strength of multiple decision trees. Through this operation it can improve accuracy and reduce overfitting in classification and regression tasks.

$$\hat{y} = \frac{1}{B} \sum_{i=1}^B h_i(x) \quad (1)$$

Where h_i is the prediction of the i^{th} tree and B is the number of trees.

XGBoost

Extreme Gradient Boosting is a streamlined code of gradient boosting, which constructs an additive model in a forward stage-wise manner. It takes several weak learners (decision trees) and makes them a strong ensemble predictor, in addition to using regularization and optimization of training to enhance speed and performance.

$$L = \sum_{i=1}^n l(y_i - y) + \sum_k \Omega(f_k) \quad (2)$$

Where, y_i is the true label of sample i , y is the prediction for sample i , f_k is the k -th decision tree for weak learner, $\Omega(f_k)$ is the tree complexity regularization term

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) performs through storing the entire dataset and producing outcomes through the most frequent class or averaging for regression tasks.

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (3)$$

Where,

x_i = Computes the distance between

y_i = every training instance

Decision Tree

A decision tree used by recursively segmenting the dataset into subset built on the feature attributes and forms a tree-like structure where each node shows one feature condition, each branch represents outcome of that condition, and each leaf node provides the final class or value.

$$Entropy(S) = \sum_{i=1}^c p_i \log_2(p_i) \quad (4)$$

Where:

S = The existing dataset

c = classes number

p_i = class i proportion of samples

Classification Rule

$$\hat{y} = \underset{y}{\operatorname{argmax}} P(Y = y) \prod_{i=1}^n P(X_i | Y = y) \quad (5)$$

Using key performance metrics the train models are assessed:

Accuracy(A):

$$A = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Where, TP, TN, FP, and FN represent true positives, true negatives, false positives, false negatives, respectively.

Precision(P):

$$P = \frac{TP}{TP+FP} \quad (7)$$

That is measure the proportion of correctly predicted users.

Recall(R):

$$R = \frac{TP}{TP+FN} \quad (8)$$

The ability evaluates to detect actual media users.

F1-Score:

$$F1 = 2 \times \frac{P \times R}{P+R} \quad (9)$$

Actually, it provides a balance between precision and recall.

SHAP

SHapley Additive exPlanations has its foundation on Shapley values of cooperative game theory. It measures how much a feature contributes to a prediction by a model by fairly sharing the credit of prediction amongst all features. The SHAP values in a CVD risk model determine the extent to which each aspect raises or lowers the predicted risk of CVD in an individual, as compared to the risk of an average person.

$$\Phi_j(f, x) = \sum_{S \subseteq P \setminus \{j\}} \frac{|S|!(|P|-|S|-1)!}{|P|!} [f_x(S \cup \{j\}) - f_x(S)] \quad (10)$$

Where:

- P = set of all features
- S = subset of features
- $f_x(S)$ = expected prediction given feature subset S
- $|S|$ = cardinality of set S

LIME

Local Interpretable Model-agnostic Explanations describes the model predictions in terms of what the complex model would do locally (around the instance of interest) by a simple interpretable model, like a linear regression. It disturbs the data around a specific point and looks at the variation of predictions, and fits a local surrogate model that behaves similarly to the complex model in that area.

$$\xi(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (11)$$

Where:

- $\xi(x)$ = explanation for instance x
- G = class of interpretable models (e.g., linear models, decision trees)
- f = original black-box model
- g = interpretable surrogate model
- $\pi_x(z)$ = proximity measure between instance z and x
- $\Omega(g)$ = complexity/regularization term for g
- $L(f, g, \pi_x)$ = loss function measuring how well g approximates f

Loss Function:

$$L(f, g, \pi_x) = \sum_{z \in Z} \pi_x(z) (f(z) - g(z'))^2 \quad (12)$$

Where:

- Z = perturbed samples around x
- z' = interpretable representation of z

Weighted Feature Fusion

Let $\mathcal{F}_C$, $\mathcal{F}_L$, and $\mathcal{F}_E$ represent the normalized feature vectors corresponding to clinical, lifestyle, and engineered feature subsets, respectively. Each subset is first scaled using RobustScaler to ensure numerical stability.

The fused feature vector $\mathbf{F}_{\text{fused}}$ is defined as:

$$\mathbf{F}_{\text{fused}} = \mathbf{W}_C \odot \mathcal{F}_C \oplus \mathbf{W}_L \odot \mathcal{F}_L \oplus \mathbf{W}_E \odot \mathcal{F}_E \quad (13)$$

where:

- $\odot$ denotes element-wise multiplication,
- $\oplus$ denotes vector concatenation,
- $\mathbf{W}_C, \mathbf{W}_L, \mathbf{W}_E$ are weight vectors assigned to each feature subset.

The weights are not heuristically assigned; they are optimized jointly with model hyperparameters using Optuna, a Bayesian optimization framework. The objective function maximizes the macro-average F1-score via stratified 5-fold cross-validation.

In ablation experiments, to isolate the contribution of Weighted Feature Fusion, we conducted ablation experiments comparing five configurations: Clinical Only ($\mathcal{F}_C$), Lifestyle Only ($\mathcal{F}_L$), Engineered Only ($\mathcal{F}_E$), Simple Concatenation ($\mathcal{F}_C \oplus \mathcal{F}_L \oplus \mathcal{F}_E$) - unweighted, Weighted Feature Fusion (WFF) - as defined above. All experiments used Random Forest with the same Optuna tuning strategy and SMOTE-ENN resampling applied inside the cross-validation pipeline.

PCA Implementation with Inverse Mapping

The engineered feature matrix $X_E \in \mathbb{R}^{n \times 6}$ was transformed into principal components:

$$Z = X_E W \quad (14)$$

where $W \in \mathbb{R}^{6 \times 3}$ contains the principal component loadings.

The inverse transformation for SHAP value interpretation is defined as:

$$\phi_{X_E} = \phi_Z W^T \quad (15)$$

where ϕ_Z are SHAP values for principal components and ϕ_{X_E} are the corresponding SHAP values mapped to original engineered features.

3.10. Real-time application of CVD Risk prediction system

Fig.14 (a) Real-time application of CVD Risk prediction system

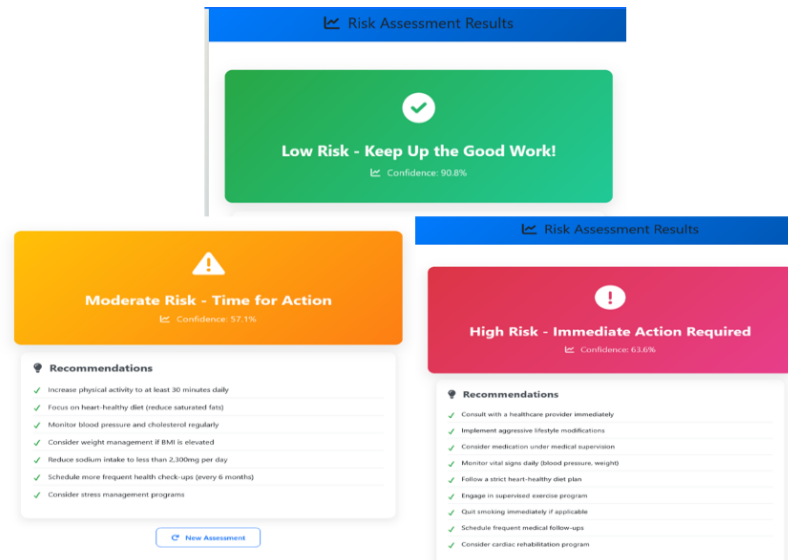


Fig.14(b) Real-time application of CVD Risk prediction results with predefined recommendations

3.11. Limitations and Future Work

Although the predictive performance and interpretability are good, this study has a number of limitations, which ought to be improved in the future.

- The dataset utilized was quite sufficient to validate the experimentally, but it was quite small and could not represent all the diversity of the global populations and regional differences in cardiovascular risk factors.
- Moreover, even though the transparency of the models can be improved with the help of Explainable AI algorithms like SHAP and LIME, they might introduce errors and inconsistency in the interpretation of the features, which makes their use in the medical field rather risky.
- Lastly, the weighted feature fusion and explainability modules pose relatively small computational burdens, which can be difficult to constrain healthcare resources.
- The focus on future working should be on clinical implementation and real-life verification through the collaboration with healthcare organizations to integrate the model into the Electronic Health Records (EHR) systems.
- Further improvement of predicting cardiovascular outcomes long-term could be achieved by expanding the set of data sources, such as genomic data, imaging biomarkers, and longitudinal patient history.
- The combination of the Weighted Feature Fusion framework with deep learning models (e.g., CNN-RF or XGBoost-LSTM hybrids, multimodal fusion) could be also considered in the future studies to be more effective at capturing nonlinear and temporal signals in physiological signals.

4. CONCLUSION

This study included a powerful, interpretable, and complete framework of cardiovascular disease (CVD) Risk Level Classification, which combines the use of Weighted Feature Fusion and Explainable Artificial Intelligence (XAI) methodologies. The proposed framework was effective in the integration of clinical lifestyle, and engineered indicators of health into a single framework, and its predictive accuracy and interpretability improved tremendously, compared to conventional approaches.

The Random Forest Fusion model was the most successful of all the models tested and had 91 percent accuracy and an AUC of 0.980447, indicating considerable classification capacity between low, Intermediary and high-risk groups. The mechanism of Weighted Feature Fusion was able to ensure that the contribution of the various feature domains was balanced so that there was an equal representation of behavioral and physiological parameters in the decision-making process of the model. The inclusion of Explainable AI tools SHAP and LIME enabled the realization of global and local interpretability which ensured clinicians could learn the contribution of each individual feature to CVD risk predictions, including Systolic Blood Pressure, LDL, Age-BMI Interaction, and Smoking Status. The framework is a scientifically based, open, and evidence-based instrument of early detection and risk-based stratification of cardiovascular disease, which can help in preventive medicine and tailored treatment patterns.

Ethical Considerations

To ensure that scientific and moral standards were taken into consideration in this study, ethical integrity of this research was maintained during all the data handling data, model development and interpretation stages.

- As the study implied using patient health information, the ethical practices were tight to ensure that privacy, confidentiality, and data security are preserved.
- Data collected in this research were de-identified and anonymized, and no personally identifiable information (PII) was available or utilized in building or analytical models.
- The proposed framework was additionally developed in fairness and transparency based on the institutional ethics of research and current data protection laws like the General Data Protection Regulation (GDPR) and local health data protection laws.
- Machine learning models may also accidentally reproduce the bias that exists in the data; thus, data preprocessing and model testing were done taking care of the identification and reduction of possible disparities among demographic or socioeconomic subsets.
- The stratified validation has ensured that the sample in the classes is well balanced, eliminating overfitting and biased classifications of levels of risk of CVD.
- Ethical transparency was also facilitated through the usage of Explainable AI (SHAP and LIME), as it allowed clinicians and other stakeholders to gain insight into the process by which the model reached the conclusion and, therefore, facilitated accountability and responsible application in healthcare settings.

REFERENCES

- [1] Ozcan, M., & Peker, S. (2022). A classification and regression tree algorithm for heart disease modeling and prediction. *Healthcare Analytics*, 3, 100130. <https://doi.org/10.1016/j.health.2022.100130>
- [2] Yahaya, L., Oye, N. D., & Garba, E. J. (2020). A comprehensive review on heart disease prediction using data mining and machine learning techniques. *American Journal of Artificial Intelligence*, 4(1), 20. <https://doi.org/10.11648/j.ajai.20200401.12>
- [3] Hossain, S., Hasan, M. K., Faruk, M. O., Aktar, N., Hossain, R., & Hossain, K. (2024). Machine learning approach for predicting cardiovascular disease in Bangladesh: evidence from a cross-sectional study in 2023. *BMC Cardiovascular Disorders*, 24(1). <https://doi.org/10.1186/s12872-024-03883-2>
- [4] Chandrasekhar, N., & Peddakrishna, S. (2023). Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization. *Processes*, 11(4), 1210. <https://doi.org/10.3390/pr11041210>
- [5] Baghdadi, N. A., Abdelaliem, S. M. F., Malki, A., Gad, I., Ewis, A., & Atlam, E. (2023). Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. *Journal of Big Data*, 10(1). <https://doi.org/10.1186/s40537-023-00817-1>
- [6] Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022). Risk prediction of cardiovascular disease using machine learning classifiers. *Open Medicine*, 17(1), 1100-1113. <https://doi.org/10.1515/med-2022-0508>
- [7] Mahmud, I., Kabir, M. M., Mridha, M. F., Alfarhood, S., Safran, M., & Che, D. (2023). Cardiac failure forecasting based on clinical data using a lightweight machine learning metamodel. *Diagnostics*, 13(15), 2540. <https://doi.org/10.3390/diagnostics13152540>
- [8] Gharib, S. A., Charafeddine, J., Dornaika, F., & Haddad, S. (2025). Hybrid Learning Framework for Explainable Cardiovascular Disease Detection. *IEEE Access*, 1. <https://doi.org/10.1109/access.2025.3591241>
- [9] Stonier, A. A., Gorantla, R. K., & Manoj, K. (2023). Cardiac disease risk prediction using machine learning algorithms. *Healthcare Technology Letters*, 11(4), 213-217. <https://doi.org/10.1049/htl2.12053>
- [10] Zhu, X., Wang, Y., Cambria, E., Rida, I., López, J. S., Cui, L., & Wang, R. (2025). RMER-DT: Robust multimodal emotion recognition in conversational contexts based on diffusion and transformers. *Information Fusion*, 123, 103268. <https://doi.org/10.1016/j.inffus.2025.103268>
- [11] Wang, R., Xu, D., Cascone, L., Wang, Y., Chen, H., Zheng, J., & Zhu, X. (2025). RAFT: Robust Adversarial Fusion Transformer for multimodal sentiment analysis. *Array*, 27, 100445. <https://doi.org/10.1016/j.array.2025.100445>
- [12] Wang, R., Guo, C., Shabaz, M., Rida, I., Cambria, E., & Zhu, X. (2025). CIME: Contextual Interaction-Based Multimodal emotion Analysis with enhanced semantic information. *IEEE Transactions on Computational Social Systems*, 1-11. <https://doi.org/10.1109/tcss.2025.3572495>
- [13] Zhu, X., Feng, H., Cambria, E., Huang, Y., Ju, M., Yuan, H., & Wang, R. (2025). EMVAS: end-to-end multimodal emotion visualization analysis system. *Complex & Intelligent Systems*, 11(8). <https://doi.org/10.1007/s40747-025-01931-8>
- [14] Chen, J., Xing, D., Shabaz, M., Zhu, Y., Wang, Y., & Zhu, X. (2025). DNLN: Image super-resolution with Deformable Non-Local attention and Multi-Branch Weighted Feature Fusion. *Image and Vision Computing*, 162, 105721. <https://doi.org/10.1016/j.imavis.2025.105721>
- [15] Zhang, Y., Chen, H., Rida, I., & Zhu, X. (2025). A generative random modality dropout framework for robust multimodal emotion recognition. *IEEE Intelligent Systems*, 40(5), 62-69. <https://doi.org/10.1109/mis.2025.3597120>

- [16] Nirob, M. a. S., Bishshash, P., Siam, A. K. M. F. K., Haque, M. A., & Assaduzzaman, M. (2025). CAIR-CVD-2025: An Extensive Cardiovascular Disease Risk Assessment Dataset from Bangladesh. *Mendeley Data*. <https://doi.org/10.17632/d9scg7j8fp.1>

BIOGRAPHY OF AUTHORS



Tasnimul Intazam Asif is student at Department of Computer Science and Engineering, International University of Business Agriculture and Technology (IUBAT), Dhaka, Bangladesh. His primary research interests lie in the fields of Machine Learning and Artificial Intelligence. He can be contacted at email: intazam2896@gmail.com.



Bishwaprotap Ray is a student at the Department of Computer Science and Engineering, International University of Business Agriculture and Technology (IUBAT), Dhaka, Bangladesh. His research areas include the Internet of Things (IoT), Machine Learning, and Artificial Intelligence. He focuses on applications in education, medical science, and agriculture. He can be contacted via email at baburay214@gmail.com.



Dr. Md. Alomgir Hossain is working as an Associate Professor in the department of Computer Science & Engineering at IUBAT-International University of Business, Agriculture & Technology. Dr. Hossain completed Master of Science in Computer Science from Jahangiragar University. Also, he has completed PhD in CSE at DUET (Dhaka University of Engineering and Technology, Gazipur, Dhaka) Under Dept. of Computer Science and Engineering. He has attended in many workshops and seminar in Bangladesh and Global. Dr. Hossain has published several numbers of Journal and conference papers at reputed SCOPUS index database. His main areas of research interest are education and research. Major interest areas are Data mining and warehousing, Machine Learning, Deep learning, AI, IOT and Digital image processing. He can be contacted at email: alomgir.hossain@iubat.edu



Faisal Imran is a Associate Professor & Head in the Department of Computer Science and Engineering (CSE) & CSIT, Shanto-Mariam University of Creative Technology. He has been working with the same department since December 2015 in numerous public as well as private universities, and is engaged in teaching and researches. He received his B.S.c and M.S.c. in Computer Science and Statistics from Rajshahi University of Engineering & Technology, Bangladesh in February 2014 and June 2020, respectively. Now he is pursuing his PhD in CSE from Jagannath University, Bangladesh. His research interests span the areas of Deep Learning with emphasis on Image Processing and Artificial Intelligence.

He can be contacted at email: faisalruetcs09@gmail.com



Prime Biswajit Barua holds a Bachelor's degree in Computer Science and Engineering from the College of Engineering and Technology (CEAT) at the International University of Business Agriculture and Technology (IUBAT) in Dhaka, Bangladesh. His research interests focus on Machine Learning and Artificial Intelligence. He can be reached via email at primebiswajit2210@gmail.com



Nishat Salsabil Anisha is currently pursuing a bachelor's degree in computer science and engineering at the International University of Business Agriculture and Technology (IUBAT), Dhaka, Bangladesh. Her research interests include Internet of Things (IoT), embedded systems, software engineering, natural language processing, and deep learning. She has worked on undergraduate projects involving Arduino-based systems and real-time sensor integration for smart safety and rider monitoring applications. She can be contacted at email: nishatsalsabilanisha@gmail.com.



Ariful Haque Minhaj is currently pursuing a bachelor's degree in computer science and engineering at the International University of Business Agriculture and Technology (IUBAT), Dhaka, Bangladesh. His research interests include Internet of Things (IoT), embedded systems, software engineering, and smart safety applications. He has worked on undergraduate projects involving Arduino-based systems and real-time sensor integration for rider safety monitoring. He can be contacted at email: arifulhaqueminhaj@gmail.com.



Amit Roy has completed bachelor's degree in computer science and engineering, College of Engineering and Technology (CEAT), International University of Business Agriculture and Technology (IUBAT), Dhaka, Bangladesh. His research area is Machine Learning artificial intelligence. He can be contacted at email: amitroy97bd@gmail.com