Biomolecules and Biomedicine
ISSN: 2831-0896 (Print) | ISSN: 2831-090X (Online)
Journal Impact Factor® (2025): 2.8
CiteScore® (2025): 4.7
www.biomolbiomed.com | blog.bjbms.org

SUPPLEMENTAL DATA

Development and temporal validation of an interpretable XGBoost model for predicting cardiovascular death in patients with diabetic foot

Xiaoling Wan[#], Ting Shi[#], Qiao Liu[#], Ying She[*], Jing Yu[*]

Department of Emergency, Xianning Central Hospital, The First Affiliated Hospital of Hubei University of Science and Technology, Xianning, Hubei, China

[*]Correspondence to Ying She: 549395825@qq.com and Jing Yu: medresearch189@163.com

[#]Equally contributed to this work: Xiaoling Wan, Ting Shi, and Qiao Liu.

Full article is available at the following link: Development and temporal validation of an interpretable XGBoost model for predicting cardiovascular death in patients with diabetic foot

Supplemental methods

This section provides detailed technical specifications of the modeling workflow to ensure full analytical reproducibility and transparency.

1. Data Preprocessing

All preprocessing steps were fitted exclusively on the training set and then applied to the internal validation set and held-out test set to eliminate data leakage.

1.1 Missing data handling

Patients with missing core predictive variables were excluded at enrollment according to the pre-specified exclusion criteria. For all 2,325 patients in the final development cohort, all included variables had complete records with no variable-level missing data; thus no missing value imputation procedure was applied.

1.2 Categorical variable encoding

Unordered categorical variables (e.g., ulcer location) were processed with one-hot encoding.

Ordinal categorical variables (e.g., foot deformity severity graded as absent/mild/moderate/severe) were encoded as ordered numerical values to preserve the rank information of the original grading.

Binary variables (e.g., sex, smoking status, medication use) were encoded as 0/1 binary values.

1.3 Feature scaling

For distance-based models (SVM, KNN, MLP), features were standardized via Z-score transformation (mean = 0, standard deviation = 1), with the mean and standard deviation calculated solely from the training set.

For tree-based models (XGBoost, LightGBM, AdaBoost, decision tree), logistic regression and Gaussian naive Bayes, no feature scaling was performed, as these algorithms are inherently insensitive to feature magnitude.

2. Random Seed Specification

A fixed random seed (random_state = 42) was uniformly applied to all stochastic procedures across the entire analysis pipeline to ensure full reproducibility, including:

Stratified random splitting of the development cohort into training, internal validation and held-out test sets;

Initialization and training of all machine learning models;

10-fold stratified cross-validation for hyperparameter tuning;

SMOTE oversampling procedure (when applied for robustness analysis).

3. Hyperparameter Tuning

All 9 machine learning models underwent standardized hyperparameter optimization using grid search with 10-fold stratified cross-validation within the training set, with AUC as the unified optimization target. This consistent tuning framework guaranteed fair head-to-head comparison across all algorithms. The core hyperparameter search spaces for each model are listed below:

Logistic regression: regularization coefficient C ∈ [0.01, 0.1, 1, 10, 100]

XGBoost: learning rate ∈ [0.01, 0.05, 0.1]; max_depth ∈ [3, 5, 7]; n_estimators ∈ [100, 200, 300]; subsample ∈ [0.7, 0.8, 0.9]; colsample_bytree ∈ [0.7, 0.8, 0.9]; reg_lambda ∈ [0.1, 1, 10]

LightGBM: learning rate ∈ [0.01, 0.05, 0.1]; max_depth ∈ [3, 5, 7]; n_estimators ∈ [100, 200, 300]

AdaBoost: n_estimators ∈ [50, 100, 200]; learning rate ∈ [0.01, 0.1, 1.0]

Decision Tree: max_depth ∈ [3, 5, 7, 10]; min_samples_split ∈ [2, 5, 10]

Gaussian Naive Bayes: var_smoothing ∈ [1e-9, 1e-8, 1e-7]

MLP: hidden_layer_sizes ∈ [(64,), (128,), (64, 32)]; L2 penalty alpha ∈ [0.0001, 0.001, 0.01]

SVM: regularization coefficient C ∈ [0.1, 1, 10, 100]; kernel ∈ ['rbf', 'linear']

KNN: n_neighbors ∈ [3, 5, 7, 9, 15]; weights ∈ ['uniform', 'distance']

The parameter set with the highest average AUC across 10 cross-validation folds was selected as the final optimal configuration for each model.

4. Classification Threshold Selection

The optimal probability cut-off for each model was determined by maximizing the Youden's J statistic, calculated as: J = Sensitivity + Specificity − 1. Critically, this threshold selection was performed exclusively on the internal validation set, not on the training set. The workflow was: (1) hyperparameters were optimized via 10-fold cross-validation on the training set; (2) with hyperparameters fixed, the model was trained on the full training set; (3) the trained model was applied to the internal validation set; (4) the threshold maximizing Youden's index on the internal validation set was selected as the optimal cut-off. The held-out test set was not used for threshold selection, and the temporal validation cohort was not involved in any

threshold determination. All cut-off values presented in the supplementary tables were derived from this method applied to the internal validation set.

5. SMOTE Implementation and Final Model Specification

SMOTE was implemented with k_neighbors = 5 and applied exclusively to the training set after data splitting, with no modification to the internal validation set, held-out test set, or temporal validation cohort. SMOTE was used only as an exploratory robustness analysis (results in Supplementary Table 5). The final optimal XGBoost model, as well as all 9 models in the comparative analysis, were trained on the original unmodified training set without SMOTE oversampling, for three reasons: (1) SMOTE did not significantly improve AUC or core classification performance; (2) training on original data avoids potential bias from synthetic samples; and (3) the model remains consistent with real-world clinical data distribution. All evaluations on held-out datasets were performed using models trained exclusively on the training set, with no retraining after any evaluation.

6. Fairness of Cross-Algorithm Comparison

We ensured unbiased comparison across all models from the following aspects:

Identical data split: All models used exactly the same training, internal validation and held-out test sets.

Identical tuning scheme: All models adopted 10-fold stratified cross-validation grid search, with AUC as the unified optimization target.

Identical evaluation criteria: All models were evaluated on the same held-out test set, with thresholds selected by the uniform Youden’s index method.

Model-appropriate preprocessing: Standardization was applied only to distance-based algorithms to avoid performance underestimation from inappropriate preprocessing, while tree-based models retained their native input format.

Table S1. DeLong test results for AUC comparisons across nine machine learning models

Name	logistic	XGBoost	LightGBM	AdaBoost	Decision Tree	GNB	MLP	SVM	KNN
logistic	NA	<0.001	0.489	0.059	0.175	0.082	0.136	0.219	0.303
XGBoost	<0.001	NA	0.001	0.001	0.005	0.001	<0.001	<0.001	<0.001
LightGBM	0.489	0.001	NA	0.268	0.210	0.402	0.406	0.253	0.362
AdaBoost	0.059	0.001	0.268	NA	0.694	0.464	0.036	0.011	0.025
Decision Tree	0.175	0.005	0.210	0.694	NA	0.566	0.032	0.010	0.043
GNB	0.082	0.001	0.402	0.464	0.566	NA	0.087	0.057	0.045
MLP	0.136	<0.001	0.406	0.036	0.032	0.087	NA	0.380	0.516
SVM	0.219	<0.001	0.253	0.011	0.01	0.057	0.380	NA	0.273
KNN	0.303	<0.001	0.362	0.025	0.043	0.045	0.516	0.273	NA

Abbreviations**:** AdaBoost, adaptive boosting; AUC, area under the receiver operating characteristic curve; GNB, Gaussian naive Bayes; KNN, k-nearest neighbors; LightGBM, light gradient boosting machine; MLP, multilayer perceptron; NA, not applicable; SVM, support vector machine; XGBoost, extreme gradient boosting.

Table S2. Pairwise DeLong comparison of AUC between XGBoost and comparator models

Comparator Model	AUC Difference (XGBoost – Comparator)	*95% CI*	*p value*
Logistic	0.129	0.064–0.194	<0.001
LightGBM	0.123	0.058–0.188	0.001
AdaBoost	0.096	0.046–0.146	0.001
Decision Tree	0.066	0.020–0.112	0.005
GNB	0.122	0.057–0.187	0.001
MLP	0.172	0.093–0.251	<0.001
SVM	0.207	0.117–0.297	<0.001
KNN	0.190	0.104–0.276	<0.001

Abbreviations**:** AdaBoost, adaptive boosting; AUC, area under the receiver operating characteristic curve; CI, confidence interval; GNB, Gaussian naive Bayes; KNN, k-nearest neighbors; LightGBM, light gradient boosting machine; MLP, multilayer perceptron; SVM, support vector machine; XGBoost, extreme gradient boosting.

Table S3. Performance metrics of XGBoost models with varied feature domain combinations in the internal validation set

Classifications	AUC	Cut-off	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 score
DBD+LBP	0.815	0.399	0.761	0.694	0.796	0.643	0.834	0.665
DBD+CC	0.705	0.361	0.676	0.587	0.722	0.522	0.77	0.551

Classifications	AUC	Cut-off	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 score
							3	
DBD+DFT	0.706	0.367	0.671	0.549	0.735	0.522	0.761	0.530
DBD+LBP+CC	0.803	0.403	0.759	0.677	0.801	0.639	0.829	0.655
DBD+LBP+DFT	0.829	0.383	0.766	0.660	0.821	0.666	0.824	0.660
DBD+CC+DFT	0.712	0.347	0.655	0.613	0.677	0.499	0.771	0.548
DBD+LBP+CC+DFT	0.819	0.379	0.751	0.689	0.783	0.626	0.830	0.653

Abbreviations**:** AUC, area under the receiver operating characteristic curve; CC, complications and comorbidities; DBD, demographic, baseline, and diabetes-related traits; DFT, diabetic foot-specific traits; LBP, laboratory biochemical parameters; NPV, negative predictive value; PPV, positive predictive value; XGBoost, extreme gradient boosting.

Table S4. Performance metrics of XGBoost models with various feature domain combinations on the held-out test set

Classifications	AUC	Cut-off	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 score
DBD+LBP	0.836	0.373	0.763	0.747	0.773	0.655	0.841	0.698
DBD+CC	0.725	0.325	0.665	0.741	0.620	0.529	0.80	0.618

Classifications	AUC	Cut-off	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 score
							6	
DBD+DFT	0.749	0.328	0.656	0.741	0.607	0.521	0.803	0.612
DBD+LBP+CC	0.817	0.383	0.729	0.747	0.719	0.605	0.831	0.668
DBD+LBP+DFT	0.844	0.404	0.781	0.782	0.780	0.672	0.861	0.723
DBD+CC+DFT	0.744	0.410	0.686	0.647	0.708	0.561	0.777	0.601
DBD+LBP+CC+DFT	0.817	0.386	0.757	0.706	0.786	0.656	0.823	0.680

Abbreviations**:** AUC, area under the receiver operating characteristic curve; CC, complications and comorbidities; DBD, demographic, baseline, and diabetes-related traits; DFT, diabetic foot-specific traits; LBP, laboratory biochemical parameters; NPV, negative predictive value; PPV, positive predictive value; XGBoost, extreme gradient boosting.

Table S5. Predictive performance metrics of XGBoost models for cardiovascular mortality in the independent temporal validation set

Classifications	AUC	Cut-off	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 score
XGBoost	0.828	0.368	0.763	0.736	0.783	0.720	0.796	0.728

Abbreviations**:** AUC, area under the receiver operating characteristic curve; NPV, negative predictive value; PPV, positive predictive value; XGBoost, extreme gradient boosting.

Table S6. Confusion matrix of the XGBoost model in the independent temporal validation set

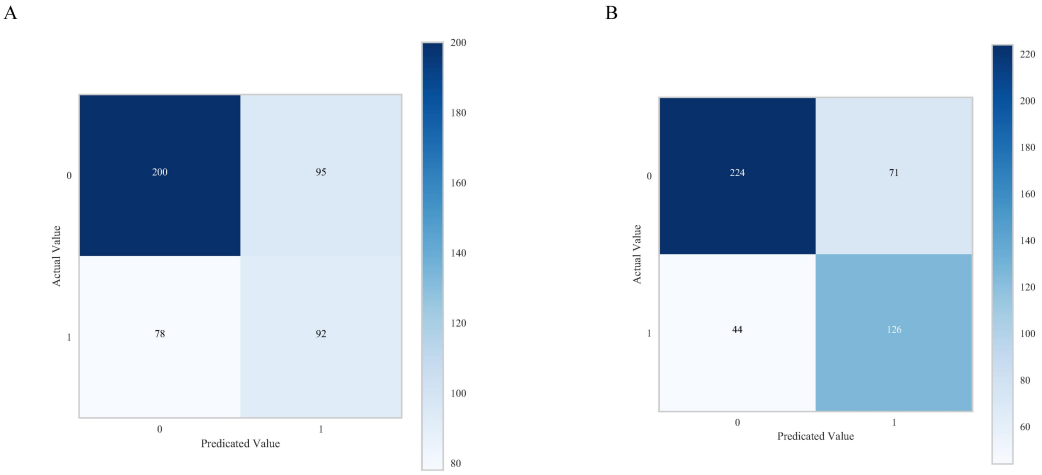
Model	Predicted CV Death	Predicted Non-CV Death
Actual CV Death	162 (TP)	58 (FN)
Actual Non-CV Death	63 (FP)	227 (TN)

Abbreviations**:** CV, cardiovascular; FN, false negative; FP, false positive; TN, true negative; TP, true positive; XGBoost, extreme gradient boosting.

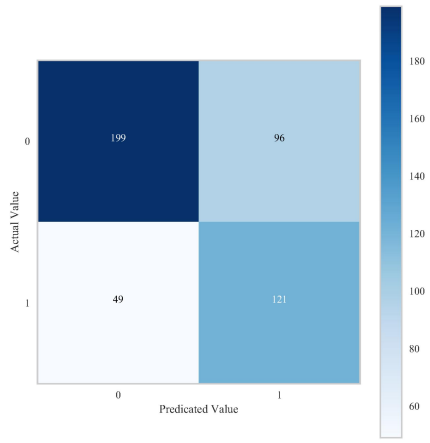
Table S7. Comparative analysis of XGBoost model performance with and without SMOTE oversampling

Model	AUC	Cut-off	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 score
No SMOTE	0.844	0.404	0.781	0.782	0.780	0.672	0.861	0.723
With SMOTE	0.821	0.593	0.768	0.815	0.742	0.652	0.872	0.724

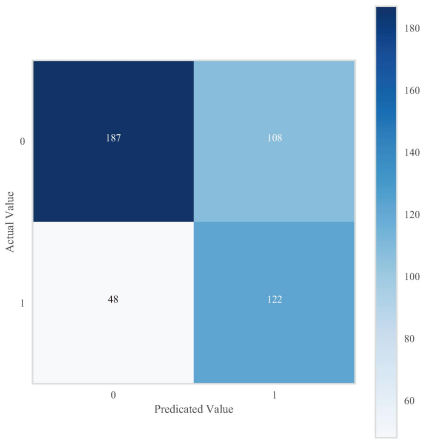
Abbreviations**:** AUC, area under the receiver operating characteristic curve; NPV, negative predictive value; PPV, positive predictive value; SMOTE, synthetic minority oversampling technique; XGBoost, extreme gradient boosting.



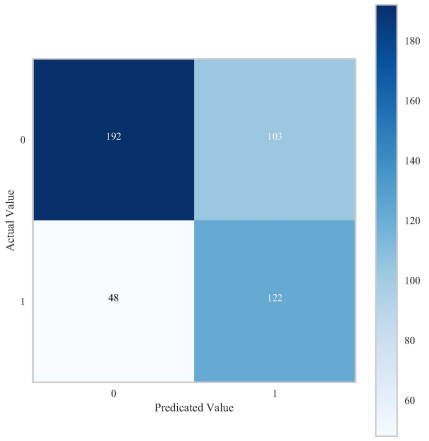
C


199
96
49
121
0
1
Actual Value
Predicated Value
180
160
140
120
100
80
60

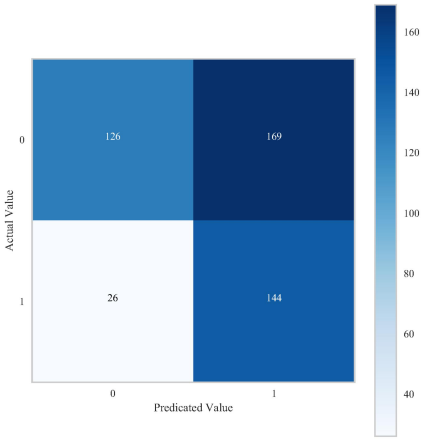
D


187
108
48
122
0
1
Actual Value
Predicated Value
180
160
140
120
100
80
60

E


192
103
48
122
0
1
Actual Value
Predicated Value
180
160
140
120
100
80
60

F


126
169
26
144
0
1
Actual Value
Predicated Value
160
140
120
100
80
60
40

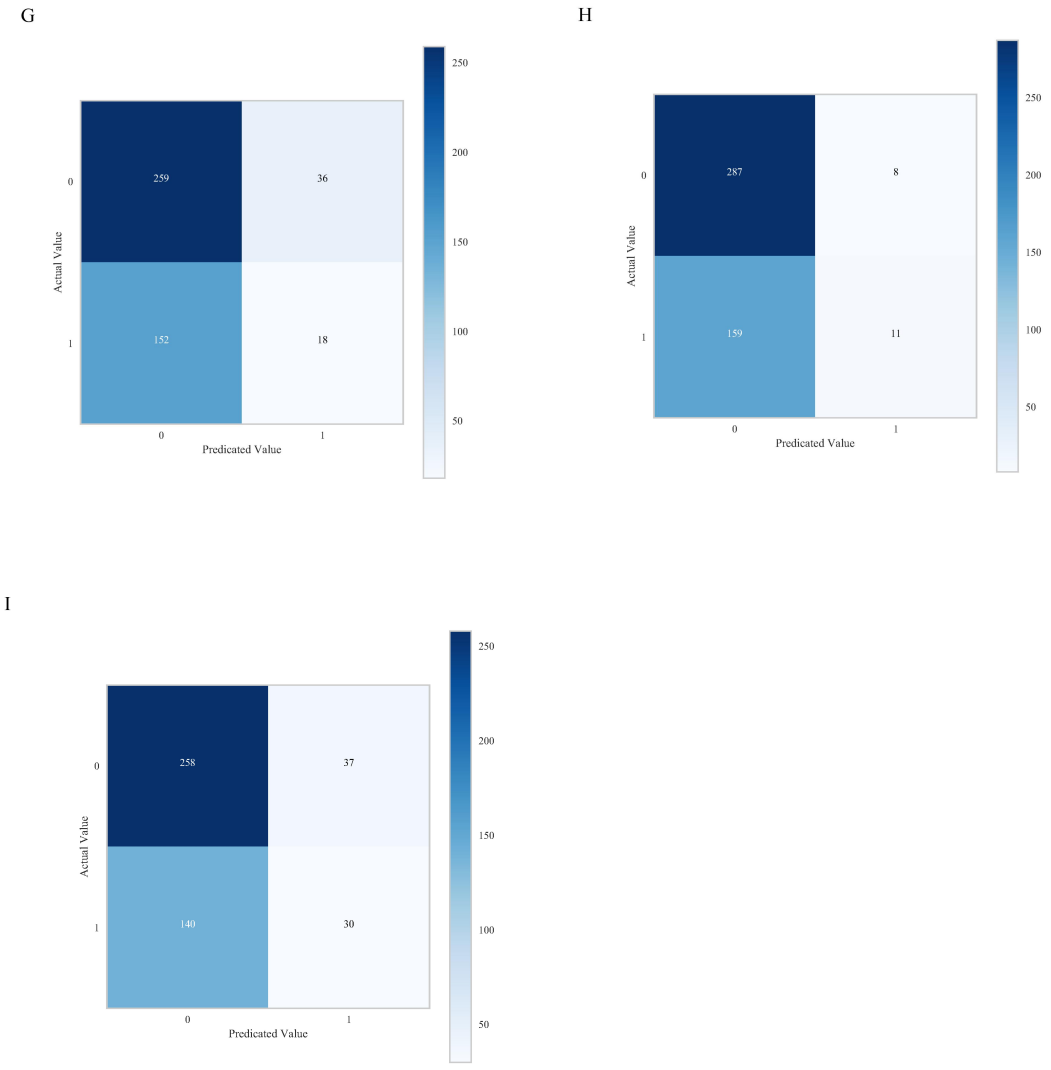
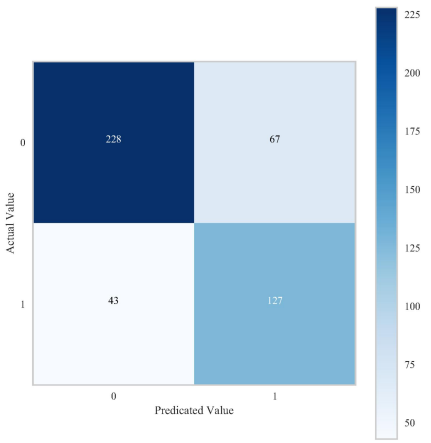
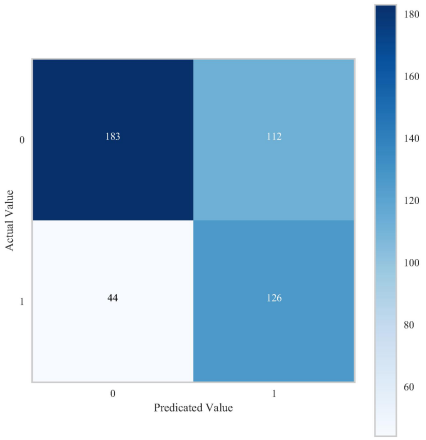


Figure S1. Confusion matrices of nine machine learning models in the held-out test set. (A) Logistic regression; **(B)** extreme gradient boosting; **(C)** light gradient boosting machine; **(D)** adaptive boosting; **(E)** decision tree; **(F)** Gaussian naive Bayes; **(G)** multilayer perceptron; **(H)** support vector machine; **(I)** k-nearest neighbors. Each panel shows the numbers of correctly and incorrectly classified cases for cardiovascular death and non-cardiovascular death in the held-out test set. **Abbreviations:** AdaBoost, adaptive boosting; GNB, Gaussian naive Bayes; KNN, k-nearest neighbors; LightGBM, light gradient boosting machine; MLP, multilayer perceptron; SVM, support vector machine; XGBoost, extreme gradient boosting.

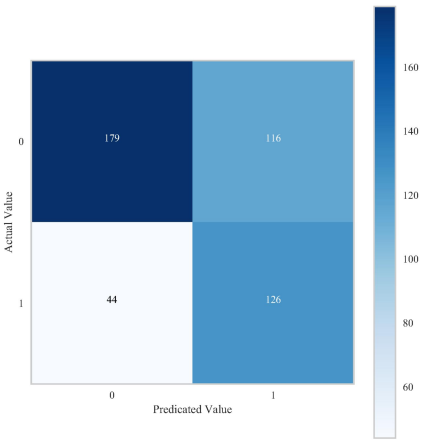
A


228
67
43
127
Actual Value
Predicated Value
0
1
225
200
175
150
125
100
75
50

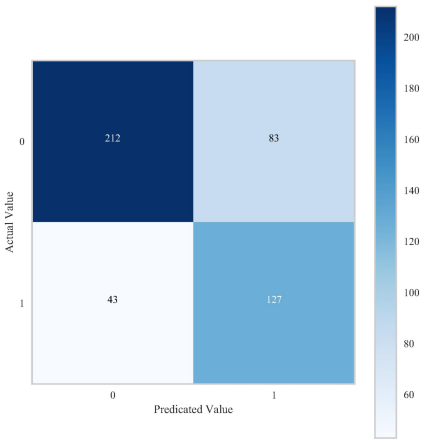
B


183
112
44
126
Actual Value
Predicated Value
0
1
180
160
140
120
100
80
60

C


179
116
44
126
Actual Value
Predicated Value
0
1
160
140
120
100
80
60

D


212
83
43
127
Actual Value
Predicated Value
0
1
200
180
160
140
120
100
80
60

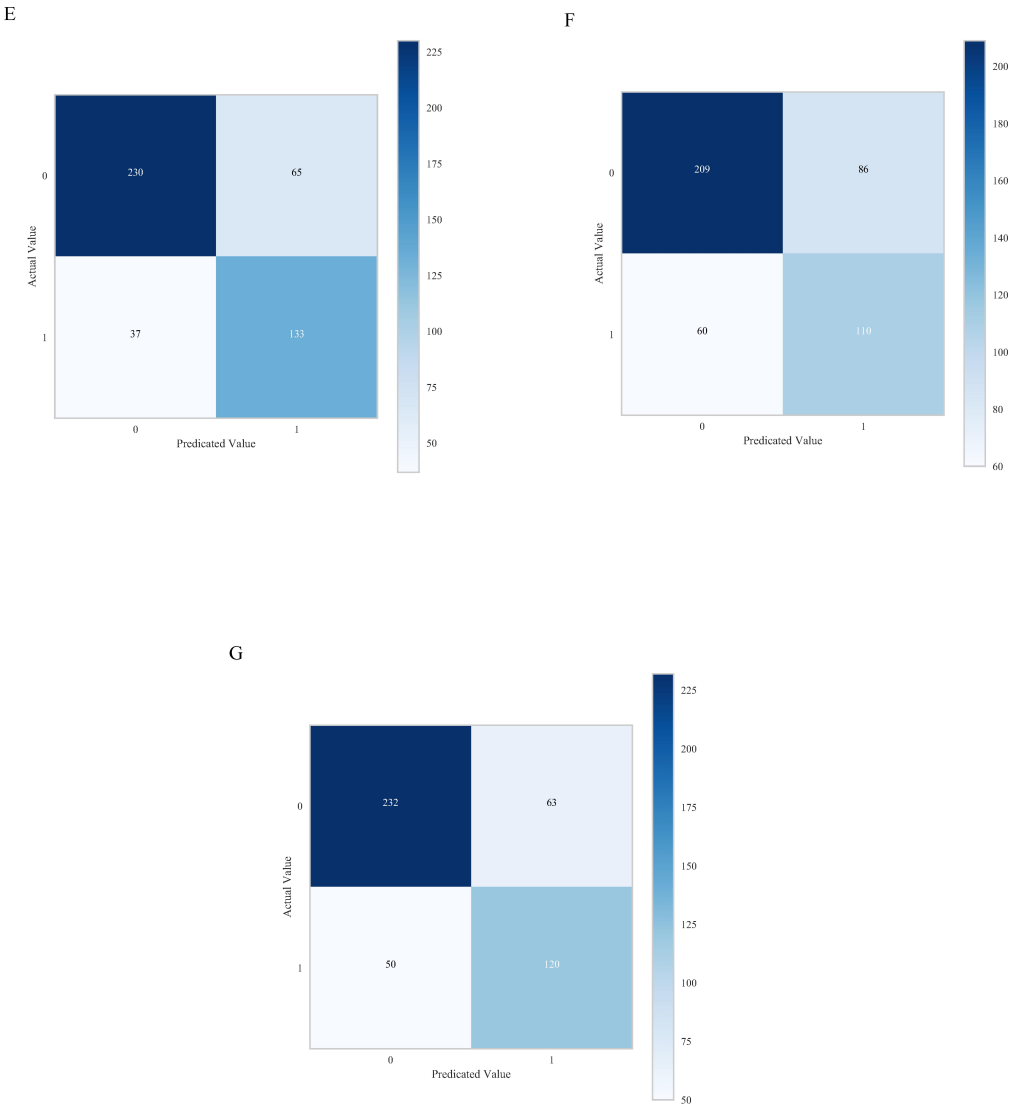


Figure S2. Confusion matrices of XGBoost models across seven feature-domain combinations in the internal validation set. (A) DBD+LBP; **(B)** DBD+CC; **(C)** DBD+DFT; **(D)** DBD+LBP+CC; **(E)** DBD+LBP+DFT; **(F)** DBD+CC+DFT; **(G)** DBD+LBP+CC+DFT. Each panel shows the numbers of correctly and incorrectly classified cardiovascular death and non-cardiovascular death cases for the corresponding XGBoost model. **Abbreviations:** CC, complications and comorbidities; DBD, demographic, baseline, and diabetes-related traits; DFT, diabetic foot-specific traits; LBP, laboratory biochemical parameters; XGBoost, extreme gradient boosting.

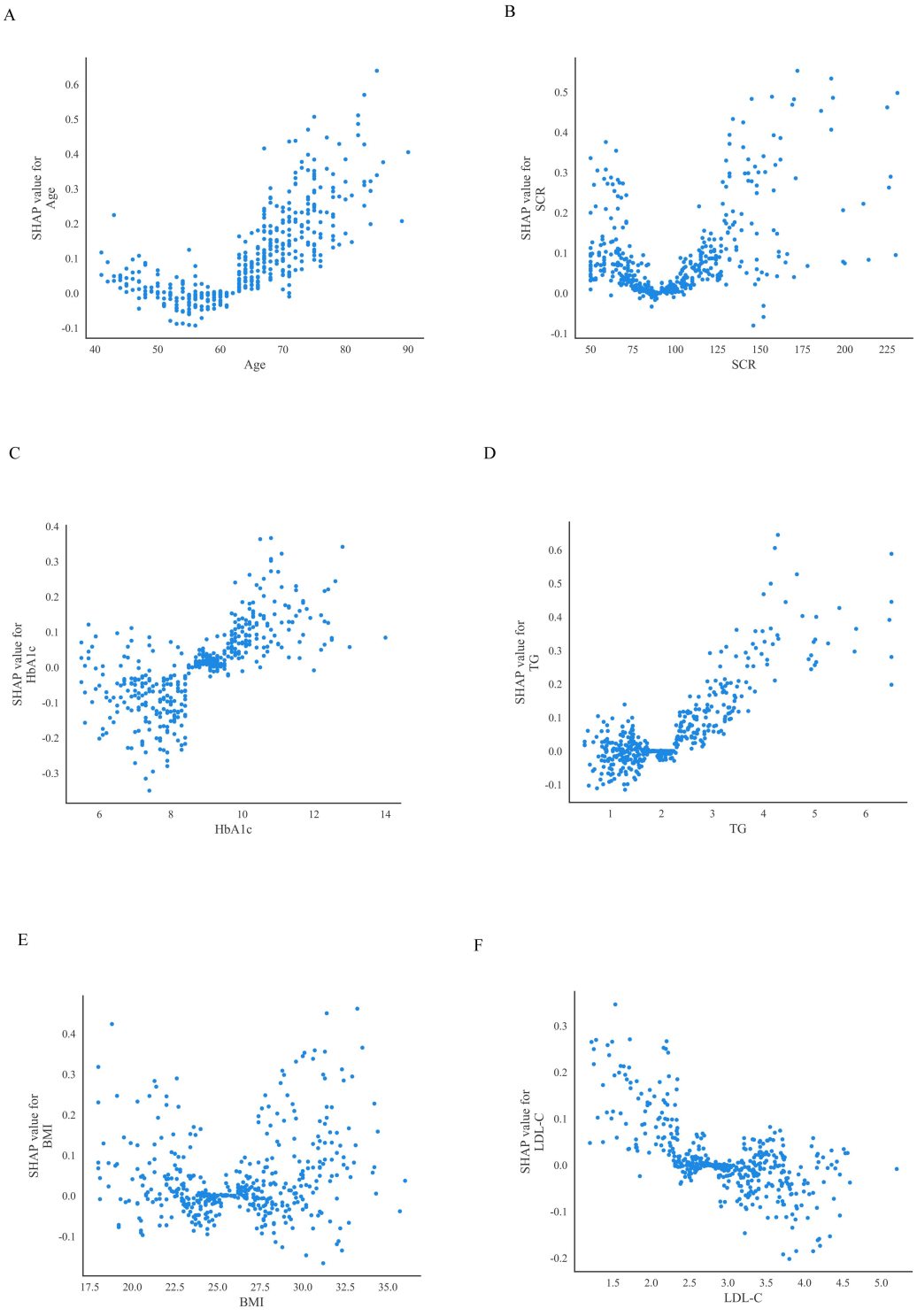


Figure S3. SHAP dependence plots for the six most influential continuous predictors in the optimal XGBoost model. (A) Age; **(B)** serum creatinine; **(C)** glycated hemoglobin; **(D)** triglycerides; **(E)** body mass index; **(F)** low-density lipoprotein cholesterol. Each plot illustrates the relationship between the feature value

and its SHAP value, showing the direction and magnitude of its contribution to predicted cardiovascular death risk. **Abbreviations**: BMI, body mass index; HbA1c, glycated hemoglobin; LDL-C, low-density lipoprotein cholesterol; Scr, serum creatinine; SHAP, SHapley Additive exPlanations; TG, triglycerides; XGBoost, extreme gradient boosting.