Multicenter Evaluation of Interpretable AI for Coronary Artery Disease Diagnosis from PET Biomarkers

Supplemental Materials

Supplementary Table 1. Missingness in the external testing set (n=1,278).

Variables	Missing %	Missing Count
Ischemic TPD ⊗	0.78	10
Stress TPD ⊗	0.70	9
MFR ▽	0.78	10
Stress MBF ▽	0.55	7
Coronary Calcium ▽	0.08	1
TID	0.16	2
Rest LVEF	0.39	5
Stress LVEF	0.78	10
RPP	0	0
Male	0	0

No missing data was observed in the training set, while the extent of missingness in the external testing set was <1%. For univariate comparisons in the external testing set, missing values of ⊗ were derived from corresponding summed scores with "$(\frac{scores}{68}) * 100\%$" according to 17-segment AHA convention; missing values of ∇ were imputed using medians.

Abbreviations: LVEF – Left Ventricle Ejection Fraction; MBF – Myocardial Blood Flow; MFR – Myocardial Flow Reserve; RPP – Rate Pressure Product; TID – Transient Ischemic Dilation; TPD – Total Perfusion Deficit.

Supplementary Table 2. Performance comparison of the models for detecting CAD on external validation set.

Model	AUC	95% CI	P-value
AI Model	0.83	0.81-0.85	Reference
Clinical Score	0.80	0.77-0.82	0.02
Ischemic TPD	0.79	0.76-0.81	<0.001
MFR	0.75	0.72-0.78	<0.001
Stress MBF	0.75	0.72-0.77	<0.001
Coronary Calcium	0.69	0.66-0.72	<0.001

Area under the receiver-operating characteristic curve (AUC) with 95% confidence interval (CI) for the AI Model, Clinical Score, Ischemic TPD, MFR, Stress MBF, and Coronary Calcium in all patients. Clinical score was defined as the summed difference score (ischemia score) from the 17-segment AHA model, reflecting standardized myocardial perfusion abnormalities.

Abbreviations: AI – Artificial Intelligence; MBF – Myocardial Blood Flow; MFR – Myocardial Flow Reserve; TPD – Total Perfusion Deficit.

Supplementary Table 3. Thresholds of AI models matched with clinical metrics on specificity.

Matching with clinical metric on specificity	Threshold of AI score	Matched Specificity	AI Sensitivity
ischemic TPD ≥ 5%	0.45	0.60	0.84 [0.81, 0.87]
MBF <1.8	0.36	0.47	0.90 [0.88, 0.92]
MFR < 2	0.37	0.50	0.89 [0.87, 0.91]

Abbreviations: AI – Artificial Intelligence; MBF – Myocardial Blood Flow; MFR – Myocardial Flow Reserve; TPD –Total Perfusion Deficit.

Supplementary Table 4. Sensitivity Comparison of AI and Quantitative Analysis in High-Risk Groups

	n	AI	Ischemic TPD ≥ 5%
Left main stenosis ≥50% or ≥70% in other arteries	674	566	530
High risk groups			
Duke category 6	224	210	188
Duke category 5	165	153	138
	n	**AI**	**MFR < 2**
Left main stenosis ≥50% or ≥70% in other arteries	674	602	514
High risk groups			
Duke category 6	224	213	190
Duke category 5	165	157	136
	n	**AI**	**MBF < 1.8**
Left main stenosis ≥50% or ≥70% in other arteries	674	607	540
High risk groups			
Duke category 6	224	213	200
Duke category 5	165	157	142

Comparison on the external testing set (n=1,278) with matched specificity. MFR and MBF represent the minimum values across the three vessels. AI model identified more true positives in each category over the quantitative analysis (e.g., ischemic TPD ≥5%). Duke category 6 is defined as left main ≥50% or 3-vessel ≥70%, or 2-vessel CAD including proximal LAD. Duke category 5 is defined as 2-vessel severe stenosis not including the proximal LAD, 1-vessel severe proximal LAD, or 3-vessel moderate stenosis ≥50%[1].

Abbreviations: AI – Artificial Intelligence; CAD – Coronary Artery Disease; LAD – Left Anterior Descending Artery; MBF – Myocardial Blood Flow; MFR – Myocardial Flow Reserve; TPD – Total Perfusion Deficit.

Supplementary Table 5. Performance assessment of AI model stratified by site and isotope.

External Site	Isotope	AI AUC [95% CI]	n
Training set Cedars-Sinai ^{82}Rb (n=386)			
Mayo	$^{13}NH_3$	0.84 [0.79-0.88]	290
BWH	$^{13}NH_3$ / ^{82}Rb		377
BWH-$^{13}NH_3$	$^{13}NH_3$	0.83 [0.77-0.87]	245
BWH-^{82}Rb	^{82}Rb	0.75 [0.66-0.83]	132
Intermountain	^{82}Rb	0.86 [0.83-0.89]	611
Total	$^{13}NH_3$ / ^{82}Rb	0.83 [0.81-0.85]	1278
Training set Cedars-Sinai ^{82}Rb (n=386) and Columbia University $^{13}NH_3$ (n=394)			
Mayo	$^{13}NH_3$	0.84 [0.79-0.88]	290
BWH	$^{13}NH_3$ / ^{82}Rb		377
BWH-$^{13}NH_3$	$^{13}NH_3$	0.82 [0.76-0.87]	245
BWH-^{82}Rb	^{82}Rb	0.75 [0.67-0.83]	132
Intermountain	^{82}Rb	0.87 [0.85-0.90]	611
Total	$^{13}NH_3$ / ^{82}Rb	0.83 [0.81-0.85]	1278

Area under the receiver-operating characteristic curve (AUC) with 95% confidence intervals (CI) was evaluated in the external testing set, stratified by site (Mayo Clinic [Mayo], Brigham and Women's Hospital [BWH], and Intermountain Medical Center [Intermountain]) and by isotope type (^{82}Rb and ^{13}N-ammonia). The Columbia University cohort included 394 patients referred to invasive coronary angiography and utilized $^{13}NH_3$ isotope.

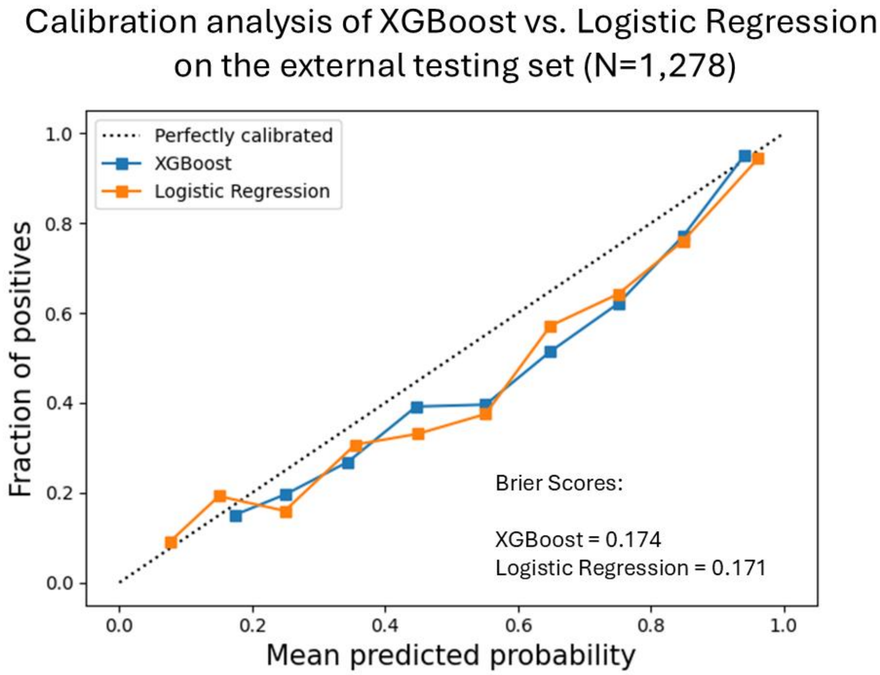
Abbreviations: AI – Artificial Intelligence.

Supplementary Table 6. Comparison of the XGBoost model and conventional logistic regression model in the external testing set (N = 1,278).

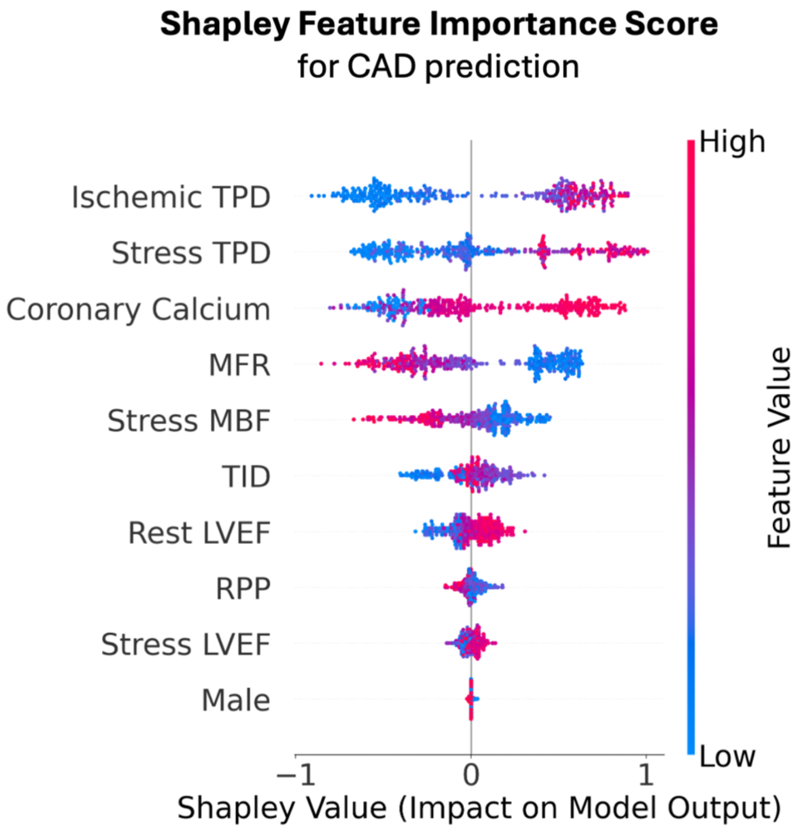
Site	AI AUC [95% CI]	Logistic Regression AUC [95% CI]	p-value	n
Mayo	0.84 [0.79-0.88]	0.84 [0.79-0.88]	0.95	290
BWH	0.77 [0.73-0.82]	0.77 [0.73-0.82]	0.96	377
Intermountain	0.86 [0.83-0.89]	0.88 [0.85-0.90]	0.05	611
Total	0.83 [0.81-0.85]	0.83 [0.81-0.85]	0.27	1278

Area under the receiver-operating characteristic curve (AUC) with 95% confidence intervals (CI) for the XGBoost and logistic regression models, reported overall and stratified by site for performance comparison. Paired DeLong tests were utilized to compare the AUCs.

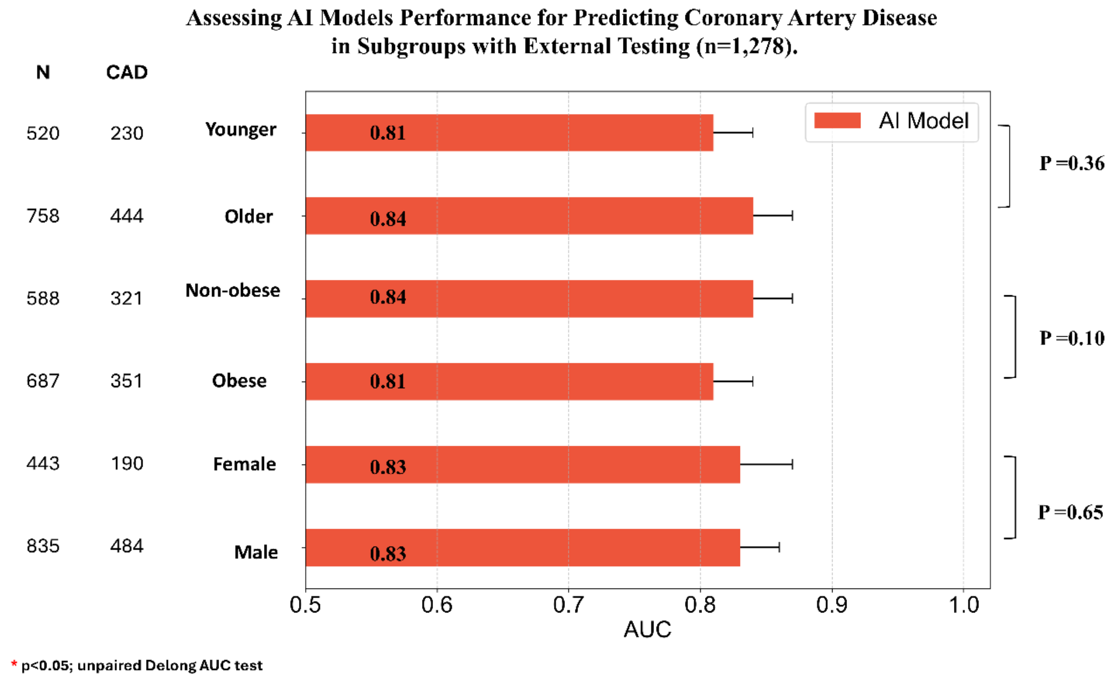
Abbreviations: AI – Artificial Intelligence; XGBoost – Extreme Gradient Boosting.



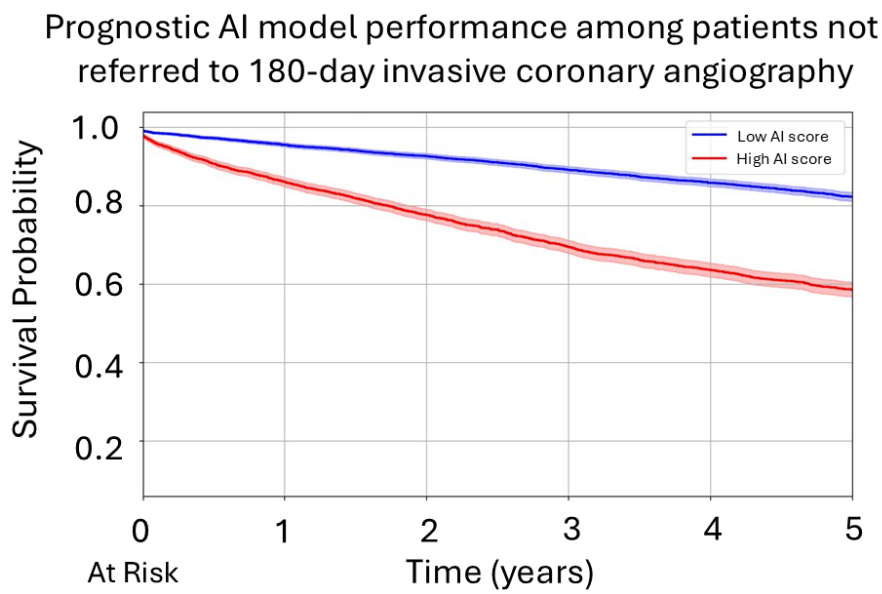
Supplementary Figure 1. Calibration analysis of XGBoost vs Logistic Regression on the external testing set. The Brier scores (XGBoost: 0.174; Logistic Regression: 0.171) suggest that the predicted risks aligned closely with the actual patient outcomes, with lower Brier scores reflecting better calibration.



Supplementary Figure 2. Shapley feature importance for CAD prediction. Each dot represents an individual patient and a specific feature. The distance of each dot from the center line (0 on the x-axis) indicates how strongly that feature drives the prediction, either positively or negatively, for that patient. Abbreviations: LVEF – Left Ventricle Ejection Fraction; MBF – Myocardial Blood Flow; MFR – Myocardial Flow Reserve; RPP – Rate Pressure Product; TID – Transient Ischemic Dilation; TPD – Total Perfusion Deficit.

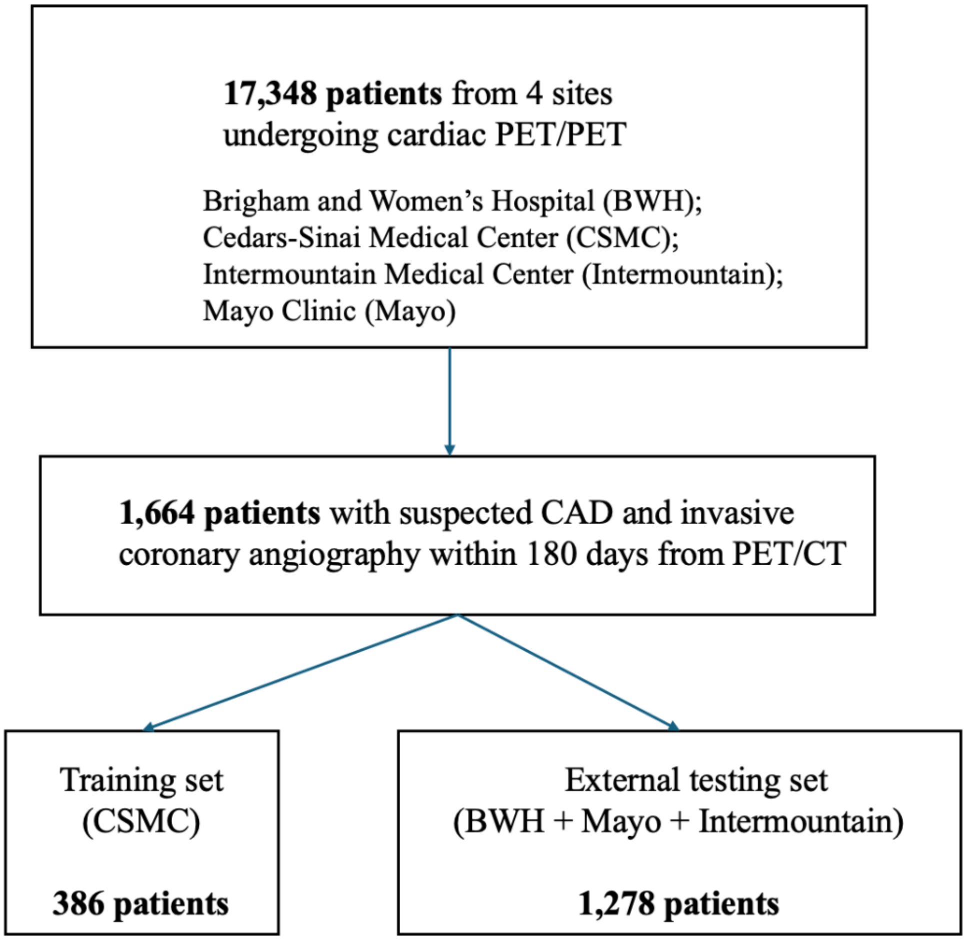


Supplementary Figure 3. Performance of artificial intelligence (AI) for prediction of coronary artery disease (CAD) in subgroups. AI models exhibit comparably robust performance in all subgroups. The younger group was defined as patients aged <65 years, while the older group was defined as patients aged ≥65 years. Similarly, the obese group was defined as patients with a BMI ≥30 kg/m², whereas the non-obese group was defined as patients with a BMI <30 kg/m². Race analysis was excluded due to the limited availability of data on minority race groups. P-values were generated using unpaired Delong test. Abbreviations: AUC – Area Under the Receiver Operating Characteristic Curve; BMI – Body Mass Index.

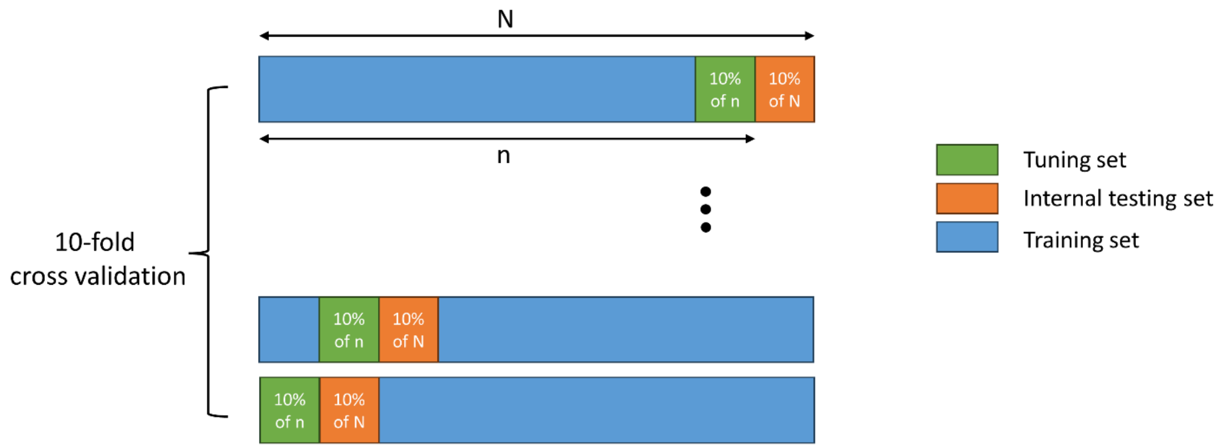


At Risk	0	1	2	3	4	5
High AI score	3547	2752	2211	1822	1528	762
Low AI score	6567	5399	4537	4024	3505	1866

Supplementary Figure 4: Kaplan-Meier curves stratified by high and low AI scores. Prognostic performance of the AI model among patients not referred to 180-day invasive coronary angiography was evaluated in external testing set. The high and low AI scores were stratified to match with clinical score≥4 (AI threshold at 0.41). The predicted likelihoods of CAD from the AI model were used as a prognostic risk score: high predicted CAD AI score might be correlated with more MACE events, and low predicted CAD might be correlated with event-free. Abbreviations: AI – Artificial Intelligence.



Supplementary Figure 5. Flowchart of the study. An illustration describes the inclusion criteria used in the analysis and how training set and external testing set were created for model development.



Supplementary Figure 6. 10-Fold Cross-Validation for Training, Tuning, and Testing. Model training is shown in blue, hyper-parameter tuning in green, and internal testing in orange. Hyper-parameters (e.g., the depth of the tree or the number of estimators) of XGBoost models are often empirically chosen and cannot be learned from model training. The recommended approach to discover the best values of hyper-parameters is by fitting the models with various combinations and assess their performance. To achieve this, a nested 10-fold cross-validation was implemented in each 10-fold subset (n): allocating 90% of this data for model fitting and 10% for tuning.

References

1 Reynolds, H. R. *et al.* Outcomes in the ISCHEMIA Trial Based on Coronary Artery Disease and Ischemia Severity. *Circulation* **144**, 1024-1038 (2021).