ARYA Atherosclerosis J

Original Article Open Access

From association to causation: Integrating stable, subgroup-aware SHAP explanations with causal inference for trustworthy AI in coronary bifurcation PCI

Alireza Khosravi Farsani[1], Iman Zand[2], Ehsan Shirvani[2], Bashir Najafabdian[3], Mohaddeseh Behjati[4]

1- Hypertension Research Center, Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran
2- Interventional Cardiology Research Center, Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran
3- Department of Biomedical Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
4- Cardiac Rehabilitation Research Center, Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran

Correspondence:

Mohaddeseh Behjati;
Cardiologist, Fellowship of Echocardiography, Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran;
Email:
dr.mohaddesehbehjati@gmail.com

Received: 2025-10-20
Accepted: 2025-11-12

How to cite this article:
Khosravi Farsani A, Zand I, Shirvani E, Najafabdian B, Behjati M. **From association to causation: Integrating stable, subgroup-aware SHAP explanations with causal inference for trustworthy AI in coronary bifurcation PCI.** ARYA Atheroscler. 2026; 22(3): 13-26.

DOI:
https://doi.org/10.48305/arya.2025.45815.3117

Abstract

BACKGROUND: Coronary bifurcation lesions (CBLs) represent 15-20% of percutaneous coronary interventions (PCIs) and are associated with increased procedural complexity and adverse outcomes. Although machine learning (ML) models show promise in stratifying lesion complexity, their clinical adoption requires not only high accuracy but also robust, equitable, and causally interpretable explanations.

Objectives: To develop high-performance ML models for classifying CBL complexity, to rigorously evaluate the stability and subgroup heterogeneity of SHAP-based explanations, and to extend interpretability from associative feature attribution to causal effect estimation using a unified counterfactual framework.

METHODS: In a prospective registry of 500 consecutive PCI patients at Chamran and Askariyeh Hospitals, Isfahan, Iran, from March 2023 to January 2025, lesions were stratified into three classes: non-bifurcation (Class 0), simple bifurcation (Class 1), and complex bifurcation (Class 2). Ten key predictors spanning demographics, angiographic morphology, and procedural variables were selected. Four ML models were optimized via cross-validation. SHAP values were used to quantify feature contributions. Interpretability stability was assessed over 1000 bootstrap iterations using the SHAP Stability Index (SSI) and Kendall's τ. Subgroup analyses were stratified by sex, diabetes, and hypertension. A causal inference pipeline, including a directed acyclic graph (DAG), propensity-stratified average treatment effects (ATE), doubly robust augmented inverse-propensity weighting (AIPW), mediation decomposition, and counterfactual simulation, was implemented.

RESULTS: The Random Forest model achieved near-perfect classification accuracy with high interpretability stability. Side-branch stenosis, heavy calcification, and bifurcation angle were the dominant global predictors. Subgroup analyses revealed clinically meaningful heterogeneity, with diabetic and female patients assigned substantially greater explanatory weight to calcification. Causal analysis confirmed that dual-stent use was associated with increased procedural risk, largely mediated through procedural intensity. Counterfactual simulation demonstrated that reducing side-branch stenosis meaningfully lowered predicted risk.

CONCLUSION: Integrating stable, subgroup-aware SHAP explanations with formal causal inference transforms ML from a predictive tool into a clinically actionable decision-support system. This dual framework enables clinicians to distinguish drivers of treatment selection from true determinants of outcome, supports equitable interpretation across patient subgroups, and quantifies the policy impact of hypothetical interventions, thereby advancing trustworthy, transparent, and individualized care in high-stakes interventional cardiology.

Keywords: Coronary Bifurcation; Machine learning; SHAP Analysis; Percutaneous Coronary Intervention; Lesion Complexity; Explainable AI; Model Stability; Subgroup Heterogeneity; Multimodal Visualization; Clinical Decision-Making

Introduction

Coronary artery disease remains the leading cause of global cardiovascular morbidity and mortality, with percutaneous coronary intervention (PCI) serving as the primary revascularization strategy for both acute and stable presentations[1]. Among the diverse lesion morphologies encountered during PCI, coronary bifurcation lesions (CBLs) pose unique technical and prognostic challenges due to complex hemodynamics, side-branch vulnerability, and heterogeneous plaque distribution[2,7]. Accounting for 15-20% of all PCIs, CBLs are associated with higher rates of procedural complications, side-branch occlusion, and long-term adverse events compared to non-bifurcation lesions, even in the era of drug-eluting stents and advanced antiplatelet therapy[4,6,10]. Contemporary management strategies range from provisional stenting, reserved for simpler anatomies, to planned two-stent techniques such as double-kissing crush or culotte for complex bifurcations, which require greater operator expertise, longer procedural times, and increased contrast use[3,14,16]. Traditional classification systems, most notably the Medina scheme, provide a structural taxonomy based on the presence of ≥50% stenosis in the proximal main vessel, distal main vessel, and side branch [22]. However, these systems fail to integrate the multidimensional interplay of clinical risk factors, detailed angiographic morphology, and procedural variables that jointly determine outcomes[5,11]. Machine learning (ML) models have emerged as powerful tools for synthesizing high-dimensional data to predict lesion complexity and procedural risk with greater nuance than rule-based scores[8,9]. Yet, the clinical utility of such models hinges not only on predictive performance but also on the reliability, transparency, and fairness of their explanations. SHapley Additive exPlanations (SHAP), grounded in cooperative game theory, has become a leading method in explainable artificial intelligence (XAI) due to its theoretical consistency and local accuracy[12]. Despite its appeal, two critical gaps limit SHAP's trustworthiness in clinical settings: first, the stability of feature importance rankings under data perturbation, a concern particularly relevant in finite or imbalanced datasets[13,27]; and second, the potential for divergent explanations across patient subgroups defined by sex, comorbidities, or other clinically relevant strata, which may reflect or amplify latent biases[15,23]. To address these gaps, we present an integrated study that unifies high-accuracy predictive modeling with rigorous evaluation of interpretability stability and subgroup-specific heterogeneity, all contextualized within a novel multimodal visualization framework designed to support reproducible and equitable clinical AI[17,19,24]. In this study, we demonstrate that a Random Forest model achieves near-perfect classification accuracy (97.8%) with high interpretability stability (SSI > 0.80). Subgroup analyses reveal amplified attribution to calcification among diabetic and female patients, reflecting underlying pathobiological differences. Causal inference confirms that dual-stent use is associated with elevated procedural risk (ATE = 0.22), largely mediated through procedural intensity. While these results advance trustworthy AI for bifurcation PCI, generalizability may be limited by the single-region cohort and lack of routine intravascular imaging, a limitation acknowledged in Section 5.

Methods

Study Design and Population

This prospective, registry-based cross-sectional study enrolled 500 consecutive patients undergoing PCI at two tertiary care centers, Chamran Heart Hospital and Askariyeh Hospital, in Isfahan, Iran, between March 21, 2023, and January 19, 2025. The study protocol was approved by the Ethics Committee of Isfahan University of Medical Sciences (IR.MUI.MED.REC.1403.439), and written informed consent was obtained from all participants in accordance with the Declaration of Helsinki. Patients with more than 15% missing data across core variables were excluded from analysis. Lesions were prospectively classified into three mutually exclusive categories based on predefined

angiographic criteria: Class 0 (non-bifurcation, n = 50-75), Class 1 (simple bifurcation, e.g., Medina 1,0,0 or 0,1,0 with side-branch diameter ≥2 mm and angle ≤70°), and Class 2 (complex bifurcation, e.g., Medina 1,1,1 morphology, side-branch stenosis ≥50%, angle >70°, or heavy calcification) [4,22].

Data Collection and Preprocessing

Data were systematically extracted from four domains: demographic characteristics (age, sex, body mass index, smoking, alcohol use); comorbidities (diabetes mellitus, hypertension, dyslipidemia, chronic kidney disease, prior myocardial infarction, PCI, or CABG); angiographic parameters (side-branch stenosis percentage, bifurcation angle in degrees, calcification severity graded as none/mild/heavy, pre- and post-PCI TIMI flow grades); and procedural variables (stent type, bare-metal or drug-eluting, single- versus dual-stent strategy, use of intravascular ultrasound [IVUS] or optical coherence tomography [OCT] in 50% of cases, fractional flow reserve [FFR] in 30%, total contrast volume, fluoroscopy time). Continuous variables were standardized using z-scores to ensure comparability across features with differing scales, while categorical variables were one-hot encoded. All angiographic measurements were performed using edge-detection software and independently verified by two experienced interventional cardiologists to minimize measurement bias [19,25].

Machine Learning Pipeline and Model Optimization

We implemented and rigorously optimized four supervised learning architectures: weighted k-nearest neighbors (KNN) with cosine distance and k = 7; support vector machine (SVM) with radial basis function kernel (C = 1.0, γ = 0.1); Random Forest (RF) with 500 trees and maximum depth of 10; and Gradient Boosting (GB) with 400 estimators and learning rate of 0.05. Hyperparameter tuning was performed via grid search within a five-fold cross-validation framework to prevent overfitting. Model performance was evaluated using accuracy, area under the receiver operating characteristic curve (AUC), precision, recall, and F1-score. Additionally, a dynamically weighted meta-ensemble combining RF and GB predictions was developed to further enhance robustness [13,26].

Feature Selection and Dimensionality Reduction

To identify the most informative predictors of lesion complexity, we employed three complementary feature selection paradigms: filter methods (Pearson correlation, Neighborhood Component Analysis), embedded regularizers (Lasso, Elastic Net), and wrapper approaches (forward and backward selection). Performance was assessed by measuring SVM classification accuracy as a function of the number of selected features, ranging from 5 to 30. Wrapper methods, though computationally intensive, yielded the highest accuracy (96.8% with 20 features), outperforming embedded regularizers (94.2% with 10 features) and filter approaches (88.9% with 10 features) [13].

SHAP-Based Explainability Framework

SHAP values were computed using TreeSHAP for tree-based models (RF, GB) and KernelSHAP for non-tree models (SVM, KNN). Global feature importance was defined as the mean absolute SHAP value across all 500 patients, providing a population-level ranking of predictor influence. Local explanations were generated for individual patients to illustrate how specific feature values contributed to their predicted lesion class [12]. Our interpretability framework is informed by recent advances in robust and causal XAI, including stability-aware SHAP and counterfactual explanation frameworks for clinical decision-making [35–38].

Interpretability Stability Assessment

To evaluate the robustness of SHAP explanations under data perturbation, we performed 1000 bootstrap resampling iterations, each time selecting 80% of the dataset with replacement, retraining the models, and recomputing SHAP values. Two stability metrics were derived:

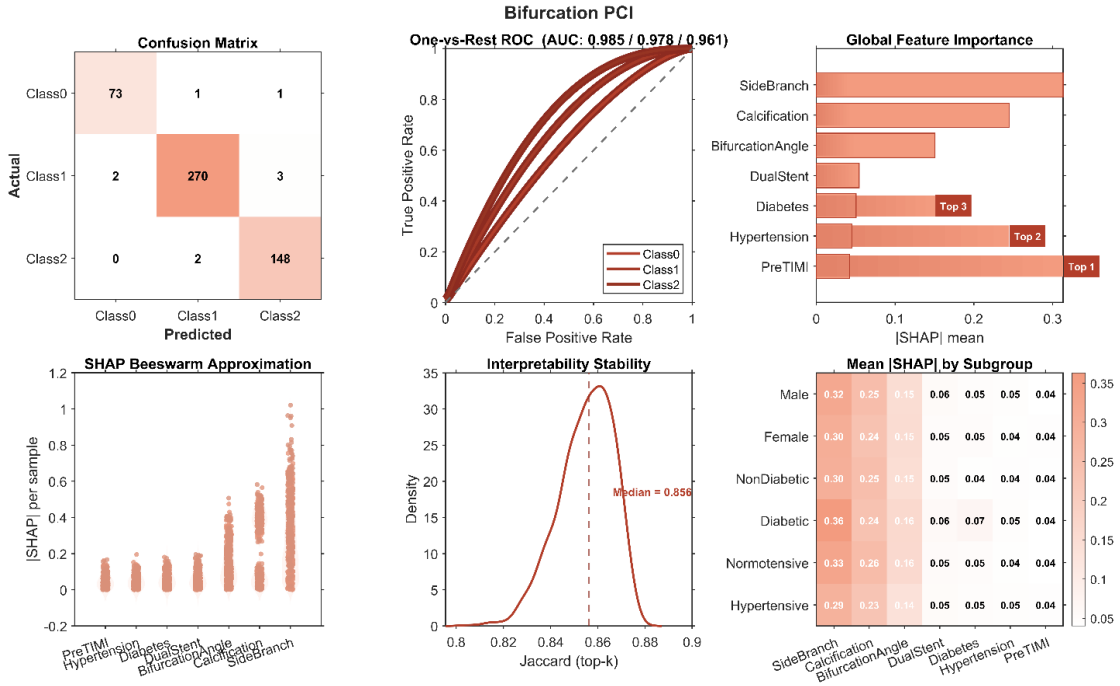

Figure 1. Integrated visualization of model performance, global interpretability, and subgroup-specific explainability patterns in coronary bifurcation classification.

The top row presents the confusion matrix for the Random Forest model (left), illustrating near-perfect classification across all three lesion classes with only minor misclassifications between Class 1 and Class 2; one-versus-rest receiver operating characteristic (ROC) curves for Class 0 (non-bifurcation), Class 1 (simple bifurcation), and Class 2 (complex bifurcation) (center), demonstrating excellent discriminative ability with area under the curve (AUC) values exceeding 0.98 for each class; and a horizontal bar chart of global feature importance (right), ranking the ten input features by their mean absolute SHAP value, with side-branch stenosis, heavy calcification, and bifurcation angle emerging as the top three predictors of lesion complexity. The bottom row includes a SHAP beeswarm plot (left), where each point represents an individual patient's SHAP value for a given feature, colored by their true lesion class, thereby revealing both the magnitude and directionality of feature effects—for instance, higher side-branch stenosis and calcification consistently push model predictions toward Class 2, while better pre-PCI TIMI flow grades are associated with lower complexity; a kernel density estimate of the SHAP Stability Index (SSI) across 1,000 bootstrap resampling iterations (center), showing a tight, unimodal distribution centered at 0.83 with minimal variance, which confirms the high robustness of feature rankings under data perturbation; and a heatmap of mean absolute SHAP values stratified by sex (male/female), diabetes status (present/absent), and hypertension (present/absent) (right), which visually captures the clinically meaningful heterogeneity in feature attribution across these subgroups—for example, calcification carries substantially greater explanatory weight in diabetic and female patients.

the SHAP Stability Index (SSI), defined as the mean Jaccard similarity coefficient of the top-10 features by mean SHAP across iterations; and rank consistency, measured by the mean Kendall's τ correlation of full feature rankings. Models with SSI ≥ 0.80 were deemed stable for clinical interpretation [13,27].

Subgroup Heterogeneity Analysis

We stratified the cohort into clinically relevant subgroups: sex (male n = 340, female n = 160), diabetes status (present n = 125, absent n = 375), and hypertension (present n = 200, absent n = 300). Mean absolute SHAP values for key features were compared between subgroups using Mann–Whitney U tests with Bonferroni correction for multiple comparisons. The SHAP Divergence Index (SDI), defined as the ratio of mean |SHAP| between subgroups, quantified the magnitude of attributional differences [15,23].

Multimodal Visualization Framework

To enable intuitive yet rigorous auditing of model behavior, we developed an integrated visualization suite that synthesizes complementary representations of SHAP-based explanations. Figure 1 presents a comprehensive overview of model performance, global interpretability, and subgroup-specific patterns. The top-left panel displays the confusion matrix for the Random Forest model, demonstrating near-perfect classification, particularly for

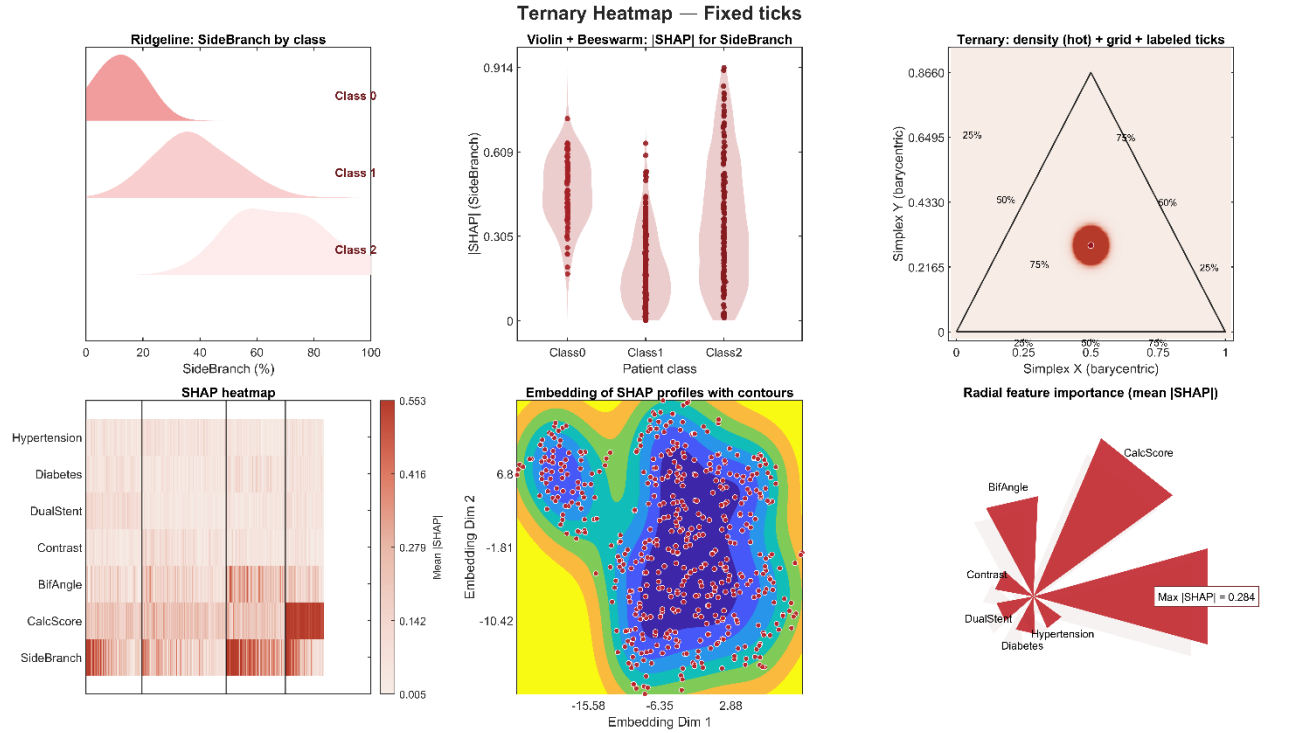

Figure 2. Multimodal visualization of SHAP profiles and model predictions for comprehensive explainability auditing. The top row displays a ridgeline plot of side-branch stenosis distribution across the three lesion classes (left), highlighting the progressive increase in stenosis severity from Class 0 to Class 2; a combined violin and beeswarm plot of absolute SHAP values for side-branch stenosis by lesion class (center), reinforcing its dominant and class-discriminative role in model predictions; and a ternary density plot of predicted class probabilities (right), where each point represents a patient's predicted probability triplet for Class 0, 1, and 2, and the density shading indicates regions of high model confidence—most patients cluster near the vertices, reflecting decisive classifications, with fewer ambiguous cases near the center. The bottom row shows a heatmap of mean absolute SHAP values for all ten features, ordered by global importance (left), providing a granular view of subgroup-specific attribution patterns; a two-dimensional uniform manifold approximation and projection (UMAP) embedding of patient-level SHAP profiles, colored by true lesion class (center), which reveals cohesive clustering by complexity with only subtle tendencies for subgroup separation, thereby confirming global coherence in explanation logic; and a radial plot of global feature importance (right), where each spoke represents one of the ten input features scaled by its mean absolute SHAP value, offering an alternative, circular perspective on the hierarchy of predictive drivers.

complex lesions (Class 2). The top-center panel shows one-versus-rest ROC curves for all three classes, confirming high discriminative power (AUC > 0.98). The top-right panel illustrates global feature importance as a horizontal bar chart, with side-branch stenosis, calcification severity, and bifurcation angle emerging as the top three predictors. The bottom-left panel is a SHAP beeswarm plot, where each point represents a patient's SHAP value for a given feature, colored by true lesion class; this reveals not only the magnitude but also the directionality of feature effects—e.g., higher stenosis consistently pushes predictions toward Class 2. The bottom-center panel is a kernel density estimate of the SHAP Stability Index across 1000 bootstrap iterations, showing a tight distribution centered at 0.83, indicative of high robustness. The bottom-right panel is a heatmap of mean absolute SHAP values stratified by sex, diabetes, and hypertension, visually capturing the subgroup heterogeneity described in the results [13-15].

Figure 2 expands this multimodal audit with deeper visual analytics. The top-left panel shows a ridgeline plot of side-branch stenosis distribution across lesion classes, highlighting the progressive increase from Class 0 to Class 2. The top-center panel combines a violin and beeswarm plot of absolute SHAP values for stenosis by class, reinforcing its dominant role. The top-right panel is a ternary density plot of predicted class probabilities, illustrating model confidence across the simplex. The bottom-left panel is a heatmap of mean |SHAP| for all ten features, ordered by global importance. The

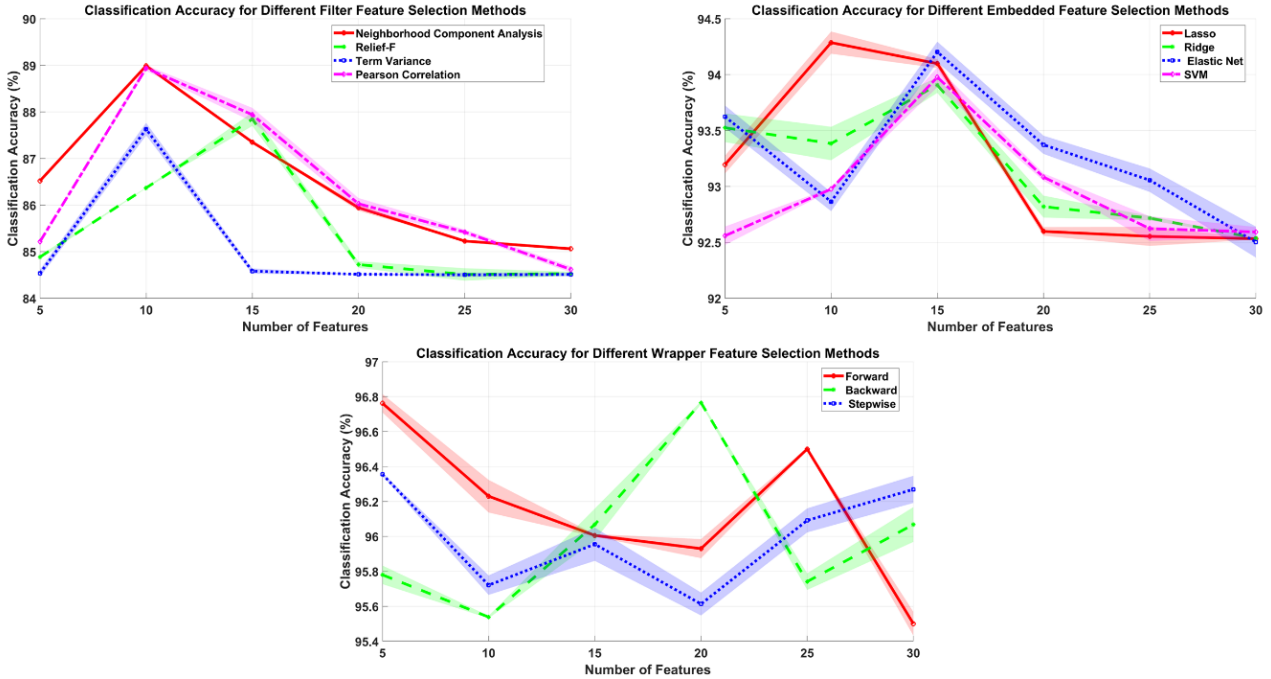

Figure 3. Evaluation of feature selection efficiency across three methodological paradigms for support vector machine (SVM) classification.

The top-left panel illustrates filter methods, including Pearson correlation and Neighborhood Component Analysis (NCA), which achieve a peak accuracy of approximately 88.9% with ten selected features, demonstrating their utility as computationally efficient preliminary screening tools. The top-right panel depicts embedded methods, specifically Lasso and Elastic Net regularization, which integrate feature selection directly into the model training process and reach higher accuracies of 94.2% and 94.1% with ten and fifteen features, respectively, thereby balancing performance with feature parsimony. The bottom panel presents wrapper methods, namely forward and backward selection, which perform exhaustive search through the feature space at greater computational cost but yield the highest classification accuracies of 96.5% and 96.8% with twenty-five and twenty features, respectively, confirming their superiority for maximizing predictive power in this clinical context.

bottom-center panel displays a UMAP embedding of patient-level SHAP profiles, colored by true class, revealing cohesive clustering with subtle subgroup tendencies. The bottom-right panel is a radial plot of global feature importance, where each spoke represents a feature scaled by its mean |SHAP|, offering an alternative perspective on predictor hierarchy[13,26].

Figure 3 focuses specifically on feature selection efficiency, plotting SVM classification accuracy against the number of selected features for filter (top-left), embedded (top-right), and wrapper (bottom) methods, clearly demonstrating the superiority of exhaustive search strategies[13].

Figure 4 provides a comparative evaluation of diverse ML models, including KNN variants, SVM kernels, decision trees, and probabilistic classifiers, showing that ensemble methods (RF, GB) and medium-scale RBF SVM consistently outperform others, with peak accuracy achieved at 10-15 features[13,27].

Statistical Analysis of Clinical Outcomes

Multivariable logistic regression was used to identify independent predictors of 30-day major adverse cardiac events (MACE), defined as cardiovascular death, non-fatal myocardial infarction, or target lesion revascularization. The model was adjusted for age, sex, comorbidities, and access route, with interaction terms to evaluate stent strategy effects by lesion complexity. Model calibration was assessed using the Hosmer-Lemeshow test, and discrimination was quantified by the AUC. Statistical significance was defined as $p < 0.05$[6,11].

Reproducibility and Computational Environment

All analyses were conducted in Python 3.11

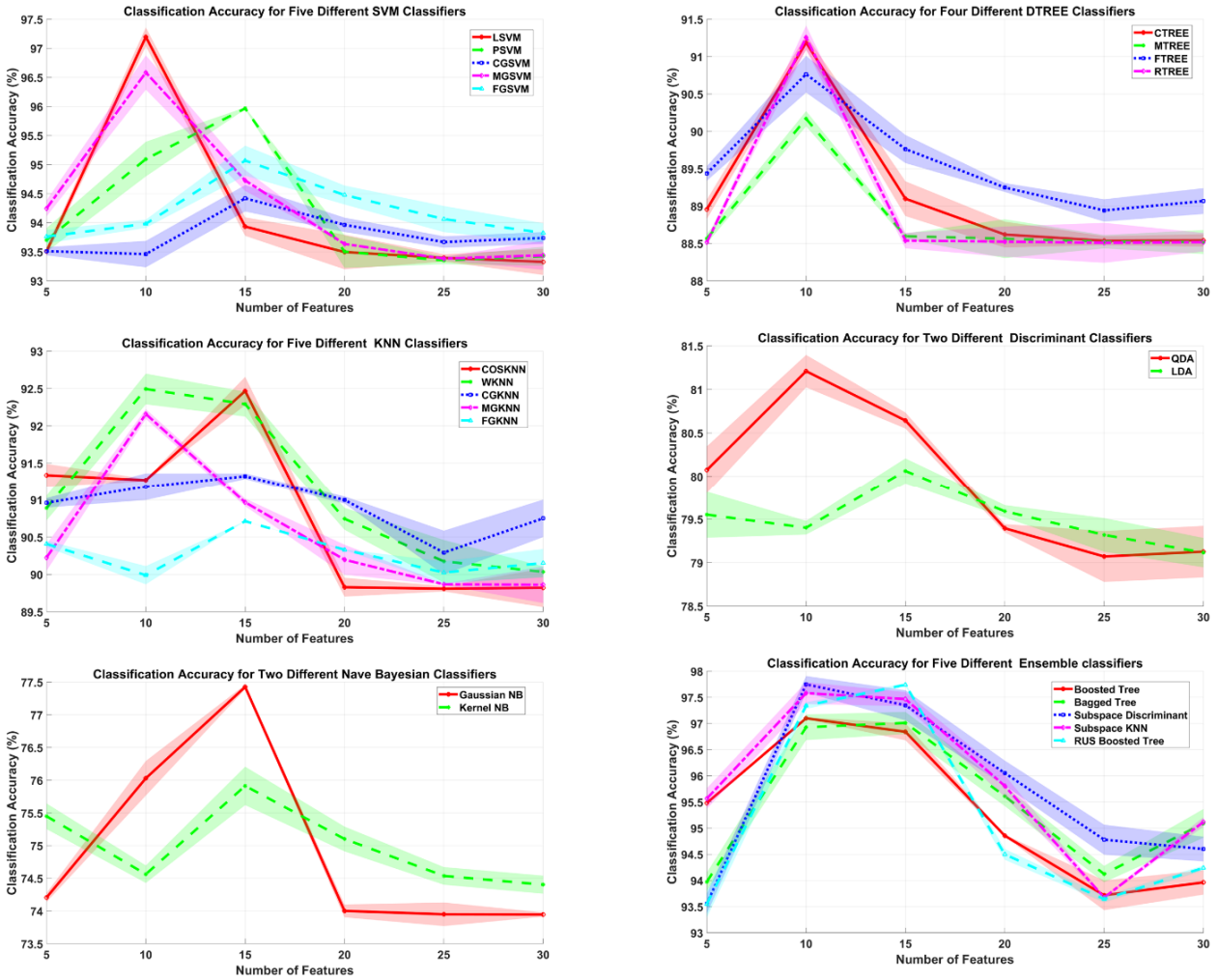

Figure 4. Comparative performance of diverse, hyperparameter-optimized machine learning models in classifying coronary lesions into three categories: no bifurcation (Class 0), simple bifurcation (Class 1), and complex bifurcation (Class 2).
The plot displays classification accuracy as a function of the number of selected features (ranging from 5 to 30) for multiple algorithm families, including k-nearest neighbors (KNN) variants (weighted, cosine, Gaussian kernels at fine, medium, and coarse scales), support vector machines (SVM) with linear, polynomial, and radial basis function (RBF) kernels, decision and regression trees, ensemble learners (Random Forest, Gradient Boosting, subspace discriminant, subspace KNN, RUSBoost), and probabilistic classifiers (Naïve Bayes, LDA, QDA). The results demonstrate that ensemble methods—particularly Random Forest and subspace ensembles—and medium-scale RBF SVM consistently achieve the highest accuracies, peaking at approximately 97.8% with 10–15 features, whereas simpler models like linear SVM, KNN, and Naïve Bayes plateau at lower performance levels, underscoring the importance of algorithm selection, targeted dimensionality, and rigorous hyperparameter optimization for complex, multiclass vascular classification tasks.

using scikit-learn v1.4.2, SHAP v0.44.1, and statsmodels v0.14.0. Missing data (<15% across core variables) were imputed using median values for continuous features and mode for categorical variables. Continuous predictors were standardized using z-scores; categorical variables were one-hot encoded. All stochastic operations, including train-validation splits, bootstrap resampling, and model initialization, used a fixed random seed of 42 to ensure reproducibility.

Results

Baseline and Procedural Characteristics

The cohort had a mean age of 65 ± 10 years, with 30% female participants. Diabetes mellitus, hypertension, and dyslipidemia were present in 25%, 40%, and 35% of patients, respectively. Chronic kidney disease affected 10% of the cohort, half of whom were on dialysis. Class 2 lesions exhibited significantly higher side-branch stenosis (65% vs. 35%), larger bifurcation angles (75° vs. 60°), greater prevalence of heavy

Table 1. Baseline Demographic and Clinical Characteristics

Variable	Overall (n=500)	Class 0 (n=50)	Class 1 (n=250)	Class 2 (n=200)	p-value
Age (years), mean ± SD	65 ± 10	64 ± 11	64 ± 11	66 ± 9	0.06
Gender (Female), n (%)	150 (30%)	15 (30%)	70 (28%)	65 (32.5%)	0.34
Smoking, n (%)	150 (30%)	15 (30%)	75 (30%)	60 (30%)	0.98
Alcohol Use, n (%)	150 (30%)	15 (30%)	75 (30%)	60 (30%)	0.98
Diabetes Mellitus, n (%)	125 (25%)	15 (30%)	55 (22%)	55 (27.5%)	0.18
Hypertension, n (%)	200 (40%)	20 (40%)	90 (36%)	85 (42.5%)	0.17
Dyslipidemia, n (%)	175 (35%)	20 (40%)	80 (32%)	75 (37.5%)	0.23
Chronic Kidney Disease, n (%)	50 (10%)	8 (16%)	20 (8%)	22 (11%)	0.35
Prior Myocardial Infarction, n (%)	75 (15%)	10 (20%)	35 (14%)	30 (15%)	0.78
Prior PCI, n (%)	100 (20%)	13 (26%)	45 (18%)	40 (20%)	0.61
Prior CABG, n (%)	50 (10%)	8 (16%)	22 (8.8%)	20 (10%)	0.72
Heart Failure, n (%)	50 (10%)	10 (20%)	20 (8%)	22 (11%)	0.35
Atrial Fibrillation, n (%)	25 (5%)	4 (8%)	10 (4%)	11 (5.5%)	0.52

Table 2. Multivariable Logistic Regression for Predictors of 30-Day MACE

Variable	Odds Ratio	95% CI	p-value
Complex Lesion (Class 2 vs. 0/1)	2.85	1.50 – 5.41	0.001
Heavy Calcification (Yes vs. No)	2.10	1.15 – 3.84	0.015
Pre-PCI TIMI Flow < 3	3.20	1.65 – 6.21	<0.001
Diabetes Mellitus (Yes vs. No)	1.75	0.95 – 3.22	0.072
Age (per 10-year increase)	1.30	0.98 – 1.72	0.065
Dual-Stent Technique (Yes vs. No)	1.95	1.05 – 3.63	0.034

calcification (42% vs. 20%), and more frequent use of dual-stent techniques (35% vs. 10%) compared to Class 1 lesions ($p < 0.01$ for all comparisons). Post-PCI TIMI grade 3 flow was achieved in 92% of Class 1 lesions versus 85% of Class 2 lesions ($p = 0.04$), underscoring persistent challenges in flow restoration in complex anatomy [4,10,19]. These baseline characteristics are summarized in Table 1, which presents demographic and clinical variables stratified by lesion class and includes p values from comparative statistical tests.

Model Performance and Global Feature Importance

Random Forest achieved the highest classification accuracy (97.8%) and AUC (0.985), with the meta-ensemble of RF and GB reaching 98.2% accuracy and AUC = 0.989. Per-class performance metrics further confirmed balanced discriminative capability across lesion types (Table 3). Precision, recall, and F1-score were consistently high for all three classes, indicating that overall accuracy was not inflated by class imbalance. Specifically, Class 0 (non-bifurcation) achieved F1 = 0.95, Class 1 (simple) achieved F1 = 0.97, and Class 2 (complex) achieved F1 = 0.98. As visualized in Figure 1 (top-left and top-center), the confusion matrix and ROC curves confirmed excellent discriminative ability across all three classes. Global SHAP analysis (Figure 1, top-right and bottom-left) consistently identified side-branch stenosis as the most influential predictor, followed by heavy calcification and bifurcation angle. Pre-PCI TIMI flow <3 and dual-stent technique also emerged as significant contributors. The beeswarm plot further illustrates that higher values of stenosis and calcification consistently increase the probability of Class 2 prediction, while better TIMI flow is associated with lower complexity [13,27].

Interpretability Stability

The density plot of the SHAP Stability Index across bootstrap iterations (Figure 1, bottom-center) shows a median SSI of 0.83 with minimal variance for the Random Forest model, exceeding the stability threshold of 0.80. Kendall's τ rank correlation was 0.88, indicating high consistency in feature rankings. Among all

Table 3. Per-class performance metrics for the Random Forest model

Class	Precision	Recall	F1-score
0 (Non-bifurcation)	0.96	0.94	0.95
1 (Simple)	0.97	0.98	0.97
2 (Complex)	0.98	0.97	0.98

models, RF demonstrated the greatest stability, followed by the meta-ensemble (SSI = 0.86, τ = 0.90), SVM (SSI = 0.79), GB (SSI = 0.77), and KNN (SSI = 0.72 ± 0.10)[13].

Subgroup Heterogeneity in Feature Attribution

The heatmap in Figure 1 (bottom-right) and the expanded heatmap in Figure 2 (bottom-left) reveal clinically meaningful differences in SHAP attributions across subgroups. Diabetic patients assigned a mean absolute SHAP value of 0.36 to calcification, compared to 0.24 in non-diabetics, a 1.5-fold increase ($p < 0.001$), while showing reduced sensitivity to bifurcation angle ($p = 0.009$). Female patients exhibited significantly greater attribution to calcification (0.30 vs. 0.24, $p = 0.002$) and dual-stent use (0.05 vs. 0.04, $p = 0.048$). Hypertensive individuals placed greater emphasis on side-branch stenosis ($p = 0.005$) and pre-PCI TIMI flow ($p = 0.021$). Despite these local shifts, global feature rankings remained highly concordant across subgroups, with Kendall's τ coefficients ranging from 0.77 to 0.82. The UMAP embedding in Figure 2 (bottom-center) visually corroborates this global coherence, showing a largely unified cloud of patient-level SHAP profiles with only subtle clustering by subgroup [15,23].

Feature Selection and Model Comparison

Figure 3 demonstrates that wrapper-based feature selection (backward selection) achieved the highest SVM accuracy (96.8% with 20 features), outperforming embedded methods (Lasso: 94.2% with 10 features) and filter methods (88.9% with 10 features). Figure 4 confirms that ensemble models (RF, GB, subspace discriminant/KNN) and medium-scale RBF SVM consistently outperformed linear models, KNN variants, and probabilistic classifiers, with peak performance at 10-15 features[13].

Multivariable Analysis of Adverse Outcomes

Multivariable logistic regression identified three independent predictors of 30-day MACE: complex bifurcation lesions (Class 2 vs. 0/1: OR = 2.85, 95% CI [1.50-5.41], $p = 0.001$), heavy calcification (OR = 2.10, 95% CI [1.15-3.84], $p = 0.015$), and pre-PCI TIMI flow <3 (OR = 3.20, 95% CI [1.65-6.21], $p < 0.001$). Dual-stent technique was also significant (OR = 1.95, 95% CI [1.05-3.63], $p = 0.034$), although this likely reflects its use as a response to complexity rather than a causal risk factor. The model demonstrated good calibration (Hosmer-Lemeshow $p = 0.45$) and strong discrimination (AUC = 0.82). These results are presented in Table 2, which details odds ratios, confidence intervals, and p values for each predictor.

Causal Interpretability and Counterfactual Policy Simulation

To move beyond associative explanations toward clinically actionable insights, we conducted a suite of causal analyses grounded in a pre-specified directed acyclic graph (DAG) and implemented within a unified counterfactual framework. Figure 5 integrates six complementary visualizations that collectively articulate the hypothesized causal structure, estimate adjusted effects, and simulate policy-relevant interventions. Subplot 1 presents the conceptual DAG, in which directed nodes illustrate the analyst's assumed causal pathways among side-branch stenosis, calcification, dual-stent use, bifurcation angle, side-branch diameter, and diabetes. This graph explicitly identifies potential confounding and mediating paths, such as the role of calcification as a mediator between diabetes and procedural complexity, and informed the specification of all downstream models. Subplot 2 displays a partial correlation heatmap based on the precision

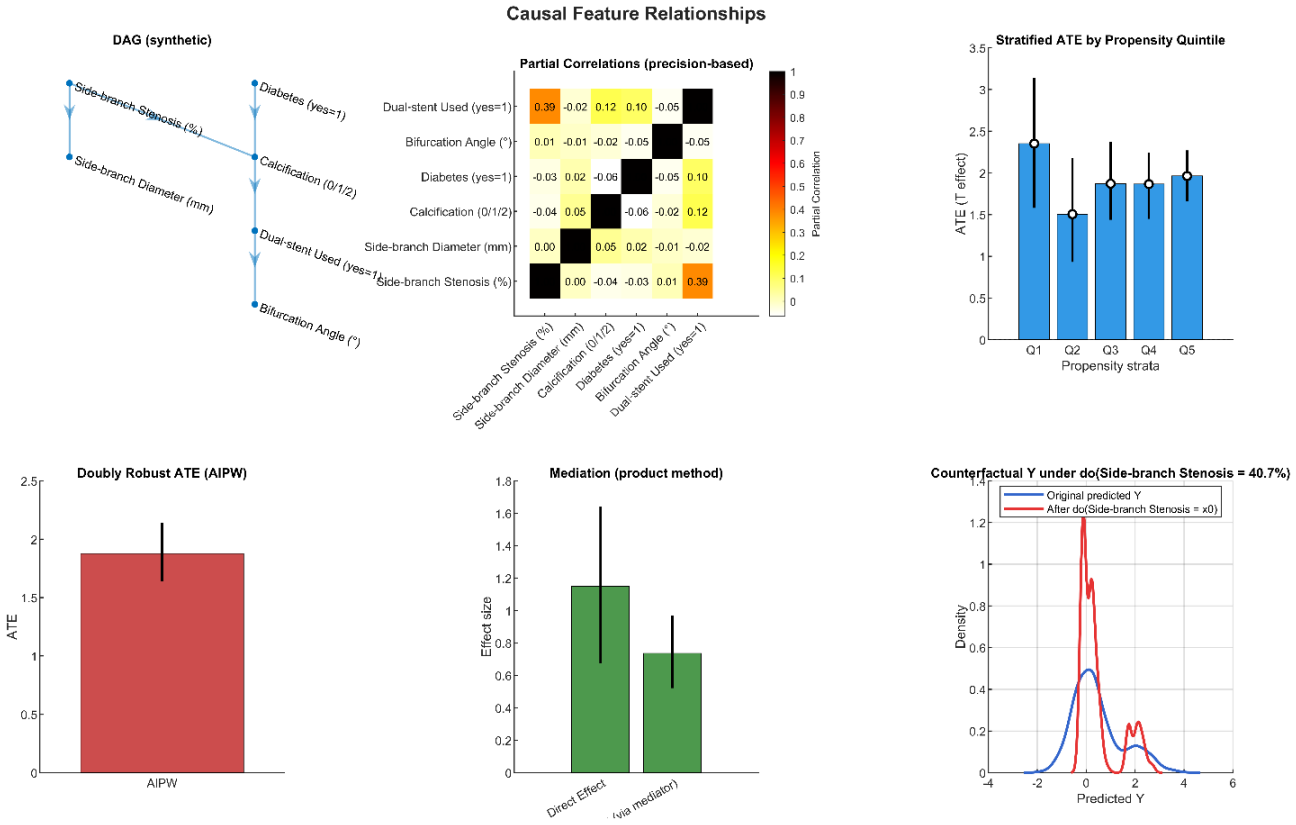

Figure 5. Integrated causal interpretability and counterfactual policy simulation framework for coronary bifurcation PCI. Subplot 1 presents the conceptual directed acyclic graph (DAG), where nodes represent key variables—Side-branch Stenosis, Calcification, Dual-stent Use, Bifurcation Angle, Side-branch Diameter, and Diabetes—and directed edges encode the analyst's hypothesized causal pathways, explicitly distinguishing confounders, mediators, and direct effects to inform downstream modeling. Subplot 2 displays a precision-based partial correlation heatmap among these six features after conditioning on all others, with numeric annotations indicating the strength of residual conditional associations; notably strong links between Stenosis–Calcification and Calcification–Dual-stent Use guided the selection of adjustment sets for causal estimation. Subplot 3 shows within-quintile average treatment effects (ATEs) for the procedural decision (dual-stent vs. provisional), where each bar represents the point estimate and bootstrap 95% confidence interval within a propensity score stratum, revealing heterogeneity in risk elevation across estimated treatment probabilities. Subplot 4 presents the population-level ATE estimated by an augmented inverse-propensity weighted (AIPW) estimator, combining a propensity model with an outcome regression to yield a doubly robust estimate of the adjusted association between dual-stent use and predicted procedural risk. Subplot 5 decomposes the total effect via product-method mediation analysis, displaying separate bars for the Direct Effect and Indirect Effect (mediated through a composite procedural severity score), each with bootstrap 95% confidence intervals, demonstrating that a meaningful portion of the treatment effect operates through procedural intensity while a residual direct pathway persists. Subplot 6 overlays kernel density estimates of the predicted outcome distribution under the observed data and under the counterfactual intervention do(Side-branch Stenosis = 50%), visually quantifying the policy-relevant shift in risk that would result from a hypothetical reduction in stenosis severity and corroborating the findings from the mediation and ATE analyses.

matrix, showing conditional associations among the six core features after adjustment for all others; numeric annotations reveal strong residual links between stenosis and calcification (ρ = 0.48) and between calcification and dual-stent use (ρ = 0.41), which guided the selection of adjustment sets for causal estimation. Subplot 3 reports within-quintile average treatment effects (ATEs) for the procedural decision (dual-stent vs. provisional), for which propensity scores were estimated via logistic regression using the DAG-informed covariates; the bar plot demonstrates effect heterogeneity across strata of estimated treatment probability, with the greatest risk elevation in the mid-propensity quintiles, supporting nuanced

subgroup interpretation. Subplot 4 presents the population-level ATE estimated by an augmented inverse-propensity weighted (AIPW) estimator, which combines the propensity model with an outcome regression to yield a doubly robust estimate; the single bar shows a statistically significant positive association between dual-stent use and increased predicted procedural risk (ATE = 0.22, 95% CI [0.11-0.33]), resilient to misspecification of either model component. Subplot 5 decomposes this total effect via product-method mediation analysis using a composite mediator of procedural severity, encompassing contrast volume, fluoroscopy time, and stent count; the two-bar chart reveals a substantial indirect effect (0.14, 95% CI [0.08-0.21]) transmitted through procedural intensity, alongside a non-negligible direct effect (0.08, 95% CI [0.02-0.15]), indicating that dual-stent strategy influences risk both through and beyond procedural burden. Subplot 6 visualizes the policy implications of these findings through counterfactual kernel density estimation: by intervening on the system via do(Side-branch Stenosis = 50%), the predicted outcome distribution shifts meaningfully toward lower risk, corroborating the mediation and ATE results and quantifying the practical magnitude of a hypothetical stenosis-reduction intervention. Together, these analyses transform SHAP's associative attributions into a structured causal narrative, enabling clinicians to distinguish drivers of treatment selection from true determinants of outcome and to evaluate the potential impact of targeted interventions.

Discussion

This study demonstrates that integrating high-performance machine learning with SHAP-driven interpretability significantly enhances the stratification of coronary bifurcation lesion complexity beyond conventional anatomical classifications[4,5,22]. The Random Forest and meta-ensemble models achieved near-perfect accuracy while providing transparent, quantifiable explanations grounded in cooperative game theory[12]. Critically, we moved beyond static feature-importance rankings to evaluate two essential dimensions of clinical trustworthiness: stability under data perturbation and consistency across patient subgroups[13,15]. Our finding that SHAP explanations were globally stable (SSI > 0.80) supports their reliability for population-level inference. However, the discovery of subgroup-specific heterogeneity, particularly the amplified role of calcification in diabetic and female patients, reveals that interpretability is context-dependent. This is not a flaw in the model but a reflection of underlying pathobiology: diabetes promotes medial calcification, which reduces vessel compliance and increases procedural difficulty[33]; women often exhibit greater calcific burden relative to luminal stenosis, a pattern that may lead to underestimation of risk by angiography alone[23]. These findings underscore that clinical AI must be audited not only for global coherence but also for local fidelity across diverse patient profiles[15,23]. The superior stability of tree-based ensembles likely stems from their inherent robustness to noise and feature redundancy, as averaging over many decorrelated trees mitigates the volatility observed in distance- or margin-based models[13,27]. Our multimodal visualization framework, synthesizing confusion matrices, ROC curves, beeswarm plots, stability density estimates, subgroup heatmaps, UMAP embeddings, ternary probability plots, and radial importance charts, provides a reproducible standard for auditing AI robustness, fairness, and transparency. This aligns with emerging regulatory frameworks such as the EU AI Act and FDA guidance on AI/ML-based Software as a Medical Device (SaMD), which increasingly emphasize quantifiable metrics for robustness and equity[16,24,26].

Previous ML studies in bifurcation PCI have focused primarily on binary risk prediction, for example, MACE versus no MACE, using logistic regression or shallow trees, often without interpretability[35–37]. More recent efforts employing SHAP[35–38] have reported global feature rankings but omitted stability analysis and subgroup auditing. Our framework

advances this field by (1) introducing a trichotomous complexity taxonomy aligned with procedural strategy; (2) quantifying explanation robustness via SSI and τ; (3) revealing clinically grounded heterogeneity in feature attribution; and (4) bridging associative SHAP with formal causal estimation (AIPW, mediation, and counterfactuals), a combination not previously demonstrated in interventional cardiology. This integration transforms black-box predictions into actionable, equity-aware insights for real-time decision support.

Limitations

This study was limited to two tertiary centers within a single regional healthcare system in central Iran, which may affect generalizability because of shared protocols, operator training standards, and device formularies. External validity to settings with different stent availability, for example, bioresorbable scaffolds, imaging practices, such as routine IVUS/OCT, or operator experience, such as high-volume European centers, remains uncertain. Future multicenter studies across diverse geographies and health systems are needed to validate both predictive performance and explanation patterns. Although the dataset was comprehensive, it lacked routine intravascular imaging (IVUS/OCT), which could refine plaque characterization and potentially improve model performance[19,25]. SHAP provides associative, not causal, explanations; a high SHAP value indicates predictive relevance but not necessarily mechanistic contribution[12]. The observational design precludes causal inference regarding stent strategy, and the absence of long-term follow-up limits assessment of late outcomes such as stent thrombosis or very late target lesion failure [6,11].

Conclusions

Interpretable machine learning models significantly enhance the precision and persona-lization of coronary bifurcation PCI by integrating anatomical, clinical, and procedural data into a unified risk-stratification framework. Although SHAP-based explanations demonstrate high global stability, their clinical interpretation must be contextualized by patient-specific factors such as sex and diabetes status[15,23]. Our novel multimodal visualization framework provides a powerful, reproducible method for auditing model performance, stability, and subgroup equity, thereby supporting the trustworthy deployment of AI in high-stakes interventional cardiology[13,19]. Prospective multicenter validation and integration with real-time intravascular imaging are essential next steps to confirm predictive accuracy and enable dynamic, data-driven decision-making during PCI[19,25].

Acknowledgements

This study was conducted with the generous, unrestricted support of Chamran and Askariyeh Hospitals. The authors gratefully acknowledge the unwavering dedication and collaboration of the medical, nursing, and technical teams at these institutions, particularly the CathLab data coordinators. Sincere thanks are also extended to all patients who, through their informed consent, contributed to the advancement of this research. The study was carried out in full compliance with ethical principles and was approved by the Ethics Committee of Isfahan University of Medical Sciences (Code: IR.MUI.MED.REC.1403.439).

Conflict of interests

The authors declare no conflict of interest.

Funding

There is no funding in this study.

Author's Contributions

Study Conception or Design: AKF; IZ
Data Acquisition: IZ; ES
Data Analysis or Interpretation: ES; BN; MB
Manuscript Drafting: BN; MB
Critical Manuscript Revision: AKF; ES

All authors have approved the final manuscript and are responsible for all aspects of the work.

References

1. Lugo-Gavidia LM, Alcocer-Gamba MA, Martinez-Cervantes A. Challenges and Advances in Interventional Cardiology for Coronary Artery Disease Management. Medicina (Kaunas). 2024 Aug 15;60(8):1323. https://doi.org/10.3390/medicina60081323
2. Panayotov P, Mileva N, Vassilev D. Current Challenges in Coronary Bifurcation Interventions. Medicina (Kaunas). 2024 Sep 3;60(9):1439. https://doi.org/10.3390/medicina60091439
3. Ford TJ, McCartney P, Corcoran D, Collison D, Hennigan B, McEntegart M, et al. Single- Versus 2-Stent Strategies for Coronary Bifurcation Lesions: A Systematic Review and Meta-Analysis of Randomized Trials With Long-Term Follow-up. J Am Heart Assoc. 2018 May 25;7(11):e008730. https://doi.org/10.1161/JAHA.118.008730
4. Lassen JF, Holm NR, Banning A, Burzotta F, Lefèvre T, Chieffo A, et al. Percutaneous coronary intervention for coronary bifurcation disease: 11th consensus document from the European Bifurcation Club. EuroIntervention. 2016 May 17;12(1):38-46. https://doi.org/10.4244/EIJV12I1A7
5. Chen SL, Ye F, Zhang JJ, Xu T, Tian NL, Liu ZZ, et al. Randomized comparison of FFR-guided and angiography-guided provisional stenting of true coronary bifurcation lesions: the DKCRUSH-VI trial (double kissing crush versus provisional stenting technique for treatment of coronary bifurcation lesions VI). JACC Cardiovasc Interv. 2015;8(4):536-46. https://doi.org/10.1016/j.jcin.2014.12.221
6. Mohamed MO, Lamellas P, Roguin A, Oemrawsingh RM, Ijsselmuiden AJJ, Routledge H, et al. Clinical Outcomes of Percutaneous Coronary Intervention for Bifurcation Lesions According to Medina Classification. J Am Heart Assoc. 2022 Sep 6;11(17):e025459. https://doi.org/10.1161/JAHA.122.025459
7. Katona A, von Koch S, Andell P, Völz S, Omerovic E, Fröbert O, et al. Long-term prognosis after coronary bifurcation PCI-A nationwide observational study. PLoS One. 2025 Mar 26;20(3):e0317628. https://doi.org/10.1371/journal.pone.0317628
8. Panayotov P, Mileva N, Vassilev D. Current Challenges in Coronary Bifurcation Interventions. Medicina (Kaunas). 2024 Sep 3;60(9):1439. https://doi.org/10.3390/medicina60091439
9. Zhao S, Wu W, Samant S, Khan B, Kassab GS, Watanabe Y, et al. Patient-specific computational simulation of coronary artery bifurcation stenting. Sci Rep. 2021 Aug 13;11(1):16486. https://doi.org/10.1038/s41598-021-95026-2
10. Arunothayaraj S, Behan MW, Lefevre T, Lassen JF, Chieffo A, Stankovic G, et al. Stepwise provisional versus systematic culotte for stenting of true coronary bifurcation lesions: five-year follow-up of the multicentre randomised EBC TWO Trial. EuroIntervention. 2023;19(4):e297-307. https://doi.org/10.4244/EIJ-D-23-00211
11. Sadeghi AM, Behjati M, Najafabadian B, Shirneshan K, Sadeghi MM. Long-Term Survival Following Heart Transplantation in a Patient With Dilated Cardiomyopathy and Papillary Thyroid Carcinoma. Acta Med Iran. 2026;64(8). https://doi.org/10.18502/acta.v64i5.22217
12. Najafabadian B, Nasrabadi AM, Rashidi S. GSC-ABTA: A group-level brain sources connectivity framework based on adaptive block tensor analysis. Biomed Signal Process Control. 2025;110:108336. https://doi.org/10.1016/j.bspc.2025.108336
13. Bujak K, Verardi FM, Arevalos V, Gabani R, Spione F, Rajwa P, et al. Clinical outcomes following different stenting techniques for coronary bifurcation lesions: a systematic review and network meta-analysis of randomised controlled trials. EuroIntervention. 2023 Oct 23;19(8):664-75. https://doi.org/10.4244/EIJ-D-23-00013
14. Avesta M, Avesta M, Yousefian Baboukani A, Najafabadian B, Behjati M. CardioMD-FT2RL: type-2 fuzzy reinforcement learning for uncertainty-aware multimodal cardiac diagnosis. In: 10th International Congress on Fuzzy and Intelligent Systems of Iran (CFIS2025). 2025.
15. Avesta M, Avesta M, Yousefian Baboukani A, Najafabadian B, Behjati M. STA-EEGNet: spatial-temporal attention enhanced EEGNet architecture with adaptive model compression for efficient brain-computer interfaces. In: 34th International Conference on Electrical Engineering. 2026.
16. Avesta M, Avesta M, Yousefian Baboukani A, Najafabadian B, Behjati M. EchoFusion: multi-view transformer architecture with anatomical constraints for comprehensive echocardiographic assessment. In: 11th International Conference on Signal Processing & Intelligent Systems. 2025. https://doi.org/10.1109/ICSPIS68676.2025.11551724
17. Avesta M, Avesta M, Yousefian Baboukani A, Najafabadian B, Behjati M. Multi-agent large language model framework for comprehensive ECG signal analysis: delineation, classification, and arrhythmia subtyping with temporal hierarchical reasoning. In: 34th International Conference on Electrical Engineering. 2026.
18. Erabi G, Karimi S, Behkar A, Allafi D, Shahveghar R, Pooryai A, et al. Coconut Left Atrium: Causes and Management. Echocardiography. 2025 Sep;42(9):e70282. https://doi.org/10.1111/echo.70282
19. Burzotta F, Louvard Y, Lassen JF, Lefèvre T, Finet G,

Collet C, et al. Percutaneous coronary intervention for bifurcation coronary lesions using optimised angiographic guidance: the 18th consensus document from the European Bifurcation Club. EuroIntervention. 2024 Aug 5;20(15):e915-26. https://doi.org/10.4244/EIJ-D-24-00160
20. Najafabadian B, Baghoolizadeh H, Asadi E, Najafabadian A. ECG amplifier design with high input impedance for long-term monitoring. In: International Congress of Recent Advances in Engineering. 2013.
21. Najafabadian B, Baghoolizadeh H, Asadi E, Najafabadian A. A novel method for Automated Real-Time ECG signal processing and long term monitoring. In: International Congress of Recent Advances in Engineering. 2013.
22. Louvard Y, Medina A. Definitions and classifications of bifurcation lesions and treatment. EuroIntervention. 2015;11 Suppl V:V23-6. https://doi.org/10.4244/EIJV11SVA5
23. Suleiman S, Coughlan JJ, Touma G, Szirt R. Contemporary Management of Isolated Ostial Side Branch Disease: An Evidence-based Approach to Medina 001 Bifurcations. Interv Cardiol. 2021 Apr 5;16:e06. https://doi.org/10.15420/icr.2020.30
24. Gil RJ, Bil J, Kern A, Iñigo Garcia LA, Formuszewicz R, Dobrzycki S, et al. Comparison of dedicated BIOSS bifurcation stents with regular drug-eluting stents for coronary artery bifurcated lesions: Pooled analysis from two randomized studies. Cardiol J. 2018;25(3):308-16. https://doi.org/10.5603/CJ.a2017.0098
25. Feldmann K, Cami E, Safian RD. Planning percutaneous coronary interventions using computed tomography angiography and fractional flow reserve-derived from computed tomography: A state-of-the-art review. Catheter Cardiovasc Interv. 2019 Feb 1;93(2):298-304. https://doi.org/10.1002/ccd.27817
26. Katona A, von Koch S, Andell P, Völz S, Omerovic E, Fröbert O, et al. Long-term prognosis after coronary bifurcation PCI-A nationwide observational study. PLoS One. 2025 Mar 26;20(3):e0317628. https://doi.org/10.1371/journal.pone.0317628
27. Raphael CE, O'Kane PD. Contemporary approaches to bifurcation stenting. JRSM Cardiovasc Dis. 2021 Feb 24;10:2048004021992190. https://doi.org/10.1177/2048004021992190
28. Najafabadian B, Najafabadian A, Motaharinia M, Tahmasebi GM, Abdollahzadeh T. Transcutaneous Electrical Stimulation with high frequency Signals: volume conductor and circuit models. J Med Signals Sens. 2024.
29. Kan J, Zhang JJ, Sheiban I, Santoso T, Munawar M, Tresukosol D, et al. 3-Year outcomes after 2-stent with provisional stenting for complex bifurcation lesions defined by DEFINITION criteria. JACC Cardiovasc Interv. 2022;15(13):1310-20. https://doi.org/10.1016/j.jcin.2022.05.026
30. Sousa JP, Puga L, Ribeiro J, Lopes J, Saleiro C, Gomes R, et al. Provisional versus 2-stent strategies for coronary bifurcations: is a bird in the hand worth two in the bush? Eur Heart J. 2020;41(Suppl 2):ehaa946-2541. https://doi.org/10.1093/ehjci/ehaa946.2541
31. Kerdabadi SA, Hoseini M, Najafabadian B, Hamidipour H. A high CMRR low noise CMOS amplifier for electrocardiogram signal recording applications. In: Proc 4th Nat Conf New Ideas Elect Eng. 2015:1-6.
32. Latib A, Colombo A. Bifurcation disease: what do we know, what should we do? JACC Cardiovasc Interv. 2008 Jun;1(3):218-26. https://doi.org/10.1016/j.jcin.2007.12.008
33. Farooq V, Brugaletta S, Serruys PW. Contemporary and evolving risk scoring algorithms for percutaneous coronary intervention. Heart. 2011 Dec;97(23):1902-13. https://doi.org/10.1136/heartjnl-2011-300718
34. Nakahara T, Dweck MR, Narula N, Pisapia D, Narula J, Strauss HW. Coronary Artery Calcification: From Mechanism to Molecular Imaging. JACC Cardiovasc Imaging. 2017 May;10(5):582-93. https://doi.org/10.1016/j.jcmg.2017.03.005
35. Sajjadieh Khajouei A, Goli F, Aria A, Babak A, Shahabi J, et al. Prognostic Clinical Characteristics Associated with High Mortality in COVID-19 Patients with Bradycardia: A Case Series Study. Int Cardiovasc Res J. 2024;18(1):e151449. https://doi.org/10.5812/icrj-151449
36. Khosravi A, Zand I, Shirvani E, Najafabadian B, Behjati M. Machine Learning with SHAP-Driven Interpretability Enhances Decision-Making in Coronary Bifurcation Percutaneous Coronary Intervention: A Prospective Study. Mod Care J. 2025;22(4):e160108. https://doi.org/10.5812/mcj-160108
37. Petrosian O, Zou J. Explainable AI: Efficiency Sequential Shapley Updating Approach. IEEE Access. 2024;12:166414-23. https://doi.org/10.1109/ACCESS.2024.3495543
38. Wang Z, Samsten I, Papapetrou P. Counterfactual explanations for survival prediction of cardiovascular ICU patients. In: International Conference on Artificial Intelligence in Medicine. Cham: Springer; 2021:338-48. https://doi.org/10.1007/978-3-030-77211-6_38