Integrating Whole Blood Gene Expression and Clinical Metadata into a Machine Learning Pipeline for Predictive Coronary Artery Disease Diagnosis

Bilgin Demir [1], Djansel Bukovec [2], and Zhilbert Tafa [3,4,*]

[1] Computer Engineering, Faculty of Engineering, International Balkan University, Skopje, North Macedonia
[2] Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, North Macedonia
[3] Department of Computer Science and Engineering, University for Business and Technology, Prishtina, Kosovo
[4] Faculty of Computer Engineering, International Balkan University, Skopje, North Macedonia
Email: bilgin.demir@ibu.edu.mk (B.D.); dzhansel.bukovec@students.finki.ukim.mk (D.B.); tafaul@t-com.me (Z.T.)
*Corresponding author

***Abstract*—Timely and accurate identification of signs of Coronary Artery Disease (CAD) remains a central issue in modern preventive cardiology. Traditional diagnostic methods overlook the molecular alterations that drive disease occurrence and progression. This study proposes a multimodal machine learning framework integrating whole-blood gene expression profiles with clinical variables to support molecularly informed CAD diagnostics. We employed a Machine Learning framework which prioritizes robustness and interpretability over maximal predictive performance. Models based on the combined microarray and clinical data achieved the highest internal discrimination (AUC ≈ 0.76). The confounding analyses identified age as the dominant predictor, whereas gene expression features contributed an independent and biologically meaningful signal. The final selected gene set was enriched in immune and inflammatory pathways relevant to CAD pathophysiology. External validation revealed a decrease in performance, consistent with expected cross-platform domain shifts. The study underscores both the potential and the challenges of transcriptomic prediction of CAD, highlighting the importance of rigorous validation and data interpretation in high-dimensional biomedical modeling.**

***Keywords*—coronary artery disease, functional enrichment, gene expression, machine learning, logistic regression, statistical analysis**

I. Introduction

Globally, Coronary Artery Disease (CAD) remains the leading cause of cardiovascular mortality, responsible for nearly half of all heart-related deaths and contributing significantly to Disability-Adjusted Life Years (DALYs) worldwide [1].

CAD begins with endothelial dysfunction—damage to the inner lining of coronary arteries, which typically prevents cell adhesion and selectively regulates permeability. Oxidative stress, turbulent shear forces, and elevated Low-Density Lipoprotein (LDL) cholesterol levels stimulate the endothelium to express adhesion molecules, enabling monocytes to infiltrate the arterial wall. Once inside, these monocytes differentiate into macrophages, ingest oxidized LDL through scavenger receptors, and transform into lipid-laden foam cells, initiating the earliest observable lesion known as the fatty streak. Foam cells, cholesterol crystals, and cytokines drive vascular inflammation, prompting smooth muscle cells to migrate, proliferate, and form a collagen-rich fibrous cap over a lipid core. Advanced plaques may calcify, haemorrhage, or degrade, weakening the cap. Atherosclerotic plaques may significantly narrow the arterial lumen and restrict oxygen supply to tissues. When these plaques rupture, they expose thrombogenic material to the bloodstream, initiating platelet activation and thrombus formation. This can lead to sudden vessel occlusion and potentially cause a myocardial infarction [2, 3].

While conventional risk factors, such as age, hypertension, dyslipidaemia, diabetes, smoking, obesity, and family history, remain essential in clinical screening, the variable timing and severity of acute coronary events suggest a more intricate interplay between genetic and molecular regulators. Understanding the molecular underpinnings of CAD is essential for developing effective therapeutic strategies [4].

Gene expression has been extensively utilized to feed pipelines of computational medicine for various purposes. Some examples of their use in neurology, immunology, and oncology are given in Refs. [5−7], respectively. Recent studies suggest that gene expression profiling in peripheral blood may offer a non-invasive window into early molecular perturbations in the coronary circulation [8]. Traditional statistical approaches show potential but, when applied alone, they struggle to handle high dimensionality

Manuscript received December 20, 2025; revised January 24, 2026; accepted March 2, 2026; published May 25, 2026.

of noisy transcriptomic data. More advanced data mining methods provide greater flexibility and power to extract clinically relevant patterns from such datasets.

This work enriches the line of inquiry by integrating transcriptomic signatures with routine clinical variables in a precision medicine framework. We propose a predictive classification framework that combines whole-blood gene expression with age, sex, and smoking history, and evaluates its diagnostic accuracy across multiple models. Using the well-curated GSE20681, profiled on GPL4133 (Agilent) microarray platform [9], we apply Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression within nested cross-validation to identify a fold-stable transcriptomic signal. This process yields an 8-gene signature which, when combined with clinical metadata, forms the input for supervised predictive classification.

Our results demonstrate that integrating whole-blood gene expression with clinical metadata yields measurable discrimination. Biological interpretation was supported through pathway enrichment analysis of the final selected gene set, providing mechanistic context for the predictive signal. External validation on an independent cohort profiled on a different microarray platform showed attenuated performance, consistent with cross-platform domain shifts commonly observed in transcriptomic modeling.

Although our models perform classification, the broader task is predictive in nature, as we aim to infer disease status from molecular and clinical inputs. To emphasize both aspects, we refer to the task as predictive classification throughout this work. The proposed framework was designed to prioritize robustness over maximal predictive performance.

The key contributions of this work span both computational and biological processes and can be summarized as follows:

- We designed a machine learning framework for microarray data processing, taking into account the specific nature of this data type—high dimensionality and limited sample size.
- We demonstrated that integrating microarray data with clinical metadata enhances the real-world relevance, surpassing the use of each of them alone.
- We performed a confounding sensitivity analysis to separate clinical and transcriptomic effects.
- Our analysis maps immune-inflammatory and ion channel pathways to identify genes for targeted monitoring and pharmacogenomics stratification, guiding personalized antiarrhythmic therapy.

We hypothesize that a well-designed machine-learning framework with gene expression and clinical variables as inputs can yield stable and biologically interpretable signatures for CAD prediction that generalize across cohorts. We aim to assess whether gene expression adds predictive value beyond demographic confounders, and whether a stable gene signature can be identified. We expect moderate predictive performance in this setting, reflecting the biological and technical heterogeneity of CAD data.

The paper is structured as follows: Section I presents the research background, objectives, and significance. Section II provides a brief literature review on this field. Section III outlines the materials and methods, including data acquisition, pre-processing, feature selection, and model construction. Section IV presents and discusses the classification performance, validation results, and functional enrichment outcomes. Section V concludes with a summary of key findings and directions for future work. Supplementary results are provided in Appendices.

II. Related Work

Recent biomarker studies for CAD diagnosis have increasingly focused on integrating blood-derived transcriptomic profiles with machine learning, demonstrating considerable promise but also exposing important limitations.

Shi *et al.* [10] conducted a meta-analysis across multiple microarray studies to derive consistent CAD gene signatures. The approach highlights the potential of gene expression data for CAD prediction. However, it neglects the importance of clinical covariates and the possibility of achieving synergistic effects through data integration which limits its clinical applicability. Moreover, the model was not tested for generalizability.

Yang and Xu [11] applied weighted gene co-expression networks to identify CAD-related genes. Although the approach expands the scope of data integration, the proposed models are complex and less interpretable. The analysis also lacks comparisons with simpler baseline models, leaving it unclear whether the performance gains justify the additional methodological complexity.

Other studies have concentrated on specific biological mechanisms. Chen *et al.* [12] built a necroptosis and immune infiltration-based diagnostic model, while Li *et al.* [13] examined the role of Neutrophil Extracellular Traps (NETs) in coronary atherosclerosis. These works provide rich biological insight but remain constrained by relatively narrow feature sets and limited benchmarking against broader machine learning pipelines.

Several other research groups have applied conventional machine learning pipelines directly. Liu *et al.* [14] used Support Vector Machine (SVM) and LASSO regression to develop diagnostic classifiers and explore potential drug interactions. Patel *et al.* [15] integrated gene expression with clinical data to predict CAD in a multimodal fashion. Although effective at handling high-dimensional data, these pipelines typically use standard feature selection and classification techniques without sufficient emphasis on interpretability, clinical integration, or rigorous baseline comparisons.

While introducing significant contributions in different aspects of this area, the existing approaches often fall short in three important ways. First, many apply high-dimensional gene expression data without robust feature selection, which risks overfitting and hinders interpretability. Second, they frequently fail to incorporate patient demographics and clinical risk factors, leaving models incomplete for real-world application. Third, they often lack rigorous comparisons against simpler baseline

classifiers, which makes it difficult to judge whether performance gains truly justify more complex algorithms.

Our framework explicitly addresses the aforementioned gaps. By integrating both transcriptomic and clinical variables into a unified predictive model, it enhances real-world relevance beyond gene-only approaches or approaches based exclusively on clinical data. While utilizing well-established, mature algorithms, we aim to perform a rigorous integration of data within a leakage-free evaluation framework, combined with systematic confounding analysis, multi-level validation, and clinically interpretable modeling. This design enables more reliable performance estimation, a robust framework, and statistical and biological interpretability.

III. Materials and Methods

The study was conducted in several structured phases: data collection, data cleaning, data visualization, statistical analysis, feature selection, model construction, and evaluation. All experiments were conducted in Python 3.12.12 on Linux-6.6.105-x86_64. Key libraries included NumPy (2.0.2), pandas (2.2.2) and scikit-learn (1.6.1).

A. Data Acquisition and Data Preprocessing

Whole-blood gene expression data and corresponding clinical metadata were obtained from the Gene Expression Omnibus (GEO) under accession number GSE20681. The dataset comprises 198 samples, including: Coronary Artery Disease (CAD) cases and control individuals. Gene expression profiling was performed using the Agilent-014850 Whole Human Genome Microarray platform (GPL4133), which initially included approximately 41,000 probe-level features.

The expression matrix was first cleaned by mapping probe identifiers to gene symbols using the official annotation file for GPL4133. Probes that lacked gene annotations or mapped ambiguously to multiple symbols were discarded. Mean expression values were computed for genes represented by multiple probes to yield a single gene-level profile, resulting in a non-redundant expression matrix with unique gene identifiers across all samples.

Clinical metadata accompanying the gene expression samples included disease status (CAD or control), age, gender, and smoking history. These variables were extracted from the GEO series matrix file, cleaned for consistency, and standardized for downstream integration. Sample identifiers were harmonized to ensure a one-to-one match between expression data and metadata records. Any samples with incomplete metadata or mismatched identifiers were left out to ensure data integrity.

All further data preparation steps, such as scaling, encoding, and feature selection, were fitted to the outer training folds of nested cross-validation to avoid data leakage.

To prevent information leakage between datasets, no joint batch correction or cross-cohort normalization was performed during external validation. Instead, each dataset was processed independently, beginning with platform-specific probe-to-gene annotation, followed by within-dataset standardization. This approach preserves the strict separation of cohorts while enabling a fair assessment of cross-platform generalization. It neither enforces artificial similarity between batches nor ignores systematic shifts between cohorts.

B. Data Visualization and Statistical Analysis

Following data pre-processing and normalization, Exploratory Data Analysis (EDA) was conducted to examine the distribution of clinical variables and global gene expression structure concerning CAD status.

The histogram in Fig. 1 provides an overview of the data distribution with respect to age across the CAD control groups. The plot visually confirms the expected trend that CAD cases are more prevalent in older age brackets.

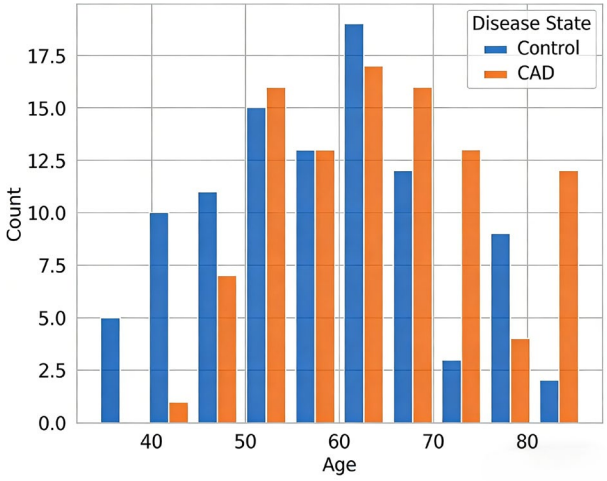

Fig. 1. Age distribution by disease state.

Principal Component Analysis (PCA) was further used to project the normalized expression data into two dimensions (Fig. 2). It shows considerable overlap between CAD and non-CAD groups, with no sharp boundary separating them. Subsequently, t-distributed Stochastic Neighbor Embedding (t-SNE) was applied to analyze the associations (Fig. 3). It reveals subtle clustering patterns, although, again, no distinct separation was evident.

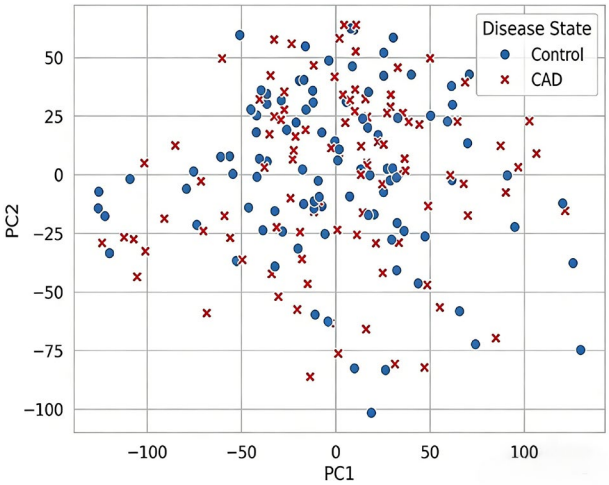

Fig. 2. PCA Plot of gene expression by disease state.

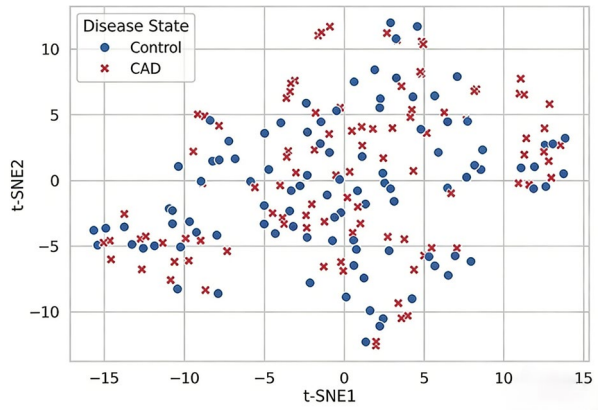

Fig. 3. T-SNE plot of gene expression by disease state.

A correlation matrix (Fig. 4) based on Pearson correlation coefficients describes co-expression relationships using the 8-gene signature. The heatmap indicates moderate pairwise correlations generally, with the strongest positive associations observed between F13B-MMP12 and RRAD-SLC2A10. This result suggests partial co-regulation among subsets of the signature. Several genes in the panel are linked to inflammatory or stress-response biology (e.g., IL18RAP, DUSP1) and extracellular matrix remodeling (e.g., MMP12), providing biological plausibility for their coordinated behavior in a CAD-relevant context.

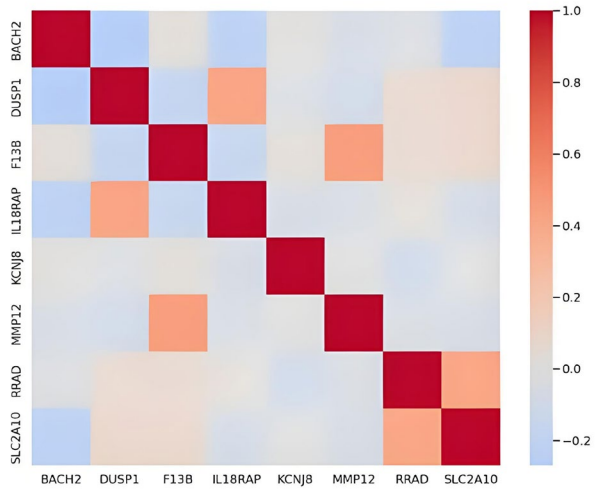

Fig. 4. Correlation matrix of 8-gene signature.

To further guide the understanding of clinical relevance and inform model design, insights on the relations between the clinical data and CAD were investigated through other statistical tests. Differences in age between CAD and control groups were evaluated using Welch's t-test due to unequal variances. Associations between categorical variables (gender and smoking status) and disease status were examined using Fisher's exact test or chi-squared tests, as appropriate, depending on the counts and distribution. These analyses were used to characterize cohort differences and to inform downstream model design, rather than for variable exclusion. The corresponding results are given in Section IV.A.

All gene expression data were z-score normalized prior to visualization and dimensionality reduction.

C. Feature Selection

LASSO was used for feature selection. In the outer cross-validation fold, the feature selection was repeated independently to prevent selection bias.

The selection of genes was tracked by selection frequency across outer folds. The genes chosen in more than 60% of the time were preserved as a stable molecular signature. Consequently, an 8-gene signature was produced via this method (BACH2, DUSP1, F13B, IL18RAP, KCNJ8, MMP12, RRAD, SLC2A10), that represented a balance of predictive capacity and robustness.

Subsequently, three different feature configurations were evaluated: Clinical-only (age, gender, and smoking), Genes-only (stable gene signature), Combined (clinical variables + stable genes). This formulation enabled the individual assessment of each factor, considered separately as clinical or molecular, while also allowing evaluation of their combined contribution to predictive value.

D. Classifiers and Baseline Models

Supervised classification models were developed to predict CAD status using three feature configurations. Logistic Regression (LogReg) was selected as the primary model due to its interpretability and suitability for biomedical inference. Other classifiers, including Support Vector Machines (RBF kernel), Random Forests (RF), k-Nearest Neighbours (kNN), and Gaussian Naïve Bayes (GNB), were evaluated as comparators.

Nested cross-validation was implemented with 10 outer folds and 5 inner folds. All pre-processing steps including scaling, categorical encoding, and gene selection were performed exclusively within training folds. Hyperparameter tuning was performed in the inner cross-validation folds using grid search with a fixed random seed (42). A fixed random seed (42) was used wherever applicable to ensure reproducibility. Hyperparameter grids for LogReg, SVM, RF, and kNN, together with the exact scoring metric used for tuning and selection of the best setting per model, are provided in Appendix A. Default configuration was used for GNB.

Performance metrics were computed using Out-of-Fold (OOF) predictions, from the outer folds exclusively, ensuring unbiased estimation. Reported metrics included AUC, accuracy, precision, recall, F1-Score, and Brier score for calibration assessment. We report 95% Confidence Intervals (CIs) only for AUC and F1. AUC is a threshold-independent measure, whereas F1—Although threshold-dependent—reflects the balance of precision and recall into a single metric. On the other hand, accuracy, precision, and recall are threshold-dependent and sensitive to class prevalence. Adding CIs for these metrics would reduce table readability without providing substantial additional interpretive value over F1. CIs in performance analysis Tables are presented in brackets.

Confounding sensitivity was explicitly evaluated by comparing performance across clinical-only, genes-only, and combined feature sets.

The workflow diagram of the proposed framework is shown in Fig. 5.

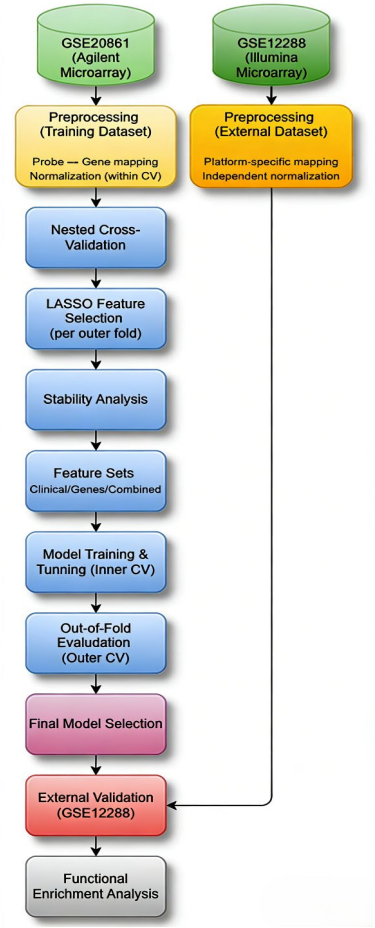

Fig. 5. Analysis pipeline.

IV. Results and Discussion

A. Statistical Grounding

Statistical tests were conducted to evaluate associations between clinical variables and Coronary Artery Disease (CAD) status as well as to characterize cohort differences and to identify a robust gene subset for downstream modeling.

Sample sizes in microarray studies are often relatively small due to limited availability of biological material, as well as the associated experimental complexity and cost. In the present study, the sample size is 198, evenly divided into two groups: a control group and a Coronary Artery Disease (CAD) group.

Among the clinical variables, age was found to have a statistically significant relationship with CAD (Welch's t-test, $p < 0.001$). On the other hand, gender and smoking status were not significantly associated with the disease status (Fisher's exact test $p = 1.0$; x^2 test $p = 0.24$, respectively). Although gender and smoking status were not statistically significant in this cohort, all clinical variables were retained to preserve clinical relevance. Demographic data are shown in Table I. Percentages may not sum to 100% due to missing metadata and rounding.

TABLE I. Demographic Table

Variable	Control (0)	CAD (1)	Test	p-value
Age (years)	57.7 ± 12.1 (median 57)	64.2 ± 10.9 (median 64)	Welch's t-test	<0.001
Gender	Female: 24.2% Male: 75.8%	Female: 24.2% Male: 75.8%	Fisher's exact	1.00
Smoking status	Current: 14.1% Never: 48.5% Quit: 34.3%	Current: 25.3% Never: 47.5% Quit: 25.3%	χ^2 test	0.24

LASSO-based feature selection was applied within each outer cross-validation training fold to identify genes with the strongest predictive signal for CAD and prevent selection bias. The gene selection frequencies over the outer folds were combined to find the most relevant and robust predictors. This process resulted in a gene signature of 8 genes that were stable, consistently selected in more than 60% of the outer folds: BACH2, DUSP1, F13B, IL18RAP, KCNJ8, MMP12, RRAD, and SLC2A10.

These results directly inform the construction of the predictive models, ensuring that included features were statistically grounded. The extracted genes are involved in immune regulation, inflammatory signaling, vascular remodeling, and cardiometabolic processes, which supports biological plausibility and interpretability.

B. Models' Performance

To assess predictive performance, multiple classification algorithms were evaluated, including a linear baseline (Logistic Regression), tree-based ensembles (Random Forest), and non-linear classifiers (SVM, kNN, and Gaussian Naive Bayes). Model performance was compared across clinical-only, genes-only, and combined feature sets using a nested cross-validation framework.

Results of logistic regression models on clinical-only, genes-only, and combined dataset are shown in Table II. Confusion matrix for the LogReg model under the combined features is given in Fig. 6.

TABLE II. Confounding Sensitivity Analysis

Feature Set	Clinical-only	Genes-only	Combined
AUC	0.69 (0.62–0.79)	0.71 (0.61–0.84)	0.76 (0.68–0.87)
Accuracy	0.66	0.66	0.69
Precision	0.65	0.66	0.70
Recall	0.70	0.65	0.66
F1	0.67 (0.60–0.74)	0.65 (0.56–0.75)	0.68 (0.58–0.76)

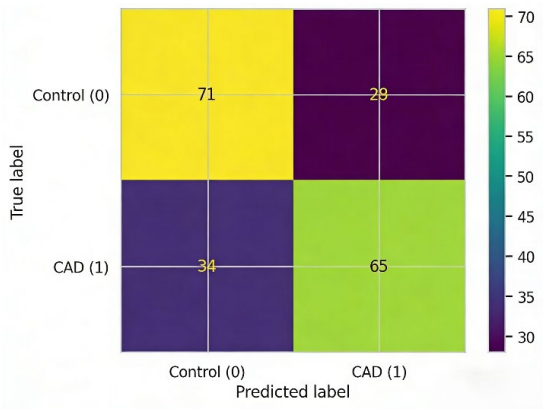

Fig. 6. Confusion matrix for combined data—LogReg model.

The analysis shows that gene expression features provided predictive information comparable to clinical variables, while their integration with clinical metadata yielded the most consistent and robust performance.

Logistic Regression was further compared against non-linear classifiers, including Random Forest, SVM (RBF), kNN, and Gaussian Naive Bayes, using the combined feature set (Table III). While non-linear models achieved comparable performance in selected metrics, Logistic Regression consistently demonstrated equal or superior discrimination (AUC) with more stable performance across cross-validation folds. Its robustness, together with its transparency and suitability for coefficient-based interpretation, motivated its selection as the primary model for downstream biological analysis.

Logistic Regression achieved the strongest and most stable discrimination on the combined feature set and performed competitively on clinical-only data. For genes-only data, Gaussian NB yielded higher AUC, suggesting that non-linear likelihood assumptions may better capture gene-only structure in this cohort. Tree-based and SVM did not provide systematic performance improvements in this cohort. This suggests that, given the moderate sample size and high-dimensional feature space, simpler linear models offer a favorable balance between discrimination, robustness, and interpretability.

Tables II and III report the average performance across the outer cross-validation folds.

TABLE III. COMPARISON OF CLASSIFICATION MODELS

Feature Set	Model	AUC	Accuracy	Recall
Clinical-only	Logistic Regression	0.69 (0.61−0.78)	0.66	0.70
	kNN	0.69 (0.61−0.78)	0.63	0.64
	Random Forest	0.67 (0.57−0.77)	0.64	0.64
	SVM (RBF)	0.59 (0.46−0.72)	0.54	0.55
	Gaussian NB	0.61 (0.54−0.68)	0.51	0.96
Genes-only	Gaussian NB	0.69 (0.60−0.79)	0.63	0.66
	Logistic Regression	0.66 (0.56−0.76)	0.61	0.59
Combined	Logistic Regression	0.76 (0.67−0.86)	0.69	0.62
	Gaussian NB	0.68 (0.58−0.79)	0.55	0.94
	Random Forest	0.65 (0.55−0.74)	0.62	0.61
	SVM (RBF)	0.60 (0.50−0.71)	0.56	0.57

C. Model Interpretation, Calibration, and Feature Importance

To support interpretability and clinical relevance, post-hoc analyses were performed on the final Logistic Regression model trained on the combined feature set. Model interpretation focused on relative feature contributions and probabilistic calibration rather than individual biomarker discovery.

Coefficient inspection and permutation-based feature importance analyses indicated that age was the most influential predictor, consistent with established epidemiological evidence. Several gene expression variables also contributed meaningfully to model predictions, demonstrating that molecular features provide complementary information beyond clinical risk factors. Importantly, predictive performance was driven by the aggregate effect of multiple genes rather than reliance on a single dominant transcript.

Model calibration was assessed using precision–recall analysis and Brier score. The resulting Brier score of 0.2 indicated reasonable agreement between predicted probabilities and observed outcomes, supporting the use of the model for risk ranking rather than binary decision-making.

Together, these analyses demonstrate that the proposed framework achieves a balance between predictive performance, interpretability, and calibration, which is essential for translational applications in cardiovascular risk stratification.

D. External Validation

To evaluate generalizability, the finalized model was tested on an independent external dataset GSE12288 (Affymetrix) [16] processed separately to prevent data leakage. The datasets for internal modeling and for external validation were processed independently: platform-specific probe identifiers were mapped to gene symbols using the corresponding platform annotation files. Probes without valid gene mapping were removed. When multiple probes mapped to the same gene, their expression values were aggregated (mean) to obtain a single gene-level value per sample. Each cohort was then standardized within-cohort prior to inference. After harmonization, only genes present in both platforms were considered eligible for evaluation. The number of retained genes after probe filtering and gene-level aggregation was 19,704 and 13,516 for GSE20681 (GPL4133) and GSE12288 (GPL96), respectively.

Performance metrics for internal evaluation and external validation are summarized in Table IV. The results reflect the differences in cohort composition and platform technology.

On the external cohort profiled on a distinct microarray platform, predictive performance was lower than internal estimates, with a ROC AUC of approximately 0.61.

TABLE IV. EXTERNAL VALIDATION

Dataset	Internal	External
GEO Accession	GSE20681	GSE12288
Platform	GPL4133	GPL96
AUC	0.76 (0.67−0.86)	~0.61 (0.53−0.68)
Accuracy	0.69	~0.54
Precision	0.70	~0.57
Recall	0.66	~0.25

E. Functional Enrichment Analysis

To investigate the biological processes and pathways potentially underlying the identified candidate genes, a functional enrichment analysis was performed on the final gene set (BACH2, DUSP1, F13B, IL18RAP, KCNJ8, MMP12, RRAD, and SLC2A10). The analysis was performed using a widely used Enrichr enrichment tool that tests whether a given gene list shows statistically significant over-representation in curated biological knowledge resources. In our case, we queried three standard reference resources: Reactome and KEGG pathway databases, and a functional ontology—Gene

Ontology Biological Process. Multiple-testing correction was applied using the Benjamini-Hochberg False Discovery Rate (FDR). The top significant pathways for the Gene Ontology Biological Process (2021) collection, Reactome (2022), and KEGG (2021, human) are provided in Appendix B.

Due to limited input size (n = 8), the results were interpreted as hypothesis generating. No term passed the FDR < 0.05 threshold, which is expected under multiple-testing correction for short gene lists. Nevertheless, the top-ranked nominal enrichments were biologically coherent and aligned with established CAD mechanisms, including inflammatory signaling (driven by IL18RAP/IL-18 pathways), coagulation/hemostasis (F13B), extracellular matrix remodeling (MMP12), and vascular/metabolic regulation (SLC2A10, KCNJ8).

Collectively, these themes offer a mechanistic framework for the candidate genes and encourage subsequent validation using larger gene sets, ranked enrichment, and independent cohorts.

V. Conclusions and Future Work

The objective of this study was to develop an efficient and effective model for predictive, biologically interpretable, and clinically actionable CAD diagnostics. Using strict nested cross-validation and confounding sensitivity analysis, we demonstrated that combining transcriptomic and clinical features yields more consistent discrimination than either data source alone, achieving an internal ROC AUC of approximately 0.76. The performed analysis enabled the identification of robust, fold-stable signatures, whereas external validation showed that discriminatory signal was retained beyond the training dataset.

The results demonstrate that the integration of gene expression with granular clinical metadata, and proper processing through a machine learning pipeline, shows strong potential for efficient CAD predictive classification.

The interpretability of the proposed framework enables additional insight into CAD biology. Statistical analysis highlights immune and inflammatory pathways as dominant drivers, giving the model a solid pathophysiological foundation. The analysis also flags ancillary pathways that may indirectly shape disease expression.

Future expansion of the model with high-quality, disease-relevant datasets will enhance precision and reliability, support a focused early-detection gene panel, and inform national health policies. Domain-informed feature selection approaches, such as pathway-guided filtering, will be incorporated to further align statistical findings with biological mechanisms. Although DL approaches are less efficient and require larger cohorts, they will be explored in future work for their potential with non-linear, high-dimensional feature spaces. Finally, to bolster clinician confidence and public transparency, explainable artificial intelligence techniques will be integrated, setting the stage for deployment in precision cardiology settings.

Appendix A. Model Configuration and Hyperparameter Grids

Model	Hyperparameter grids
Logistic Regression	Solver: saga Maximum iterations: 2000 Penalty: l1, l2 Regularization strength (C): [0.1, 1, 10]
SVM (RBF)	Kernel: rbf C: [0.1, 1, 10] Gamma: ["scale", "auto"] Probability estimates enabled
RF	Number of trees: [200, 400] Maximum depth: [5, None] Minimum samples per leaf: [1, 2] Random seed: 42
KNN	Number of neighbors: [3, 5, 11] Weighting: ["uniform", "distance"]

Appendix B. Top Enrchment Tables for GO Biological Process 2021, Reactome 2022, and KEGG 2021 Human

Source	Enriched term	Genes	Adjusted p-value	Main CAD-related theme
GO Biological Process	import into nucleus (GO:0051170)	MMP12, BACH2	0.0192	Nuclear transport / transcriptional regulation
GO Biological Process	vascular transport (GO:0010232)	KCNJ8, SLC2A10	0.0192	Vascular function / transport homeostasis
GO Biological Process	interleukin-18-mediated signaling pathway (GO:0035655)	IL18RAP	0.0192	Inflammation / immune signaling
GO Biological Process	membrane repolarization during ventricular cardiac muscle cell action potential (GO:0098915)	KCNJ8	0.0192	Cardiac electrophysiology
Reactome 2022	Interleukin-18 Signaling (R-HSA-9012546)	IL18RAP	0.0649	Inflammation / plaque activity
Reactome 2022	Common Pathway of Fibrin Clot Formation	F13B	0.0649	Coagulation / thrombosis
Reactome 2022	Collagen Degradation	MMP12	0.0649	Extracellular matrix remodeling / plaque instability
Reactome 2022	Cellular Hexose Transport	SLC2A10	0.0649	Metabolic / endothelial stress
KEGG 2021 Human	Complement and coagulation cascades	F13B	0.1082	Thrombotic processes
KEGG 2021 Human	Fluid shear stress and atherosclerosis	DUSP1	0.1082	Atherosclerosis / vascular stress
KEGG 2021 Human	Cytokine-cytokine receptor interaction	IL18RAP	0.1121	Immune-inflammatory signaling
KEGG 2021 Human	MAPK signaling pathway	DUSP1	0.1121	Stress-response / inflammatory regulation

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Bilgin Demir conducted the main part of the research; Zhilbert Tafa provided inputs on technical aspects, analyzed the data and the results, and substantially participated in writing and revising the manuscript; Djansel Bukovec provided inputs on biological aspects, with a focus on biological statistics and interpretation; all authors had approved the final version.

REFERENCES

[1] T. Vos *et al*., "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019," *The Lancet*, vol. 396, no. 10258, pp. 1204–1222, 2020.

[2] M. A. Gimbrone and G. G. Cardeña, "Endothelial cell dysfunction and the pathobiology of atherosclerosis," *Circ. Res.*, vol. 118, no. 4, pp. 620–636, 2016.

[3] P. Libby, "Inflammation in atherosclerosis—No longer a theory," *Clin. Chem*., vol. 67, no. 1, pp. 131–142, 2021.

[4] X. Chang *et al.* "Identification of hub biomarkers in coronary artery disease patients using machine learning and bioinformatic analyses" *Sci. Rep*., vol. 15, no. 17244, 2025.

[5] M. Sarma and S. Chatterjee, "Machine learning' multiclassification for stage diagnosis of Alzheimer's disease utilizing augmented blood gene expression and feature fusion," *Discov. Appl.* Sci 7, 2025.

[6] K. M. Verrou *et al.*, "Machine learning-based identification of a transcriptomic blood signature discriminating between systemic autoimmunity and infection," *Med,* vol. 6, no. 11, 2025.

[7] Z. Wang, Y. Fu, H. Zhang, N. Liu, and N. Lei, "Diagnostic potential of the B9D2 gene in colorectal cancer based on whole blood gene expression data and machine learning," *Discov Oncol.,* vol. 16, no. 1554, 2025.

[8] A. Q. Nawabi and L. Chen, "Shared genetic characteristics of coronary artery disease and peripheral artery disease: Insights from integrated bioinformatics analysis of RNA-sequencing data," *Bioinformatics and Biology Insights*, vol. 19, pp. 1–14, 2025.

[9] GSE20681: Whole blood gene expression profiling of coronary artery disease patients and controls. (2010). [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20681

[10] Y. Shi, S. Yang, M. Luo, W. D. Zhang, and Z. P. Ke, "Systematic analysis of coronary artery disease datasets revealed the potential biomarker and treatment target," *Oncotarget*, vol. 8, no. 33, pp. 54583–54591, 2017.

[11] Y. Yang and X. Xu, "Identification of key genes in coronary artery disease: an integrative approach based on weighted gene co-expression network analysis and their correlation with immune infiltration," *Aging*, vol. 13, no. 6, pp. 8306–8323, 2021.

[12] Q. Chen, J.-L. Zhang, J.-S. Yang, Q. Jin, J. Yang, Q. Xue, and X. Guang, "Novel diagnostic biomarkers related to necroptosis and immune infiltration in coronary heart disease," *J. Inflamm. Res.,* vol. 17, pp. 4525–4548, 2024.

[13] Z. Li, W. Zhao, W. Ji, Z. Li, K. Wang, and T. Jiang, "A neutrophil extracellular traps-related gene trait revealed the prospective therapy strategy of coronary atherosclerosis," *J. Inflamm. Res.*, vol. 17, pp. 9925–9951, 2024.

[14] C. Liu, J. Liu, Y. Zhang, X. Wang, and Y. Guan, "Immune-related potential biomarkers and therapeutic targets in coronary artery disease," *Front. Cardiovasc. Med.*, vol. 9, 1055422, 2023.

[15] S. Patel, M. Singh, K. Sharma, and D. R. Jones, "Multi-modal prediction of coronary artery disease using expression and clinical data," *J. Biomed. Inform.*, vol. 130, 104082, 2022.

[16] GSE12288: Whole blood gene expression profiling of coronary artery disease patients and controls. (2009). [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12288

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).