Noname manuscript No.
(will be inserted by the editor)

Unlocking the Potential of microRNAs: Machine Learning Identifies Key Biomarkers for Myocardial Infarction Diagnosis

Mehrdad Samadishadlou · Reza Rahbarghazi · Zeynab Piryaei · Mahdad Esmaeili · Çığır Biray Avcı · Farhad Bani · Kaveh Kavousi ·

Received: date / Accepted: date

Abstract MicroRNAs (miRNAs) play a crucial role in regulating adaptive and maladaptive responses in cardiovascular diseases, making them attractive targets for potential biomarkers. However, their potential as novel biomarkers for diagnosing cardiovascular diseases requires systematic evaluation. In this study, we aimed to identify a key set of miRNA biomarkers using integrated bioinformatics and machine learning analysis. We combined and analyzed three

Mehrdad Samadishadlou
Medical Nanotechnology gruop, Tabriz University of Medical Sciences, Tabriz, Iran

Reza Rahbarghazi
Medical Nanotechnology gruop, Tabriz University of Medical Sciences, Tabriz, Iran

Zeynab Piryaei
Laboratory of Complex Biological Systems and Bioinformatics (CBB), Department of Bioinformatics, Institute of Biochemistry and Biophysics (IBB), University of Tehran, Iran.

Mahdad Esmaeili
Medical Bioengineering Department, School of Advanced Medical Sciences, Tabriz University of Medical Sciences, Tabriz, Iran

Çığır Biray Avcı
Medical Biology Department, School of Medicine, Ege University, İzmir, Turkey

Farhad Bani
Medical Nanotechnology gruop, Tabriz University of Medical Sciences, Tabriz, Iran
E-mail: bainf@tbzmed.ac.ir

Kaveh Kavousi
Laboratory of Complex Biological Systems and Bioinformatics (CBB), Department of Bioinformatics, Institute of Biochemistry and Biophysics (IBB), University of Tehran, Iran
E-mail: kkavousi@ut.ac.ir

gene expression datasets from the Gene Expression Omnibus (GEO) database, which contains peripheral blood mononuclear cell (PBMC) samples from individuals with myocardial infarction (MI), stable coronary artery disease (CAD), and healthy individuals. Additionally, we selected a set of miRNAs based on their area under the receiver operating characteristic curve (AUC-ROC) for separating the CAD and MI samples. We designed a two-layer architecture for sample classification, in which the first layer isolates healthy samples from unhealthy samples, and the second layer classifies stable CAD and MI samples. We trained different machine learning models using both biomarker sets and evaluated their performance on a test set. We identified miR-21, miR-186, and miR-32 as the only miRNAs among the differentially expressed genes, and a set including miR-186, miR-21, miR-197, miR-29A, and miR-296 as the optimum set of miRNAs selected by their AUC-ROC. Both biomarker sets could distinguish healthy from not-healthy samples with complete accuracy. The best performance for the classification of CAD and MI was achieved with an SVM model trained using the biomarker set selected by AUC-ROC, with an AUC-ROC of 0.96 and an accuracy of 0.94 on the test data. Our study demonstrated that miRNA signatures derived from PBMCs could serve as valuable novel biomarkers for cardiovascular diseases.

Keywords microRNA · Machine Learning · Myocardial Infarction · Bioinformatics · Biomarker ·

1 Introduction

Cardiovascular diseases (CVDs) are the leading cause of human mortality, accounting for 32% of all global deaths. It is estimated that approximately 85% of CVD mortality is due to myocardial infarction (MI) ("Cardiovascular Diseases (CVDs)" n.d.). MI is an acute coronary syndrome characterized by sudden blockage and stenosis of the coronary artery and subsequent myocardial ischemia, leading to extensive cardiomyocyte damage and necrosis (Yap et al. 2023).

Over the last 50 years, numerous attempts have been made to use biomarkers to facilitate diagnosis, assess the risk, follow-up therapy, and determine therapeutic efficacy in CVD candidates. Based on released guidelines, cardiac troponins (cTns) are used as a highly sensitive and accurate approach for detecting MI. Despite these inherent advantages, the high sensitivity of cTn-based assays has also led to more false-positive results (Thygesen et al. 2018), necessitating the advent and development of new modalities with pathological value. To improve the diagnostic value of existing MI biomarkers, a combination of complementary biological markers, such as microRNAs (miRNAs) and other genetic factors, has been proposed. Previous research supports the notion that miRNAs exhibit great potential as alternative biomarkers for CVD detection and follow-up (Schulte et al. 2020). It has been suggested that miRNAs possess 18-22 nucleotides and play a crucial role in the regulation of gene

expression. Evidence indicates that miRNAs are involved in the pathogenesis of cardiac tissue injury (Schulte, Karakas, and Zeller 2017). Several biological processes, such as angiogenesis, cardiomyocyte growth and contractility, lipid metabolism, plaque formation, and cardiac rhythm, are regulated by miRNAs (Kalayinia et al. 2021). These elements can easily circulate in biofluids and could be considered theranostic targets in terms of CVDs (Schulte, Karakas, and Zeller 2017). It has been postulated that the function and diagnostic properties of miRNAs are beyond the myocardium in patients with CVD. Specifically, the expression of miRNAs can vary in different biofluids and cell components such as serum and peripheral blood mononuclear cells (PBMCs) (Soler-Botija, Gálvez-Montón, and Bayés-Genís 2019).

PBMCs are a fraction of white blood cells, including monocytes, lymphocytes, macrophages, and other cells of the immune system (Gao et al. 2020). Emerging data indicate that PBMCs can be used as a valid source of biomarkers for monitoring various pathological conditions. Of note, the alteration of mRNAs and miRNAs under pathological conditions provides valuable information about different kinds of disorders. PBMCs can recapitulate the conditions of target tissues, thus providing a highly sensitive and specific source of biomarkers (Mosallaei et al. 2022). Combined with these conditions, these cells are repositories of dysregulated genes and miRNA expression profiles in CVDs (Gao et al. 2020; Mosallaei et al. 2022).

In recent years, the advent and application of machine learning (ML) has been an exciting prospect for advancing scientific research. Although the concept of ML and its initial algorithms were conceived many years ago, recent improvements in computing power and access to vast amounts of data have demonstrated that ML techniques outperform classical statistical methods in various fields. Furthermore, the progress made in omics technologies has enabled the analysis of massive and intricate biological datasets, consisting of hundreds to thousands of samples, which makes it possible for ML to extract valuable biological insights and information from such data (Torun et al. 2023). Consequently, ML provides innovative methods for merging and interpreting diverse types of omics data, leading to the identification of new biomarkers. These biomarkers can aid in precise disease prediction, patient stratification, and the development of novel therapeutic approaches (Reel et al. 2021).

In this study, we aimed to identify potential miRNA biomarkers in patients with MI by combining and analyzing three different microarray datasets from PBMCs. The integration of omics data with bioinformatics and ML techniques could be a promising tool in the discovery of new and more accurate biomarkers for monitoring MI. Additionally, this approach can deepen our understanding of the underlying mechanisms of MI and aid in the development of valid diagnostic biomarkers and patient stratification.

2 Materials and Methods

2.1 Microarray data collection

Microarray datasets were obtained from the Gene Expression Omnibus (GEO) database (`https://www.ncbi.nlm.nih.gov/geo/`). To obtain sufficient classification power between MI, healthy, and CAD samples, a relatively large sample size is needed Therefore, GSE59867 for the MI and CAD samples and GSE56609 and GSE54475 for the healthy samples were selected. All samples were produced using Affymetrix Human Gene 1.0 ST Array platform (GPL6244). Only healthy, CAD, and early-stage MI samples were selected from these datasets for further analysis. The basic information for the three datasets evaluated in this study is provided in Table @ref(tab:datasets). Bioinformatics analyses were conducted using R, ver. 4.2.0 (R Core Team 2022), RStudio (RStudio Team 2020). All plots and graphics of these sections were created using the ggplot2 R package (Wickham 2016).

Table 1: Sample information on the GEO microarray datasets.

Dataset	Platform	Healthy	CAD	MI	Refrence
GSE59867	GPL6244	-	46	111	(Maciejak et al. 2015)
GSE56609	GPL6244	46	-	-	(Matone et al. 2015)
GSE54475	GPL6244	5	-	-	(Canali et al. 2014)

2.2 Preprocessing

The raw data in the form of CEL files from all datasets were obtained from GEO. To prepare the data for analysis, we utilized the fRMA package (M. N. McCall, Bolstad, and Irizarry 2010) to facilitate preprocessing of individual microarray samples and their consistent combination. For each dataset, background correction was applied using the RMA algorithm, followed by quantile normalization based on the reference distribution. To account for probe-specific effects, batch effects were eliminated during summarization and gene expression variances were estimated accordingly. In cases where multiple probe sets matched the same gene, the mean log-fold change was retained. Consequently, fRMA can serve as a technique to remove batch effects across diverse datasets generated by identical microarray platforms (Lazar et al. 2013). To ensure the effectiveness of the batch effect removal, we employed principal component analysis (PCA) and relative log expression (RLE) plots to visualize the data before and after applying fRMA.

2.3 Differential expression analysis

The barcode algorithm was introduced by McCall et al. (Matthew N. McCall et al. 2011), aimed to convert actual expression values into binary barcode values. Extensive sample collections were gathered and normalization was performed using fRMA across multiple platforms, including the Affymetrix Human Gene 1.0 ST Array (GPL6244) platform. By utilizing these normalized datasets, the distribution of the observed intensities for both the expressed and unexpressed genes was estimated. The determination of whether a gene was expressed or not was based on the following equation, where a value of 1 indicates expression and a value of 0 indicates non-expression:

$$\hat{x}_{ij} = \begin{cases} 1 \text{ if } x_{ij} >= \mu^{ne} + C \times \sigma^{ne} \\ 0 \text{ otherwise} \end{cases}$$

In the barcode algorithm, the normalized intensity of gene i in sample j is denoted as x_{ij}. A user-defined parameter, C, was introduced along with the standard deviation (σ^{ne}) and mean (μ^{ne}) of the non-expressed distribution. Based on these values, the barcode representation of a sample was generated as a vector consisting of ones and zeros, representing the estimated expression (ones) and non-expression (zeros) of each gene. The barcode function within the R fRMA package was employed to implement the barcode algorithm, utilizing the default value of C.

To assess the differences in expressed ratios between the MI and healthy control groups, Fisher's exact test was performed on the barcode values of individual genes. Genes that exhibited a false discovery rate (FDR) below 0.05, calculated using the Benjamini-Hochberg procedure to account for multiple testing issues were identified as differentially expressed genes (DEGs). The same procedures were applied to the CAD versus healthy control comparison, as well as to the MI versus CAD group, to identify DEGs specific to each comparison.

2.4 Functional and pathway enrichment analyses

The R clusterProfiler package (Yu et al. 2012) was utilized to perform Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis and Gene Ontology (GO) functional annotation on the set of DEGs. GO analysis encompassed three categories: biological process (BP), cellular component (CC), and molecular function (MF). For statistical significance, an adjusted p-value threshold of less than 0.05 was employed. Enrichment analyses were conducted separately for DEGs specific to the MI-healthy and CAD-healthy comparisons. All the default parameters provided by the package were used in the analyses.

2.5 ML procedure

ML analysis was performed using Python software, ver. 3.9, Numpy (Harris et al. 2020), Pandas (McKinney 2010), and Scikit-Learn packages (Pedregosa et al. 2011). Whenever hypertuning was needed, the Scikit-opt package (Head et al. 2021) was used. In all ML analyses, the datasets were divided into training and test sets at a 0.7:0.3 ratio, and all reported results are the average of 10-fold cross-validation.

Two different approaches were used to select miRNAs for model training. The first approach was to use differentially expressed miRNAs. In the second approach, miRNAs with individual AUC-ROCs over 0.8 for separating MI from CAD were selected. The results of these two approaches can provide an informative comparison between the predictive capabilities of sets of miRNAs selected with different logics.

2.5.1 miRNAs in DEGs

In this approach, a two-layer architecture is deployed to the data to maximize the prediction values. The first layer predicted whether a sample was healthy or not, and the second layer separated MI from CAD in the samples that were predicted as not healthy in the first layer. To this end, a distinct ML model was trained for each layer. Because, there were a limited number of miRNAs in the DEGs, both layers were trained with all of them. For further comparison with the models' performance, the ROC curve of each miRNA for classifying healthy and not-healthy, as well as CAD and MI, was generated using a logistic regression model.

First layer for the isolation of healthy and not-healthy samples: A support vector machine (SVM) model using RBF kernels was trained and hypertuned using all miRNAs in the DEGs. To handle the severe imbalance in the number of samples (51 for the healthy group and 157 for the not-healthy group), the sample weights for the healthy and the not-healthy samples were set to 1 and 0.5, respectively. The ROC curve and confusion matrix for the model are reported.

Second layer for separating the MI and CAD samples: Different models were investigated to achieve the highest classification performance. To do so, SVM (with linear, polynomial, and RBF kernels), logistic regression (LR), random forests (RF), k-nearest neighbor (kNN), gradient boosting (GB), XGBoost (XGB) and decision tree (DT) models were trained. All models were trained with their preset parameters using 10-fold cross-validation. The criteria for selecting the best model were the highest accuracy and AUC-ROC for the test set. The best model was hypertuned using the scikit-opt package (Head et al. 2021) for the best classification performance. The ROC curve and confusion matrix for the best model are reported.

2.5.2 miRNAs with the highest AUC-ROC

As in the previous approach, a two-layer strategy was employed. The first layer classified samples into healthy and not-healthy, and the second layer separated the MI and CAD samples. However, to keep the number of miRNAs as low as possible, miRNAs were selected from the second layer and their performance was evaluated in the first layer. The AUC-ROC of all miRNAs for classifying MI and CAD samples was calculated, and miRNAs with AUC-ROC > 0.8 were selected. ROC curves for each selected miRNA for separating healthy samples from not-healthy samples and MI from CAD samples were also plotted for further comparison.

First layer for the isolation of healthy and not-healthy samples: An SVM model with an RBF kernel is trained using the selected set of miRNAs. Additionally, the model was hypertuned to find the hyperparameters for the highest AUC-ROC and accuracy. The same sample weights as in the previous approach (1 for healthy and 0.5 for not-healthy samples) were used. The ROC curve and confusion matrix for the model were reported.

Second layer for separating the MI and CAD samples: The selected miRNA set was used to train different algorithms to determine the best model. Similar to the previous approach, the SVM (with linear, polynomial, and RBF kernels), LR, RF, kNN, GB, XGB, and DT models were trained. All models were trained with their preset parameters using 10-fold cross-validation. The models with the highest AUC-ROC and accuracy on the test set were selected and hypertuned using the scikit-opt package (Head et al. 2021). The ROC curve and confusion matrix for the best model were reported.

3 Results

3.1 Preprocessing

The PCA plots of the samples are shown in Figures @ref(fig:PCA)A and B. Healthy samples were separated from the CAD or MI samples in the primary data and after conducting fRMA. In the RLE plot, there was a distinct difference between the dataset means for all samples before fRMA was performed (Figure @ref(fig:PCA)C). All datasets were rearranged to approximately 0 in the RLE plot after fRMA was conducted (Figure @ref(fig:PCA)D). Moreover, there was an apparent change in the interquantile distances, but the values were still greater than 0.1.

3.2 Differential expression analysis

According to the cutoff criterion of $FDR < 0.05$, there were 860 DEGs between MI and healthy samples. Among them, 323 were up-regulated, and 537 were

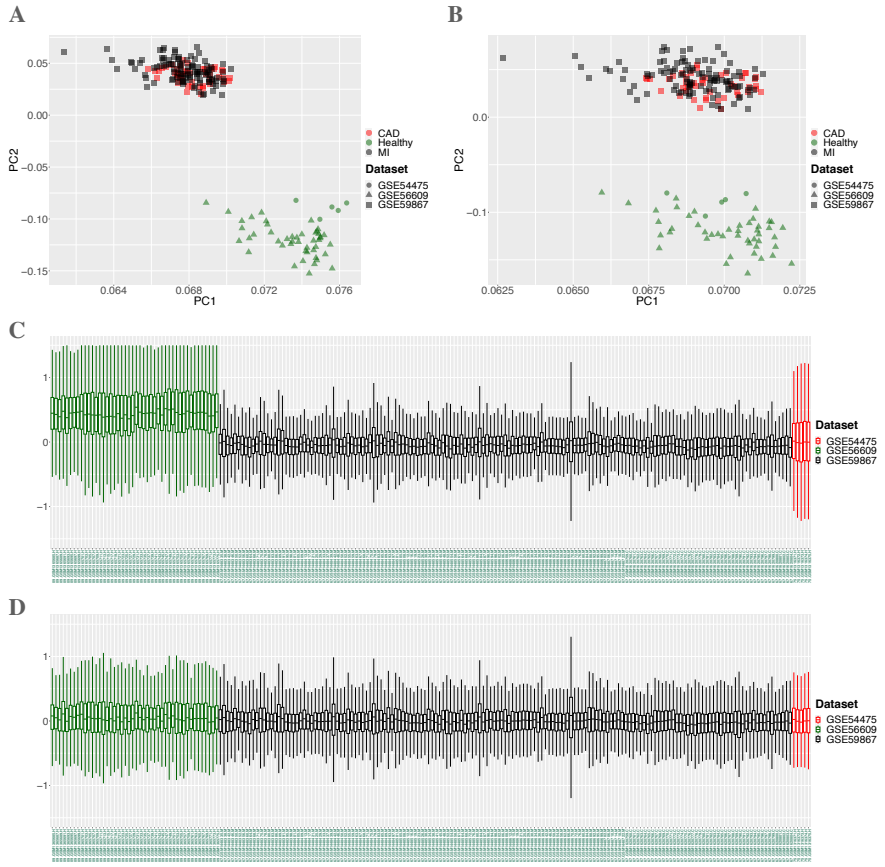

Fig. 1 Principal component analysis plots for (A) primary data and (B) the data after fRMA, and the relative log expression plots for (C) primary data and (D) the data after fRMA.

Table 2 Total, up-, and down-regulated DEGs and differentially expressed miRNAs.

	Total DEGs	Up-regulated DEGs	Down-regulated DEGs	miRNAs
MI vs. Healthy	860	323	537	hsa-miR-186, miR-21, miR-32
CAD vs. Healthy	670	262	408	hsa-miR-186, miR-21, miR-32
MI vs. CAD	260	144	116	hsa-miR-186

down-regulated in the MI group compared to the healthy group. In the CAD and healthy group comparison, we found 670 DEGs, of which 262 and 408 DEGs were up- and down-regulated, respectively, in CAD samples. In the MI and CAD groups, the number of DEGs was 260, and the numbers of up- and down-regulated genes in MI samples were 144 and 116, respectively, compared to CAD samples. The data are summarized in Table @ref(tab:DEGstab).

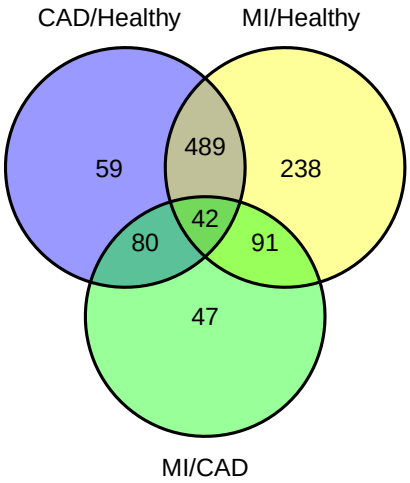

Fig. 2 Venn diagram for DEGs in CAD/Healthy, MI/Healthy, and MI/CAD.

The Venn diagram in Figure @ref(fig:venn) shows that the CAD and MI samples shared most of their DEGs. From 860 DEGs of MI/healthy and 670 DEGs of CAD/healthy, 531 genes were common, which is 62% of MI/healthy DEGs and 79% of CAD/healthy DEGs.

3.3 GO and KEGG enrichment analyses of the DEGs

To explore the biological classification of the DEGs, we performed GO and KEGG pathway enrichment analyses on the MI/healthy and CAD/healthy DEGs. For MI/healthy, GO enrichment analysis in the BP category suggested that the DEGs were enriched in "immune response-regulating signaling pathway," "lymphocyte differentiation," "immune response-regulating cell surface receptor signaling pathway," and "leukocyte activation involved in immune response" (Figure @ref(fig:MIHEnrich)A). In the CC category, DEGs were enriched in "secretory granule membrane," "azurophil granule," "ficolin-1-rich granule," "tertiary granule," and "ficolin-1-rich granule membrane" (Figure @ref(fig:MIHEnrich)B). In the MF category, DEGs were involved in "cadherin binding" and "MHC class I protein binding" (Figure @ref(fig:MIHEnrich)C). KEGG pathway analysis indicated that the DEGs were related to the following pathways: "Chemokine signaling pathway," "Lipid and atherosclerosis," and "Hematopoietic cell lineage" (Figure @ref(fig:MIHEnrich)D).

The enrichment results for the CAD/healthy DEGs were as follows. In the BP category, GO enrichment suggested that the DEGs were enriched in "positive regulation of defense response," "positive regulation of innate immune response," "mononuclear cell differentiation," and "positive regulation of response to external stimulus" (Figure @ref(fig:CADHEnrich)A). In the CC category, DEGs were enriched in "azurophil granule," "ficolin-1-rich granule,"

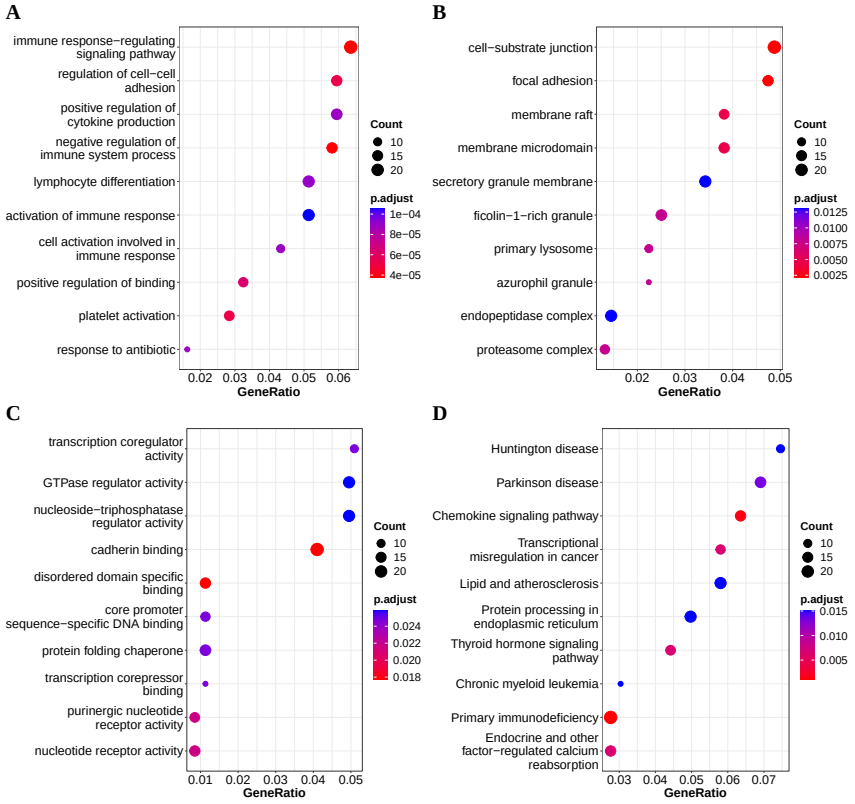

Fig. 3 Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enriched with the MI and healthy DEGs. (A) Biological process terms. (B) Cellular component terms. (C) Molecular function terms. (D) KEGG analysis.

and "ficolin-1-rich granule membrane" (Figure @ref(fig:CADHEnrich)B). In the MF category, DEGs were involved in "lipoprotein particle receptor binding" and "NF-κB binding" (Figure @ref(fig:CADHEnrich)C). KEGG pathway analysis showed that the DEGs were related to the following pathways: "Chemokine signaling pathway," "Lipid and atherosclerosis," and "Hematopoietic cell lineage" (Figure @ref(fig:CADHEnrich)D).

3.4 Machine Learning

3.4.1 miRNAs in DEGs

Among the DEGs, miR-186, miR-32, and miR-21 were identified as differentially expressed miRNAs. The expression profiles of the three miRNAs are shown in Figure @ref(fig:Expall). The ROC curves of each miRNA in each layer are presented in Figure @ref(fig:miRROC). Using the logistic regression

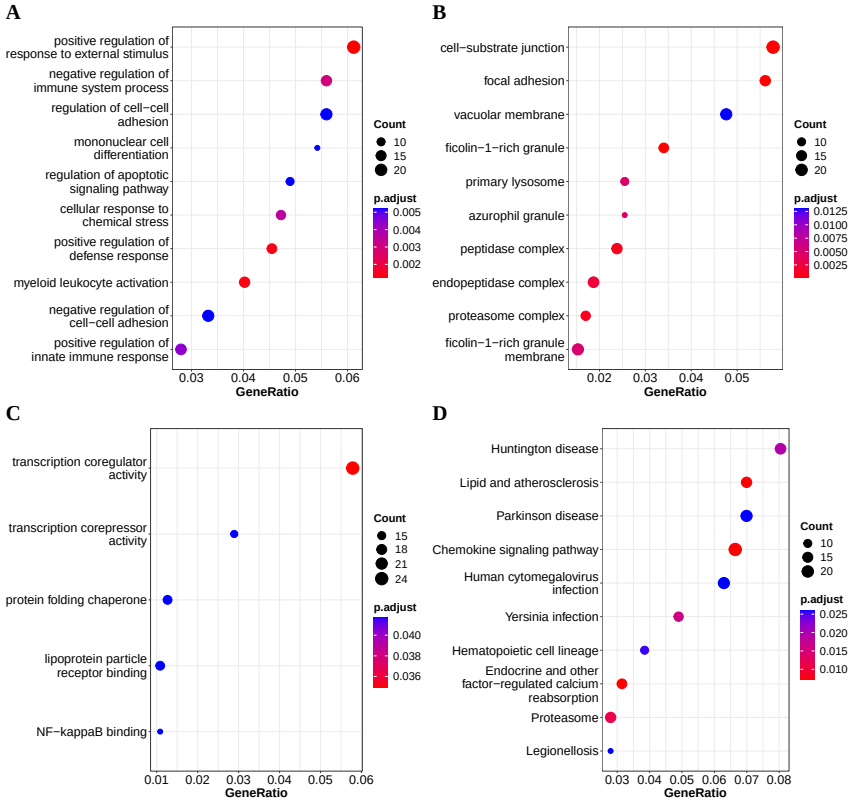

Fig. 4 Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enriched with the CAD and healthy DEGs. (A) Biological process terms. (B) Cellular component terms. (C) Molecular function terms. (D) KEGG analysis.

model, the AUC-ROC values of miR-21, miR-32, and miR-186 for separating healthy and not-healthy samples were 0.98, 0.99, and 0.90, respectively (Figure @ref(fig:miRROC)A). The accuracy of each miRNA for classifying the samples into healthy and not-healthy groups on the test set for miR-21, miR-32, and miR-186 was 0.92, 0.98, and 0.89, respectively. The ROC curve of each miRNA for classifying MI and CAD samples is presented in Figure @ref(fig:miRROC)B. The AUC-ROC and accuracy for miR-21, miR-32, and miR-186 in the test set were 0.85; 0.70; and 0.86, and 0.78; 0.67; and 0.74, respectively.

First layer for the isolation of healthy and not-healthy samples: Although single miRNAs had an acceptable performance for this layer, their predictive value could be further improved by using them as a set. The ROC curve for the SVM model with an RBF kernel trained with all three miRNAs is presented in Figure @ref(fig:DEMROC)A. The model had a better performance in classification than single miRNAs.The AUC-ROC for the model was 1, and

Table 3 Investigated miRNAs log fold-change and adjusted p-values for CAD samples relative to healthy, MI samples relative to healthy, and MI samples relative to CAD.

	CAD/Healthy		MI/Healthy		MI/CAD	
	logFC	adj. p-value	logFC	adj. p-value	logFC	adj. p-value
miR-186	1.4	3.60e-25	0.9	6.76e-20	-0.5	1.05e-09
miR-21	1.4	1.31e-17	2.3	2.07e-47	0.8	2.96e-11
miR-32	2.5	8.39e-43	2.2	3.10e-59	-0.3	7.60e-04
miR-197	0.5	2.95e-20	0.7	1.59e-47	0.2	8.58e-09
miR-29A	0.7	7.76e-29	0.1	1.70e-01	-0.5	2.14e-10
miR-296	-0.1	5.00e-02	0.1	2.00e-02	0.2	6.15e-06

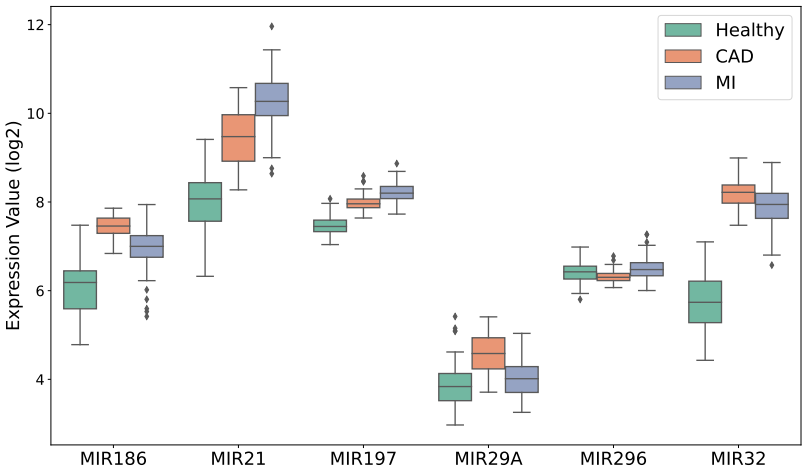

Fig. 5 Expression profile of all miRNAs in two approaches in different sample classes.

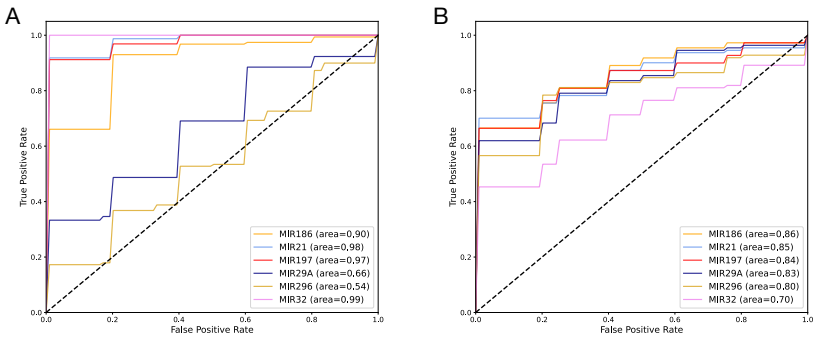

Fig. 6 ROC curve for single miRNAs on test set classification for (A) healthy and not-healthy samples and (B) CAD and MI samples.

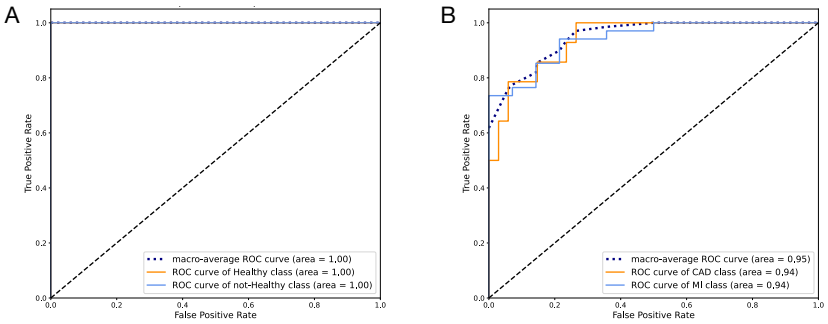

Fig. 7 ROC curve for the model trained with miRNAs in DEGs on test set classification; (A) An SVM model with RBF kernel for healthy and not-healthy and (B) An SVM model with linear kernel for CAD and MI sample classification.

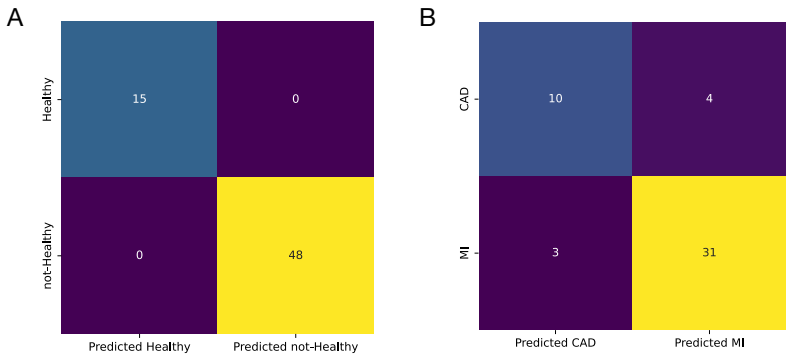

Fig. 8 Confusion matrix for the model trained with miRNAs in DEGs on test set classification; (A) An SVM model with RBF kernel for healthy and not-healthy and (B) An SVM model with linear kernel for CAD and MI sample classification.

its accuracy on the test set was also 1. In Figure @ref(fig:DEMCF)A, the confusion matrix for the model is presented.

Second layer for separating the MI and CAD samples: Different models were trained using the expression values of three differentially expressed miRNAs. The models' AUC-ROC and the accuracy of the test set are shown in Figure @ref(fig:DEMmodels). The best model from both the AUC-ROC and accuracy points of view was the SVM model with a linear kernel. The AUC-ROC and accuracy for this model with its preset values were 0.93 and 0.82, respectively. The model was hypertuned for C and gamma hyperparameters, and therefore the model showed better performance. The ROC curve of the hypertuned model is presented in Figure @ref(fig:DEMROC)B. For this model, the AUC-ROC reached 0.95, and the accuracy was improved to 0.85 (Table @ref(tab:DEGsML)). Moreover, the sensitivity and specificity for the model

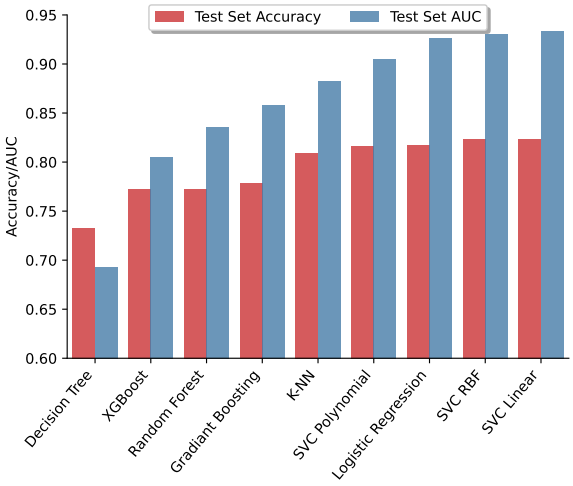

Fig. 9 Area under the receiver operating characteristic curve and accuracy of different models trained with three miRNAs in DEGs.

Table 4 AUC-ROC and accuracy for SVM with a linear kernel as the best model trained with differentially expressed miRNAs on the training and test sets before and after hypertuning

		Preset parameters		Hypertuned	
Model	Metrics	train	test	train	test
SVM-linear	AUC-ROC	0.91	0.93	0.92	0.95
	Accuracy	0.83	0.82	0.84	0.85

on the test set were 0.91 and 0.71, respectively. The confusion matrix for the hypertuned model is illustrated in Figure @ref(fig:DEMCF)B.

3.4.2 AUC-ROC approach

After calculating the AUC-ROC for each miRNA to classify of MI and CAD samples, the miRNAs with AUC-ROC > 0.8 were selected. The miRNAs selected were miR-29a, miR-197, miR-186, miR-21, and miR-296. The expression levels of these miRNAs in healthy, CAD, and MI samples are presented in Figure @ref(fig:Expall). The ROC curves of the selected miRNAs in both layers are shown in Figure @ref(fig:miRROC).

First layer for the isolation of healthy and not-healthy samples: Using the selected set, an SVM model with an RBF kernel was trained to separate healthy and not-healthy samples. The ROC curve for the model is presented in Figure @ref(fig:AUCROC)A, and the confusion matrix is illustrated in Figure @ref(fig:AUCCF)A. Both the AUC-ROC and accuracy of the model on the test set were 1.

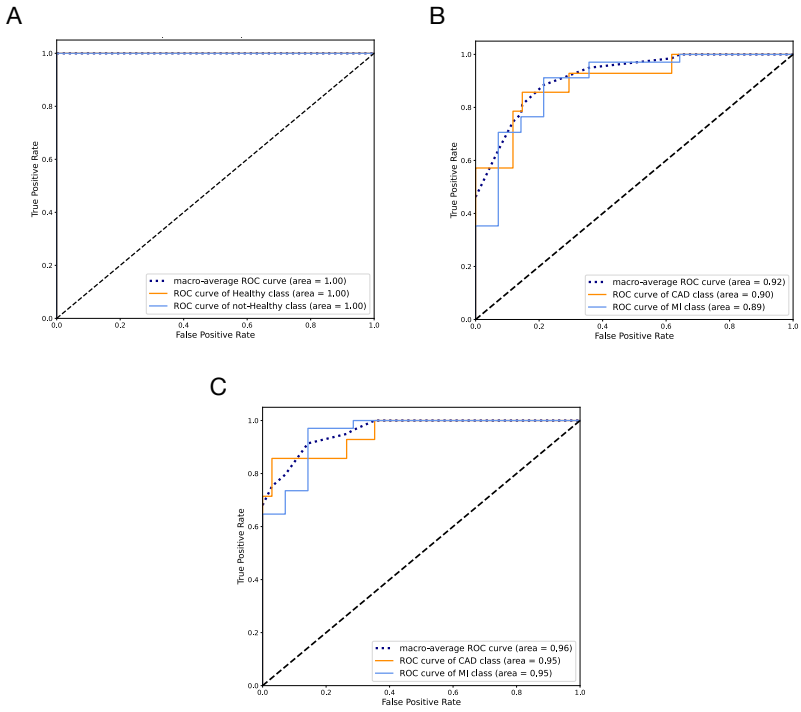

Fig. 10 ROC curve for models trained with the set of miRNAs selected by AUC-ROC on test set classification; (A) SVM with RBF kernel for healthy and not-healthy samples classification. (B) Logistic regression model for CAD and MI sample classification. (C) SVM with polynomial kernel for CAD and MI sample classification.

Second layer for separating the MI and CAD samples: To find the best model for this set of miRNAs, different models were trained using their preset values. The AUC-ROC and accuracy results for the test set are presented in Figure @ref(fig:AUCmodels). The best model from the AUC-ROC point of view was the SVM with a linear kernel, and from the accuracy point of view, it was the SVM model with an RBF kernel. For the SVM-linear model, the AUC-ROC and accuracy were 0.93 and 0.82, respectively; and for the SVM-RBF, the values were 0.92 and 0.84, respectively. Both models were hyper-tuned, and the ROC curve for their best performance is presented in Figure @ref(fig:AUCROC)B and C. The AUC-ROC and accuracy for the SVM-linear model were modified to 0.92 and 0.88, respectively. For the SVM-RBF, these values increased to 0.96 and 0.94, respectively (Table @ref(tab:AUCML)). The sensitivities for the SVM-linear and SVM-RBF models were 0.91 and 0.97, respectively; and their specificities were 0.79 and 0.86, respectively. The confusion matrix for both models is illustrated in Figure @ref(fig:AUCCF)B and C.

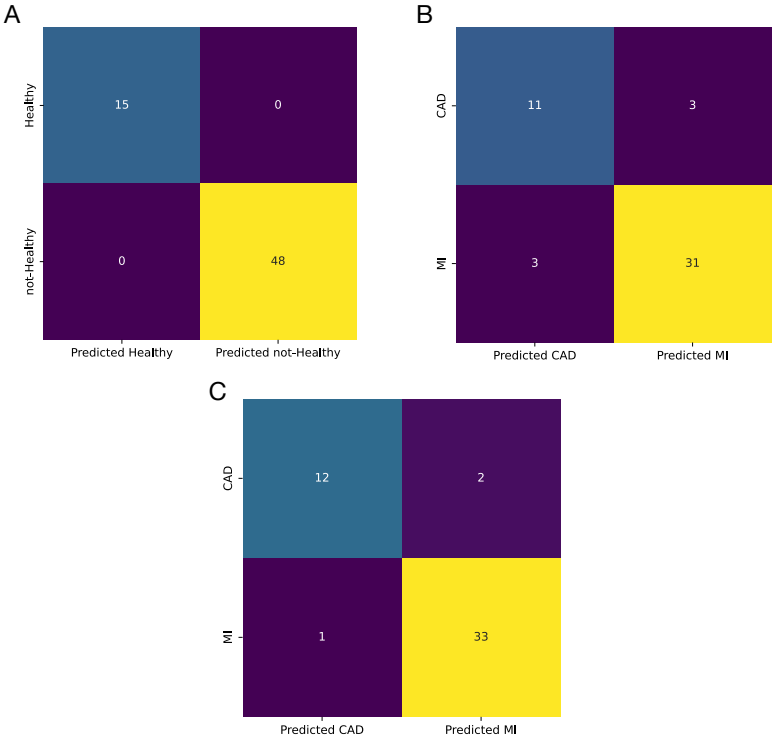

Fig. 11 Confusion matrix for models trained with the set of miRNAs selected by AUC-ROC on test set classification; (A) SVM with RBF kernel for healthy and not-healthy samples classification. (B) Logistic regression model for CAD and MI sample classification. (C) SVM with polynomial kernel for CAD and MI sample classification.

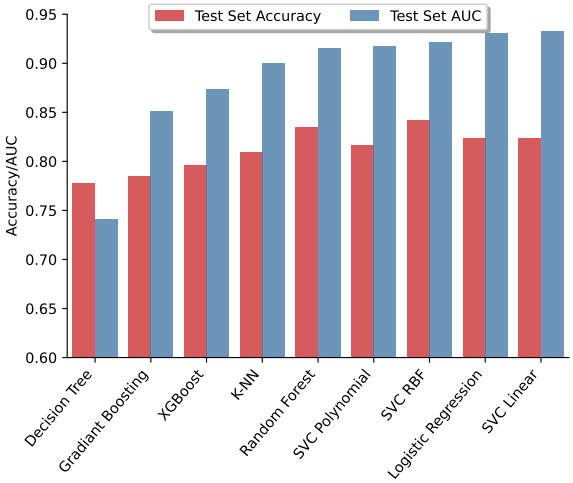

Fig. 12 Area under the receiver operating characteristic curve and accuracy of different models trained with AUC-selected miRNAs.

Table 5 AUC-ROC and accuracy for SVM with the linear kernel as the best model trained with miRNAs selected based on their AUC-ROC on the train and test set before and after hyper-tuning

		Preset parameters		Hypertuned	
Model	Metrics	Train	Test	Train	Test
SVM-linear	AUC-ROC	0.91	0.93	0.93	0.92
	Accuracy	0.85	0.82	0.90	0.88
SVM-RBF	AUC-ROC	0.90	0.92	0.96	0.96
	Accuracy	0.86	0.84	0.96	0.94

4 Discussion

The prevalence of MI can lead to high mortality rates in the clinical setting. However, early diagnosis and the application of suitable treatment protocols can reduce mortality and improve MI prognosis ("Cardiovascular Diseases (CVDs)" n.d.; Thygesen et al. 2018; Tsao et al. 2022). Studies have suggested that changes in miRNA expression may play a significant role in the progression of MI and the subsequent remodeling (Laggerbauer and Engelhardt 2022). It is believed that miRNA expression is altered during the various biological processes correlated with MI within the myocardium or other related tissues (Khan, Gupta, and Mahapatra 2022). Although several studies have focused on examining free circulating miRNAs in serum samples for the detection of cardiac tissue injuries (Kaur et al. 2020), more information is needed to fully comprehend the miRNAs found in different blood subcomponents, such as plasma, platelets, and PBMCs. Based on previous findings, PBMCs play a crucial role in the destabilization and rupture of plaques as well as in the initial inflammatory reactions in individuals experiencing myocardial infarction (MI). (Mosallaei et al. 2022; Hapke et al. 2022). Moreover, PBMCs have specific miRNA profiles that are altered under certain pathological conditions, making them great candidates as disease biomarkers (Mosallaei et al. 2022).

PBMCs can respond to several insulting conditions, such as MI, in the shortest possible time with notable changes in their miRNA profile (Mosallaei et al. 2022). Considering their regulatory roles, subtle changes in the transcription of miRNAs can be monitored even before alterations in mRNA and protein levels(Schulte et al. 2020). These features make miRNAs a valid early-stage diagnostic tool for the detection of minor and major cell injuries. To date, few studies have compared the miRNA profiles in PBMCs from patients with MI and other CADs and healthy samples to find a robust set of identical miRNAs to differentiate these pathological conditions.

In this study, we combined three GEO datasets for healthy, CAD, and MI samples. Having these sample sets alongside bioinformatics analysis and ML methods enabled us to identify potential biomarker sets and effective therapeutic targets. The results of the DEG analysis (Table @ref(tab:DEGstab) and Figure @ref(fig:venn)) prove the close relationship between the MI and CAD samples. Interestingly, functional enrichment analysis demonstrated that

DEGs in both CAD/healthy and MI/healthy were strongly correlated with the immune cell response, which is a major part of PBMCs. Two sets of miRNAs were selected as biomarker sets for sample classification. miR-21; miR-32; and miR-186 were selected as differentially expressed miRNAs, and miR-186; miR-21; miR-29a; miR-197; and miR-296 were selected based on their AUC-ROC values. As shown in Figure @ref(fig:miRROC), all miRNAs selected with both approaches had AUC-ROCs > 0.9 for isolating healthy and not-healthy samples except for miR-296 and miR-29a. The data confirmed that the real challenge was to classify CAD and MI samples because of the close overlap. Of the six miRNAs under investigation in both approaches, except for miR-32, all miRNAs had an AUC-ROC > 0.8 for the discrimination of CAD and MI samples. As expected, the high AUC-ROC values of the miRNAs confirmed their high potential as biomarkers.

ML models trained with miRNA sets selected by both DEG and AUC-ROC approaches, showed better classification performance than each miRNA. To avoid unwanted complexity and poor predictive values, a two-layer architecture was designed. The first layer was used to discriminate between healthy and not-healthy samples, and the second layer was was used to separate CAD from MI candidates. As expected, in both approaches, a hypertuned SVM model could flawlessly separate healthy and not-healthy samples using distinct miRNA sets. ML models are also capable of effectively separating CAD from MI patients. Although both miRNA sets had nearly the same AUC-ROC using the best model, their accuracy, sensitivity, and specificity were different. The model trained with AUC-selected miRNAs showed better performance in all predictive values, which is logical because of the higher number of miRNAs in the set.

Numerous studies have reported that different biological processes can affect the miRNA expression in PBMCs. However, the exact role of miRNAs in the function of immune cells and the correlation between specific pathological conditions and miRNA profiles remain controversial. Several studies have proven the activation of particular miRNA types in PBMCs under cardiovascular events H. Li et al. (2018). For instance, there is evidence that elevation of miR-186 suppresses the expression of cystathionine-γ-lyase, leading to the subsequent secretion of pro-inflammatory cytokines and cellular lipid accumulation. In addition, macrophage-derived miR-186 may promote atherosclerotic plaque formation (Yao et al. 2016). In line with this claim, we found that miR-186 was up-regulated in both CAD and MI candidates compared to their control counterparts. Surprisingly, the obtained data indicated that the expression of miR-186 was higher in patients with CAD than in patients with MI (Figure @ref(fig:Expall)). Specifically, miR-186 was the only differentially expressed miRNA between CAD and MI, with a clear up-regulation in CAD, indicating its main role in the promotion of atherosclerosis.

As mentioned before, miR-21 was also up-regulated in both MI and CAD patients compared to healthy controls. Moreover, the expression value of miR-21 was significantly higher in the MI group than in the CAD group (Table @ref(tab:mirExptable)). It is thought that the up-regulation of miR-21 in

PBMCs is a compensatory reaction to reduce the T_{reg} lymphocyte number in response to the reduction in TGFβ1 secretion into the plasma through a TGFβ1/smad-independent pathway. In line with the previous and present data, miR-21 can modulate the activity of PBMCs following the occurrence of cardiovascular diseases (S. Li et al. 2015).

Recent data have supported the elevation of miR-32 levels in CAD samples with calcification of the coronary artery. Notably, miR-32 promotes vascular smooth muscle calcification in mice by controlling the activity of several proteins, including bone morphogenetic protein-1, runt-related transcription factor-2 (RUNX2), osteopontin, and bone-specific phosphoprotein matrix GLA protein. (Liu et al. 2017). Likewise, some reports are associated with the activity of miR-32 in PBMCs in several pathologies (Zeng et al. 2021; Wang et al. 2020). The exact role of miR-32 in PBMCs after cardiovascular events remains unclear.

Molecular analyses have indicated the regulatory role of miRNAs selected using the AUC-ROC approach in PBMCs after a cardiovascular event. The biological importance of two common miRNAs in the DEGs and AUC-ROC approaches, miR-21 and miR-186, have already been discussed. Based on numerous reports, miR-29a can be activated in different diseases (Horita, Farquharson, and Stephen 2021). Data analysis indicated that miR-29a was significantly up-regulated in CAD patients compared to the healthy and MI groups (Table @ref(tab:mirExptable)). Increased miR-29a is associated with the progression of atherosclerosis, and the combination of miR-29a and ox-LDL has been suggested as a valid biomarker set for paraclinical classification (Huang et al. 2016). However, the role of miR-29a in the function of PBMCs from patients with CAD has not been thoroughly examined.

The data indicated that miR-197 was significantly up-regulated in both the CAD/healthy and MI/healthy groups. Previous studies have demonstrated that miR-197 may play a crucial role in controlling the anti-inflammatory response of IL-35 by influencing the secretion of cytokines that can either promote or suppress inflammation, the ratio of M1/M2 macrophages, and the proliferation of T_{reg} lymphocytes, which are responsible for suppressing immune responses (Bhansali et al. 2022). Alongside our findings, it can be concluded that miR-197 could be a useful diagnostic tool for predicting adverse cardiovascular events.

The findings of this study demonstrate the potential of miR-296 as a biomarker with high discriminatory power to distinguish between samples from individuals with MI and CAD. MiR-296 has been identified as a key regulator in the development and advancement of atherosclerosis by controlling the expression of target genes associated with various biological processes, including angiogenesis, cholesterol metabolism, inflammation, cellular proliferation, hypertension, and apoptosis (H. Li et al. 2018). In a previous study, miR-296 expression levels were found to be significantly increased in the PBMCs of CAD patients compared to healthy controls, suggesting its involvement in regulating proinflammatory cytokines such as IL-6 and TNF-α (Fard et al. 2020). These findings suggested that miR-296 may have a significant impact

on the pathogenesis of atherosclerosis and could potentially serve as a diagnostic biomarker for CAD or MI.

5 Conclusion

In summary, we derived a set of miRNA biomarkers by comparing MI samples with both healthy and CAD samples. We found that the SVM model performed best in both the first layer, which separated healthy and unhealthy samples, and the second layer, which classified the MI/CAD samples. The set of miRNAs selected based on their AUC-ROC values performed better in the second layer. Overall, our two-layer structure achieved an accuracy of 0.96. This demonstrates the potential of combining bioinformatics and machine learning techniques to identify novel biomarkers and gain a deeper understanding of myocardial infarction.

References

Bhansali, Shipra, Amit Kumar Yadav, Chetan Bakshi, and Veena Dhawan. 2022. "Interleukin-35 Mitigates Ox-LDL-Induced Proatherogenic Effects via Modulating miRNAs Associated with Coronary Artery Disease (CAD)." *Cardiovascular Drugs and Therapy*, April. https://doi.org/10.1007/s10557-022-07335-x.

Canali, Raffaella, Lucia Natarelli, Guido Leoni, Elena Azzini, Raffaella Comitato, Oezgur Sancak, Luca Barella, and Fabio Virgili. 2014. "Vitamin C Supplementation Modulates Gene Expression in Peripheral Blood Mononuclear Cells Specifically Upon an Inflammatory Stimulus: A Pilot Study in Healthy Subjects." *Genes & Nutrition* 9 (3): 390. https://doi.org/10.1007/s12263-014-0390-x.

"Cardiovascular Diseases (CVDs)." n.d. Accessed March 12, 2023. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

Fard, Toktam Kazemi, Samareh Tavakoli, Reza Ahmadi, Nariman Moradi, Reza Fadaei, Asghar Mohammadi, and Soudabeh Fallah. 2020. "Evaluation of Ip10 and miRNA 296-a Expression Levels in Peripheral Blood Mononuclear Cell of Coronary Artery Disease Patients and Controls." *DNA and Cell Biology* 39 (9): 1678–84. https://doi.org/10.1089/dna.2020.5650.

Gao, Jie, Jia Liu, Ying Zhang, BaoYi Guan, Hua Qu, Hua Chai, WenTing Wang, XiaoJuan Ma, and DaZhuo Shi. 2020. "PBMCs-Derived microRNA Signature as a Prethrombotic Status Discriminator in Stable Coronary Artery Disease." *Thrombosis and Haemostasis* 120 (01): 121–31. https://doi.org/10.1055/s-0039-1700518.

Hapke, Nils, Margarete Heinrichs, DiyaaElDin Ashour, Elena Vogel, Ulrich Hofmann, Stefan Frantz, and Gustavo Campos Ramos. 2022. "Identification of a Novel Cardiac Epitope Triggering T-Cell Responses in Patients

with Myocardial Infarction." *Journal of Molecular and Cellular Cardiology* 173 (December): 25–29. https://doi.org/10.1016/j.yjmcc.2022.09.001.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585 (7825): 357–62. https://doi.org/10.1038/s41586-020-2649-2.

Head, Tim, Manoj Kumar, Holger Nahrstaedt, Gilles Louppe, and Iaroslav Shcherbatyi. 2021. *Scikit-Optimize/Scikit-Optimize* (version v0.9.0). Zenodo. https://doi.org/10.5281/zenodo.5565057.

Horita, Masahiro, Colin Farquharson, and Louise A Stephen. 2021. "The Role of miR-29 Family in Disease." *Journal of Cellular Biochemistry* 122 (7): 696–715. https://doi.org/10.1002/jcb.29896.

Huang, Yu-Qing, An-Ping Cai, Ji-Yan Chen, Cheng Huang, Jie Li, and Ying-Qing Feng. 2016. "The Relationship of Plasma miR-29a and Oxidized Low Density Lipoprotein with Atherosclerosis." *Cellular Physiology and Biochemistry* 40 (6): 1521–28. https://doi.org/10.1159/000453202.

Kalayinia, Samira, Fateme Arjmand, Majid Maleki, Mahshid Malakootian, and Chandra Pal Singh. 2021. "MicroRNAs: Roles in Cardiovascular Development and Disease." *Cardiovascular Pathology* 50 (January): 107296. https://doi.org/10.1016/j.carpath.2020.107296.

Kaur, Amanpreet, Sharon T Mackin, Kenny Schlosser, Fui Lin Wong, Malik Elharram, Christian Delles, Duncan J Stewart, Natalie Dayan, Tara Landry, and Louise Pilote. 2020. "Systematic Review of microRNA Biomarkers in Acute Coronary Syndrome and Stable Coronary Artery Disease." *Cardiovascular Research* 116 (6): 1113–24. https://doi.org/10.1093/cvr/cvz302.

Khan, Abrar A., Vinayak Gupta, and Nitish R. Mahapatra. 2022. "Key Regulatory miRNAs in Lipid Homeostasis: Implications for Cardiometabolic Diseases and Development of Novel Therapeutics." *Drug Discovery Today* 27 (8): 2170–80. https://doi.org/10.1016/j.drudis.2022.05.003.

Laggerbauer, Bernhard, and Stefan Engelhardt. 2022. "MicroRNAs as Therapeutic Targets in Cardiovascular Disease." *Journal of Clinical Investigation* 132 (11): e159179. https://doi.org/10.1172/JCI159179.

Lazar, C., S. Meganck, J. Taminau, D. Steenhoff, A. Coletta, C. Molter, D. Y. Weiss-Solis, R. Duque, H. Bersini, and A. Nowe. 2013. "Batch Effect Removal Methods for Microarray Gene Expression Data Integration: A Survey." *Briefings in Bioinformatics* 14 (4): 469–90. https://doi.org/10.1093/bib/bbs037.

Li, Heng, Heng Li, Xin-Ping Ouyang, Ting Jiang, Xi-Long Zheng, Xi-Long Zheng, Ping-Ping He, and Guo-Jun Zhao. 2018. "MicroRNA-296: A Promising Target in the Pathogenesis of Atherosclerosis?" *Molecular Medicine* 24 (1): 12–12. https://doi.org/10.1186/s10020-018-0012-y.

Li, Sihui, Qian Fan, Shaolin He, Tingting Tang, Yuhua Liao, and Jiangjiao Xie. 2015. "MicroRNA-21 Negatively Regulates Treg Cells Through a TGF-β1/Smad-Independent Pathway in Patients with Coronary Heart Disease."

Cellular Physiology and Biochemistry 37 (3): 866–78. https://doi.org/10.1159/000430214.

Liu, Jianghua, Xinhua Xiao, Yingying Shen, Ling Chen, Canxin Xu, Heng Zhao, Ying Wu, et al. 2017. "MicroRNA-32 Promotes Calcification in Vascular Smooth Muscle Cells: Implications as a Novel Marker for Coronary Artery Calcification." Edited by Yin Tintut. *PLOS ONE* 12 (3): e0174138. https://doi.org/10.1371/journal.pone.0174138.

Maciejak, Agata, Marek Kiliszek, Marcin Michalak, Dorota Tulacz, Grzegorz Opolski, Krzysztof Matlak, Slawomir Dobrzycki, Agnieszka Segiet, Monika Gora, and Beata Burzynska. 2015. "Gene Expression Profiling Reveals Potential Prognostic Biomarkers Associated with the Progression of Heart Failure." *Genome Medicine* 7 (1): 26. https://doi.org/10.1186/s13073-015-0149-z.

Matone, Alice, Colm M. O'Grada, Eugene T. Dillon, Ciara Morris, Miriam F. Ryan, Marianne Walsh, Eileen R. Gibney, et al. 2015. "Body Mass Index Mediates Inflammatory Response to Acute Dietary Challenges." *Molecular Nutrition & Food Research* 59 (11): 2279–92. https://doi.org/10.1002/mnfr.201500184.

McCall, M. N., B. M. Bolstad, and R. A. Irizarry. 2010. "Frozen Robust Multiarray Analysis (fRMA)." *Biostatistics* 11 (2): 242–53. https://doi.org/10.1093/biostatistics/kxp059.

McCall, Matthew N., Karan Uppal, Harris A. Jaffee, Michael J. Zilliox, and Rafael A. Irizarry. 2011. "The Gene Expression Barcode: Leveraging Public Data Repositories to Begin Cataloging the Human and Murine Transcriptomes." *Nucleic Acids Research* 39 (suppl_1): D1011–15. https://doi.org/10.1093/nar/gkq1259.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, 56–61. https://doi.org/ 10.25080/Majora-92bf1922-00a .

Mosallaei, Meysam, Naeim Ehtesham, Shima Rahimirad, Mostafa Saghi, Nasim Vatandoost, and Sharifeh Khosravi. 2022. "PBMCs: A New Source of Diagnostic and Prognostic Biomarkers." *Archives of Physiology and Biochemistry* 128 (4): 1081–87. https://doi.org/10.1080/13813455.2020.1752257.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reel, Parminder S., Smarti Reel, Ewan Pearson, Emanuele Trucco, and Emily Jefferson. 2021. "Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review." *Biotechnology Advances* 49 (July): 107739. https://doi.org/10.1016/j.biotechadv.2021.107739.

RStudio Team. 2020. *RStudio: Integrated Development Environment for r*. Boston, MA: RStudio, PBC. http://www.rstudio.com/.

Schulte, Christian, Temo Barwari, Abhishek Joshi, Tanja Zeller, and Manuel Mayr. 2020. "Noncoding RNAs Versus Protein Biomarkers in Cardiovascular Disease." *Trends in Molecular Medicine* 26 (6): 583–96. https://doi.org/10.1016/j.molmed.2020.02.001.

Schulte, Christian, Mahir Karakas, and Tanja Zeller. 2017. "microRNAs in Cardiovascular Disease – Clinical Application." *Clinical Chemistry and Laboratory Medicine (CCLM)* 55 (5). https://doi.org/10.1515/cclm-2016-0576.

Soler-Botija, Carolina, Carolina Gálvez-Montón, and Antoni Bayés-Genís. 2019. "Epigenetic Biomarkers in Cardiovascular Diseases." *Frontiers in Genetics* 10 (October): 950. https://doi.org/10.3389/fgene.2019.00950.

Thygesen, Kristian, Joseph S. Alpert, Allan S. Jaffe, Bernard R. Chaitman, Jeroen J. Bax, David A. Morrow, Harvey D. White, and The Executive Group on behalf of the Joint European Society of Cardiology (ESC)/American College of Cardiology (ACC)/American Heart Association (AHA)/World Heart Federation (WHF) Task Force for the Universal Definition of Myocardial Infarction. 2018. "Fourth Universal Definition of Myocardial Infarction (2018)." *Circulation* 138 (20). https://doi.org/10.1161/CIR.0000000000000617.

Torun, Furkan M., Sebastian Virreira Winter, Sophia Doll, Felix M. Riese, Artem Vorobyev, Johannes B. Mueller-Reif, Philipp E. Geyer, and Maximilian T. Strauss. 2023. "Transparent Exploration of Machine Learning for Biomarker Discovery from Proteomics and Omics Data." *Journal of Proteome Research* 22 (2): 359–67. https://doi.org/10.1021/acs.jproteome.2c00473.

Tsao, Connie W., Aaron W. Aday, Zaid I. Almarzooq, Alvaro Alonso, Andrea Z. Beaton, Marcio S. Bittencourt, Amelia K. Boehme, et al. 2022. "Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association." *Circulation* 145 (8). https://doi.org/10.1161/CIR.0000000000001052.

Wang, Dan, Ting Zeng, Zhi Lin, Lu Yan, Fenglin Wang, Lanlan Tang, Leyuan Wang, Daolin Tang, Pan Chen, and Minghua Yang. 2020. "Long Non-Coding RNA Snhg5 Regulates Chemotherapy Resistance Through the miR-32/Dnajb9 Axis in Acute Myeloid Leukemia." *Biomedicine & Pharmacotherapy* 123 (March): 109802. https://doi.org/10.1016/j.biopha.2019.109802.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Yao, Yan, Xin Zhang, Hai-peng Chen, Liang Li, Wei Xie, Gang Lan, Zhenwang Zhao, Xi-Long Zheng, Zong-bao Wang, and Chao-ke Tang. 2016. "MicroRNA-186 Promotes Macrophage Lipid Accumulation and Secretion of Pro-Inflammatory Cytokines by Targeting Cystathionine γ-Lyase in THP-1 Macrophages." *Atherosclerosis* 250 (July): 122–32. https://doi.org/10.1016/j.atherosclerosis.2016.04.030.

Yap, Jonathan, Jason Irei, Javier Lozano-Gerona, Selena Vanapruks, Tianmai Bishop, and William A. Boisvert. 2023. "Macrophages in Cardiac Remodelling After Myocardial Infarction." *Nature Reviews Cardiology*, January. https://doi.org/10.1038/s41569-022-00823-5.

Yu, Guangchuang, Li-Gen Wang, Yanyan Han, and Qing-Yu He. 2012. "clusterProfiler: An r Package for Comparing Biological Themes Among Gene Clusters." *OMICS: A Journal of Integrative Biology* 16 (5): 284–87. https://doi.org/10.1089/omi.2011.0118.

Zeng, Zl, Qingyun Zhu, Zhibo Zhao, Xuyu Zu, and Jianghua Liu. 2021. "Magic and Mystery of microRNA-32." *Journal of Cellular and Molecular Medicine* 25 (18): 8588–8601. https://doi.org/10.1111/jcmm.16861.