The First Draft

2023-02-07

Contents

1 Materials and Methods

1.1 Microarray data collection

Microarray datasets were obtained from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). In order to come by sufficient classification power between MI samples and others, the sample size of the dataset should be relatively large. Therefore, GSE59867 for MI and CAD samples, and GSE56609, GSE54475, and GSE23832 for healthy samples were selected. All datasets have been conducted using Affymetrix Human Gene 1.0 ST Array (GPL6244) platform. Only healthy, stable CAD and early stage MI samples were selected from these datasets for further analysis. The basic information for the four GEO datasets evaluated in the current study is provided in Table 1.

Table 1: Basic information of the 3 GEO microarray datasets.

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

1.2 Pre-processing

Raw data (CEL files) of the four datasets were downloaded from the GEO and preprocessed using the fRMA package (M. N. McCall, Bolstad, and Irizarry 2010). fRMA allows to preprocess individual microarray

samples and combine them consistently for analysis. For each dataset, background correction is performed using RMA algorithm and then it is quantile normalized based on the reference distribution. During summarization, batch effects are removed and variances of the gene expressions are estimated by taking into account these probe-specific effects. For those multiple probe sets matched to the identical gene, the mean log fold change was retained. This way fRMA can be seen as a batch effect removal technique for different datasets produced using identical microarray platform. Thus, In order to ensure about batch effect removal, the principal component analysis and the relative log expression of train samples were plotted before and after fRMA (Lazar et al. 2013).

1.3 Differential expression analysis

The barcode algorithm proposed by McCall et al. (Matthew N. McCall et al. 2011) transforms the actual expression values into binary barcode values. Huge sets of samples were collected and normalized using fRMA for several platforms. The distribution of the expressed and non-expressed observed intensities for each gene is estimated using these normalized sets. Genes are deemed expressed (and their value coded to 1) or unexpressed (and their value coded to 0) according to the following equation:

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

where x_{ij} is the normalized intensity of gene i in sample j, C is a user-defined parameter, σ^{ne} is the standard deviation of the non-expressed distribution and μ^{ne} is the mean of the non-expressed distribution. The barcode representation of a sample is a vector of ones and zeros denoting which genes are estimated to be expressed (ones) and unexpressed (zeros). The barcode algorithm was implemented by the barcode function in the R fRMA package, and the default value of C was used.

To determine if the expressed ratios differed in the MI group versus the healthy control group, Fisher's exact test for individual genes was carried out upon the barcode values. Genes with a false discovery rate (FDR) of < 0.05, which was calculated through the Benjamini-Hochberg (BH) procedure to adjust for multiple testing issue, were considered as differentially expressed genes.

Also, the same procedure were conducted on CAD versus healthy controls as well as MI versus CAD group to find the DEGs in between them.

1.4 Functional and pathway enrichment analyses

Using the R clusterProfilter package, the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis and Gene Ontology (GO) functional annotation were carried out on the differentially expressed genes. The GO analysis included biological process (BP), cellular component (CC) and molecular function (MF) categories. An adjusted $P < 0.05$ was considered to indicate a statistically significant difference. Enrichment were counducted on the MI-healthy and CAD-healthy DEGs. In these analyses, all default parameters were used.

1.5 Machine Learning

The machine learning 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). In all ML analysis, the datasets where divided into train and test set by 0.7:0.3 ratio and the reported results are the average of a 10-fold cross-validation.

In the first step, different models has been trained on all microRNAs in DEGs. Moreover, since the number of microRNAs in DEGs are limited, in order to reaching the highest predicting value, different models also trained on microRNAs selected by their individual AUC-ROC. These approach can provide an informative comparison between the predictive capabilities of set of microRNAs selected with two different above-mentioned approaches.

1.5.1 microRNAs in DEGs

In this approach a two layer architecture has been deployed to the data in order to maximize the prediction values. The first layer will predict whether a sample is healthy or not, and the second layer will separates MI from CAD in the samples which were predicted as not healthy in the first layer. To this end, a distinct ML model was trained for each layer. Since there are limited numbers of microRNAs in DEGs both layers where trained with all microRNAs available. For both layers the receiver operating characteristic (ROC) curve were generated and area under curves (AUC) were calculated for each microRNAs in DEGs for further comparison with the models performance.

1.5.1.1 First layer for seperating healthy and not-healthy samples: An SVM models using RBF kernels were hyper-tunned and trained using all microRNAs in DEGs. In order to handle the imbalance number of samples in groups (51 for healthy and 157 for not-healthy group), sample weight for not-healthy samples were set to 0.5.

1.5.1.2 Second layer for seperating MI and CAD samples: For the sake of reaching the highest classification performance using the set of three microRNAs different models were investigated. 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) has been trained using the expression profile of the miRNAs. All models were trained with their pre-set parameters with 10-fold cross-validation.

Criteria for choosing the best model was the highest accuracy and AUC on the test set. The best model was hyper-tuned with scikit-opt package (Head et al. 2021) to get the best predictive performance.

1.5.2 microRNAs with the highest AUC-ROC

Like the previous approach, two different models will be trained. The first layer for classifying samples to healthy and not-healthy and the second layer for separating MI and CAD samples. However, for keeping the number of microRNAs as low as possible microRNAs will be selected from the second layer (which are the microRNAs with the best performance in MI/CAD separation) and then thier performance will be evaluated in the first layer.

1.5.2.1 Second layer for separating MI and CAD: AUC-ROC of all microRNAs for classifying MI and CAD samples has been calculated. For finding the number of microRNAs with the highest predictive values, the microRNAs with the highest individual AUC-ROC is adding to the set one-by-one and the AUC-ROC for the set is calculated. The set with the highest AUC-ROC has been selected as the set for the following steps.

The selected number of microRNAs has been used to train different algorithms for the sake of finding the best model. As previous approach, 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) has been trained using the expression profile of the selected miRNAs set. All models were trained with their pre-set parameters with 10-fold cross-validation. The models with the highest AUC-ROC and accuracy on the test set were selected and hyper-tuned with scikit-opt package (Head et al. 2021).

1.5.2.2 First layer for finding healthy and not-healthy samples: The set of microRNAs selected for differentiating MI and CAD samples also used to diagnose not-healthy samples and the ROC curve were ploted for each microRNA. Moreover, expression values of the microRNAs in the set were used to train an SVM model with RBF kernel and its ROc curve were plotted also.

2 Results and Discussion

2.1 Pre-processing

The PCA plot of the train samples were shown in Fig1. As it is clear, there is a complete separation between healthy samples and CAD and MI samples in primary data which also remains after conducting fRMA on the data.

The RLE plot presented in Fig1 validates batch effect removal. For an efficient batch effect removal method, the individual boxplots will be all distributed around 0 in RLE plot, and inter-quantile distances would be grater than 0.1 (Lazar et al. 2013). The mentioned criteria is not met in primary data, but has been met after conducting fRMA algorithm.

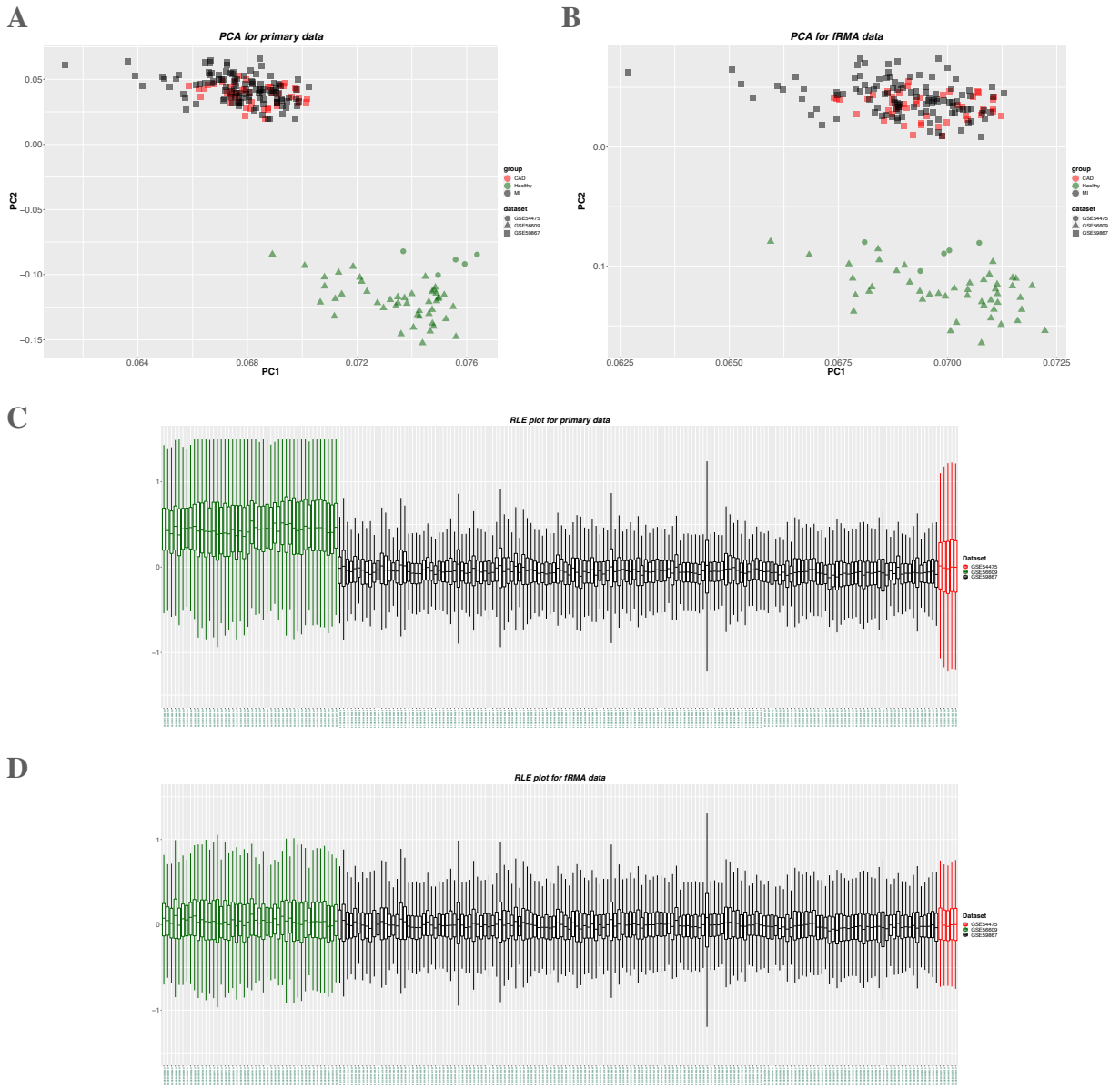

Figure 1: PCA and RLE plot for all samples before and after fRMA.

2.2 Differential expression analysis

According to the cutoff criterion of $FDR < 0.05$, there are 860 DEGs between the MI patients and the healthy controls. Among them, 323 are up-regulated in MI, and 537 are down-regulated in MI compared to the healthy controls. For CAD and healthy groups there are 670 DEGs, 262 of them were up-regulated and 408 of them were down-regulated in CAD samples in comparison with healthy samples. For MI and CAD group the number of DEGs are 260, and the number of up- and down-regulated genes in MI samples in comparison with CAD samples are 144 and 116, respectively. All these data is summarized in table 2.

Table 2: Total DEGs and name of microRNAs in DEGs.

	No. of DEGs	no. of up-regulated DEGs	no. of down-regulated DEGs	microRNAs
MI vs. Healthy	860	323	537	hsa-miR-186, hsa-miR-21, hsa-miR-32
CAD vs. Healthy	670	262	408	hsa-miR-186, hsa-miR-21, hsa-miR-32
MI vs. CAD	260	144	116	hsa-miR-186

2.3 Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses of the DEGs.

To explore the biological classification of the DEGs, we performed GO and KEGG pathway enrichment analyses on MI-healthy and CAD-healthy DEGs. For both set of DEGs, many biological functions enriched were associated with the immune cells, as expected.

For MI versus healthy samples, GO enrichment analysis in the biological process (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" (Fig2A). In the cellular component (CC) category the DEGs were enriched in "secretory granule membrane", "azurophil granule", "ficolin-1-rich granule", "tertiary granule", and "ficolin-1-rich granule membrane" (Fig2B). In the molecular function (MF) category, the DEGs were involved in "cadherin binding" and "MHC class I protein binding" (Fig2C). KEGG pathway analysis indicated that the DEGs were related to the following pathways: "Chemokine signaling pathway", "Lipid and atherosclerosis", and "Hematopoietic cell lineage" (Fig2D).

The enrichment results for DEGs of CAD versus healthy samples are as follows. In the biological process (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" (Fig3A). In the cellular component (CC) category the DEGs were enriched in "azurophil granule", "ficolin-1-rich granule", and "ficolin-1-rich granule membrane" (Fig3B). In the molecular function (MF) category, the DEGs were involved in "lipoprotein particle receptor binding" and "NF-κB binding" (Fig3C). KEGG pathway analysis indicated that the DEGs were related to the following pathways: "Chemokine signaling pathway", "Lipid and atherosclerosis", and "Hematopoietic cell lineage" (Fig3D).

2.4 Machine Learning

2.4.1 microRNAs in DEGs

Among all DEGs, just hsa-miR-186, hsa-miR-32, and hsa-miR-21 are differntially expressed miRNAs. The expression profile of these three miRNAs are presented in Fig4. The ROC curves of each miRNA for each

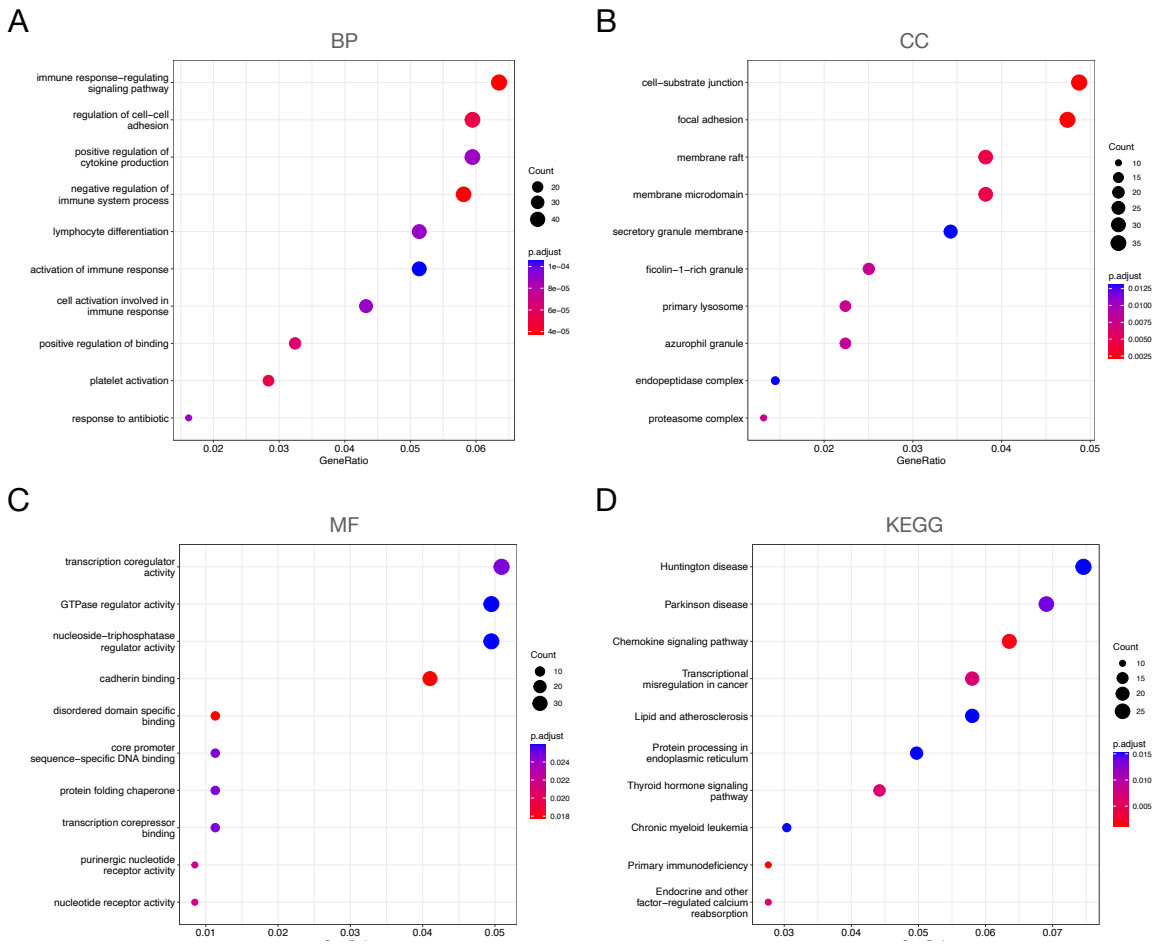

Figure 2: Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enriched with the MI and healthy DEGs. (A) Biological process terms. (B) Cellular component terms. (C) Molecular function terms. (D) KEGG analysis.

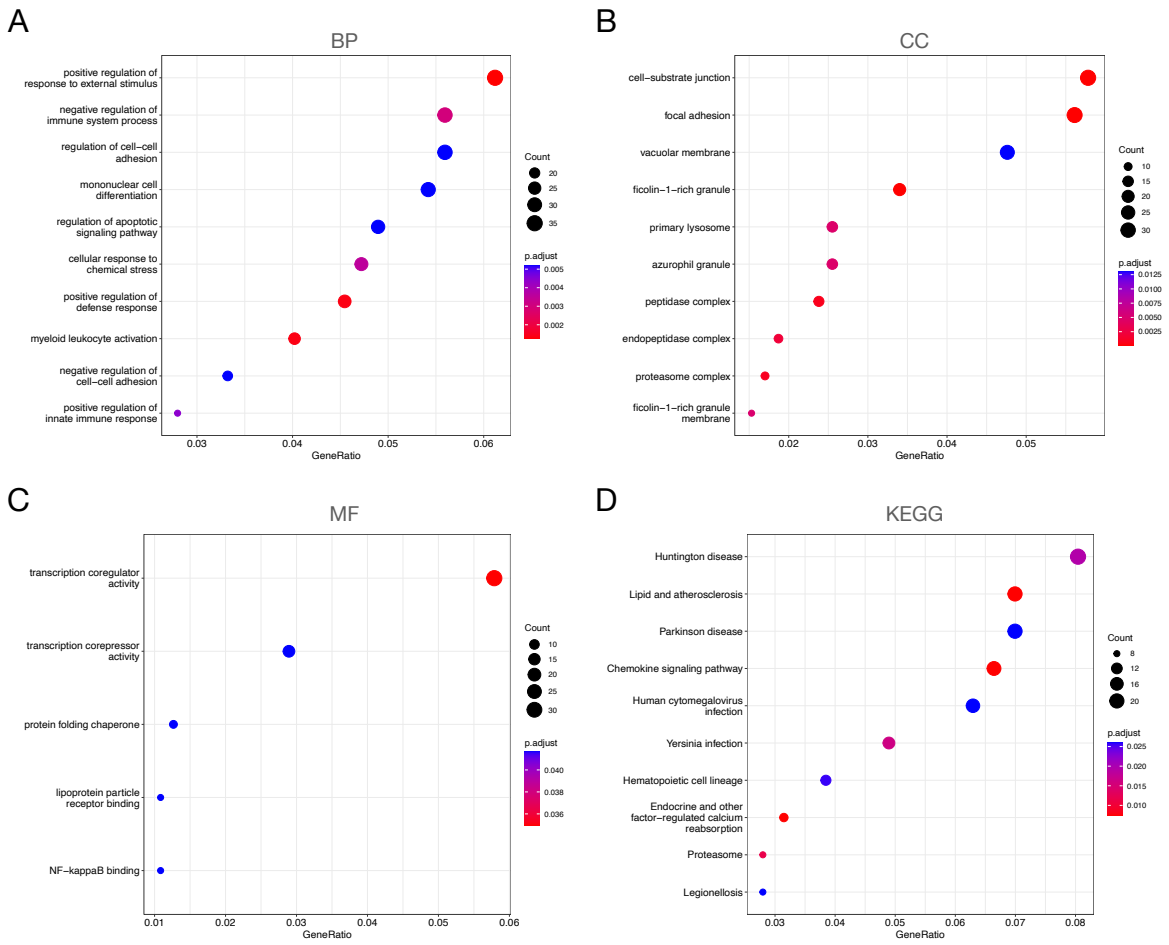

Figure 3: Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enriched with the CAD and healthy DEGs. (A) Biological process terms. (B) Cellular component terms. (C) Molecular function terms. (D) KEGG analysis.

layer is presented in Fig5. As it is clear from the Fig5A, all these three microRNAs has an acceptable performance fro separating healthy and not-healthy samples. AUC for hsa-miR-21, hsa-miR-32, and hsa-miR-186 is 0.99, 1, and 0.91 respectively. Also the accuracy of each microRNA for classifying the samples to healthy and not-healthy groups on test set using a simple logistic regression model is 0.92, 0.98, and 0.89 for hsa-miR-21, hsa-miR-32, and hsa-miR-186 respectively. Moreover, In Fig5B the ROC curve of each microRNA for classifying MI and CAD samples were presented. For hsa-miR-21, hsa-miR-32, and hsa-miR-186 AUC on test set is 0.85; 0.7; and 0.82, and accuracy on test set is 0.78; 0.67; and 0.74.

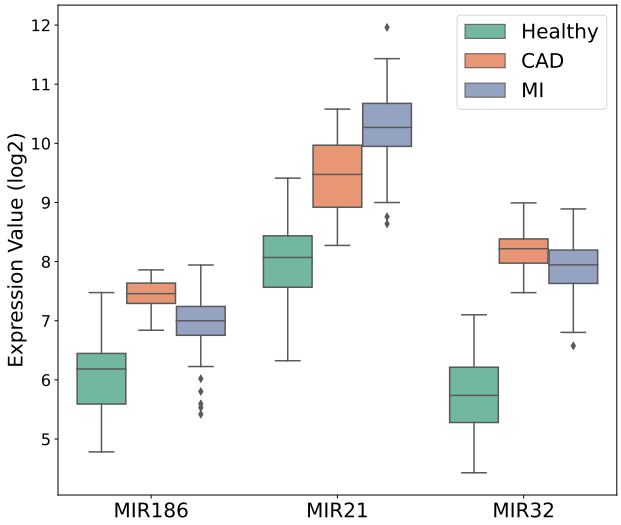

Figure 4: Expression profile of hsa-miR-186, hsa-miR-21, and hsa-miR32 in Healthy, CAD, and MI samples.

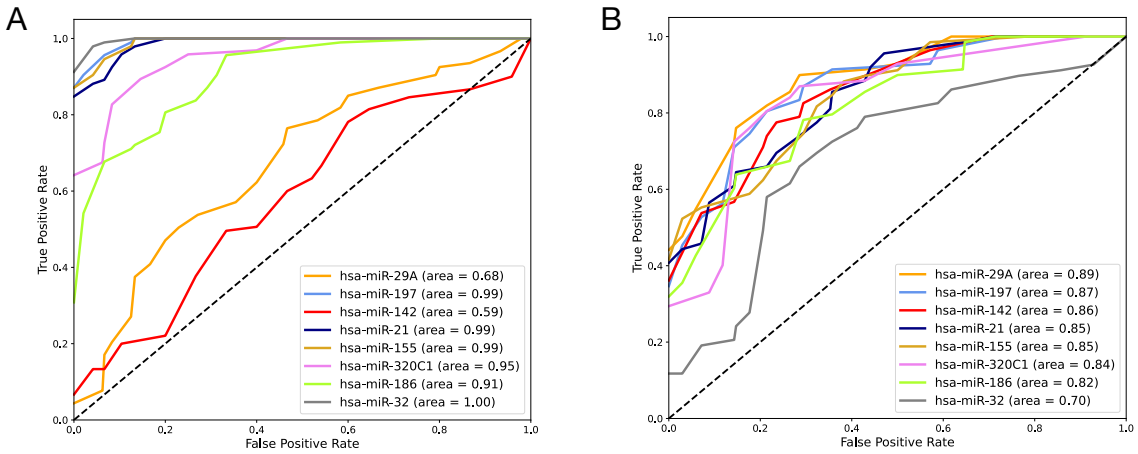

Figure 5: ROC curve for single microRNAs on test set classification for (A) healthy and not-healthy samples and (B) CAD and MI samples.

2.4.1.1 First layer for healthy not-healthy seperation: Although single microRNAs have acceptable performance, but their prediction value can improve even more using them as a set. The ROC curve for the SVM model with rbf kernel trained with all three microRNAs is presented in Fig6A. The model has better performance in classification than single microRNAs.The AUC for the model is 1, and its accuracy on test set is 1 as well. The confusion matrix for the model is presented in Fig7A.

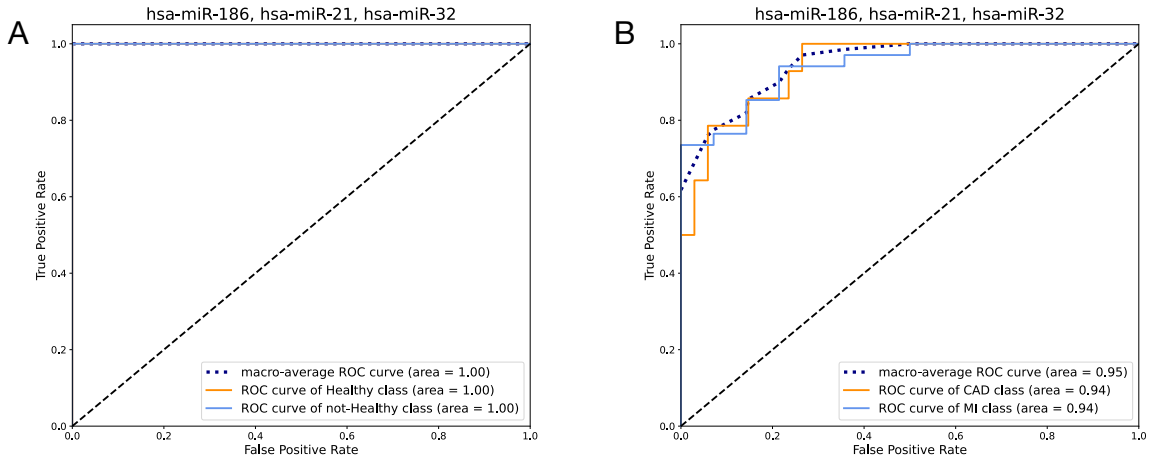

Figure 6: ROC curve for microRNAs in DEGs on test set classification for (A) healthy and not-healthy samples and (B) CAD and MI samples.

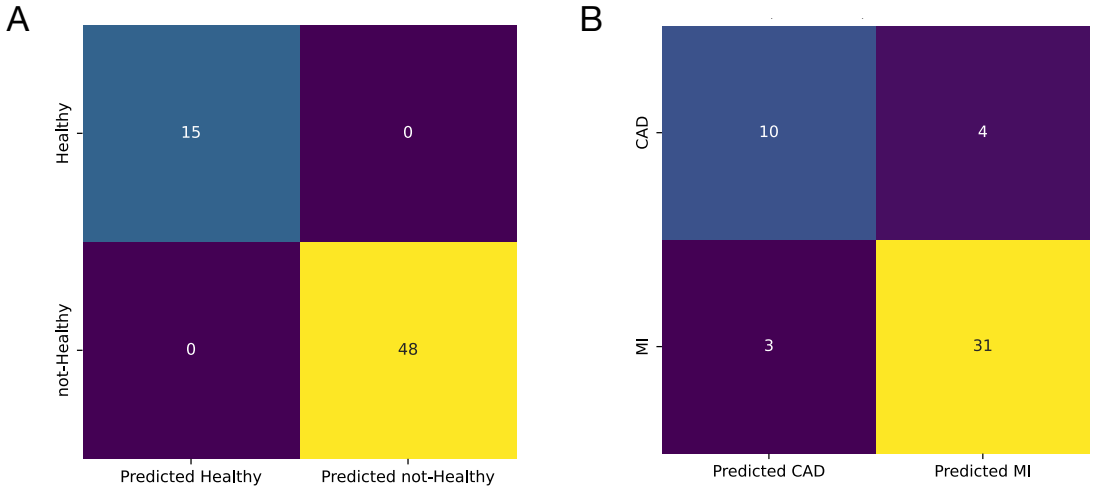

Figure 7: Confusion matrix on the test set for (A) An SVM model with RBF kernel for healthy and not-healthy samples classification. and (B) An SVM model with linear kernel for CAD and MI samples classification.

2.4.1.2 Second layer for separating MI samples from CAD: Different models has been trained using expression values of three differentially microRNAs. The models 10-fold cross-validate AUC and accuracy on the test set is reported on Fig8. The best model from both AUC and accuracy point-of-view is an SVM model with linear kernel. The AUC and accuracy for this model with its preset values is 0.93 and 0.82 respectively. The model is hyper-tuned for C and gamma hyperparameters, and therefore the model showed a better performance. The ROC curve of the hyper-tuned model as the best model is presented in Fig6B. For this model the AUC reached to 0.95 and the accuracy improved to 0.85. Moreover, the sensitivity and specificity for the model on the test set is 0.91 and 0.71 respectively. The confusion matrix for the hyper-tuned model is illustrated in Fig7B.

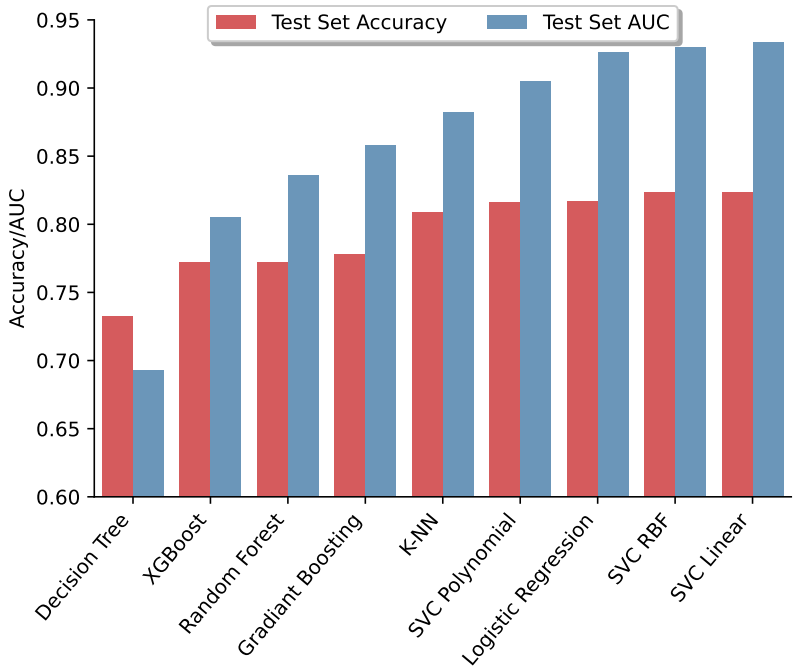

Figure 8: Area under curve (AUC) and accuracy of different models trained with three microRNAs in DEGs on the test set.

2.4.2 AUC approach

2.4.2.1 Second layer; MI form CAD: After calculating AUC for each microRNA, they are sorted and their performance as a set were investigated. The metric for finding the best set was AUC. The diagram for AUC against the number of microRNAs in the set is presented in Fig9. As it is clear, the AUC increase until the number of microRNAs in the set reaches tho 6 and after that it drops. The AUC for separating MI samples from CAD using these microRNAs is 0.93. The microRNAs in the set are has-miR-29A, has-miR-197, has-miR-142, has-miR-21, has-miR-155, and has-miR-320C1. The expression values of these microRNAs in healthy, CAD, and MI samples is presented in Fig10. The ROC curve of the selected microRNAs for MI and CAD samples classification is illustrated in Fig5B.

For finding the most model for training the best set, different models were trained using their pre-set values. Their AUC and accuracy results on the test set is presented in Fig12. The best model from AUC point of view is the LR and from accuracy point-of-view is the SVM model with polynomial kernel. For LR model the AUC and accuracy were 0.92 and 0.81; and for SVM model with polynomial kernel the values were 0.91 and 0.84 respectively. Both models were hyper-tuned and ROC curve for their best performance presented in Fig13A and B. The AUC and accuracy for LR model increased to 0.94 and 0.88 respectively; And for SVM model with polynomial kernel these values increased to 0.95 and 0.88 respectively. The sensitivity for LR and SVM models are 1 and 0.97; and the specificity for them is 0.57 and 0.64, respectively. The confusiom

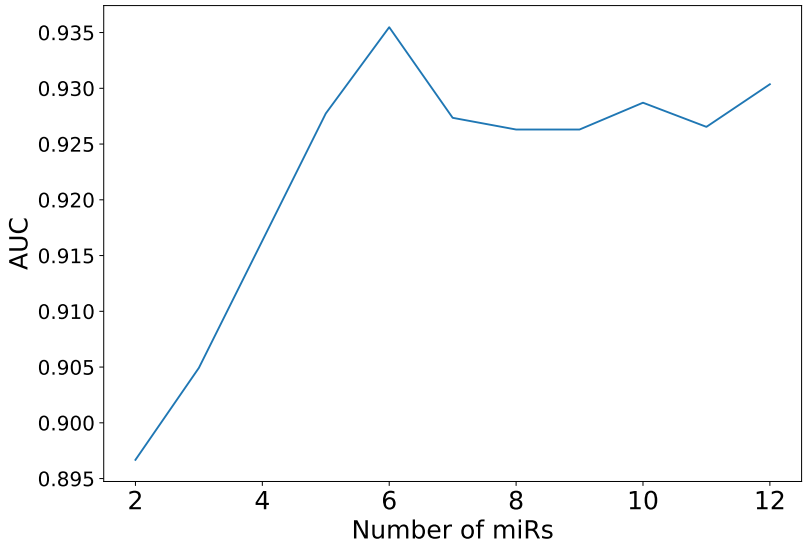

Figure 9: Area under curve (AUC) for sets containing increasing number of microRNAs with the highest individual AUC in MI/CAD separation.

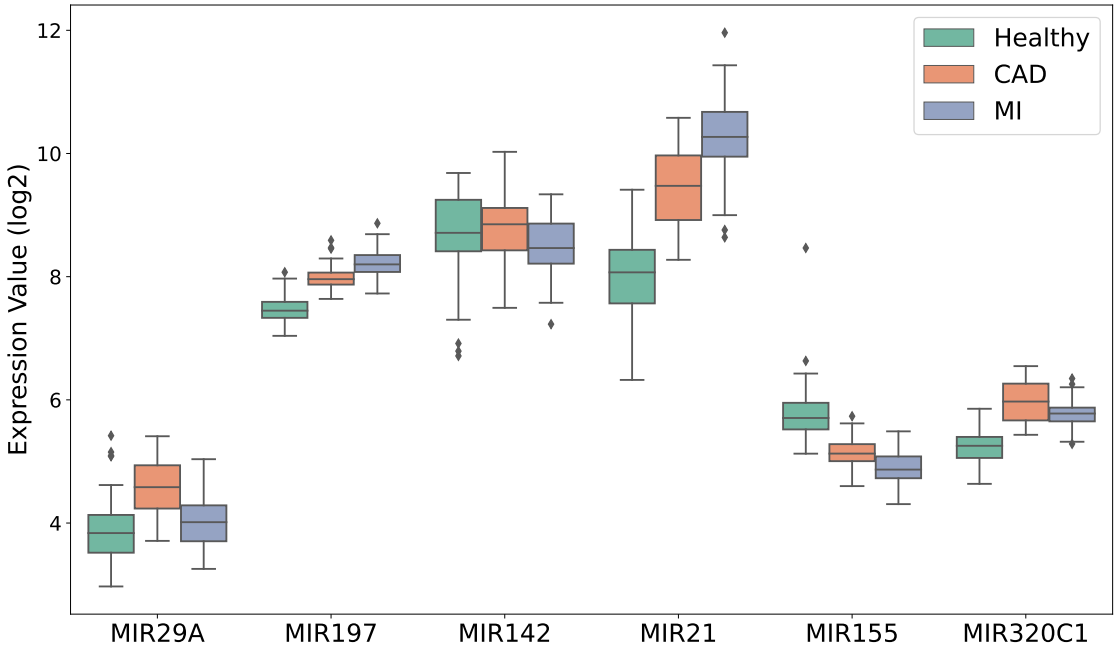

Figure 10: Expression profile of has-miR-29A, has-miR-197, has-miR-142, has-miR-21, has-miR-155, and has-miR-320C1 in Healthy, CAD, and MI samples.

matrix for both models is illustrated in Fig11A and B.

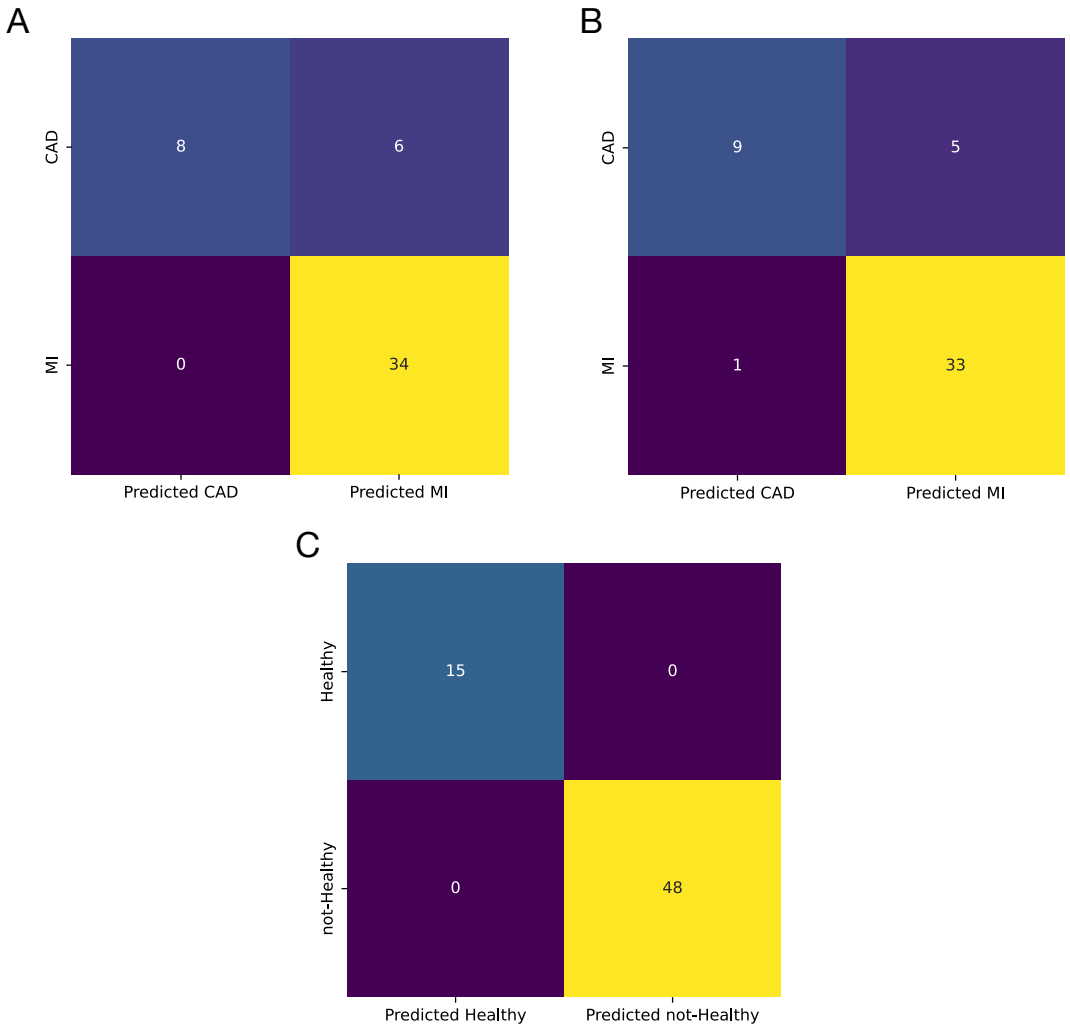

Figure 11: Confusion matrix on the test set for (A) Logistic regression model for CAD and MI samples classification. (B) SVM with polynomial kernel for CAD and MI samples classification. (C) SVM with RBF kernel for healthy and not-healthy samples classification.

2.4.2.2 First layer: Using the microRNAs set selected for classification of MI and CAD samples, an SVM model with RBF kernel has been trained for separating healthy from not-healthy samples. The ROC curve for the model is presented in Fig13C and confusion matrix is illustrated in Fig11C. Both AUC and accuracy for the model on the test set is equal to 1.

Refrences

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.

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.

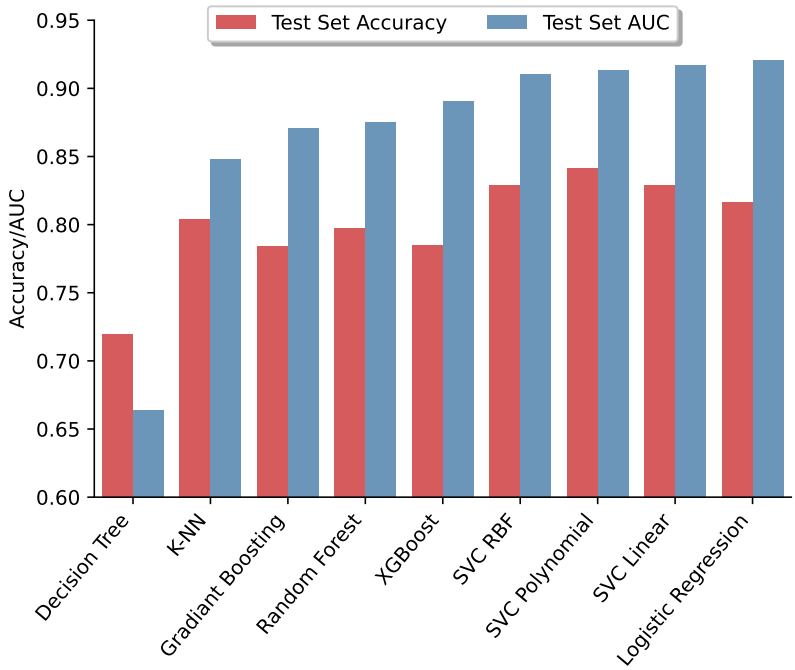

Figure 12: Area under curve (AUC) and accuracy of different models trained with three microRNAs in DEGs on the test set.

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.

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.

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 .

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.

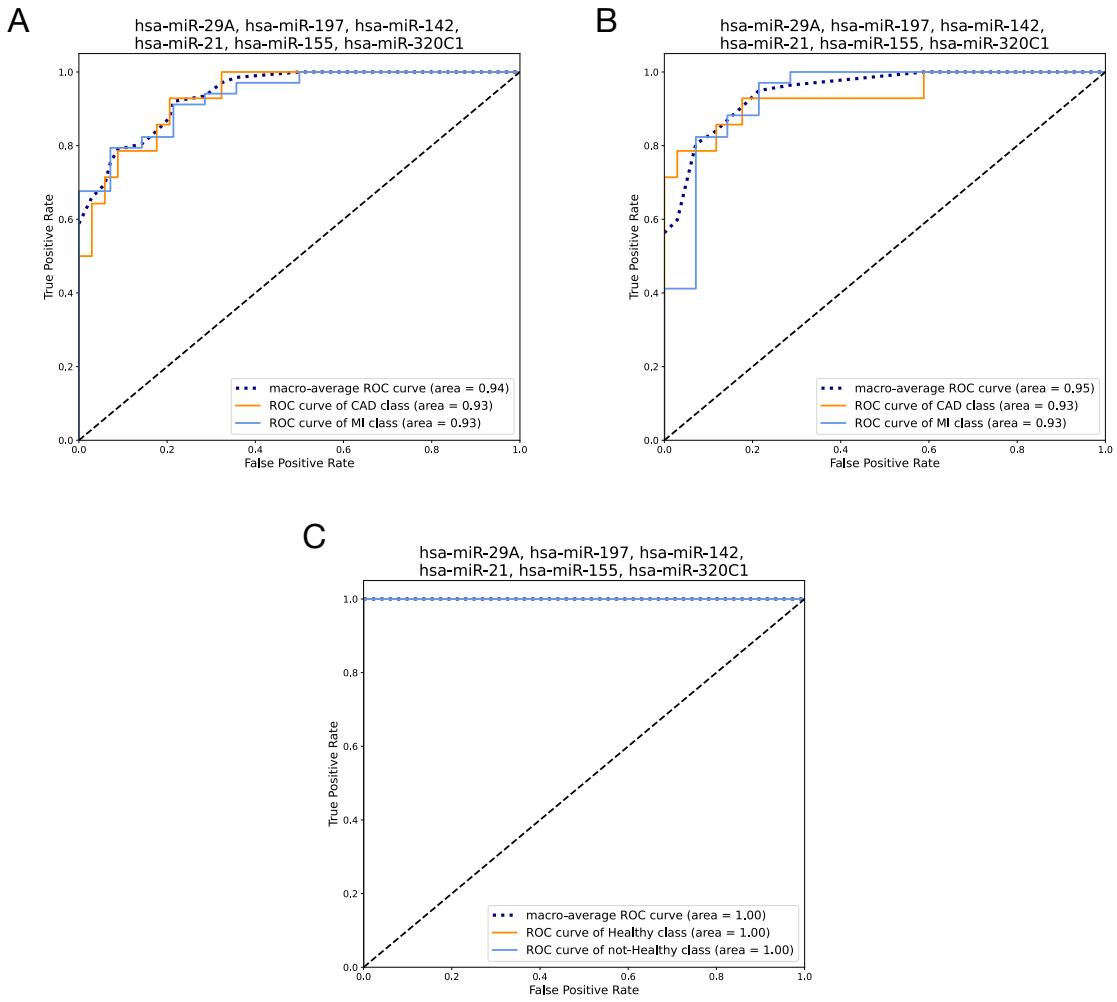

Figure 13: ROC curve for the set of microRNAs selected by AUC on test set classification. (A) Logistic regression model for CAD and MI samples classification. (B) SVM with polynomial kernel for CAD and MI samples classification. (C) SVM with RBF kernel for healthy and not-healthy samples classification.