Designing Machine Learning Methods for RNA-Sequencing Analysis of Atherosclerosis

By Jainam H. Modh

Group 9

Advisors:
Dr. Andrew Warren
Dr. Chunhong Mao

Affiliation:
Biocomplexity Institute and Initiative at the University of Virginia

Word Count: 3,365
Number of Figures and Tables: 8
Number of Equations: 1
Number of Supplements: 0
Number of Citations: 10

Signature of Approval: _____ Date 05/13/22

Signature of Approval: _____ Date 05/13/22

Designing Machine Learning Methods for RNA-Sequencing Analysis of Atherosclerosis

Jainam H. Modh[1a]. Colin Price[2], James Hixson[3], Do-Kyun Kim[3], Fannie Jackson[4], Dana Troxclair[4], Richard Vander Heide[4], Yue Wang[5], Jennifer Van Eyk[6], David Herrington[7], Andrew Warren[2]*, Chunhong Mao[2]*

[1]Biomedical Engineering, University of Virginia, Charlottesville, VA 22903, USA
[2]Biocomplexity Institute and Initiative, University of Virginia, Charlottesville, VA 22911, USA
[3]Human Genetics Center, UTHealth School of Public Health, Houston, TX 77030, USA
[4]Department of Pathology, Louisiana State University Health Science Center, New Orleans, LA 70112, USA
[5]Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA 22203, USA
[6]Heart Institute and Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA
[7]Department of Internal Medicine, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA
*Correspondence: Andrew Warren: asw3xp@virginia.edu and Chunhong Mao: cm4su@virginia.edu, contributed equally.

Abstract

Atherosclerosis is one of the primary causes of cardiovascular disease, which accounts for 32% of global deaths and costs the United States $312.6 billion each year. Heritability estimates for atherosclerosis and its associated diseases vary between 40% to 70%, suggesting a strong genetic contribution to the disease. Investigations need to be conducted on how genetic variations manifest as changes in gene expression profiles at the early stages of atherosclerotic development. Thus, this project aims to leverage machine learning methods to analyze tissue-level RNA-seq data of coronary and aortic atherosclerosis in young adults and predict a sample's pathology given its transcriptomic profile. Various feature selection methods were implemented to extract biologically relevant gene sets that were then used to train Extreme Gradient Boosting (XGBoost) classifiers to differentiate between early and late stages of atherosclerosis. The best performing models were trained on the top 90 and 120 differentially expressed genes yielding an accuracy of 85.96% and 86.54% for aortic and coronary samples respectively. Non-linear dimensionality reduction and gene ontology enrichment analysis were used to characterize the transcriptomic profiles representative of the gene sets determined by the feature selection methods. The results shown in this project lay the groundwork for the application of machine learning methods in the analysis of RNA-seq data at the tissue level which could eventually lead to applications in personal and preventative treatments for atherosclerotic development in young adults.

Keywords: Atherosclerosis, Machine Learning, RNA-Seq, Feature Selection.

Introduction

Development of Atherosclerosis in Young Adults

Cardiovascular disease is the leading cause of death, accounting for 32% of all deaths worldwide.[1] According to the Centers for Disease Control and Prevention (CDC), one person dies from heart disease every 36 seconds in the United States and about 655,000 Americans die from it each year due to heart attacks and strokes. Healthcare services, medications, and loss of productivity caused by cardiovascular diseases cost the United States $312.6 billion each year.[2] One of the main causes for these diseases is atherosclerosis; the hardening and narrowing of arteries due to the buildup of plaque derived from cholesterol, cellular waste products, loose smooth muscle cells, and macrophages. At the molecular level, hundreds of intra- and extracellular proteins jointly alter their cellular processes to remodel the local vascular environment of the artery, causing the development of lesions and plaques alongside arterial walls.[3] Acute rupture

of these lesions causes local thrombosis, leading to partial or total occlusion of the blood vessel and eventual vascular death. Severe clinical outcomes of this disease include ischemic heart disease, coronary artery disease, ischemic stroke, and peripheral arterial disease.[4]

Since atherosclerosis is predominantly a long-term asymptomatic condition until a plaque rupture occurs, it is difficult to diagnose and characterize in young adults. Atherosclerotic cardiovascular disease is commonly diagnosed for men older than 45 years and women older than 55 years of age. Current diagnostic and preventative methods focus mostly on antecedent risk factors such as hypertension, diabetes mellitus, smoking, and dyslipidemia, which are not deterministic of the molecular and genetic changes that occur during the early stages of atherosclerosis. To improve early disease detection in younger adults and prevent the progression of the disease prior to the manifestation of clinical symptoms, it is vital to investigate and identify the specific molecular patterns and pathways that constitute healthy versus atherosclerotic arterial tissues.[3]

Specifically, it is necessary to identify the key changes in gene expression that are associated with the initial development of atherosclerotic tissue in young adults. Heritability estimates for atherosclerosis and its associated diseases vary between 40% to 70%, suggesting a strong genetic contribution to the disease. Genome-wide association studies have identified 175 loci associated with an increased risk in atherosclerosis, affecting common risk factors such as blood lipid levels, nitric oxide signaling, and blood pressure. While some of these loci have been individually identified for their functionality, investigations still need to be conducted on how these genetic variations manifest as changes in gene expression profiles at the early stages of atherosclerotic development.[5]

Machine Learning for Tissue-Level RNA-Sequencing Analysis

Identification of differentially expressed genes and characterization of the different transcriptomic profiles between phenotypic disease states can be used to understand gene function and the molecular mechanisms underlying different biological processes associated with the development of the disease. The most popular method for analyzing differentially expression genes is RNA-seq because of its remarkable power and accuracy and the low cost of next generation sequencing technologies. RNA-seq analyzes transcriptomic profiles by comparing the abundance of RNA sequences extracted from target locations between treatment groups. However, RNA-seq can be highly variable as it is affected by cell types, proliferation, and differentiation status in the samples when investigating its changes at the level of entire tissues.[6] When obtaining information about a large number of genes from various cell types in the tissue, traditional statistical and correlational analyses are unable to capture the intertwining relationships between them. Hence, more advanced methods such as machine learning can be applied to better investigate these complex signatures hidden in the data.[7]

Machine learning (ML) is a multidisciplinary field that employs computer science, artificial intelligence, computational statistics, and information theory to build algorithms that can learn from large-scale datasets and make predictions on a new data set. It is increasingly becoming a key tool for biological studies including image analysis, robust phenotyping, and gene discovery. Therefore, ML methods can assist in the investigation of transcriptomic profiles and the identification of differentially expressed genes that are missed by regular RNA-seq analyses.[6] The transcriptomic patterns obtained from the systematic collection of arterial samples can lead to the development of personalized treatment strategies for the prevention of atherosclerosis in young adults with similar molecular sub-types.

In this study, we modified and optimized existing ML classifiers to analyze tissue-level genetic datasets of human coronary and aortic atherosclerosis using supervised ML algorithms and feature selection methods. The prediction models were designed to discriminate between early and late stages of atherosclerosis using RNA-seq data. Next, we analyzed and compared the top gene features selected by different feature selection methods on the performance of the ML model. Lastly, we identified differentially expressed genes in each of the methods and analyzed their functionality associated with early and late stages of atherosclerosis.

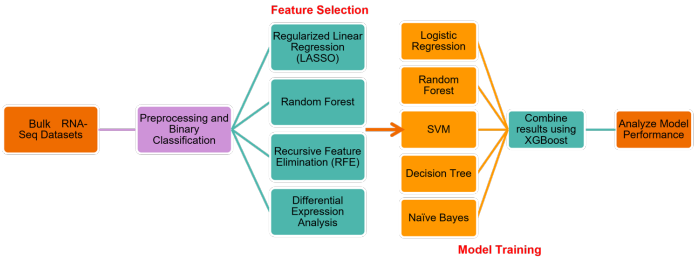

Fig. 1. ML-based RNA-seq analysis pipeline for atherosclerosis. This figure describes the design workflow for the entire project. Analysis of this pipeline was conducted through UMAP analysis of the dataset based on pathology score and model results, and gene ontology enrichment analysis of the feature selection methods.

Results

Data Acquisition and Preprocessing

The ML-based RNA-seq analysis pipeline is shown in Figure 1. The tissue-level RNA-seq datasets representing 131 coroner's autopsies of young adults were obtained from the Genomic and Proteomic Architecture of Atherosclerosis (GPAA) project. Separate analyses were conducted for data sampled from the left anterior descending coronary artery (LAD) and the distal abdominal aorta (AA). Table 1 shows the distribution of samples across each dataset according to disease pathology, sex, and race. Each of the samples in the original datasets consisted of 19,710 genes. After converting to FPKM values, removing near-zero variance features, and filtering out features with high collinearity ($r > 0.8$), the final AA dataset consisted of 243 samples x 13,696 genes and the LAD dataset consisted of 89 samples x 12,373 genes. These datasets are much larger than those used to conduct similar analysis in previous runs by Price et al (2021). By increasing the number of samples used to train the models, we hope to better generalize the ML model and improve its performance.

Table 1. Metadata counts for AA and LAD datasets. Binary class label was assigned based on pathology grading. Normal or early-stage atherosclerosis was defined by non-lesional (NL) or fatty streaks (FS) pathology. Diseased or late-stage atherosclerosis was defined by fibrous plaques (FP) and complex fibrous plaques (FC). Demographic factors shown below include age, sex (male/female), and race (white/black/Hispanic/Asian).

	AA	LAD
Class Label (Normal/Diseased)	162/81	63/26
Pathology Grading (NL/FS/FP/FC)	98/64/45/36	33/30/24/2
Total	243	89
Age	16-59	18-59
Sex (M/F)	164/79	54/35
Race (W/B/H/A)	180/54/9/0	63/22/3/1

Feature Selection

First, this project aimed to implement and characterize different feature selection methods which attempt to reduce the high dimensionality of the datasets and identify key features in the transcriptomic profile. The normalized datasets were randomly stratified and split into five 80% training sets and 20% testing sets for training, validating, and testing various feature selection methods. Different feature selection methods include L1 regularized linear regression (LASSO), random forest, and recursive feature elimination (RFE). Each algorithm provides unique, individual rankings to gene features. Figure 2 compares the stability of the feature selection methods by identifying the number of similar gene features between the top 240 ranked genes selected for each of the 5 dataset splits. For both the AA and LAD datasets, LASSO and RFE selected more than 86% of the same genes across dataset splits. However, random forest only selected 6.67% (16/240) and 1.67% (4/240) of the same gene features across splits for the AA and LAD datasets respectively. These results indicate that

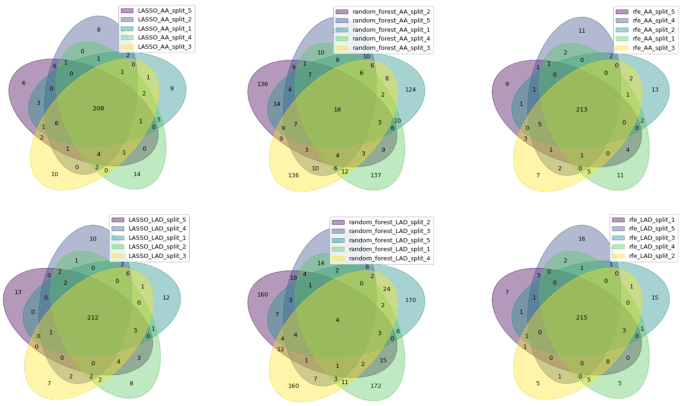

Fig. 2. Stability of feature selection methods. Venn diagrams show similarity between gene features across 5 splits of the AA (top row) and LAD (bottom row) datasets for LASSO, random forest, and RFE feature selection methods. Random forest is the least stable feature selection method.

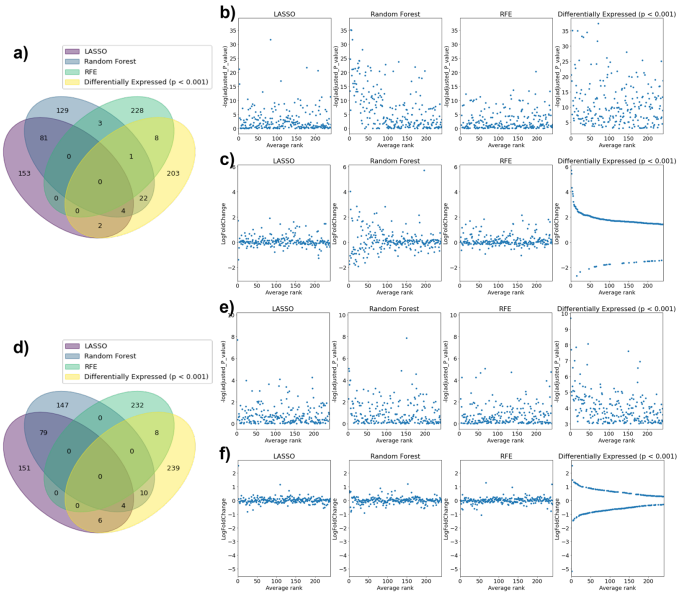

Fig 3. Comparison of feature selection methods to differentially expressed genes for AA (a, b, c) and LAD (d, e, f) datasets. Figures **a** and **d** compare similar genes across the top 240 ranked gene features for each feature selection method. Figures **b** and **e** describe the adjusted P-values for differentially expressed genes across the feature selection methods. Figures **c** and **f** describe the fold change values for differentially expressed genes across the feature selection methods.

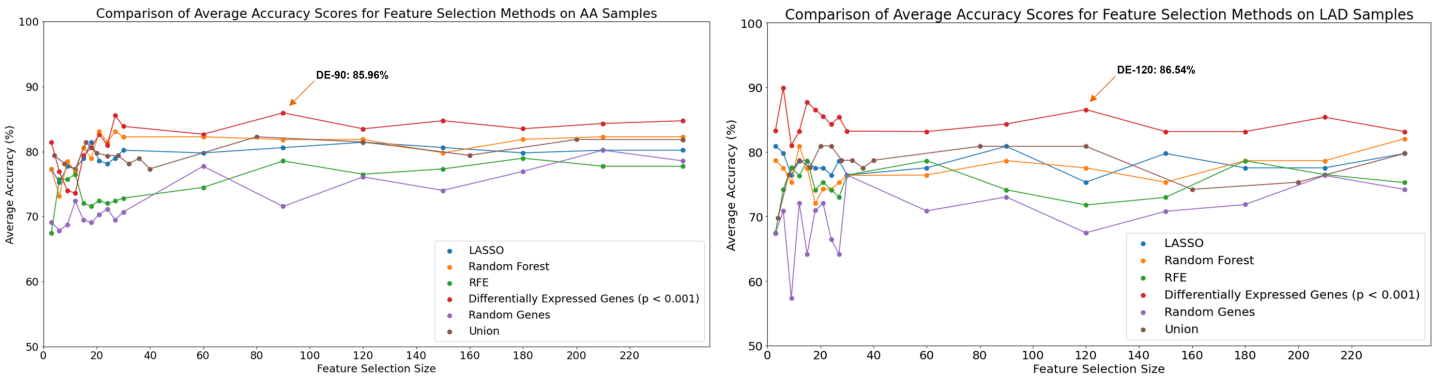

Fig. 4. Average accuracy scores for XGBoost models trained from gene sets of variable sizes selected by various feature selection methods on AA (left) and LAD (right) samples. The XGBoost model trained from top 90 and top 120 differentially expressed (DE) genes for AA and LAD datasets had the highest accuracy scores and overall model performance. Model performance of gene sets from other feature selection methods were not significantly different from a random set of genes; however, random forest did perform slightly better than other feature selection methods.

the random forest is the least stable of the feature selection methods.

Next, we conducted differential expression analysis using DESeq2 on the entire datasets and obtained 3,975 genes for AA samples and 267 genes for LAD samples that are differentially expressed between early and late stages of atherosclerosis (adjusted p-value = 0.001). The genes were sorted were ranked in descending order of the fold change magnitudes to obtain the top 240 differentially expressed features that were used to also train the ML classifier. Figure 3a and Figure 3d show the comparison of number of similar features between the top 240 ranked gene features of each feature selection method averaged across the five folds of the datasets. Gene features selection using random forest has the greatest number of similar features with the top 240 differentially expressed genes for both datasets. Figure 3b,c and Figure 3e,f show the adjusted p-value and fold change for the top 240 average ranked gene features selected by the different feature selection methods. Only random forest showed lower p-values and higher fold changes for higher ranked genes, a pattern similar to that of the top 240 differentially expressed genes. These results indicate that the random forest is more likely to identify differentially expressed genes.

XGBoost Model Performance

The project then aimed to implement and increase performance of an existing supervised ensemble learning classifier which differentiates between early and late stages of atherosclerosis. The classifier consisted of 5 shallow learners – Logistic Regression, Random Forest, Standard Vector Machine (SVM), Decision Tree, and Naïve Bayes – being trained on the top 3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 60, 90, 120, 150, 180, 210, or 240 of the selected features from each feature selection method using the same 5 folds of the dataset used earlier. The results from the shallow learners were then combined using Extreme Gradient Boosting (XGBoost) to increase model performance. The best average accuracy scores for each of the feature selection methods were obtained with the top 90 ranked gene features for the AA dataset and the top 120 ranked gene features for the LAD dataset as shown in Figure 4. The best performing classifiers were from the top 90 differentially expressed genes (accuracy = 85.96%) for the AA dataset and the top 120 differentially expressed genes (accuracy = 86.54%) from the LAD dataset.

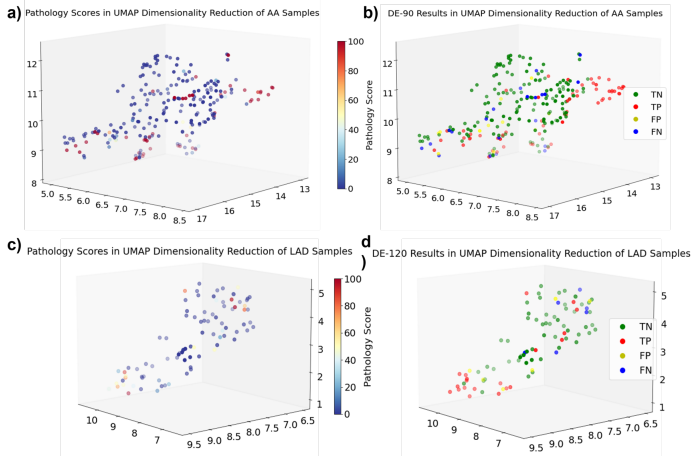

Fig. 5. UMAP dimensionality reduction and clustering based on pathology score (a, d) and confusion matrix of best performing model (b, d) for AA and LAD samples. Pathology score is defined as % involvement of FC in tissue samples, indicating that the pathology of the sample worsens as the score increases. Pathology-based clustering indicates that there are differences in gene expression profiles between early and late stages of atherosclerosis. Clustering on the confusion matrix allows for visual comparisons with pathology scores on the accuracy of the model in detecting non-linear associations.

UMAP (Unifold Manifold Approximation and Projection for Dimension Reduction)

UMAP was used to visualize the high-dimensional data from the gene expression datasets into three-dimensional space. The project aim was to improve model performance by identifying clusters of data which separate different atherosclerotic pathologies and compare those clusters with the outputs from the model. Such clusters were observed when evaluating pathology score (Figure 5a,c) and a few similarities were noted with the outputs of the best model, the top 90 and 120 differentially expressed genes for AA and LAD datasets respectively (Figure 5b,d).

Gene Ontology Enrichment Analysis

Finally, the project aimed to identify and characterize the functionality of the enriched genes obtained from the best models of each of the feature selection methods in their association with the development of atherosclerosis. Gene ontology enrichment analysis was conducted on the highest performing gene sets from each feature selection method and analyzed for its functionality in biological processes. Figure 6 shows the analysis for AA data and Figure 7 shows the analysis for LAD data. There were no statistically significant enriched terms ($p < 0.05$) for the top 120 ranked gene features for random forest and RFE feature selection on the LAD dataset.

Discussion

This project aimed to investigate various feature selection methods and leverage them to improve performance of supervised, ensemble machine learning classifiers that differentiate between early and late stages of atherosclerosis in young adults. The best feature selection method was found to be the ranked set of 90 and 120

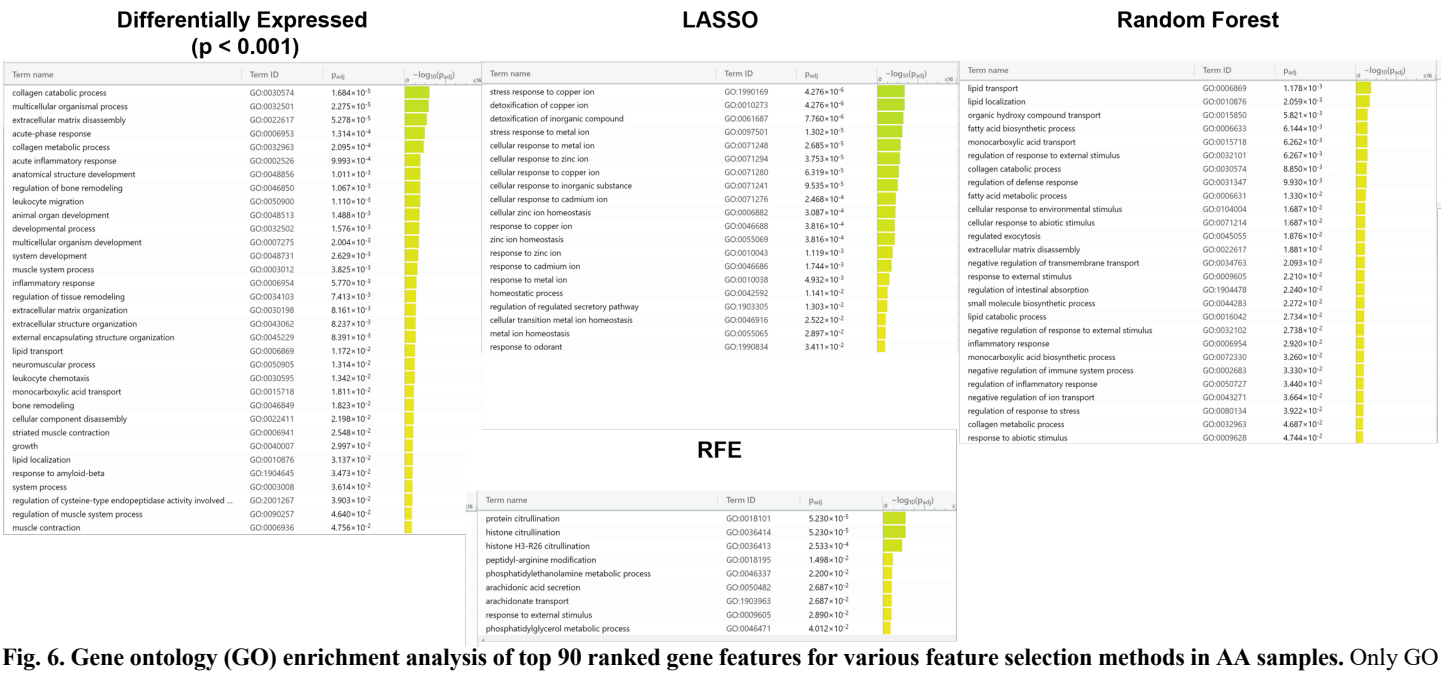

Fig. 6. Gene ontology (GO) enrichment analysis of top 90 ranked gene features for various feature selection methods in AA samples. Only GO biological processes are listed for enriched terms (adjusted P-value < 0.05).

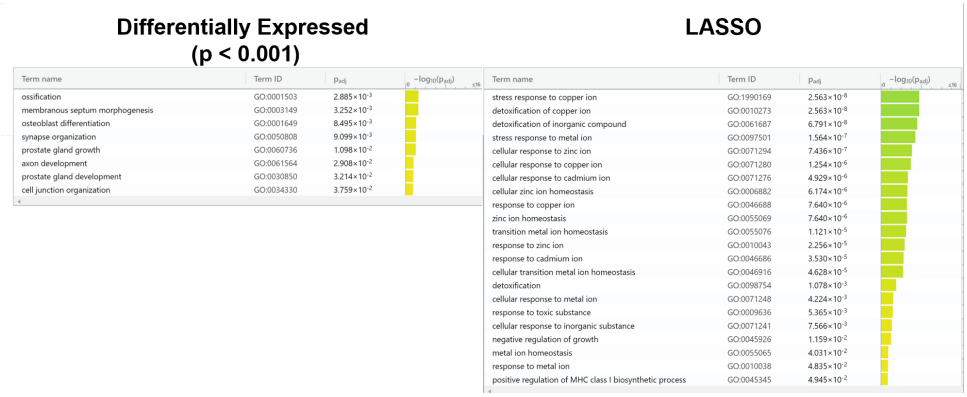

Fig. 7. Gene Ontology (GO) enrichment analysis of top 120 ranked gene features for various feature selection methods for LAD samples. Only GO biological processes are listed for enriched terms (adjusted P-value < 0.05). As such, there are no significant enriched terms for random forest and RFE feature selection.

differentially expressed genes ($p < 0.001$) for AA and LAD samples respectively. These findings indicate that traditional differential expression analysis is still able to differentiate transcriptomic profiles of RNA-seq data at the tissue-level and functions better than the ML-based feature selection methods. Additionally, the other feature selection methods show no difference in model performance compared to a set of random genes, which indicates that these methods are not able to identify functionally important sets of genes associated with early and late stages of atherosclerosis. A possible explanation for these results could be due to the extensive heterogeneity in the dataset caused by the analysis of RNA sequences from tissue-level samples. To obtain a balanced dataset that allows machine learning methods to account for this heterogeneity, a similar number of samples are needed relative to the number of genes. Thus, more data needs to be collected to train the ML-based feature selection methods and XGBoost models.

From the ML-based feature selection methods, the gene sets from random forest feature selection had the highest performance for the XGBoost classifiers. The average gene set was also the most closely associated with differentially expressed genes. However, stability analysis indicates random forest feature selection method to be the least stable with the fewest number of similar features between the folds of the dataset. A combination of high performance and a lack of stability indicates that the random forest algorithm is overfitting to each training dataset and is not able to generalize across datasets. Overfitting causes significantly lower model performance when the model predicts differences in atherosclerosis pathology based on external, novel data.

Non-linear dimensionality reduction and clustering using UMAP showed that the transcriptomic profiles of tissue-level samples are separable by atherosclerosis pathology as distinct groups of similar samples were observed in the three-dimensional UMAP space. Comparison of the pathology scores with the performance of the best XGBoost model showed similarities in clusters as expected. However, it is important to note the location of certain false positive samples that lie within clusters of true positive samples and certain false negative samples that lie within clusters of true negative samples. These samples may indicate the presence of labeling errors in the original dataset in which samples may have been assigned an incorrect pathology. Supervised UMAP clustering can be a potential method to use in the future to correct for these labeling errors and improve model performance.

Lastly, gene ontology enrichment analysis of gene sets in AA and LAD samples revealed unique functionality for each feature selection method. For both AA and LAD samples, the top 90 and 120 differentially expressed genes revealed biological processes involved in tissue and bone development and remodeling. Genes selected from LASSO feature selection revealed biological processes involved in inorganic ion homeostasis. Genes selected from random forest feature selection revealed processes involved in lipid metabolism and inflammatory response in AA samples. Genes selection from RFE feature selection revealed processes involved in protein citrullination, an irreversible post-translational modification that can lead to the production of autoantigens and an autoimmune reaction that could lead to the development of atherosclerosis.[8] Random forest and RFE feature selection methods did not reveal any significant enriched terms in LAD samples. It is likely that this is caused by the smaller sample size of the LAD dataset as there is not enough data for the feature selection methods to extract non-random information. These functionalities extracted by enrichment analysis are closely related with the development of atherosclerosis suggesting that the feature selection methods do identify functionally relevant genes from the dataset.[8] Further investigation must still be done to fully understand and characterize the mechanism for feature selection and why it reveals functionally significant information. It is important to note that LASSO and RFE feature selection methods identified relevant functional processes for atherosclerosis but were not associated with differentially expressed genes.

While this study is preliminary in nature, future availability of samples from a greater number of young adults, especially those samples representing late stages of atherosclerosis, may help in building more effective feature selection methods and ML classifiers to differentiate between transcriptomic profiles of atherosclerosis. Further investigation into the transcriptomic basis for the severity of disease pathology is necessary to characterize these genetic profiles and eventually develop personalized, preventative strategies for atherosclerosis in the future.

Materials and Methods

This project employed the workflow shown in Figure 1 to achieve the overall goal of designing machine learning methods for tissue-level RNA-seq analysis of atherosclerosis in young adults.

Data Acquisition

RNA-sequencing data was acquired from previous research conducted for the GPAA project.[3,9] Samples were extracted from the left anterior descending coronary artery (LAD) and abdominal aorta (AA) through coroner's autopsies of 131 young adults (Table 1). Samples were

graded by pathologists and scored as non-lesional or normal (NL), fatty streaks (FS), fibrous plaques (FP), or complex fibrous plaques (FC). These classifications increase in size of the plaques within the arterial samples and thus increase in the severity of atherosclerosis. RNA was extracted from the samples and sequenced at the tissue level.

Data Preprocessing

In processing the data for feature selection, the samples were grouped into a binary label based on its pathological classification: normal/early stage atherosclerosis (NL and FS) and diseased/late stage atherosclerosis (FP and FC). Then, the RNA-sequencing data was processed and quantified to produce fragments per kilobase of exon per million mapped fragments (FPKM) values which represent the abundance of each gene within a sample. Feature dimensionality reduction techniques were applied to increase readability of the dataset and remove noise. These techniques included removing features with near-zero variance and filtering out features with high collinearity ($r > 0.8$). The resulting values were then standardized across samples using a z-score normalization. Preprocessing steps were implemented using Python 3.6 and R 4.1.1.

Feature Selection

Feature selection is an advantageous step in handling large datasets and reducing its dimensionality before training a machine learning model. Feature selection helps in searching for a subset of relevant gene features associated with normal or diseased classification of atherosclerosis. To reduce the dimensionality of the dataset and extract functionally relevant genetic information, three different feature selected algorithms were implemented that provide individual rankings to gene features: LASSO, random forest, and RFE. The least absolute shrinkage and selection operator (LASSO) conducts regression analysis to simultaneously estimate parameters and select important variables. Random forest uses decision tree-based strategies to rank features based on a feature importance attribute. Recursive feature elimination (RFE) utilizes recursion for feature extraction where smaller and smaller sets are considered as features until a desired number of features is returned. Initial implementations of these feature selection methods were provided by Colin Price using the Python 3.6 scikit-learn library. The methods were trained on stratified 5-fold splits of the preprocessed FKPM dataset with an 80/20 train-test split. To obtain a consensus "average" ranking of gene sets between splits, the overall mean ranking of each feature was calculated.

Differential Expression Analysis

Differential expression analysis is a common statistical method used to analyze gene counts from RNA-seq data and identify overexpressed or under-expressed genes between treatment groups. This analysis was implemented on the raw gene counts of the datasets using the DESeq2 package in R 4.1.1. Differentially expressed genes were filtered at an adjusted p-value < 0.001 and the genes were ranked with decreasing magnitude of the fold change.

Training and Evaluation of XGBoost Classifiers

Ensemble learning methods such as XGBoost involve the consolidation of performance results from individual, "shallow" learners to increase classification accuracy while preventing overfitting on the dataset. For each split of the dataset, the top 3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 60, 90, 120, 150, 180, 210, or 240 ranked gene features from each feature selection method were used to train five shallow learners: Logistic Regression, Random Forest, Standard Vector Machine (SVM), Decision Tree, and Naïve Bayes. The results of these individual models were then used to train the XGBoost model to classify between normal and diseased states of atherosclerosis in samples. Performance of the model was analyzed using confusion matrices and accuracy scores (Equation 1). Initial implementations of the shallow learners and XGBoost classifier were provided by Colin Price using the Python 3.6 scikit-learn and xgboost libraries.

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN} \quad [1]$$

UMAP (Unifold Manifold Approximation and Projection for Dimension Reduction)

UMAP is a general-purpose non-linear dimensionality reduction algorithm for the visualization of high-dimensional datasets in three-dimensional space. The algorithm searches for a low-dimensional projection of the data that has the closest possible equivalent fuzzy topological structure. Pathology scores and confusion matrices obtained from XGBoost models were then visualized to observe clustering of samples in the low-dimensional projection of the dataset. UMAP was implemented using the Python 3.6 umap-learn and scikit-learn libraries.

Gene Ontology Enrichment Analysis

Gene ontology (GO) enrichment analysis was performed on the list of genes returned by feature selection

methods to characterize its biological functionality and how the transcriptomic profile plays a role in the development of atherosclerosis. Statistically significant ($p < 0.05$) enriched GO terms were displayed for biological processes. Enrichment analysis was performed using the web-based g:Profiler tool using a background gene list of the *Homo sapiens* genome from Ensembl (GRCh38.p13).[10]

End Matter

Author Contributions and Notes

J.H.M, A.S.W, and C.M. designed research, J.H.M performed research, J.H.M wrote software and analyzed data; C.P. developed initial ML pipeline. A.S.W and C.M. suggested strategies, interpreted data, and advised on analyses, and J.H.M wrote the paper. F.J, D.T. and R.H. provided the graded samples. J.H., D.K. generated RNA-Seq data. D.H., Y.W., R.H., J.E. and C.M. led the GPAA project.

Acknowledgments

We thank Dustin Machi for IT support. We thank Shannon Barker and Vignesh Valaboju for guidance on improving feasibility of the project for Capstone.

This work was supported by R01HL111362 from the National Heart Lung and Blood Institute of the National Institutes of Health.

References

1. Cardiovascular diseases. https://www.who.int/health-topics/cardiovascular-diseases.
2. Pahwa, R. & Jialal, I. Atherosclerosis. in *StatPearls* (StatPearls Publishing, 2022).
3. Herrington, D. M. *et al.* Proteomic Architecture of Human Coronary and Aortic Atherosclerosis. *Circulation* **137**, 2741–2756 (2018).
4. Herrington, W., Lacey, B., Sherliker, P., Armitage, J. & Lewington, S. Epidemiology of Atherosclerosis and the Potential to Reduce the Global Burden of Atherothrombotic Disease. *Circ. Res.* **118**, 535–546 (2016).
5. Aherrahrou, R. *et al.* Genetic regulation of human aortic smooth muscle cell gene expression and splicing predict causal coronary artery disease genes. 2022.01.24.477536 (2022) doi:10.1101/2022.01.24.477536.
6. Wang, L., Xi, Y., Sung, S. & Qiao, H. RNA-seq assistant: machine learning based methods to identify more transcriptional regulated genes. *BMC Genomics* **19**, 546 (2018).
7. Roy, S., Kumar, R., Mittal, V. & Gupta, D. Classification models for Invasive Ductal Carcinoma Progression, based on gene expression data-trained supervised machine learning. *Sci. Rep.* **10**, 4113 (2020).
8. Stachowicz, A. *et al.* Abstract 11607: Protein Citrullination Landscape of Human Coronary Atherosclerosis. *Circulation* **144**, A11607–A11607 (2021).
9. Price, C. *et al.* Abstract 13172: Using Machine Learning to Identify Transcriptomic Biomarkers That Differentiate Early vs Late Stage of Atherosclerosis. *Circulation* **144**, A13172–A13172 (2021).
10. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).