# A Transcriptome-Based Machine Learning Approach to Classify Mitochondrial Health in Single Cells


## Abstract

*Mitochondrial dysfunction is a hallmark of numerous diseases and cellular stress responses. In this project, I am developed a machine learning-based classifier to distinguish between healthy and dysfunctional mitochondria using single-cell RNA sequencing data from LPS-treated and fresh peripheral blood mononuclear cells (PBMCs). I used the Seurat Package in R to process the data to select only the mitochondrial features from the data. These features were then used to train and evaluate multiple machine learning models. Our results demonstrate that specific mitochondrial genes are key discriminators of dysfunctional states. This approach provides a scalable framework for leveraging scRNA-seq and ML to detect mitochondrial dysfunction in heterogenous cell populations*


## Introduction

Mitochondria, known as the "powerhouse of the cell" are crucial organelles responsible for producing ATP through oxidative phosphorylation. Beyond this, they play key roles in calcium signalling, apoptosis and reactive oxygen species (ROS) regulation. Mitochondrial dysfunction is implicated in numerous diseases including neurodegenerative diseases, metabolic syndromes and cancer. Common features of mitochondrial dysfunction include reduced ATP production, increased ROS generation, altered mitochondrial membrane potential among others.

Detecting and classifying mitochondrial health status can be challenging, especially at the single cell level. Recent advancements in single-cell RNA sequencing (scRNA-seq) provides a transcriptome-wide snapshot of individual cells, offering a high-resoltion view into mitochondrial gene expression patterns. However, due to the complexity and volume of scRNA-seq data, machine learning (ML) methods are increasingly used for this purpose.

ML models can learn patterns in gene expression that correlate with healthy or dysfunctional mitochondria. Algorithms such as random forests, support vector machines (SVMs) or deep neural networks (DNNs) are trained on labelled datasets and are therefore "supervised learning". On the other hand, algorithms like clustering reveal natural groupings in the data without labels and are therefore called "unsupervised learning".

For this study, I used the data generated by Derbois et.al.,2023. They sequenced peripheral blood mononuclear cells (PBMCs) which are target to lipopolysaccharide (LPS) treatment. PBMCs are a diverse mixture of immune cells isolated from blood. When treated with LPS, which is a potent immune stimulator, the cell avtivates inflammatory pathways such as NF-κB and MAPK which leads to mitochondrial membrane depolarisation, ROS production

etc. This makes LPS-treated PBMCs a useful model for evaluating mitochondrial health and for training ML models to classify them.

## Methodology

### Data preprocessing

The datasets used to correspond to samples from GEO Series GSE226488 (PMID: 37414801), specifically GSM7077866 (LPS-treated) and GSM7077867 (control). Raw gene matrices were downloaded in the 10x genomics format and processed using the Seurat package. Cells were filtered to retain those with 100-7500 detected genes and <30% mitochondrial content, removing potential doublets and low-quality reads. Mitochondrial genes were with the help of MitoCarta dataset.

### Differential Expression Analysis

Unsupervised clustering was performed on the mitochondrial gene-only dataet using PCA-based dimensionality reduction followed by neighbourhood graph construction. Clusters 1 and 3 were isolated and used to asses differential expression of OXPHOS and mitochondrial encoded genes between control and LPS-treated cells. To infer cell identities, the SingleR package was used in conjunction with the Human Primary Cell Atlas as a reference.

### Machine learning model construction

The mitochondrial gene expression matrix was exported and loaded into Google Colab for machine learning using Python and scikit-learn. The following ML classifiers were trained: support vector machine, logistic regression (elasticnet), decision tree classifier and random forest classifier. The data was split into training/test sets with 80/20 split. Hyperparameters for each model was tuned using GridsearchCV with 3-fold cross validation. Model performance was evaluated on the test set using accuracy, precision, recall, F1 score and ROC-AUC. Feature importance was computed using model-specific attributes: '.feature_importances_' for tree-based models and absolute coefficients for linear models. The top 20 genes from each model were aggregated, and the top 10 consensus genes (by average importance) were exported for gene ontology analysis.

### Functional Enrichment Analysis

The top 10 genes identified from the ML models were analysed using Gene Ontology (GO) enrichment in R, using the package. GO terms related to biological processes, molecular functions, cellular components and reactome enrichment were identified, with p-values adjusted using the Benjamini-Hochberg method. They were used to infer the potential functional consequences of mitochondrial gene expression in LPS-treated PBMCs.

## Results

The following results explore the computational workflow and analytical outcomes of our study, starting with quality control and mitochondrial gene expression in scRNA-seq data. Subsequent machine learning classification performance and functional enrichment analyses. Figure 1 depicts the analysis pipeline, outlining the steps taken to classify mitochondrial dysfunction from scRNA-seq data. The raw data underwent quality control and preprocessing using Seurat, followed by extraction of mitochondrial gene expression data.

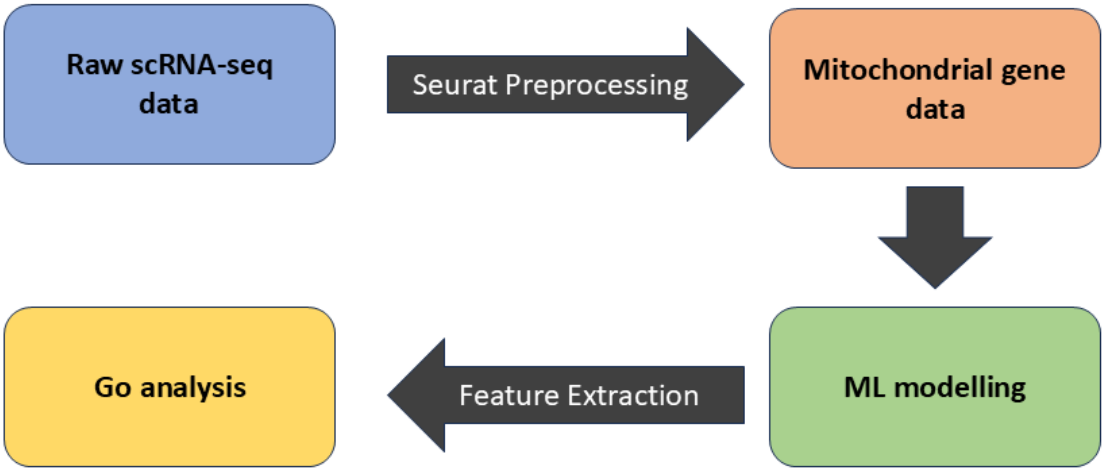


*Figure 1* Workflow

Figure 2 represents violin plots comparing the total genes, total RNA counts and mitochondrial gene expression percentage between control and LPS-treated cells. Slight differences between them were observed for the total unique genes per cell with LPS-treated cells showing slightly lower values as opposed to the control. This could be because of the immune reactions activated due to the treatment. However, the total number of molecules was higher for LPS-treated sample, indicating a possible increased expression of some genes. The percentage of total counts that come from mitochondrial genes are slightly lower for LPS-treated cells as opposed to the control cells. However, since we are only extracting the mitochondrial gene expression, we will be able to explore the mitochondrial genes that are key to mitochondrial dysfunction caused due to LPS-treatment.

First, the filtered dataset was normalised using LogNormalize, after which the top 2000 variable features were extracted. The dataset was then scaled for all features. MitoCarta3.0 was used to specifically get the mitochondrially-encoded and nuclear-encoded mitochondrial genes. For nuclear-encoded genes, I used the genes involved in the oxidative phosphorylation (OXPHOS) pathway. From here on, oxphos genes would used to indicate nuclear-encoded mitochondrial genes. The scaled dataset was filtered to included only the mitochondrial genes (all of them). This data was then exported for machine learning model training.

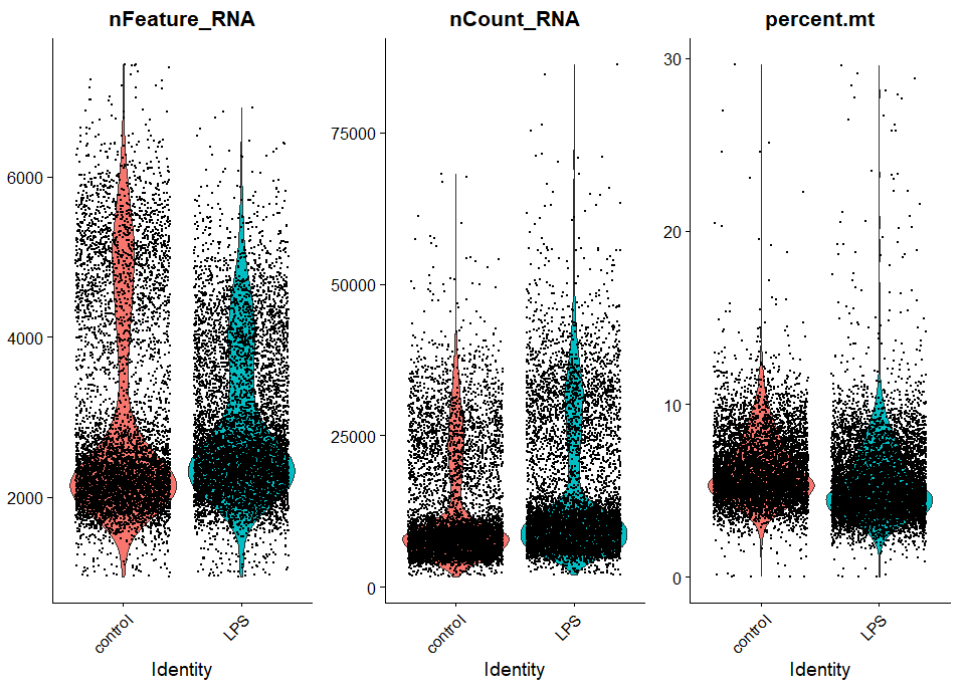


*Figure 2* Gene expression quality control

Four models: Random Forest, Decision Tree, Logistic Regression (elasticnet) and Support Vector Machine, were trained on the exported mitochondrial dataset. Condition (control vs LPS) was used as the label. The models were trained with hyperparameter tuning using gridsearchCV with 3-fold cross-validation. Owing to limited computational resources, the hyperparameters to be checked were kept to the basic values.

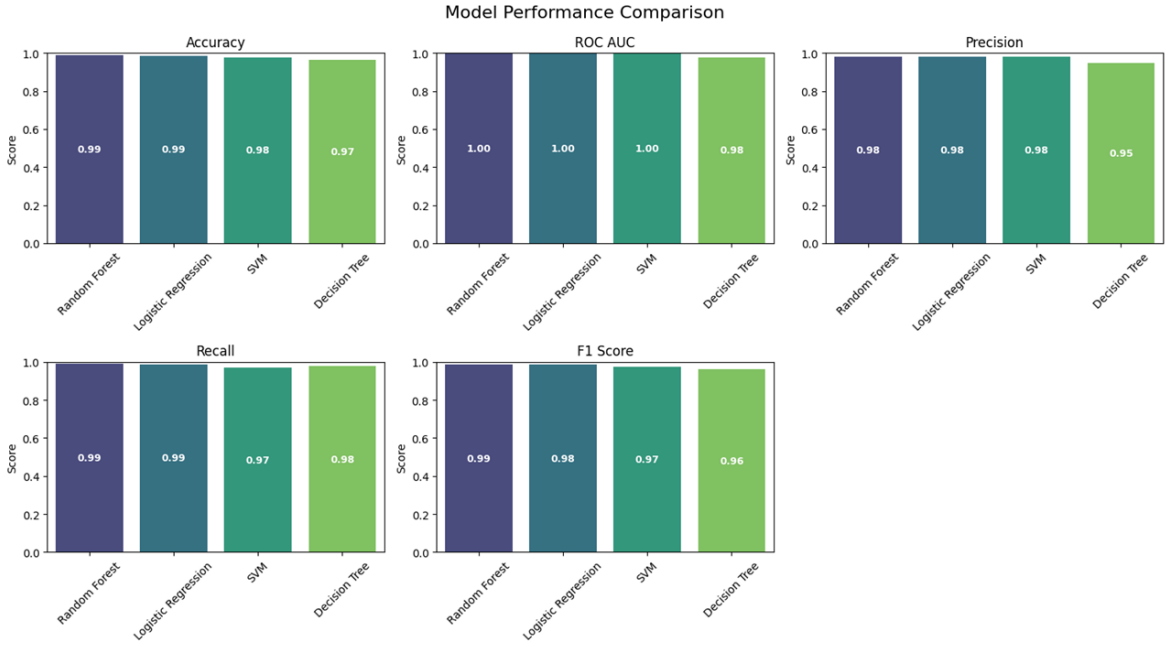


*Figure 3* Evaluation Metrics of the Models

Model performance was evaluated based on the test set using accuracy, precision, recall, F1 score and receiver operating characteristic curve (ROC-AUC) shown in figure 3. I observed

superior results for all the models which could one of two scenarios. The first one is that the classes are truly distinct and the features used very of the highest quality. This possibility is highly unlikely. Thus, it could potentially be due to data leakage, overfitting, class imbalance or improper test-train split. The confusion matrices (Figure 4) and ROC-AUC curve (Figure 5) indicate that the split is not perfect. This leaves with potential overfitting issues. I intend to correct this with regularisation or checking with random shuffling of the labels to see if the results still hold. Furthermore, this could also be a result of just using the mitochondrial genes alone with a combination of high feature-to-sample ration with many features (genes) and few samples (cells).

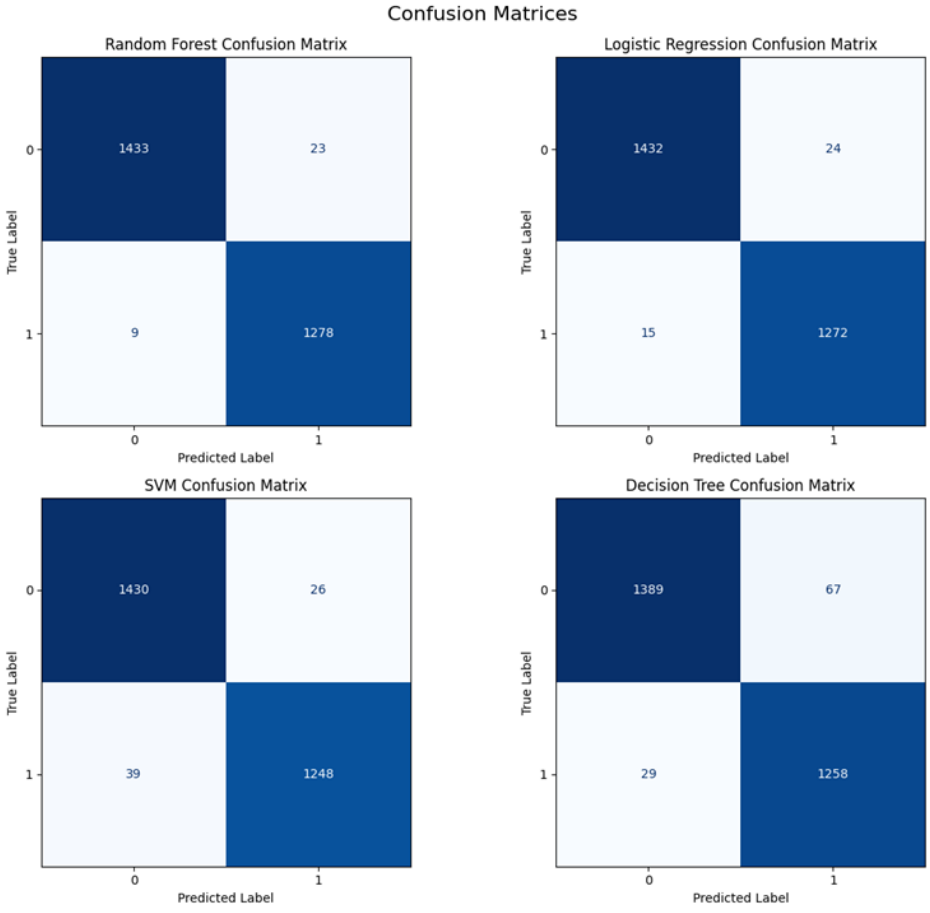


*Figure 4* Confusion Matrices of the four models

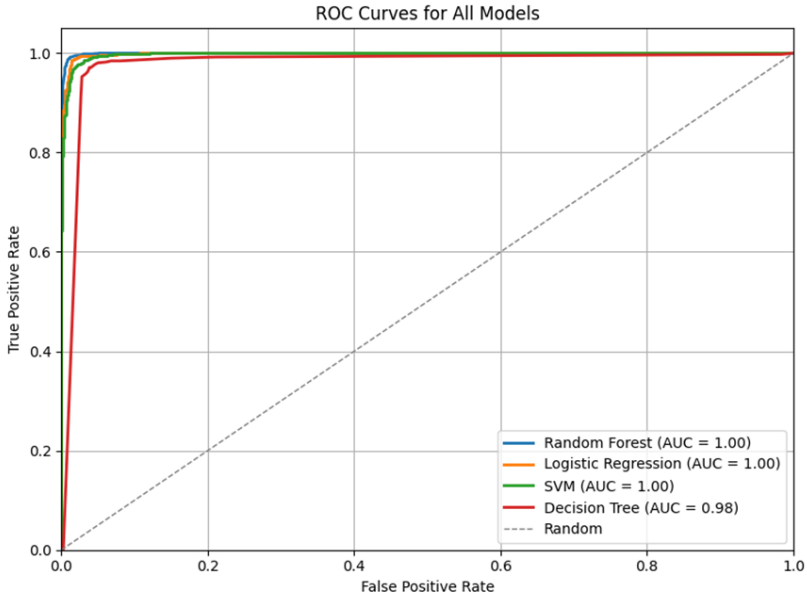


*Figure 5* ROC-AUC curves for all models

The top 10 features that contributed to the prediction of each model was calculated (Figure 6). For random forest and decision trees, it was done using the ".feature_importances_" and for logistic regression and SVM, it was calculated using the magnitude of the model coefficients. Some common features are seen for all models. These include CMPK2 (cytidine/uridine monophosphatase kinase 2), PNPT1 (Polyribonucleotide nucleotidyltransferase 1) and PPM1K (Protein phosphatase, Mg2+/Mn2+ dependent 1K). All three are important for mitochondrial homeostasis with roles in immunomodulation as well. This corroborates with the LPS-treated immune activation.

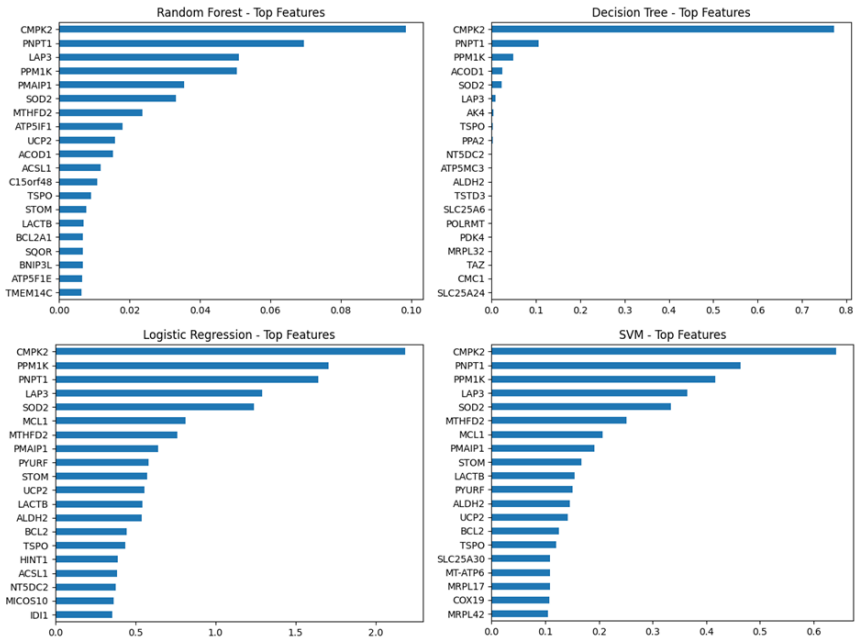


*Figure 6* Top features for each model

To gain insights into the biological relevance of the top ranked mitochondrial genes, I performed gene ontology (GO) enrichment analysis across the three domains: Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). Additionally, I also carried out the Reactome pathway enrichment to uncover associated signalling and metabolic pathways. For Biological process, the significant GO terms were for apoptosis and related process (Figure 7). This could also be because of the inclusion dead cells due to improper filtration or a result of immune activation in the LPS-treated cells. Further analysis with different filters for mitochondrial gene percentage could offer insights.

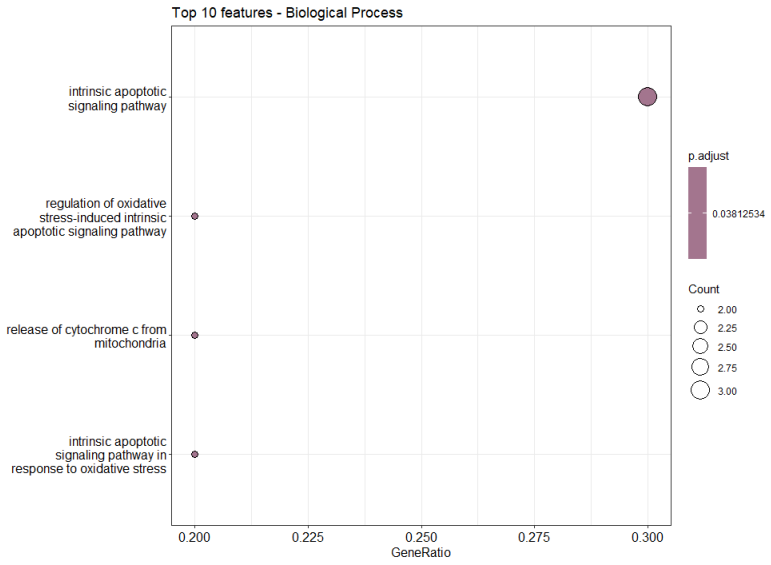


*Figure 7* GO terms for Biological Process

Similarly, for molecular function only “manganese binding” got any hits among the top 10 features (Figure 8). This is from SOD2 (Superoxide Dismutase 2), that protects cells from oxidative stress and is essential for aerobic life. It catalyzes the conversion of superoxide radicals into hydrogen peroxide, thus controlling ROS levels. Altered SOD2 levels have been implicated in various diseases making it a potential candidate.

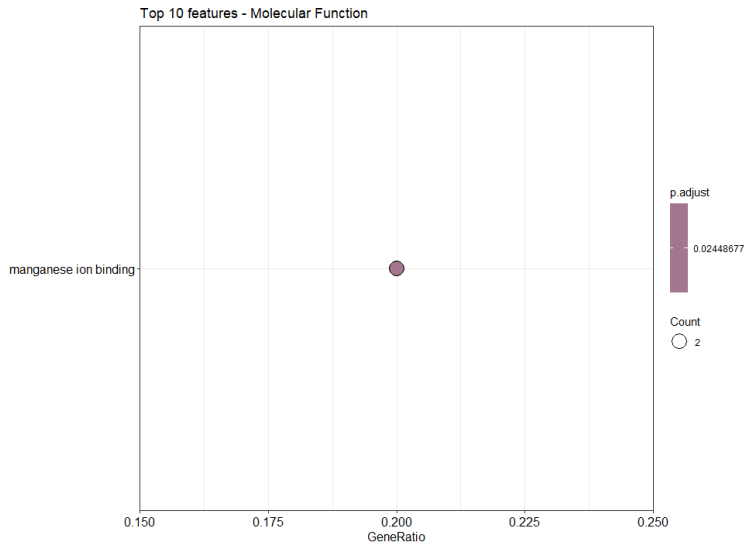


*Figure 8* GO terms for Molecular Function

For cellular components, as most of the top features were present in the mitochondrial matrix, or the outer membrane (Figure 8). This could be because proteins in these compartments are most affected by membrane depolarisation caused by LPS-treatment and a hallmark for mitochondrial dysfunction.

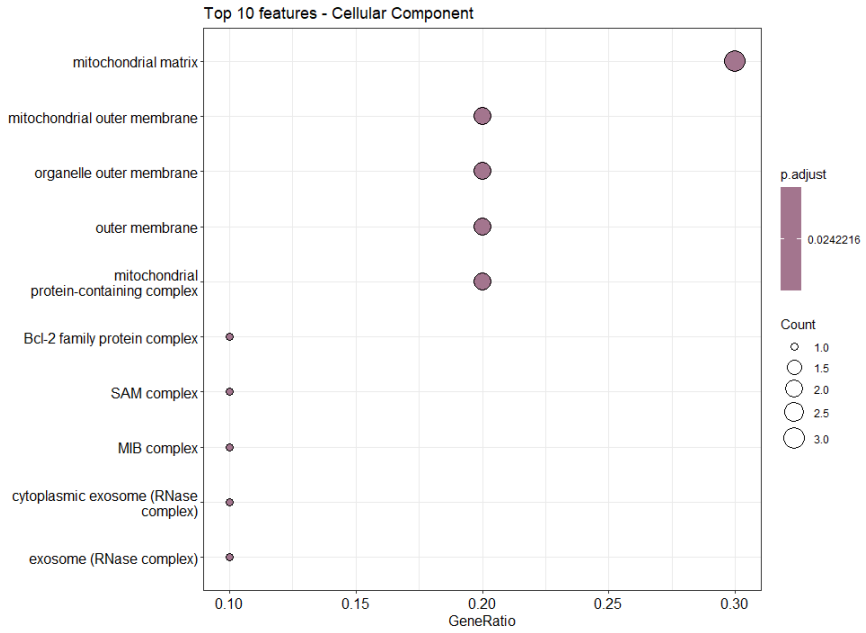


*Figure 9* GO terms for Cellular Compartment

For reactome enrichment, I observed hits primarily in mitochondrial biogenesis (Figure 10). This makes sense as during stress, one of the primary signals is to stop biogenesis of mitochondrial proteins and activate degradation machineries.

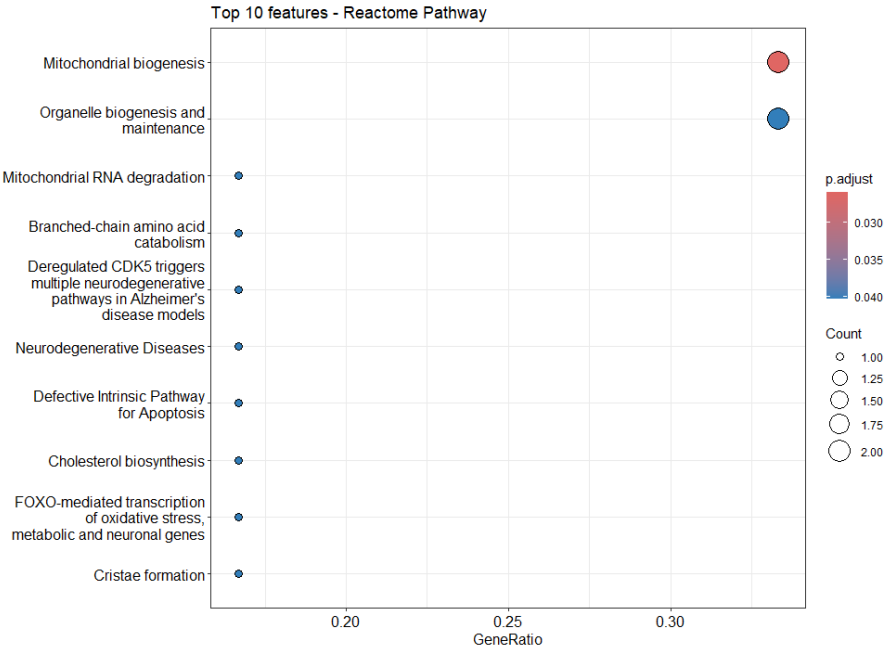


*Figure 10* GO terms for Reactome Enrichment

I was curious to look into how the mitochondrial and nuclear encoded gene expression in control vs LPS-treated cells. For this I first performed clustering to observe how well both samples separated. I observed that there was 2 completely separated clusters while one cluster contained a mixture of both (Figure 11). I isolated the genes present in clusters 1 and 3 for further processing.

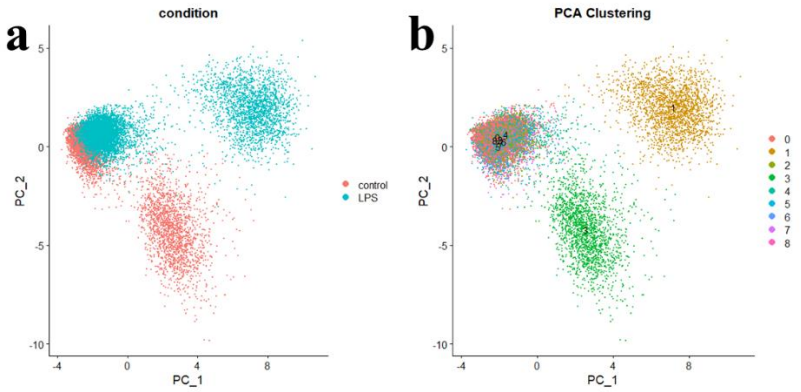


*Figure 11* Clustering based on a) condition b) distinct clusters

For the genes in clusters 1 and 3, I checked for the top 5 most variable genes that encoded for the oxphos proteins and the mitochondrially encoded genes. What I observed was that in LPS-treated cells, oxphos gene expression was significantly reduced (statistical tests yet to be performed), while the mitochondrial genes were upregulated. For control cells, both sets of genes showed reasonable expression (Figure 11).

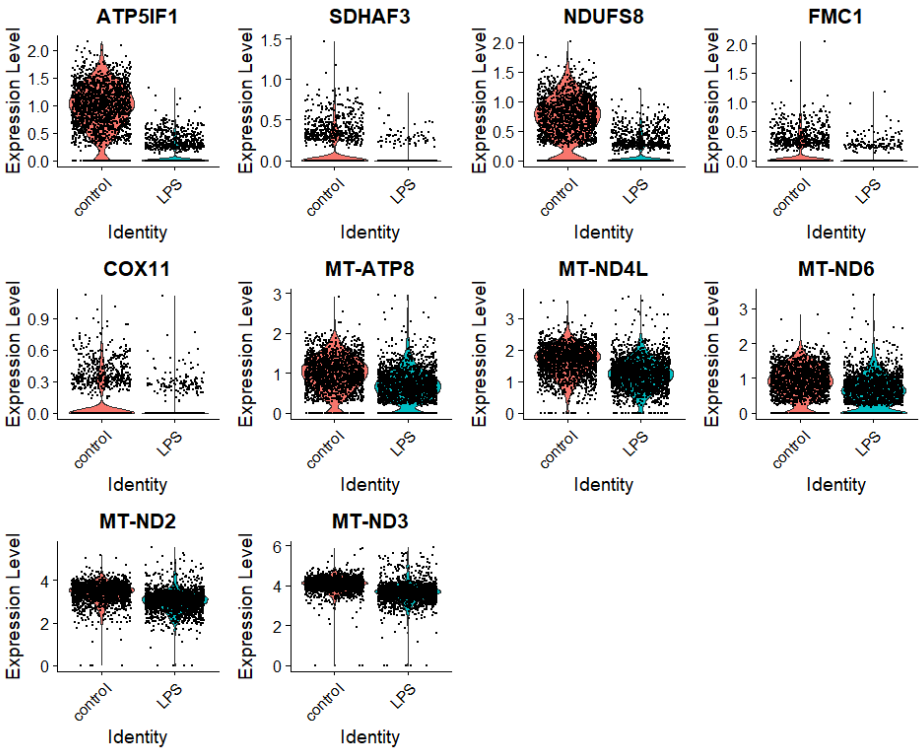


*Figure 12* Differential expression of 5 oxphos and mitochondrially encoded genes in control vs LPS-treated cells

This could imply that upon mitochondrial dysfunction, the gene expression of mitochondrial encoded genes is upregulated, while that of nuclear-encoded mitochondrial genes (speficially the ones coding for OXPHOS proteins) are downregulated.

Furthermore, out of sheer curiosity, I checked for which cell types among PBMCs were present in cluster 1 and 3 (indicating LPS and control cells, respectively). For this I annotated clusters 1 and 3 specifically with SingleR package from Bioconductor in R. I observed that Natural killer cells, Neutrophils, dendritic cell and macrophages were the cells in LPS-treated cluster. This is expected considering the immune pathway activation that these cells employ upon treatment. For control cells, Myelocyte, Endothelial cells, CD34- cells were the ones represented among others.

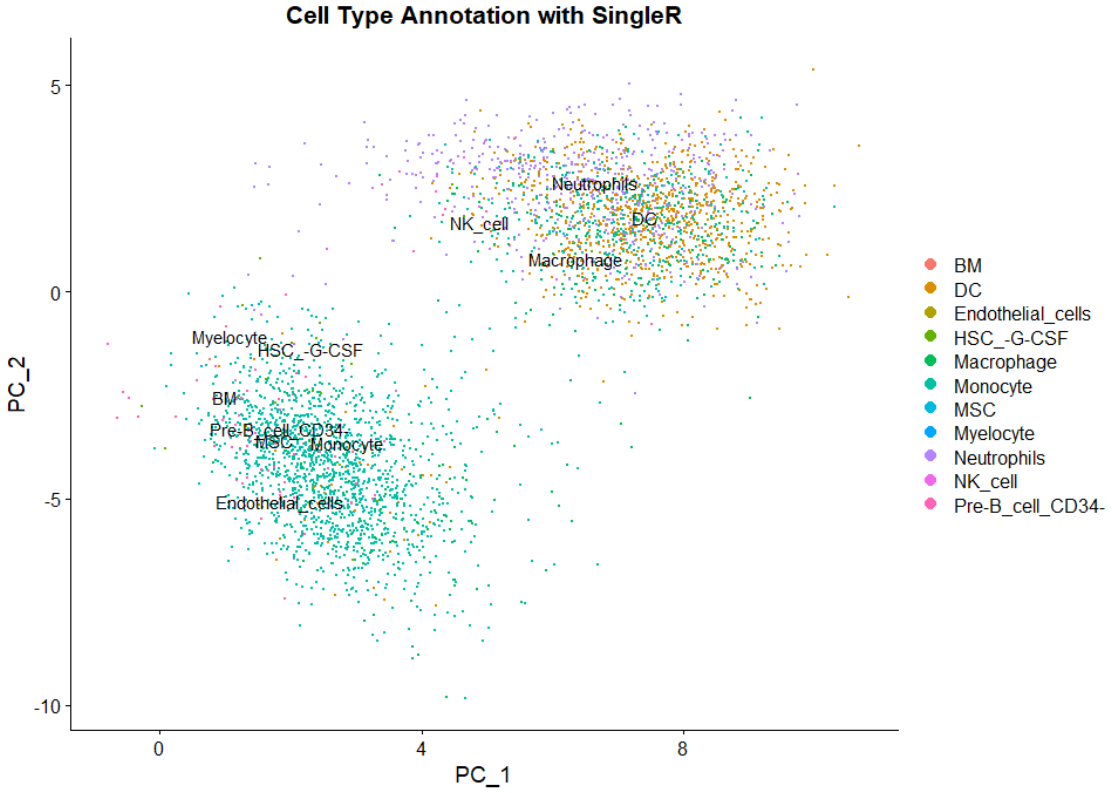


*Figure 13* Cell types associated with clusters 1 and 3

## Conclusions and Future Directions

Through this study, I aimed to classify healthy and dysfunctional mitochondria using a machine learning approach on single-cell RNA-seq data, focusing on mitochondrial gene expression profiles. Four ML classifiers were trained and evaluated. The results demonstrated a high performance across models suggesting either high discriminatory power or issues with data leakage, dataset size or overfitting. These are issue to addressed in the future with an extended dataset included non-mitochondrial genes as well. High performance of the models could also be tested with a 5- or 10-fold cross validation as opposed to the 3-fold done in this study. Feature importance analysis revealed key genes associated with mitochondrial dysfunction which were subjected to gene ontology and reactome pathway enrichment analysis. The results highlighted significant process, functions involved in mitochondrial homeostasis and cell survival. Furthermore, a differential gene expression study revealed interesting insights into the expression of OXPHOS genes and mitochondrial-encoded genes in control vs LPS treated cells. Future studies will be aimed at exploring the omitted clusters and understand it's biological significance.

While the initial findings are promising, it is important to acknowledge the limitations of the current study, including the small dataset size and potential overfitting. Future work will be foucsed on increasing the dataset size to improve the generalizability of the machine learning models. Additional datasets, particularly, those with diverse experimental conditions and cell types, could enhance the robustness of the classifiers. Furthermore, integrating multi-omics data such as proteomics and metabolomics could provide a more comprehensive understanding of mitochondrial dysfunction.

In addition to all this, a deeper investigation needs to performed to understand the specific roles of the identified genes through experimental validation and functional assays. Finally, developing a user-friendly platform or tool that integrates these findings into a resource for researchers could facilitate the broader application of these findings in both research and clinical settings.

## References


1. Derbois C, Palomares MA, Deleuze JF, Cabannes E, Bonnet E. Single cell transcriptome sequencing of stimulated and frozen human peripheral blood mononuclear cells. Sci Data. 2023 Jul 6;10(1):433. doi: 10.1038/s41597-023-02348-z. PMID: 37414801; PMCID: PMC10326076.
2. Goni L, Qi L, Cuervo M, Milagro FI, Saris WH, MacDonald IA, Langin D, Astrup A, Arner P, Oppert JM, Svendstrup M, Blaak EE, Sørensen TI, Hansen T, Martínez JA. Effect of the interaction between diet composition and the *PPM1K* genetic variant on insulin resistance and β cell function markers during weight loss: results from the Nutrient Gene Interactions in Human Obesity: implications for dietary guidelines (NUGENOB) randomized trial. Am J Clin Nutr. 2017 Sep;106(3):902-908. doi: 10.3945/ajcn.117.156281. Epub 2017 Aug 2. PMID: 28768654.
3. Pawlak JB, Hsu JC, Xia H, Han P, Suh HW, Grove TL, Morrison J, Shi PY, Cresswell P, Laurent-Rolle M. CMPK2 restricts Zika virus replication by inhibiting viral translation. PLoS Pathog. 2023 Apr 19;19(4):e1011286. doi: 10.1371/journal.ppat.1011286. PMID: 37075076; PMCID: PMC10150978.
4. Guan C, Zou X, Yang C, Shi W, Gao J, Ge Y, Xu Z, Bi S, Zhong X. Polyribonucleotide nucleotidyltransferase 1 participates in metabolic-associated fatty liver disease pathogenesis by affecting lipid metabolism and mitochondrial homeostasis. Mol Metab. 2024 Nov;89:102022. doi: 10.1016/j.molmet.2024.102022. Epub 2024 Aug 31. PMID: 39218215; PMCID: PMC11414560.
5. Zhuang A, Yang C, Liu Y, Tan Y, Bond ST, Walker S, Sikora T, Laskowski A, Sharma A, de Haan JB, Meikle PJ, Shimizu T, Coughlan MT, Calkin AC, Drew BG. SOD2 in skeletal muscle: New insights from an inducible deletion model. Redox Biol. 2021 Nov;47:102135. doi: 10.1016/j.redox.2021.102135. Epub 2021 Sep 14. PMID: 34598016; PMCID: PMC8487078.