Building A Classifier For MoTrPAC Endurance Training Data Based on Lipoprotein Receptor Levels in the Gastrocnemius

I. INTRODUCTION

In an effort to "generate a molecular map of the effects of exercise and training", the Molecular Transducers of Physical Activity Consortium has created the MoTrPAC Data Hub[1]. This data repository contains publicly available endurance training data that contains phenotype, transcriptomics, metabolomics, and proteomics data for 20 different tissues in young adult rats, across five time points. The endurance training group samples were collected after one week, two weeks, four weeks, and eight weeks. The control group was a group of untrained, sedentary rats.

In the associated paper, "Temporal dynamics of the multi-omic response to endurance exercise training", the researchers note that the results of the endurance training on the rats were greatly positive, expressing that there were effects such as a decrease in inflammation and a decrease in the expression of transcripts that indicate genetic risk for IBD[2]. In terms of the metabolic response, it was found that the liver, heart, and lung had the largest numbers of significantly enriched metabolite classes. The paper also notes that the levels of metabolites related to ether lipid and glycerophospholipid metabolism increased in the gastrocnemius skeletal muscle. The gastrocnemius, which runs along the back of the lower leg, is essential for the propulsion and stability required for running[3].

Given that the gastrocnemius is fundamental to running and given that the previous research done highlights significant metabolite enrichment related to lipid metabolism in this tissue, I explored creating a classifier based on the levels of lipoprotein receptors in the gastrocnemius. Based on how accurately the classifier is able to perform, this would allow for identification of key metabolites that help characterize the different stages of endurance training, from completely sedentary to the full eight weeks of training.

II. LITERATURE REVIEW

As mentioned above, MoTrPAC researchers have previously analyzed multi-omic data in the gastrocnemius. In the paper, "Temporal dynamics of the multi-omic response to endurance exercise training", the researchers relay their results found from pathway enrichment on the samples taken after the eight weeks of training[2]. In the gastrocnemius, in addition to the previously mentioned enrichment of lipid-related metabolites, the analysis showed that terms related specifically to peroxisome proliferator-activated receptors signalling and lipid synthesis and degradation were enriched at the protein level. Through clustering analysis on up-regulated features in the gastrocnemius, it was also found that multiple muscle adaptation processes were enriched, showing how endurance training has the power to change skeletal muscle lipid composition. Multi-omic data was important to the study, as these conclusions were the results of analyzing metabolic, proteomic, and transcriptomic data.

Another MoTrPAC study found that the gastrocnemius underwent lipid remodelling throughout the eight weeks of endurance training. In the paper, "Endurance Exercise Training Alters Lipidomic Profiles of Plasma and Eight Tissues in Rats: a MoTrPAC study", the researchers studied non-targeted and targeted lipidomics to specifically understand how lipid composition changed throughout endurance training[4]. It was found that in the gastrocnemius, mainly phospholipids increase with endurance training, with males having a more modified lipidomic profile compared to females. This study again shows how the transformation of the lipidomic profile of the gastrocnemius throughout endurance training is significant enough to be used to identify the length of time spent training.

In relation to the methods that will be discussed below, research was done on best practices for applying machine learning to metabolomics data. The paper "Statistical Analysis and Modeling of Mass Spectrometry-Based Metabolomics Data" provides an overview of four machine learning models that employ metabolomic data and also includes information on pre-processing data, variable selection, and steps for the initial analysis[5]. The paper emphasizes the importance of accurately assessing a classification model, as the results can confirm that the selected metabolites are able to classify the samples and are therefore highly important to the study. Due to the nature of biologic datasets often being small, the paper recommends cross validation to "show the variation of the model performance". The recommendations in this paper will be taken into consideration throughout the methods of this study.

III. METHODS

Due to the small amount of data points available, classical machine learning methods were applied on normalized metabolite data from the MoTrPAC Data Hub. Exploratory data analysis was done to better understand the distribution of the dataset. Visuals were created to view the distribution of the metabolites in each class (sedentary, one week of training, two weeks of training, four weeks of training, and eight weeks of training). These visuals, along with the findings in the paper mentioned in the Introduction, helped pinpoint specific metabolites that characterize the changes in lipoprotein receptors in the gastrocnemius from the sedentary class to the eight weeks of training class. Specifically, the averages of the metabolite levels were calculated for each class and the ten metabolites with the largest change from the control class to the eight weeks class were identified.

For the first classifier, a multinomial logistic regression model was built to classify a sample into one of the five classes, based on its levels of lipoprotein receptors in the gastrocnemius. The data was split into 30% testing data and 70% training data. The logistic regression model was initialized using all 131 metabolites as parameters. The initial accuracy for this model was calculated based on the testing set.

Then, recursive feature elimination with 5-fold cross validation was performed to recursively determine the most optimal number and selection of metabolites to include as parameters. Once the model was updated with the optimal selection of metabolites, the confusion

matrix and overall accuracy were calculated based on the held out testing set, as well as the precision and recall for each class.

To better understand how the logistic regression model makes its classifications, the coefficients for each parameter were identified and plotted. In addition, the metabolites selected to be used in the final logistic regression model were compared to the top 10 metabolites found during the initial exploratory data analysis to identify any overlap.

For the second classifier, a random forest model was built to classify a sample into one of the five classes, based on its levels of lipoprotein receptors in the gastrocnemius. The data was split into 30% testing data and 70% training data. The initial random forest model was initialized using all 131 metabolites as parameters and using 100 decision trees. The initial accuracy for this model was calculated based on the testing set.

Then, recursive feature elimination with 5-fold cross validation was performed to recursively determine the most optimal number and selection of metabolites to include as parameters. To further tune the model, hyperparameter grid search was used to sweep over different values for four different hyperparameters: (1) number of trees (25, 50, 100, 150); (2) max features (square root, log2, None); (3) max depth (3, 6, 9); and (4) max leaf nodes (3, 6, 9). Once the model was updated with both the optimal selection of metabolites and the optimal hyperparameters, it was evaluated using 5-fold cross validation with the training set and the overall accuracy was calculated with the held out testing set. The confusion matrix and the precision and recall for each class were also calculated for the final model, based on the testing set.

To better understand how the random forest model makes decisions, the feature importances for each of the included metabolites were calculated and plotted in a horizontal bar graph. In addition, the metabolites selected to be used in the final random forest model were compared to the top 10 metabolites found during the initial exploratory data analysis to identify any overlap.

IV. RESULTS

A. DATASET PREPROCESSING

MoTrPAC's associated R package[6] was used to extract the normalized, untargeted lipoprotein receptor metabolite level data from the MoTrPAC package. The values for each metabolite are the log-fold change from the control group, and the values were KNN-imputed. The original dataframe downloaded had the vial label numbers as the columns and the metabolite names as the rows. The data frame was transposed so that the vial labels were the rows and the metabolites were the columns. This format allowed each row of metabolite levels to correspond to one sample. The "PHENO" dataframe provided was used to match each vial label to the corresponding participant ID, training or control group, training time, and sex. The final dataframe was exported to a csv file to be read into a Python notebook.

The final dataset contains 131 metabolites measured for 48 different samples. The number of samples per group is balanced, with nine samples for the two weeks and four weeks groups and ten samples for the control, one week, and eight weeks groups (Figure 1).

B. EXPLORATORY DATA ANALYSIS

The average values for each metabolite in each of the five groups were calculated to perform exploratory data analysis on. A heatmap showing the change in average value for each metabolite across the five groups was created (Figure 2). Then, the top ten metabolites with the largest magnitude of change between the control group and eight weeks group were identified and a heatmap of those ten was created as well (Figure 3). Most of the metabolites in this group of ten had a decrease in levels from the control group to the eight weeks group. The one metabolite that increased from the control group to the eight weeks group was PE(36:3)>PE(18:1_18:2)_feature1.

C. LOGISTIC REGRESSION MODEL

The initial logistic regression model includes all 131 metabolites as parameters and classifies a sample into one of five groups (control, one week, two weeks, four weeks, eight weeks). The metabolite level data and class labels were split into 30% for testing and 70% for training. 5-fold cross validation showed accuracies of 0.143, 0.423, 0.571, 0.333, and 0.667 for each of the folds. The average accuracy for the 5-fold cross validation was 0.429. Then, using the testing set to make predictions, the accuracy for the initial model was 0.667.

To try to improve the performance of the model, recursive feature elimination with 5-fold cross validation was done. The optimal number of features found was 63 (Figure 4). These 63 metabolites were then used as the parameters in a new version of the model. The model with 63 metabolites has an accuracy of 0.667, based on predictions made for the testing set.

After finding the intersection between the 63 metabolites used as parameters in this model and top ten metabolites found during the exploratory data analysis, seven common metabolites were found (Figure 5).

D. RANDOM FOREST MODEL

The initial random forest model includes all 131 metabolites as parameters and contains 100 decision trees. The model classifies a sample into one of the five groups. The metabolite level data and class labels were split into 30% for testing and 70% for training. Using the testing set to make predictions after it was fit with the training data, the initial model had an accuracy of 0.6.

Similarly to the logistic regression model, recursive feature elimination with 5-fold cross validation was done. The optimal number of features found was 35. These 35 metabolites were then used as the parameters of a new version of the model. This new model was evaluated using 5-fold cross validation. The accuracy scores for each of the folds were 0.571, 0.714, 0.571,

0.833, and 0.833. The average accuracy for the 5-fold cross validation was 0.705. Then, using the testing set to make predictions, the accuracy for this model was 0.733.

Next, hyperparameter tuning using grid search with 5-fold cross validation was run on the random forest model with 35 metabolites. The most optimal parameters found were maximum depth as 3, maximum number of leaf nodes as 9, and number of trees as 150. The model rebuilt with these new parameters was then evaluated using 5-fold cross validation. The accuracy scores for each of the folds were 0.571, 0.714, 0.571, 0.833, and 1. The average accuracy for the 5-fold cross validation was 0.738. Finally, the model was tested using the testing set. The testing accuracy for the final model was 0.667.

Again, the 35 metabolites used as parameters in this model were compared to the top 10 metabolites found during exploratory data analysis. There were three metabolites found that were in common between the two sets (Figure 6).

V. DISCUSSION

A. LOGISTIC REGRESSION MODEL

The final logistic regression model contains 63 metabolites as the parameters and has a testing accuracy of 0.667. This final model's testing accuracy is the same as the initial model's testing accuracy, showing that all 131 metabolites were not necessary for maintaining this accuracy.

As seen in the confusion matrix and the classification report, the control, four weeks, and eight weeks groups had the highest precision out of the five classes (Figures 7, 8). One reason for this performance may be due to the structure of how the data was collected. The five groups (control, one week, two weeks, four weeks, and eight weeks) have varying lengths of time between the samples collected for each of them. The control, four weeks, and eight weeks groups each have four weeks between the next group. This allows these groups to be more distinct. Particularly, the eight weeks group has the most amount of time between it and any other group, allowing more time for larger changes in the metabolites to potentially occur.

The ten metabolites with the coefficients of the largest magnitude are shown in Figure 9. The metabolite coefficient with the largest magnitude is PS(40:5)>PS(18:0_22:5)_feature5, with a positive coefficient of 0.595. This metabolite was the most influential in determining the classification of the samples, and the positive coefficient indicates that the higher the levels of this metabolite, the more likely the sample is to be in the eight weeks group.

As mentioned in the results, there were seven metabolites, or about 11% of the 63, found in common between the logistic regression parameters and the top ten metabolites found earlier.

B. RANDOM FOREST MODEL

The final random forest model contains 35 metabolites as the parameters and 150 trees. It has a final testing accuracy of 0.667 and an average accuracy of 0.738 for the 5-fold cross validation. The tuned model with 35 parameters has a higher testing accuracy compared to the initial model, which had a testing accuracy of 0.6. This improvement shows that all 131

metabolites were not necessary in the random forest model and that including all 131 may have potentially led to some noise in the classification that led to a lower accuracy.

When looking at the classification report, the two weeks, four weeks, and eight weeks groups had perfect precision (Figures 10, 11). The precision of the eight week group was high for both the logistic regression model and the random forest model. Again, this finding makes sense due to the eight weeks group having the most amount of time to allow for changes and to have a more distinct metabolic makeup compared to the other groups.

The ten features with the highest relative importances are shown in Figure 12. The feature with the largest relative importance is PS(40:5)>PS(18:0_22:5)_feature5, with a relative importance of about 0.08. This relative importance is fairly low, indicating that the random forest model relies on the values of the entire group of metabolites rather than just highly weighing the value of a specific metabolite.

As mentioned in the results, there were three metabolites, or about 8.5% of the 35, found in common between the random forest parameters and the top ten metabolites found earlier. When comparing the metabolites found during the exploratory data analysis, the logistic regression parameters, and the random forest parameters, there were three in common (Figure 13). The common metabolites were PE(36:3)>PE(18:1_18:2)_feature1, HexCer(d42:1 M+H2CO2_feature2, and HexCer(d40:1)_M+H2CO2_feature2.

Overall, the final random forest model outperformed the final logistic regression model based on its accuracy scores during cross validation and testing. One reason for this may be that the relationship between the metabolite levels and the classes is not linear. The random forest model may be able to pick up on different aspects of the relationships between the features and classes that a logistic regression model cannot, as it is limited to linear relationships.

VIII. FUTURE WORK

To expand upon this project, I would implement a multi-omic classification model. The MoTrPAC data hub emphasizes the collection and analysis of multi-omic data. Specifically, for the gastrocnemius, the data hub includes targeted and untargeted metabolomics, targeted and untargeted proteomics, and transcriptomics data. It would be interesting to add onto one of the existing metabolite classification models to see if including other proteins and genes in the gastrocnemius lead to a higher accuracy.

The corresponding paper to this MoTrPAC data emphasized the multi-omic response to endurance training, noting that their expansive data across tissues and omes “highlights the multi-faceted, organism-wide nature of molecular adaptations to endurance training”[2]. This addition of multi-omic data would then allow us to analyze how the metabolites, proteins, and genes included change throughout training. By identifying the most important, or influential, features in the multi-omic model, we could then see if and how these various factors are related to each other. In particular, we could identify any common pathways or processes that they are related to, which would allow us to better understand the changes in the gastrocnemius throughout endurance training.

However, it would be necessary to ensure that the pathway analysis results are meaningful by ensuring that the group of features being analyzed is large enough to pinpoint specific pathways. If the group of features is too small, the results from pathway analysis will be too noisy and will lead to an abundance of pathways that may or may not be related to what is occurring in the gastrocnemius. One paper on recommendations for pathway analysis in metabolomics suggests visualizing the curve of the number of significant pathways vs. the number of compounds of interest[7]. This method would help identify what significant threshold is best to use when selecting the significant metabolites to feed into pathway analysis. The current models implemented (including 63 metabolites and 35 metabolites) do not have enough features to lead to a meaningful pathway analysis. By expanding the model to include multi-omic data, it may lead to a larger group of features included in the final model, which would allow for pathway analysis to be performed.

IX. CONCLUSION

The main goal of building a classifier for endurance training groups based on lipoprotein receptor levels in the gastrocnemius was accomplished through the final random forest model (cross training average accuracy of 0.738 and testing set accuracy of 0.667). This classifier shows that the changes occurring in the gastrocnemius relating to lipid metabolism are distinct based on the length of time spent endurance training. The random forest model outperformed the logistic regression model due to its higher accuracy score. In addition, the random forest model used only 35 metabolites for its features, while the logistic regression model used 63 metabolites, showing that random forest is able to make more accurate decisions based on less information. Both the logistic regression and random forest models did well in classifying the eight week class, most likely due to this group having the largest amount of time to produce a change in metabolite levels.

There were three metabolites found in common between the exploratory data analysis, the logistic regression model, and the random forest model. These metabolites were the most influential to classification in both models and were also found to have the largest changes from the control group to the eight weeks group.

The changes in the makeup of lipid-related metabolites in the gastrocnemius throughout endurance training were significant enough to lead to the ability to identify distinct classes based on training duration. This finding aligns with the previous research done on this data and strengthens the conclusion that endurance training has the power to transform lipid composition in the gastrocnemius. Overall, the analysis done shows that the gastrocnemius, which plays a large role in the action of running, is highly impacted by increased endurance training duration.

REFERENCES

[1]*MoTrPAC Data Hub*, motrpac-data.org/. Accessed 29 Nov. 2024.

[2]MoTrPAC Study Group. 2022. Temporal dynamics of the multi-omic response to endurance exercise training across tissues. bioRxiv doi: 10.1101/2022.09.21.508770

[3]Bordoni B, Varacallo M. Anatomy, Bony Pelvis and Lower Limb, Gastrocnemius Muscle. [Updated 2023 Apr 17]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK532946/

[4]Eric Ortlund, Zhenxin Hou, Chih-Yu Chen et al. Endurance Exercise Training Alters Lipidomic Profiles of Plasma and Eight Tissues in Rats: a MoTrPAC study, 21 November 2024, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-5263273/v1]

[5]Xi B, Gu H, Baniasadi H, Raftery D. Statistical analysis and modeling of mass spectrometry-based metabolomics data. Methods Mol Biol. 2014;1198:333-53. doi: 10.1007/978-1-4939-1258-2_22. PMID: 25270940; PMCID: PMC4319703.

[6]MoTrPACRatTraining6moData R Package, motrpac.github.io/MotrpacRatTraining6moData/. Accessed 29 Nov. 2024

[7]Wieder C, Frainay C, Poupin N, Rodríguez-Mier P, Vinson F, Cooke J, Lai RP, Bundy JG, Jourdan F, Ebbels T. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol. 2021 Sep 7;17(9):e1009105. doi: 10.1371/journal.pcbi.1009105. PMID: 34492007; PMCID: PMC8448349.

APPENDIX

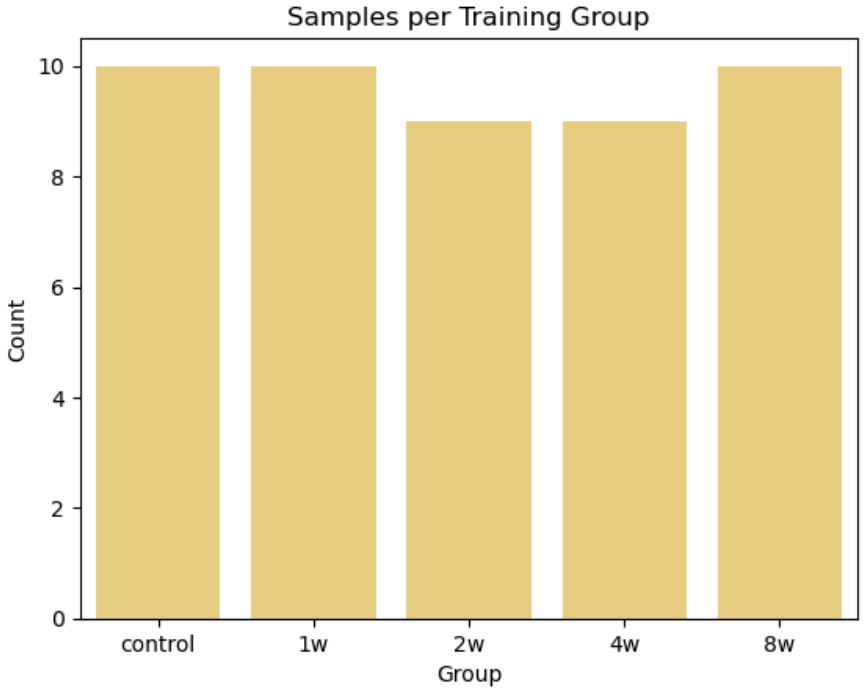

Figure 1. Distribution of the samples per training group in the final dataframe.

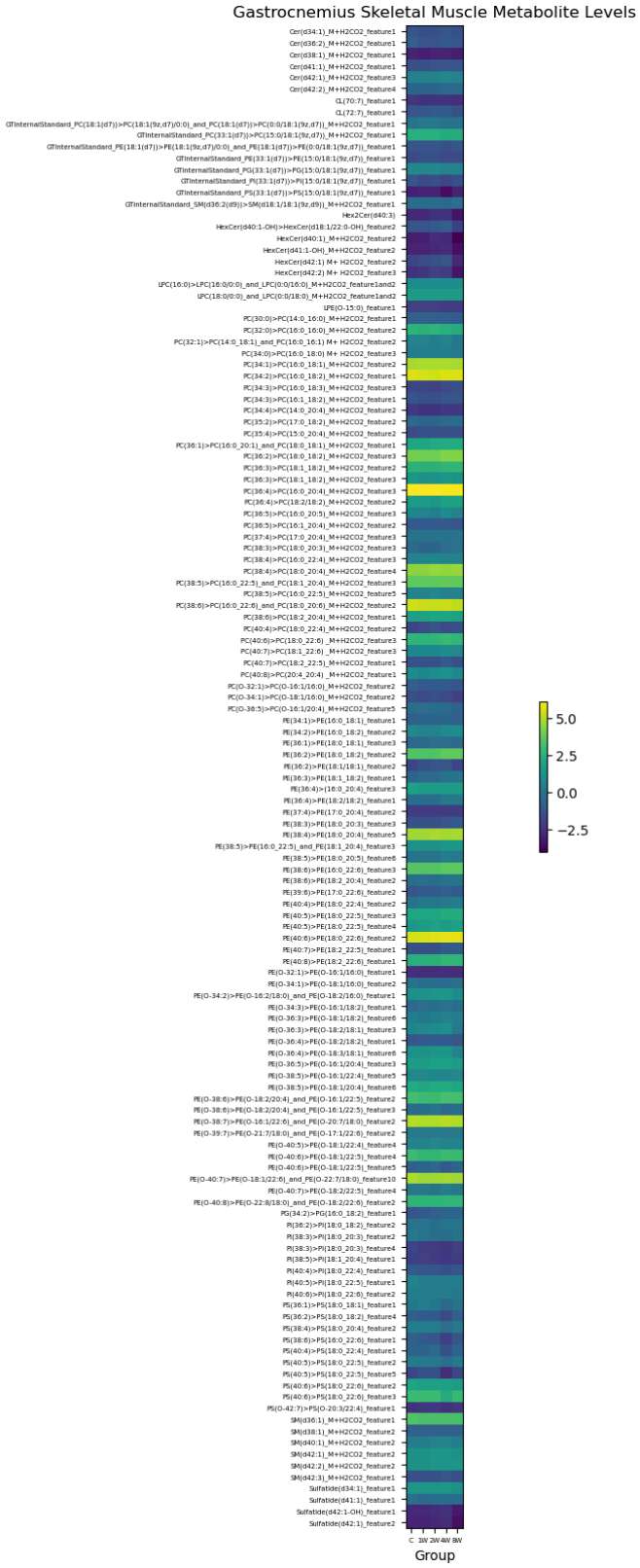

Figure 2. Full heatmap of the changes in average levels of all 131 metabolites across the five timepoints.

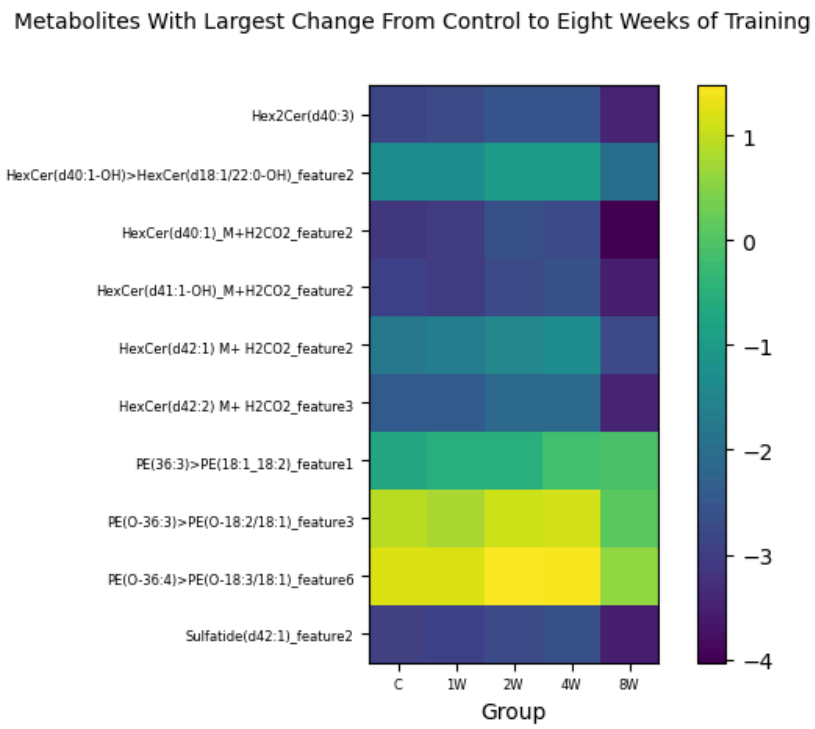

Figure 3. Heatmap of the changes in average levels for the ten metabolites with the largest change from the control group to the eight weeks group.

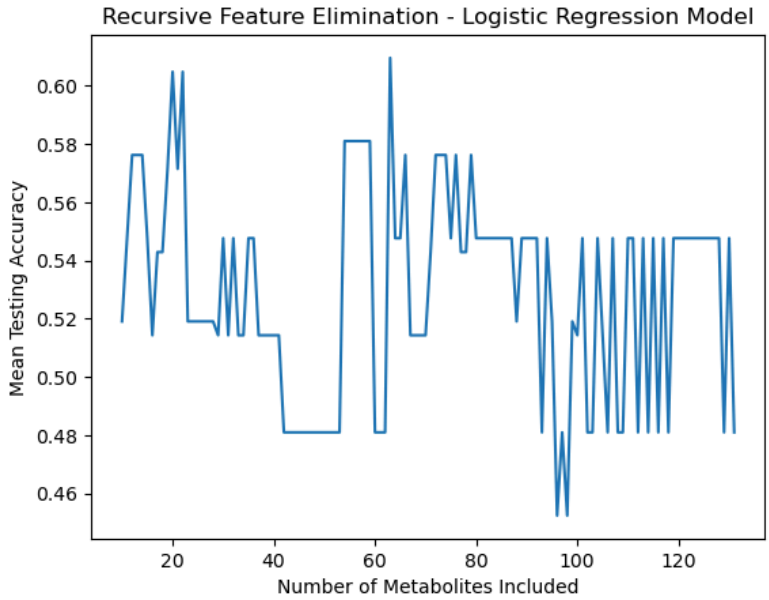

Figure 4. The optimal number of metabolites included was found to be 63.

	Feature	Coefficient	Metabolite	Change
0	HexCer(d40:1-OH)>HexCer(d18:1/22:0-OH)_feature2	-0.143056	HexCer(d40:1-OH)>HexCer(d18:1/22:0-OH)_feature2	-0.65
1	HexCer(d40:1)_M+H2CO2_feature2	0.112158	HexCer(d40:1)_M+H2CO2_feature2	-0.89
2	HexCer(d41:1-OH)_M+H2CO2_feature2	-0.198934	HexCer(d41:1-OH)_M+H2CO2_feature2	-0.62
3	HexCer(d42:1) M+ H2CO2_feature2	-0.294455	HexCer(d42:1) M+ H2CO2_feature2	-0.98
4	HexCer(d42:2) M+ H2CO2_feature3	0.179244	HexCer(d42:2) M+ H2CO2_feature3	-0.95
5	PE(36:3)>PE(18:1_18:2)_feature1	0.024490	PE(36:3)>PE(18:1_18:2)_feature1	0.63
6	PE(O-36:3)>PE(O-18:2/18:1)_feature3	-0.310166	PE(O-36:3)>PE(O-18:2/18:1)_feature3	-0.80

Figure 5. The common metabolites between the logistic regression parameters and the metabolites found during exploratory data analysis.

	Feature	Importance	Formula	Metabolite	Change
0	PE(36:3)>PE(18:1_18:2)_feature1	0.032979	C41 H76 N O8 P	PE(36:3)>PE(18:1_18:2)_feature1	0.63
1	HexCer(d42:1) M+ H2CO2_feature2	0.020231	C49 H95 N O10	HexCer(d42:1) M+ H2CO2_feature2	-0.98
2	HexCer(d40:1)_M+H2CO2_feature2	0.010936	C47 H91 N O10	HexCer(d40:1)_M+H2CO2_feature2	-0.89

Figure 6. The common metabolites between the random forest parameters and the metabolites found during exploratory data analysis.

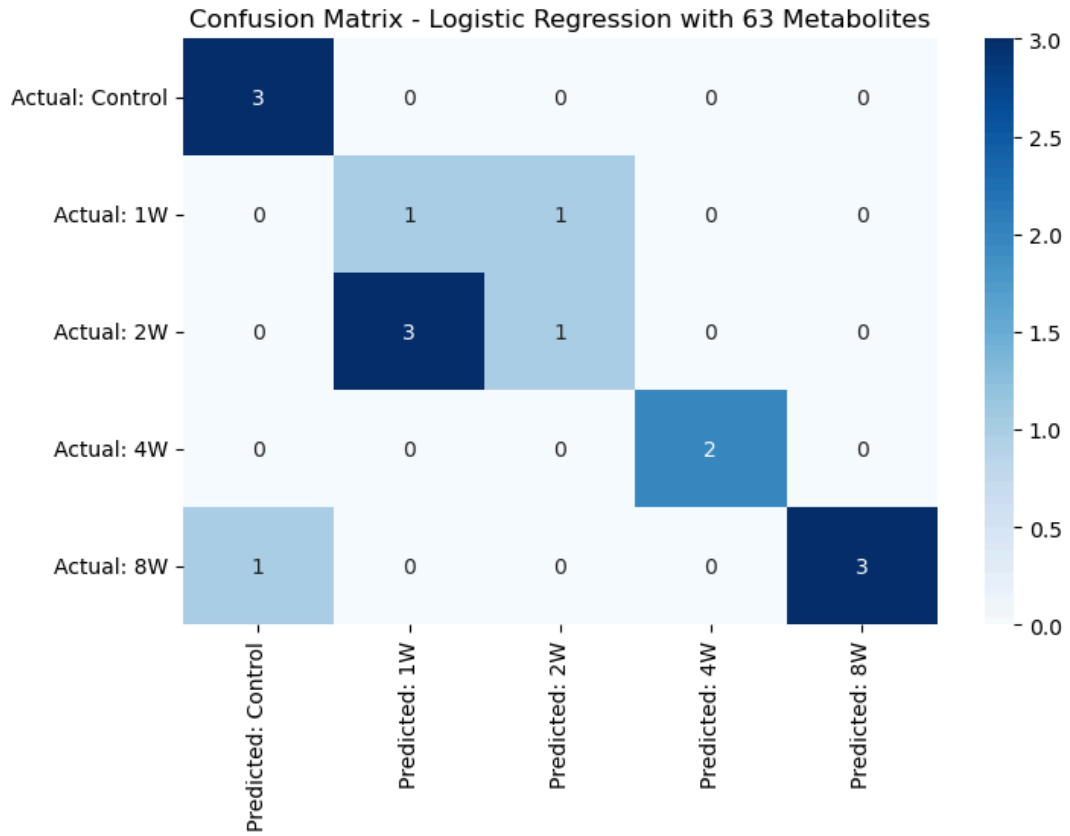

Figure 7. The control, four weeks, and eight weeks groups have the best performance in classification.

	precision	recall	f1-score	support
1w	0.25	0.50	0.33	2.00
2w	0.50	0.25	0.33	4.00
4w	1.00	1.00	1.00	2.00
8w	1.00	0.75	0.86	4.00
control	0.75	1.00	0.86	3.00
accuracy	0.67	0.67	0.67	0.67
macro avg	0.70	0.70	0.68	15.00
weighted avg	0.72	0.67	0.67	15.00

Figure 8. The control, four weeks, and eight weeks groups have the highest precision.

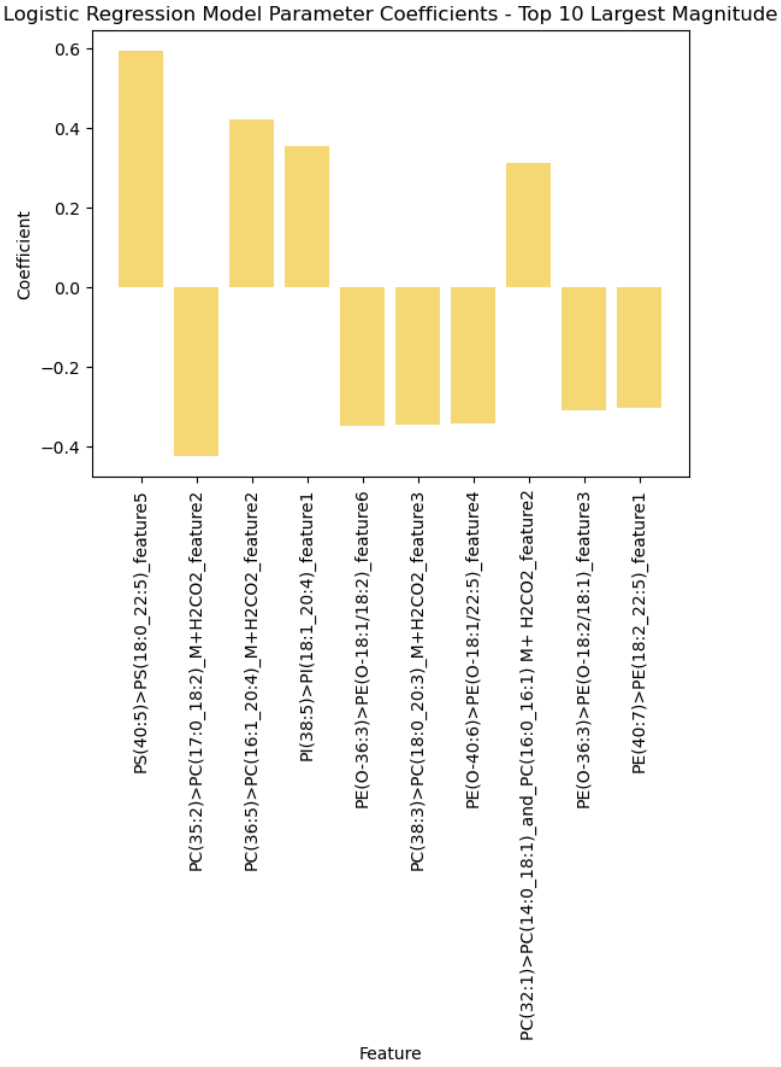

Figure 9

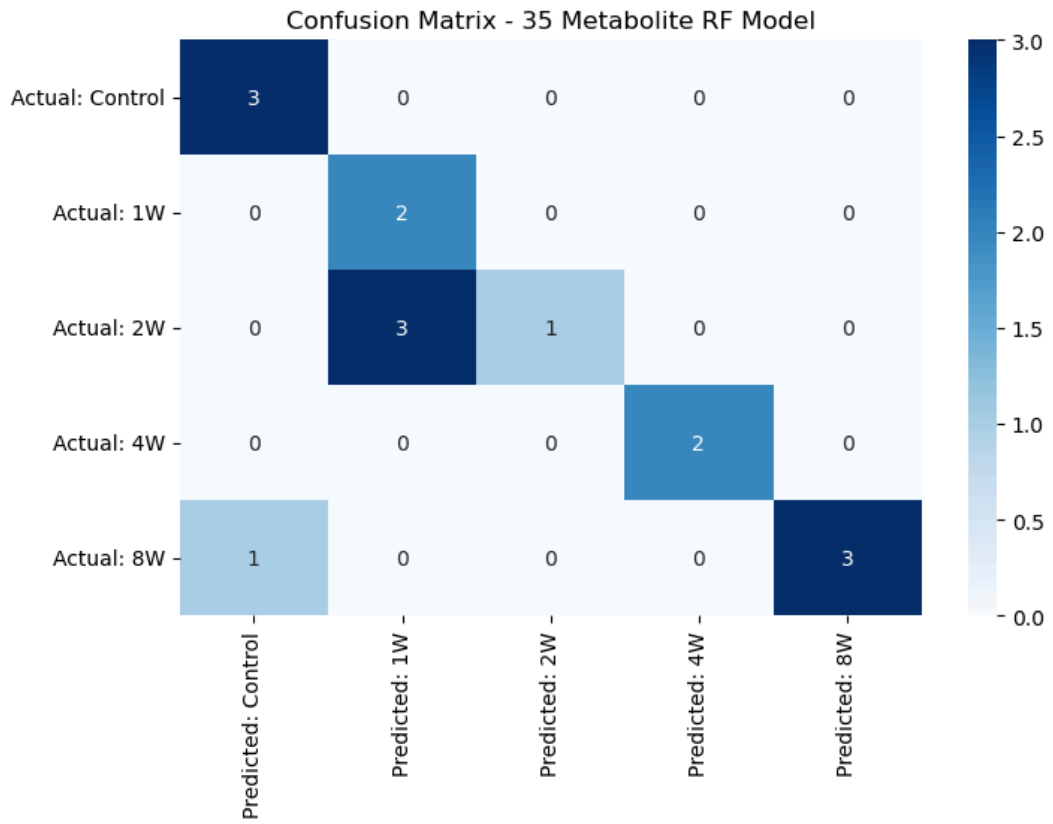

Figure 10

	precision	recall	f1-score	support
1w	0.33	1.00	0.50	2.00
2w	1.00	0.25	0.40	4.00
4w	1.00	1.00	1.00	2.00
8w	1.00	0.75	0.86	4.00
control	0.67	0.67	0.67	3.00
accuracy	0.67	0.67	0.67	0.67
macro avg	0.80	0.73	0.68	15.00
weighted avg	0.84	0.67	0.67	15.00

Figure 11. The two weeks, four weeks, and eight weeks groups have perfect precision.

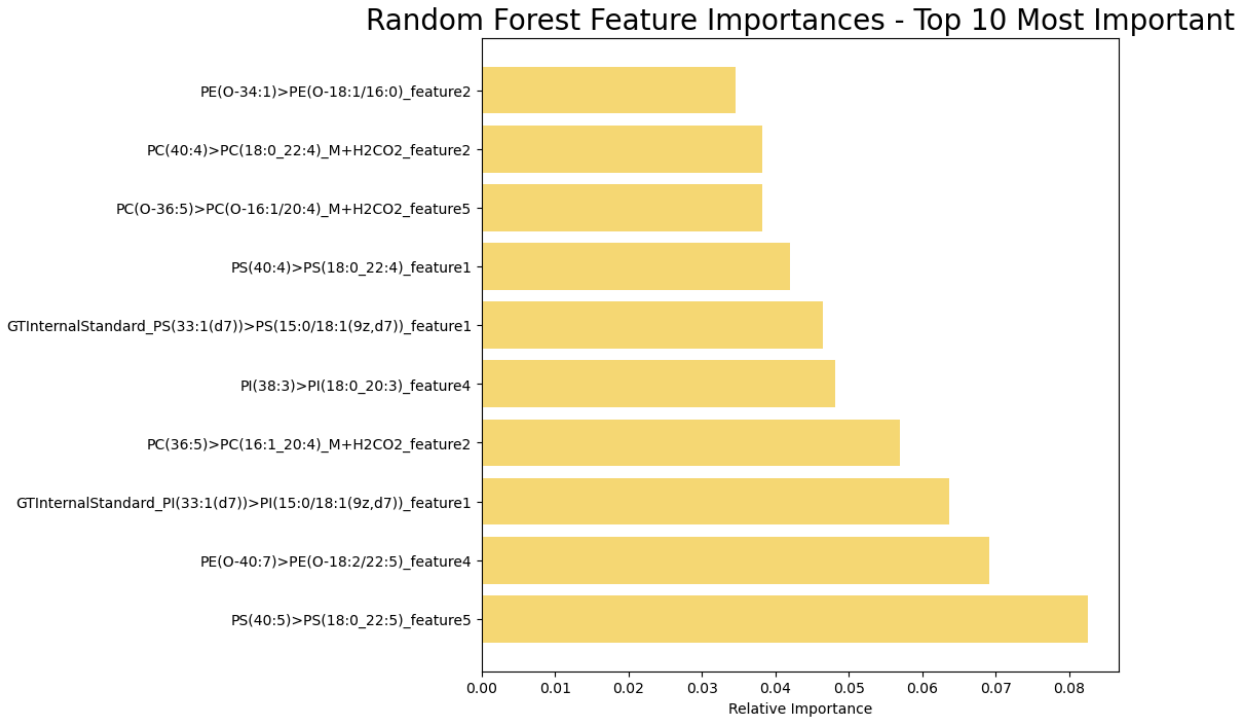

Figure 12

	Feature	Importance	Formula	Metabolite	Change	Coefficient
0	PE(36:3)>PE(18:1_18:2)_feature1	0.032979	C41 H76 N O8 P	PE(36:3)>PE(18:1_18:2)_feature1	0.63	0.024490
1	HexCer(d42:1) M+ H2CO2_feature2	0.020231	C49 H95 N O10	HexCer(d42:1) M+ H2CO2_feature2	-0.98	-0.294455
2	HexCer(d40:1)_M+H2CO2_feature2	0.010936	C47 H91 N O10	HexCer(d40:1)_M+H2CO2_feature2	-0.89	0.112158

Figure 13. Common metabolites between the metabolites found during exploratory data analysis, the parameters of the final logistic regression model, and the parameters of the final random forest model.