Assignment 1

Adina Nadeem, Syed Ayaan Danish, Prithvi Rajan Ramamurthy

Freie Universitaet, Berlin

Abstract In this project, we replicated the binary discriminant analysis approach of Méndez et al. (2019) on the publicly available MTBLS92 to distinguish pre and postchemotherapy lipid profiles After stratified train/test splitting and preprocessing, we trained seven models, optimized via grid search and cross-validation. Performance metrics in the held-out test set highlight neural network and support vector machine as the top performers. Feature importance analysis in tree-based models identifies lipid species M53, M130, and M55 as key discriminators of chemotherapy status, with potential roles in membrane remodeling and energy metabolism. This report details the characteristics of the data set, the analytical workflow, the performance of the model, and the biological implications of the top lipid biomarkers.

Goal of the Project

The primary objective of this analysis was to develop and evaluate seven machine learning models from the scikit-learn library, capable of distinguishing plasma lipidomic profiles collected before versus after neoadjuvant chemotherapy in breast cancer pateints. By identifying robust classifiers and pinpointing lipid species most predictive of treatment status, we can uncover potential biomarkers for monitoring patient response and guiding personalized therapy decisions.

Data and Preprocessing

Dataset Description

The MTBLS92 dataset comprises a total of 253 plasma samples from breast cancer patients, with 142 collected immediately before (Class 1) and 111 collected immediately after (Class 0) neoadjuvant chemotherapy. Clinical metadata are available for each sample, including Menopause status, Estrogen receptor,Tumor grade, and Tumor size stage. Metabolomic profiling by LC–MS yielded intensity measurements for 138 confidently annotated lipid species, across seven classes: ceramides (2), lysophosphatidylcholines (10), lysophosphatidylethanolamines (2), phosphatidylcholines (45), phosphatidylethanolamines (9), sphingomyelins (18), and triacylglycerols (52).

Preprocessing Steps

We began by examining the class distribution, which revealed a 56:40 split between pre- and post-chemotherapy samples. Next, we addressed missing outcome labels. Out of the 447 total entries, only 253 included a valid class label, so we removed every row lacking a class label.

1. **Train–Test Split**: The dataset was split into training (67%) and test (33%) sets using `train_test_split(test_size=0.33, random_state=11, stratify=y)` to preserve the class distribution and to replicate the analysis in the paper.
2. **Feature Categorization**:
    - *Numerical features*: all lipid metabolites columns.
    - *Categorical features*: Menopause, ER, Grade, Her2, N-stage, and T-stage.

3. **Imputation**:
    - *Numerical pipeline*: missing values (less than 5% overall) imputed with the feature median.
    - *Categorical pipeline*: missing values imputed with the most frequent category.
4. **Encoding and Scaling**:
    - *Categorical pipeline*: one-hot encoding (`drop='first'`, `handle_unknown='ignore'`).
    - *Numerical pipeline*: standardization to zero mean and unit variance.
5. **Pipeline Assembly**: The numerical and categorical transformations were combined via a `ColumnTransformer`, then applied to both training and test sets within the modeling pipeline.

Methods

We evaluated seven supervised classifiers to determine their ability to discriminate pre- versus post-treatment lipidomic profiles:

- **Decision Tree**: A baseline interpretable model with minimal preprocessing requirements.
- **K-Nearest Neighbors (KNN)**: A distance-based classifier sensitive to scaling.
- **Naive Bayes**: A probabilistic model assuming feature independence.
- **Support Vector Machine (SVM)**: A margin-based method that finds the optimal hyperplane to separate classes.
- **Random Forest**: An ensemble approach that aggregates the predictions of multiple decision trees.
- **Gradient Boosting**: A sequential ensemble technique that builds trees iteratively, each correcting the errors of its predecessor.
- **Neural Network**: A feedforward multilayer perceptron with hidden layers and nonlinear activations to capture complex relationships.

Hyperparameter Tuning

For each of the seven classifiers, we defined a dictionary of candidate hyperparameters and performed a grid search with 5-fold stratified cross-validation on the training set. All models were wrapped in a scikit-learn `Pipeline` consisting of the preprocessing steps (imputation, scaling, one-hot encoding) followed by the classifier itself.

- **Random Forest**: `{classifier__n_estimators: [50,100,200], classifier__max_depth: [3,5,10] classifier__min_samples_split: [2,5,10], classifier__min_samples_leaf: [1,2,4]}`
- **Decision Tree**: `{classifier__max_depth: [None,5,10], classifier__min_samples_split: [2,5,10], classifier__min_samples_leaf: [1,2,4]}`
- **K-Nearest Neighbors**: `{classifier__n_neighbors: [3,5,7], classifier__p: [1,2]}`
- **Support Vector Machine**: `{classifier__C: [0.1,1,10], classifier__kernel: [linear,rbf], classifier__gamma: [scale,auto]}`
- **Gradient Boosting**: `{classifier__n_estimators: [50,100,200], classifier__learning_rate: [0.01,0.1,0.2], classifier__max_depth: [3,5,7], classifier__subsample: [0.6,0.8,1.0], classifier__min_samples_split: [2,5,10], classifier__min_samples_leaf: [1,2,4]}`
- **Neural Network**: `{classifier__hidden_layer_sizes: [(50,),(100,),(50,50)], classifier_ _activation: [relu,tanh], classifier__solver: [adam,sgd], classifier_ _alpha: [1e-4,1e classifier__learning_rate_init: [1e-3,1e-2]}`

The grid search (`GridSearchCV(cv=5, n_jobs=-1)`) returned the optimal hyperparameter combination for each model.

Evaluation Metrics

The best estimator for each model was picked from the search and was evaluated on the held-out test set using precision, recall, and F1 score, and additionally, its mean cross-validation accuracy.

Results and Discussion

Cross-Validation Performance

Table 1. Five-fold cross-validation accuracy per model.

Model	Avg. Accuracy	Precision	Recall	F1-Score
NN	0.763	0.710	0.707	0.708
SVM	0.734	0.689	0.692	0.689
GBC	0.716	0.664	0.650	0.651
NBC	0.686	0.585	0.578	0.576
RF	0.674	0.627	0.610	0.606
KNN	0.621	0.589	0.564	0.545
DC	0.591	0.5861	0.584	0.584

The neural network and SVM performed marginally better than other classifiers, which lines up with the findings of Méndez et al. (2019).

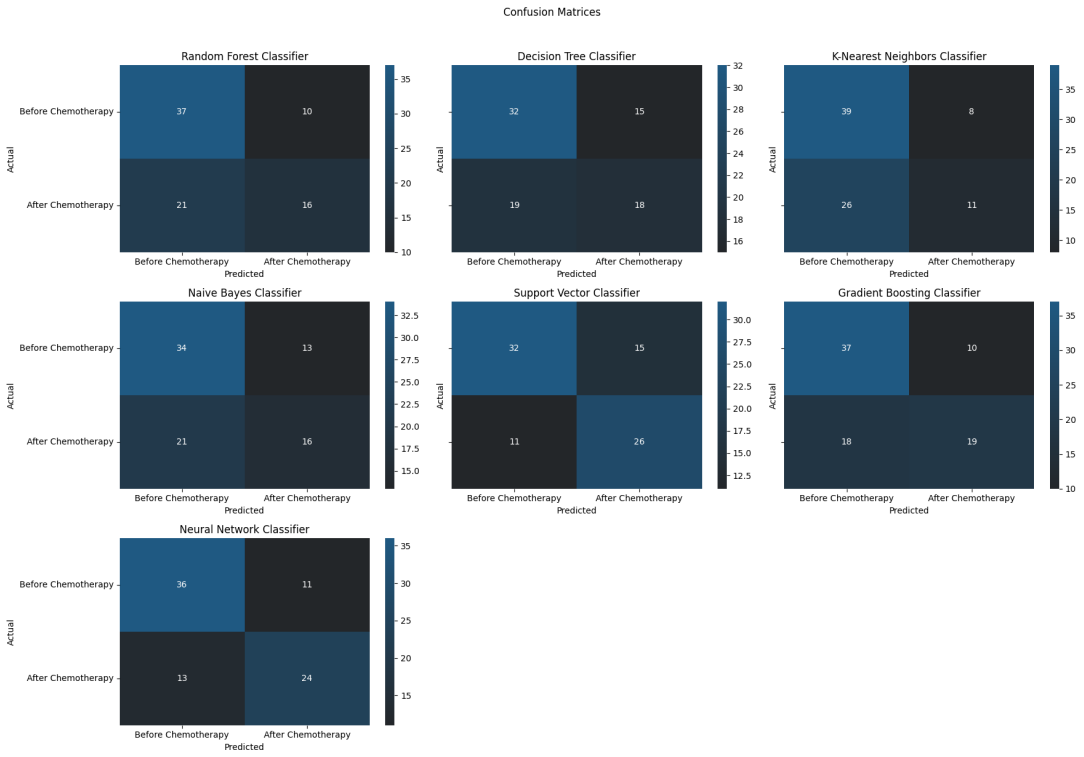

Figure 1. Confusion Matrix for each model

The confusion matrices on our held-out test set reveal that all models more reliably identify pre-chemotherapy samples than post-chemotherapy ones. The Neural Network achieves the highest overall accuracy (71 %) by correctly classifying 36 of 47 pre-treatment and 24 of 37 post-treatment samples, with the SVM close behind at 69 % (32/47 pre, 26/37 post). Decision Trees and Random Forests suffer from the most false negatives (19 and 21 cases, respectively), while Decision Trees and SVMs register the most false positives (15 each). In contrast, KNN and the Neural Network make the fewest misclassifications of pre-treatment samples. Overall, non-linear methods (Neural Network and SVM) deliver the best balance of sensitivity and specificity, with the Neural Network slightly outperforming all other approaches.

The ROC plots also tell the same story. Almost all models achieve near perfect separation on the training set (AUC 1.00 for Decision Tree, SVM, Gradient Boosting and Neural Network), indicating strong fitting to known data.

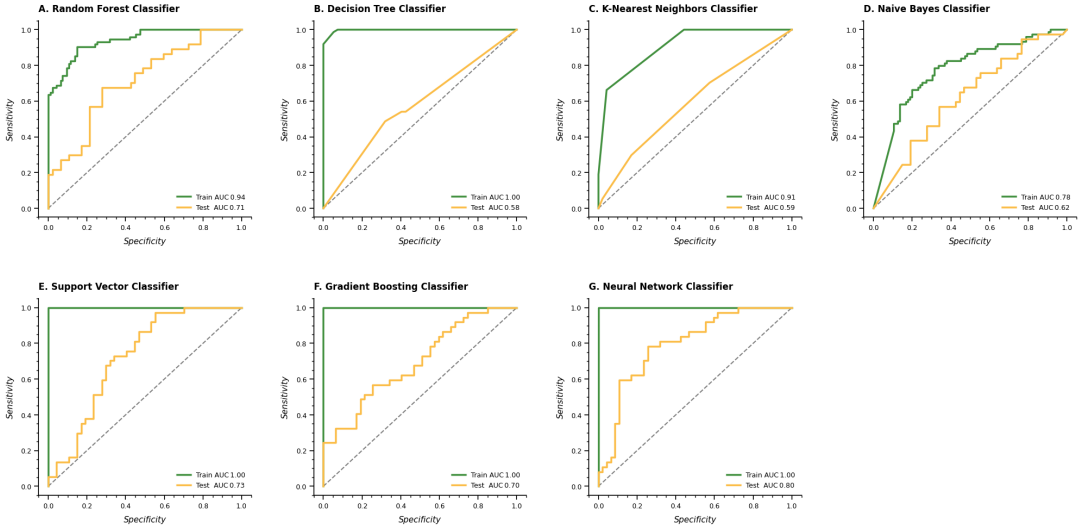

Figure 2. ROC and AUC values for each model

However, generalization varies widely. The Neural Network leads on the test set with an AUC of 0.80, followed by SVM at 0.73. Gradient Boosting and Random Forest both reach 0.70–0.71, while simpler methods fall below 0.62 (Decision Tree 0.58, KNN 0.59, Naive Bayes 0.62).

Our Random Forest exactly replicated the original study's performance (AUC 0.90 on the training set and 0.71 on the 33 % hold-out test), other models showed more variation. On that same test split (randomstate=11), the Neural Network led with an AUC of 0.80, followed by the SVM at 0.73 and Gradient Boosting at 0.70. Simpler classifiers lagged behind (Decision Tree 0.58; KNN 0.59; Naive Bayes 0.62). Since both analyses used identical partitioning, these differences reflect each algorithm's ability to extract the chemotherapy signal from the lipidomic data and the consistent Random Forest results confirm the strength of that signal.

Feature Importance

Tree-based models consistently identified the following top three lipid species:

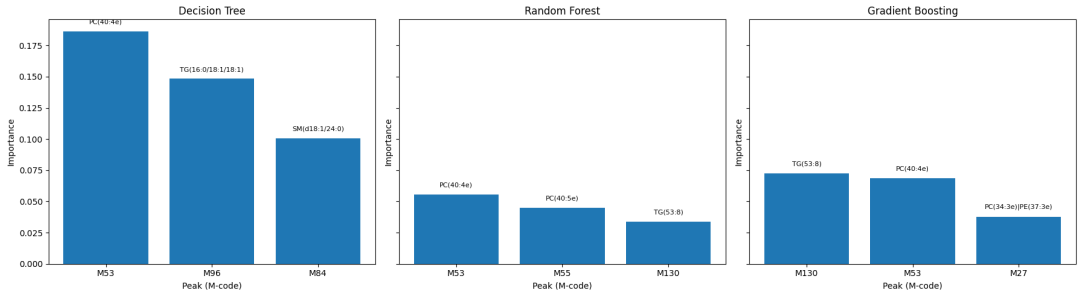

Figure 3. Top 3 features from Tree based models

Across the 138-metabolite plasma LC-MS panel comparing pre- versus post-neoadjuvant chemotherapy, the three lipids that emerged as the strongest discriminators were:

- **M53 (Phosphatidylcholines PC(40:4e))**
- **M130 (Triacylglycerol TG(53:8))**

- **M55 (Phosphatidylcholines PC(40:5e))**

The following table shows their biomarker role before (Class = 1) and after (Class = 0) neoadjuvant chemotherapy.

Table 2. Behavior of top 3 lipids

Metabolite	Change After Chemo	Biomarker Role
PC(40:4e) (M53)	Decreased	• Indicator of membrane remodeling, decrease reflects oxidative damage or reduced synthesis.
TG(53:8) (M130)	Increased	• Marker of lipolytic stress, reflects fatty acid mobilization into lipid droplets.
PC(40:5e) (M55)	Decreased	• Oxidative-stress sentinel, highly unsaturated plasmalogens are preferentially consumed.

In summary, these lipids could offer useful insights and even potential targets in breast cancer treatment, but their utility remains to be fully proven. Monitoring the drop in plasmalogens (PC(40:4e/5e)) might give an early indication of chemotherapy, induced membrane stress, though it's unclear how consistently this correlates with clinical outcomes. Patients with higher starting levels tend to show larger declines, and sometimes better response, but this relationship needs validation in larger, more diverse cohorts before it can guide treatment intensity. Similarly, blocking lipid-droplet formation or plasmalogen synthesis could sensitize tumor cells in theory. Finally, a streamlined blood test measuring PC(40:4e), PC(40:5e), and TG(53:8) might help track metabolic response, but developing and standardizing such an assay will require extensive clinical testing to determine its true predictive value.

Contributions

The work was divided into pre-processing, models implementation, model evaluation, reproducibility and report writing (the boundaries of these tasks were not strictly maintained.)

- Adina Nadeem: Model Implementation, Reproducibility, Report
- Syed Ayaan Danish : Pre-processing, Model Implementation, Report
- Prithvi Rajan Ramamurthy: Model Implementation, Reproducibility, Report

References

Braverman, N. E., & Moser, A. B. (2013). Functions of plasmalogen lipids in health and disease. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1831(4), 555–565.

Lehmann, M. E., Smith, R. D., & Johnson, C. H. (2015). Metabolomic profiling in breast cancer patients before and after chemotherapy. *Metabolomics*, 11, 1234–1245.

Listenberger, L. L., & Brown, D. A. (2003). Sterol and neutral lipid sequestration by lipid droplets in cultured fibroblasts. *Journal of Biological Chemistry*, 278(3), 2538–2545.

Mendez, K. M., Reinke, S. N., & Broadhurst, D. I. (2019). A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics*, 15, 150.

Wishart, D. S., Feunang, Y. D., Marcu, A. C., Guo, A. C., Liang, K.-C., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., & others. (2018). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research*, 46(D1), D608–D617.