Project Flow: Cooperative Learning for Multiclass Classification

Data Preprocessing

Input Datasets:

X (Platelet gene expression data): Rows are patients, columns are gene normalized counts.

Z (WBC RNA gene expression data): Rows are patients, columns are gene normalized counts.

Y (Outcome): The disease subtype, with four classes: PV, ET, MF, and CTRL.

Process flow

1. Extract Common Patients: Patient IDs from X, Z, and Y were intersected to ensure alignment.
2. Subset Data for Common Patients: Filtered rows in X, Z, and Y based on common patient IDs.
3. Ensure Consistent Row Order: Ordered rows in all datasets to match the patient IDs.
4. Normalize Features: Standardized X and Z (mean = 0, variance = 1) to ensure features are on the same scale.
5. Rename Columns to Avoid Confusion: Columns in X were prefixed with X\_ (e.g., X\_gene1), and columns in Z were prefixed with Z\_ (e.g., Z\_gene1).
6. Split Data: Split the data into training (70%), validation (15%), and test (15%) sets.
7. Binary Classification for Each Class (One-vs-Rest Approach)

Target Classes: PV, ET, MF, and CTRL.

Process for Each Class:

1. Binary Labeling: The outcome variable Y was converted into binary labels where 1 represents the target class and 0 represents all other classes.
2. Cross-Validation to Find Optimal Lambda: Performed cross-validation using cv.multiview by exploring different values of the regularization parameter (lambda) and using the “class” error metric to evaluate performance. Selected lambda\_min (the value of lambda minimizing validation error) and lambda\_1se (the largest lambda within 1 standard error of the minimum error for a simpler model).
3. Model Fitting: Trained the cooperative learning model using the training data and lambda\_1se and extracted the optimal agreement penalty (rho) used during training.
4. Predictions: Predicted probabilities for the test set using the fitted model and converted probabilities to binary predictions using a threshold of 0.5.
5. Evaluation for Each Class

Test Accuracy: Calculated the proportion of correct predictions on the test set.

Confusion Matrix: Generated to evaluate performance on true positive, false positive, and other metrics.

Feature Importance:

Extracted coefficients from the final model using lambda\_1se, identified the top 20 features (by absolute coefficient values) contributing to the predictions, and visualized feature importance using bar plots.

6. Multiclass Classification

One-vs-Rest Combination: Combined predictions from all binary classifiers into a multiclass classification by assigning each sample to the class with the highest predicted probability.

Evaluation: Calculated overall multiclass accuracy and generated a confusion matrix to summarize performance across all classes.

7. Outputs Generated

Model Details: For each class (PV, ET, MF, CTRL), stored optimal lambda\_min and lambda\_1se, optimal rho, and test accuracy in a CSV file (results/model\_details.csv).

Feature Importance: Saved bar plots of the top 20 features for each class

Validation Plot: Saved cross-validation plots for each class in images/{Class}\_CVPlot.png.

Key Results

Test Accuracy (Binary Classifiers): Accuracy computed separately for PV, ET, MF, and CTRL.

Top Features: Identified top genes contributing to each class, separately for platelet and WBC datasets.