Network Data Analysis project

Human temporal cortical tissue samples with varying degree of Alzheimer's Disease (AD)-related neurofibrillary pathology

Leombruni Virginia

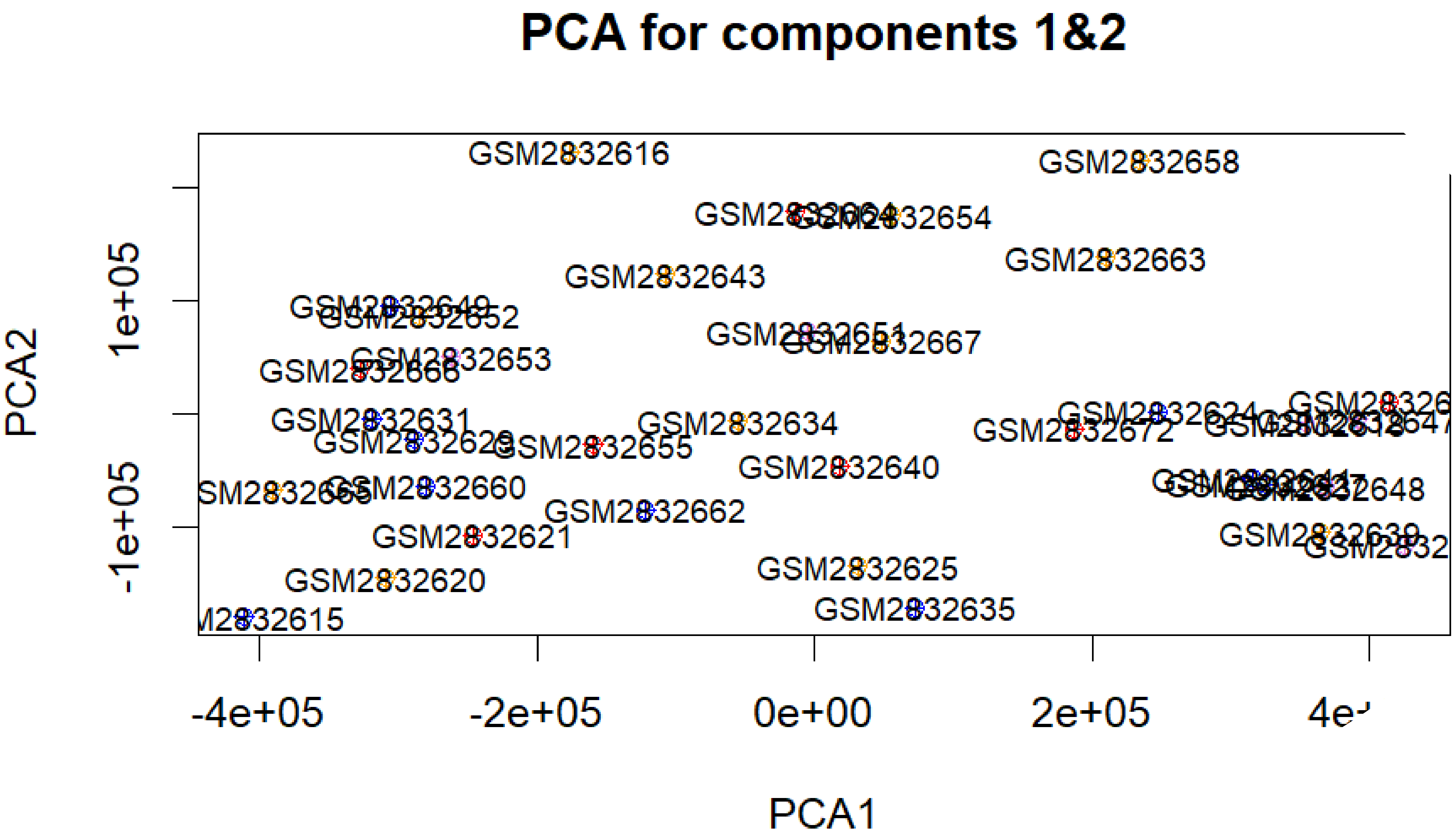
13/06/2024

Introduction

Objective: Analyze **GSE106241** dataset with 71 autopsied temporal cortical samples categorized into seven groups based on Braak's staging for Alzheimer's disease (AD)-related neurofibrillary pathology.

Grouping: Stages 0 and 1 combined as control group; stages 5 and 6 as affected group (19 affected, 16 control).

PCA (Principal Component Analysis)



Purpose: Reduce dimensionality of gene expression data.

Outcome: Mixed colors without clear separation between groups. suggest less distinction between the stages based on the data captured by PCA1 and PCA2.

Advantages:

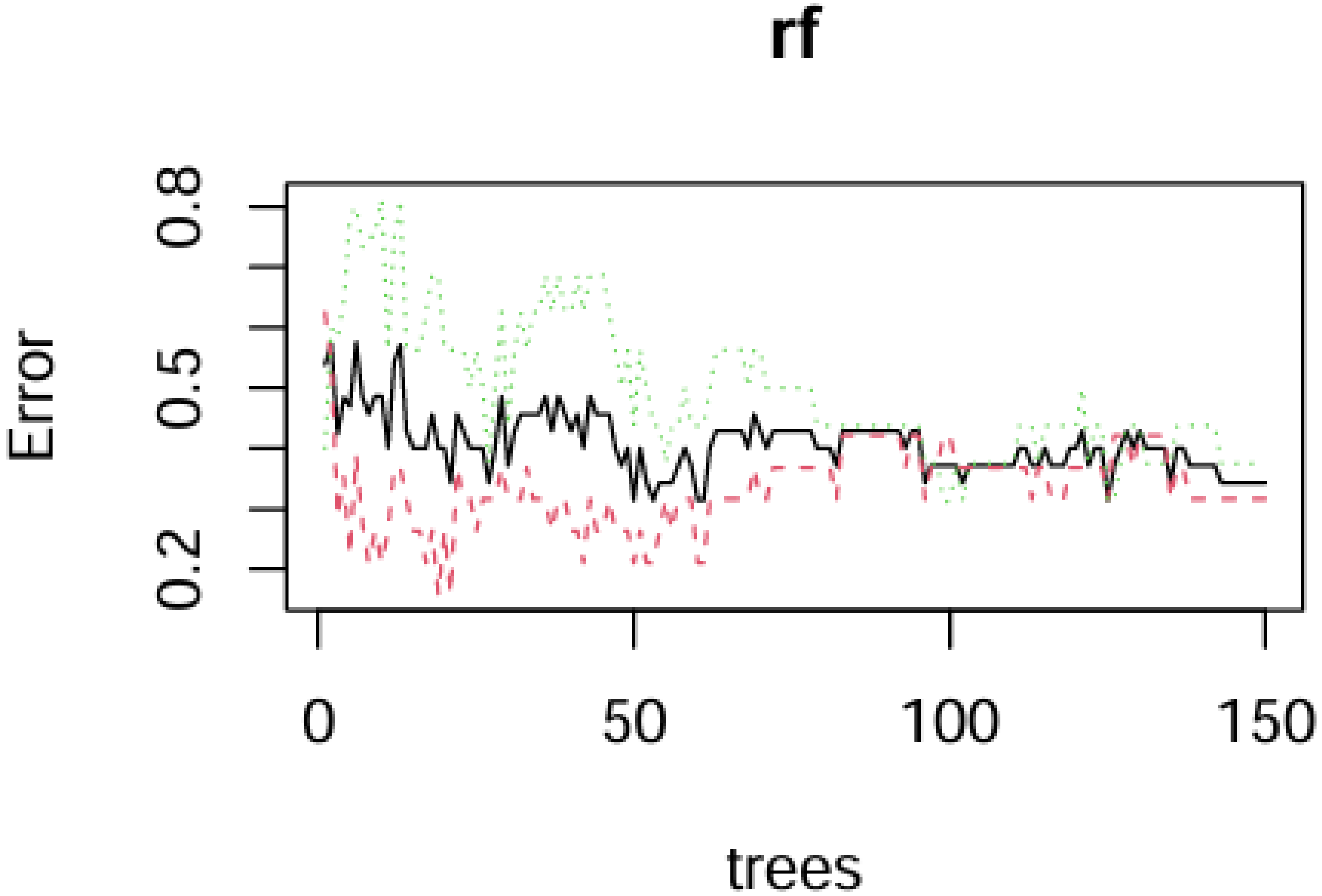
Reduces collinearity

Simplifies analysis with fewer variables

Enhances visualization

Increases computational efficiency.

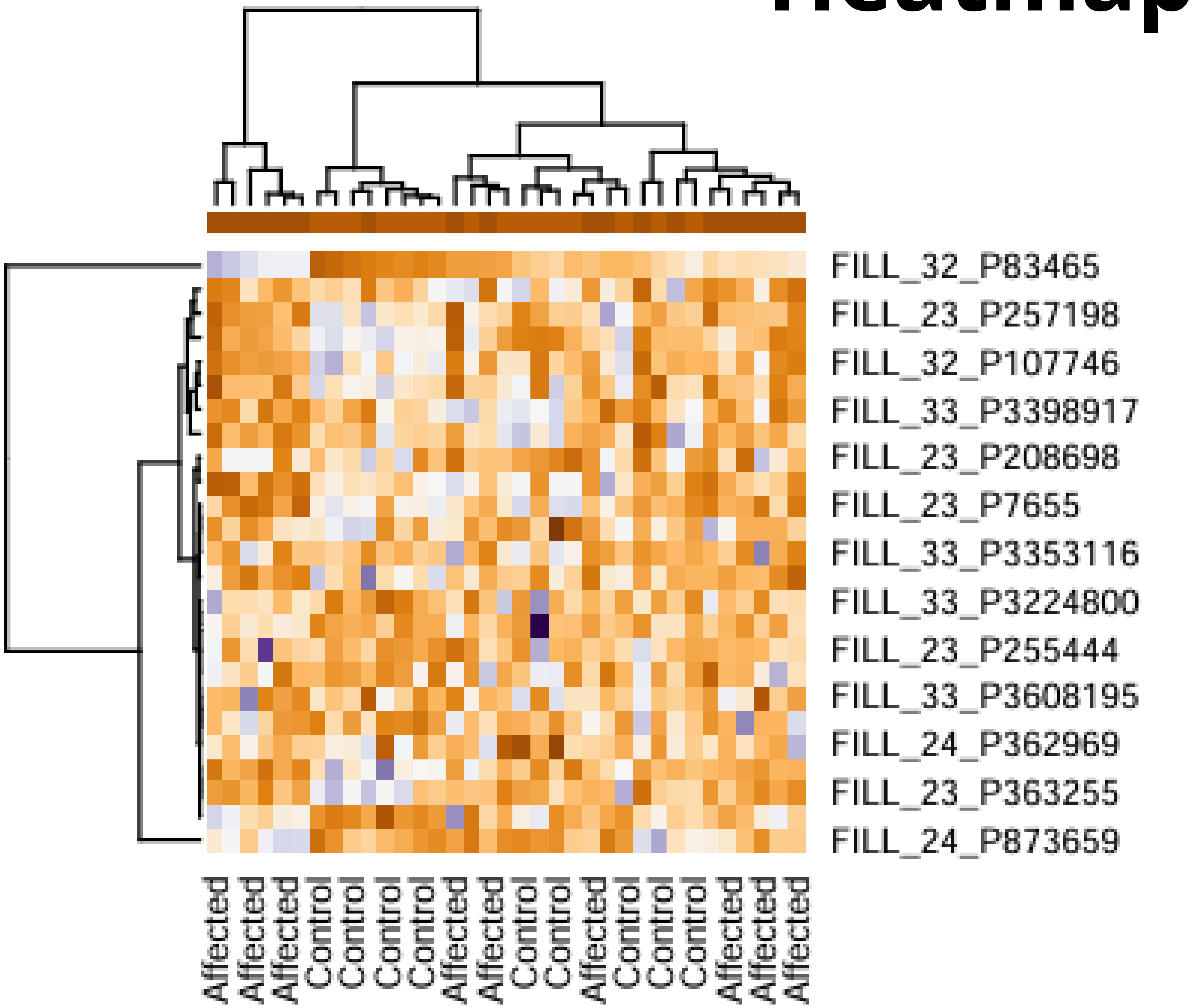
Random Forest



Purpose: Identify most important genes distinguishing 'Affected' and 'Control' groups.
Outcome: 150 trees provided stable error rate and good generalization. Identified 200 most important genes for further analysis.

- **Black Line:** OOB error rate.
- **Red Line**: Affected group
- **Green Line**: Control group

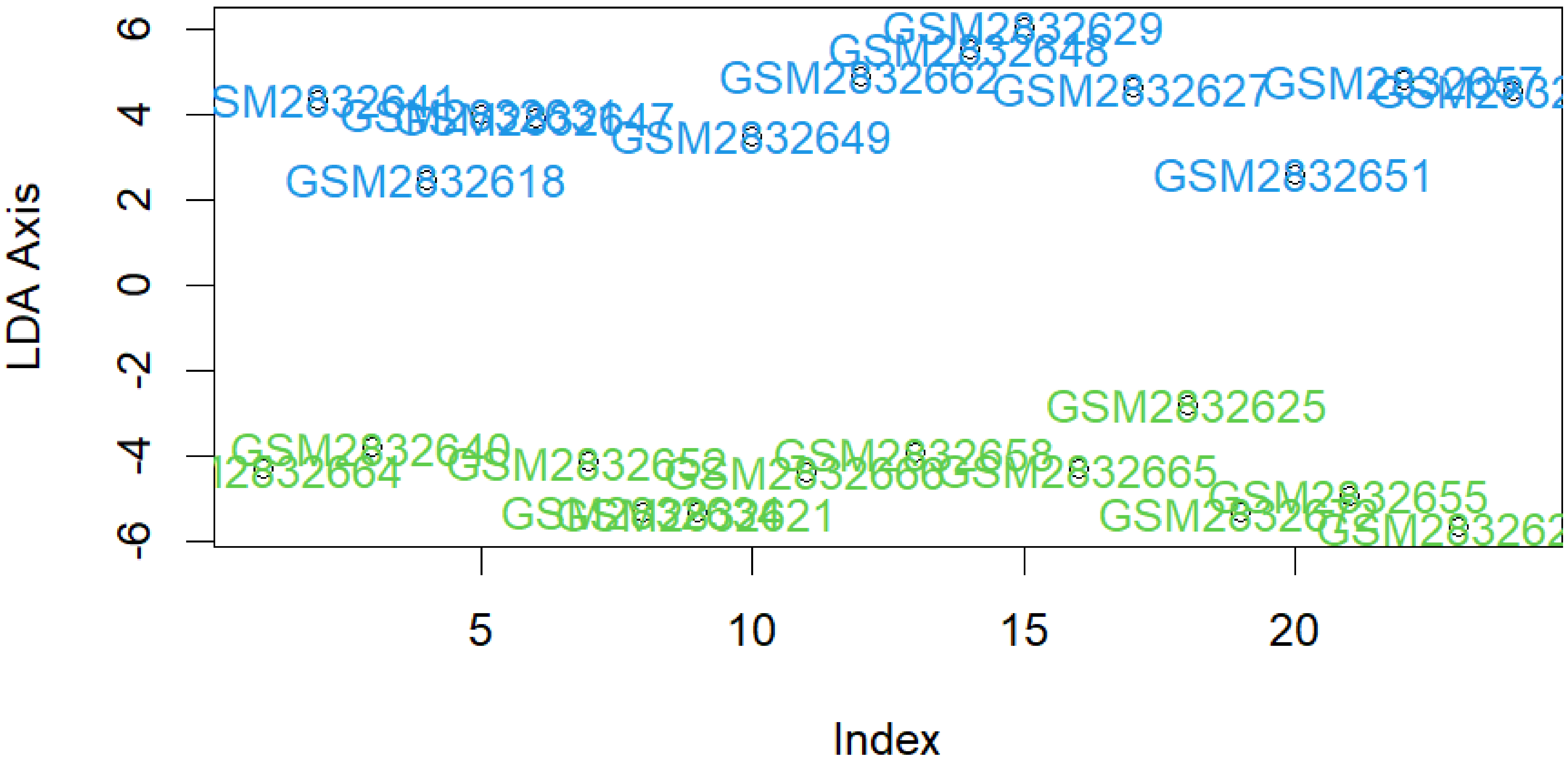
Heatmap



Purpose: Visualize differences in gene expression between two groups.

Outcome: Certain genes highly expressed in one group compared to the other.
 Consistent expression patterns suggest distinct molecular profiles.

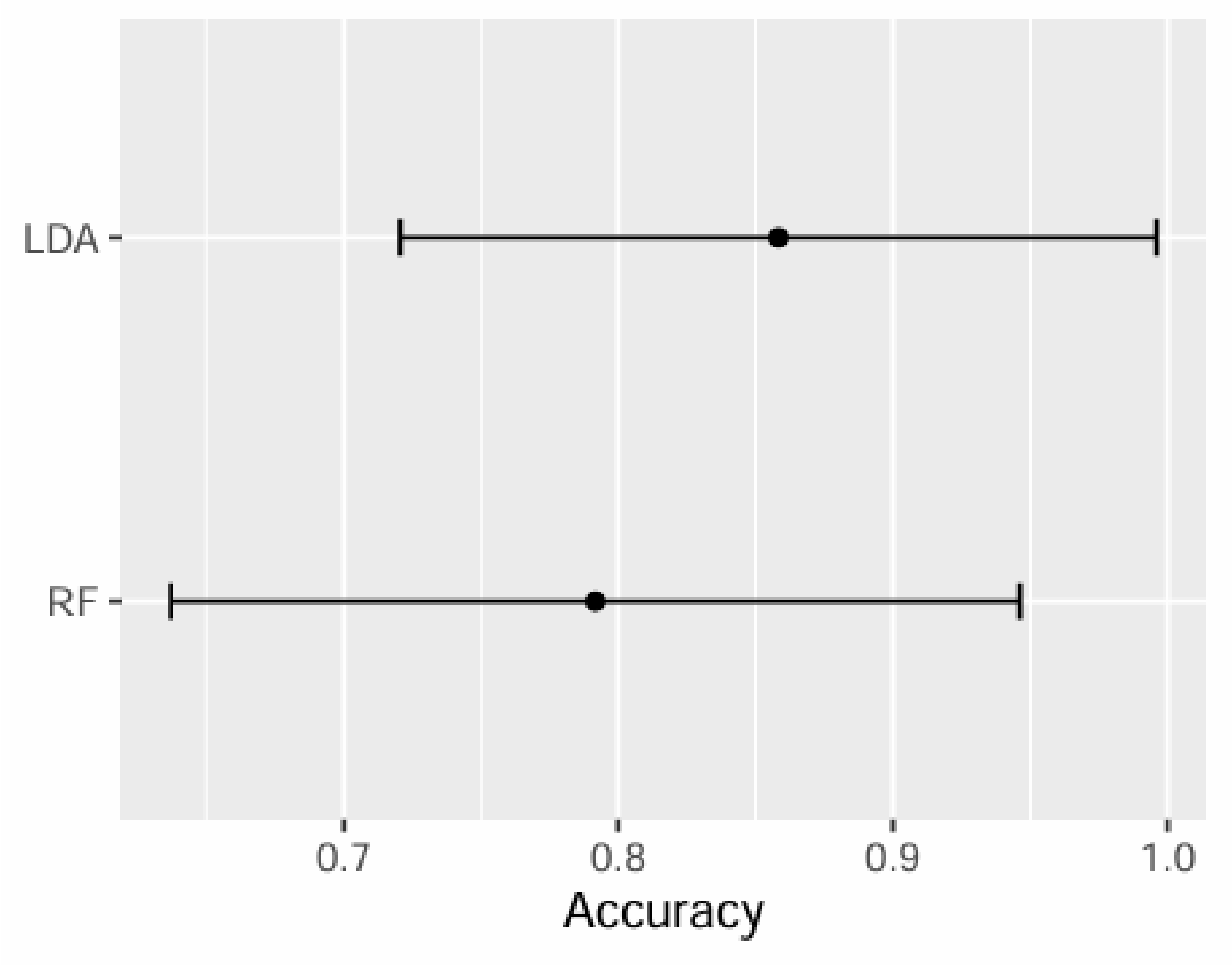
LDA (Linear Discriminant Analysis)



Purpose: Improve stability and interpretability by reducing collinearity.

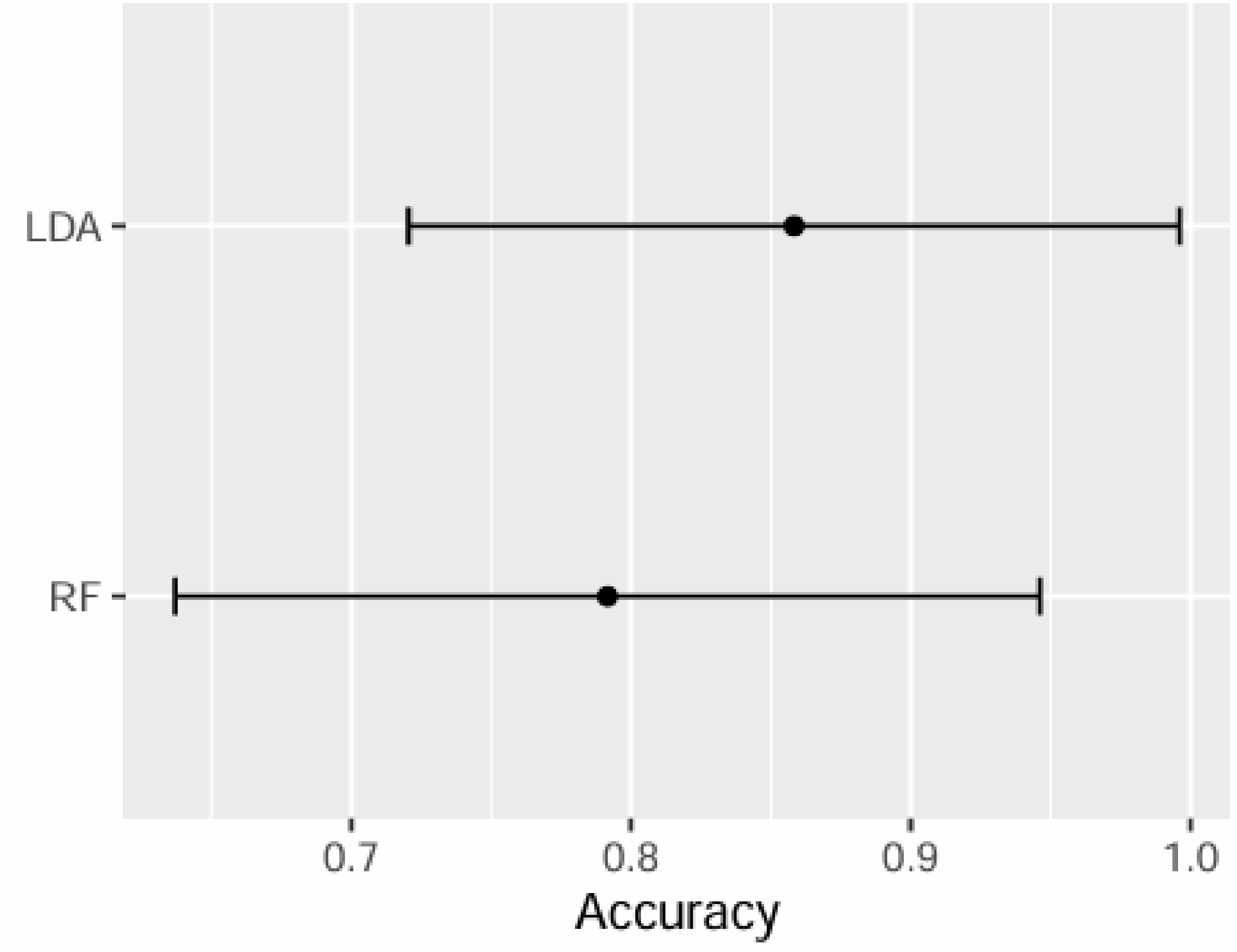
Outcome: Excellent discriminating ability, no misclassification errors. Cross-validation required for robustness.

Cross Validation



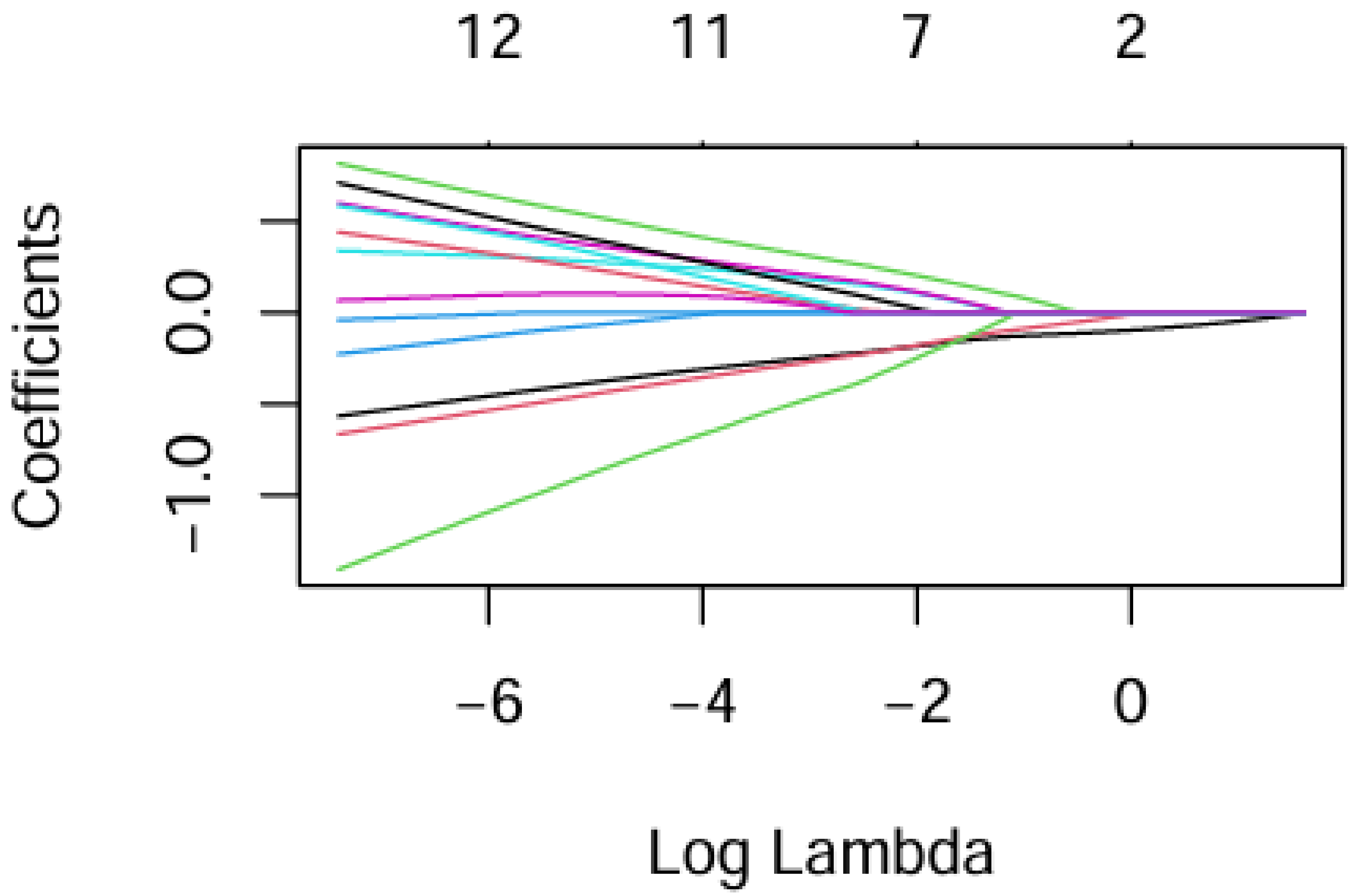
Repeated Cross Validation

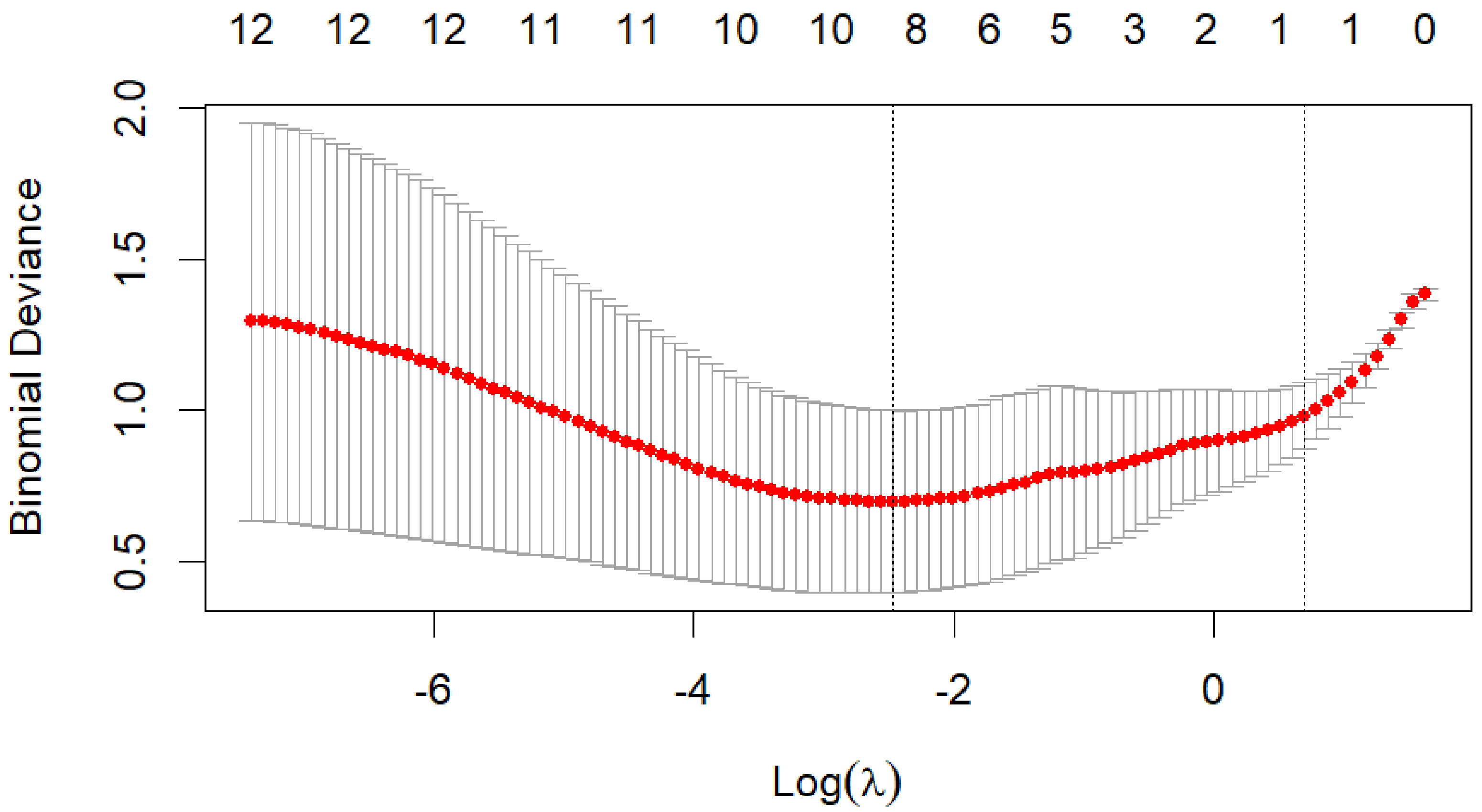
provides more reliable estimates of model performance



Outcome: Both LDA and RF models showed strong performance with high accuracy and consistent results across folds.

Lasso model



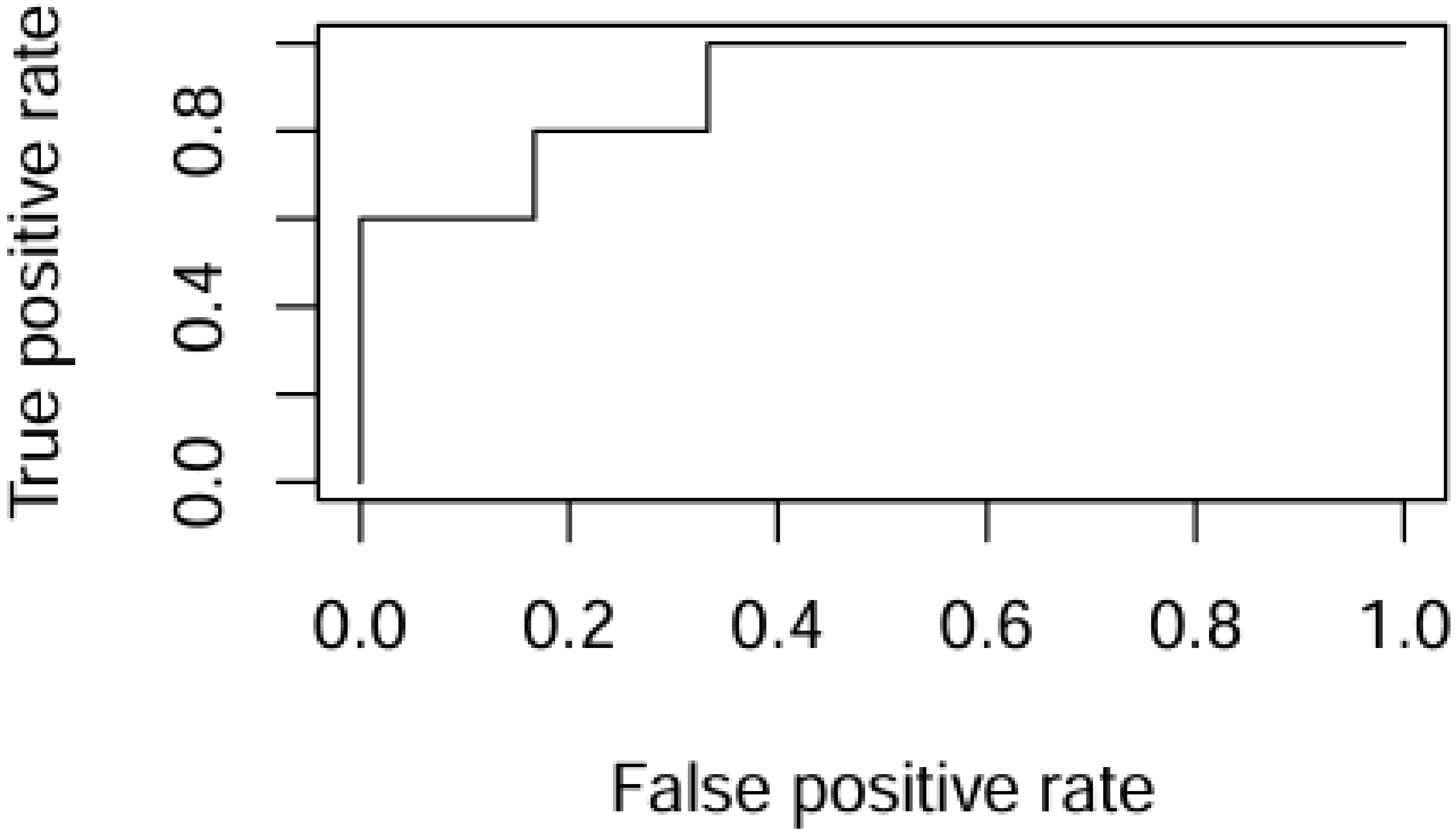


The graph is used to observe how regularization (λ) affects the coefficients of the variables in the Lasso model, helping to determine an optimal value of λ that balances the model's complexity and its predictive capability.

The graph shows binomial deviance versus the logarithm of the regularisation parameter (log(λ)) in a Lasso regression model. Red dots indicate deviance values and gray bars show confidence intervals. The optimal λ is chosen where deviance is lowest, balancing model simplicity and performance.

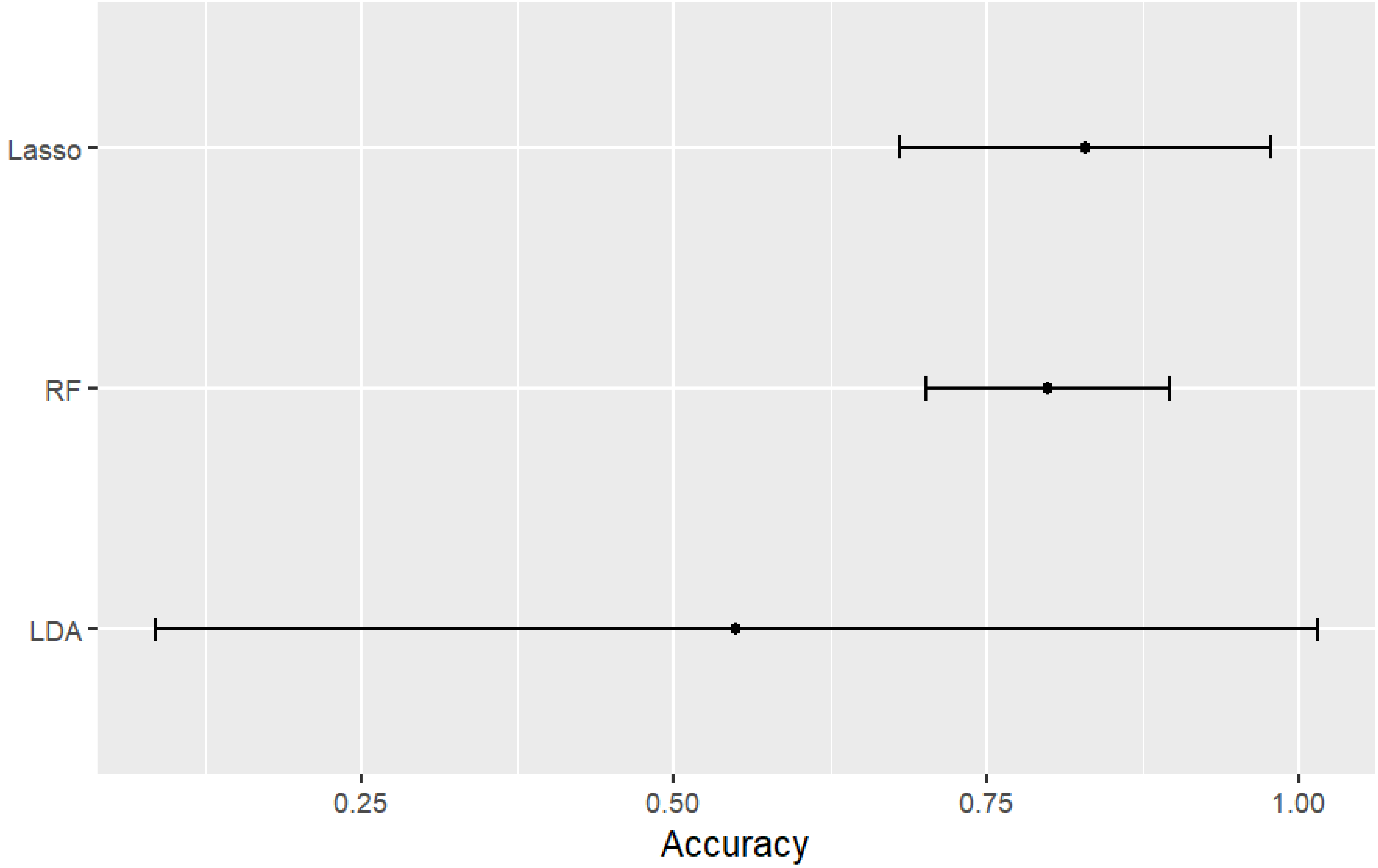
ROCR (Receiver Operating Characteristic)

is used for visualizing the performance of scoring classifiers.



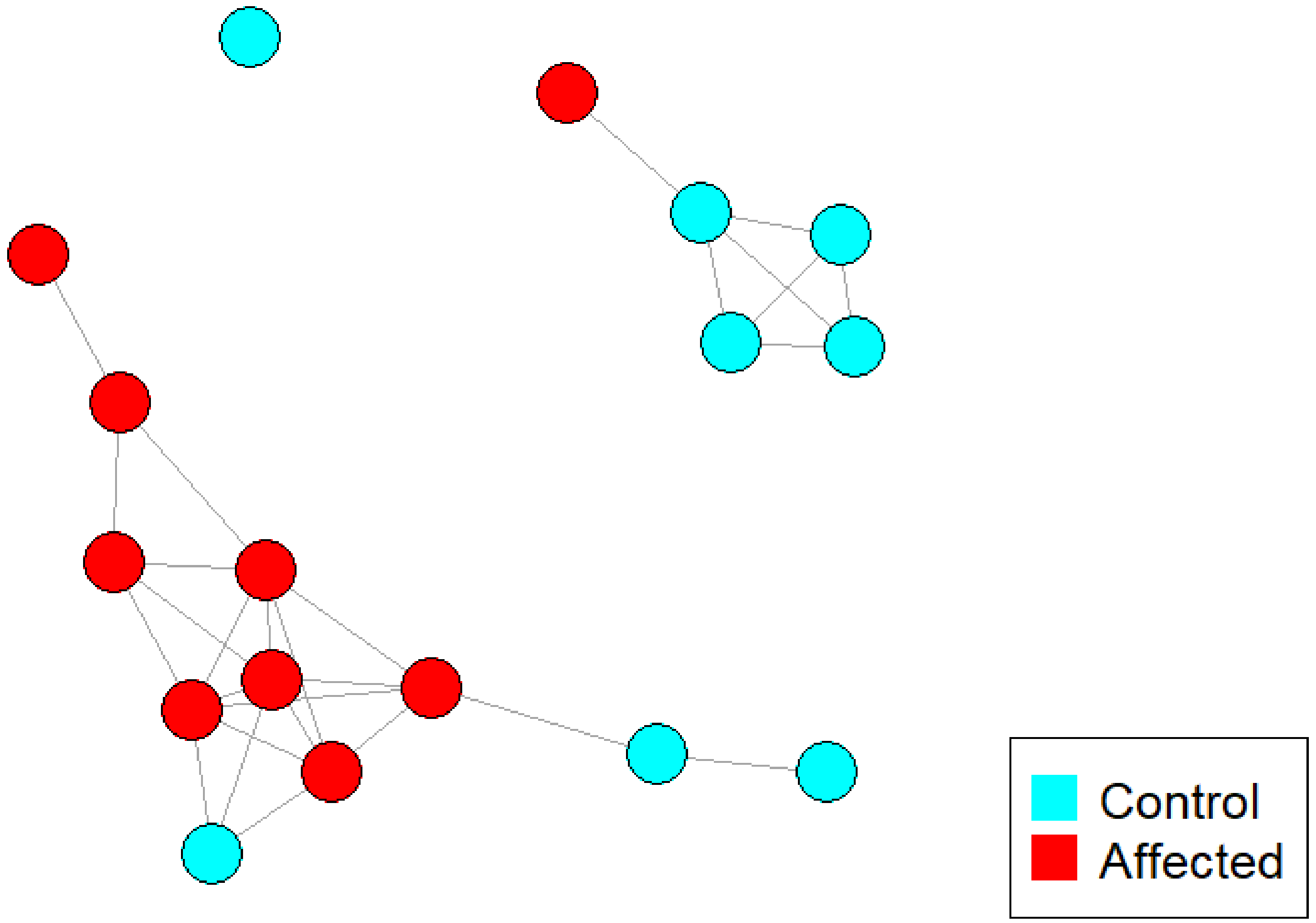
Lasso model most effective, high accuracy (AUC=0.9).
Reliable distinction between "affected" and "control" groups.

Comparison with other classification methods



The **Lasso model** was the most effective, with high accuracy value, reliably distinguishing between "affected" and "control" groups.

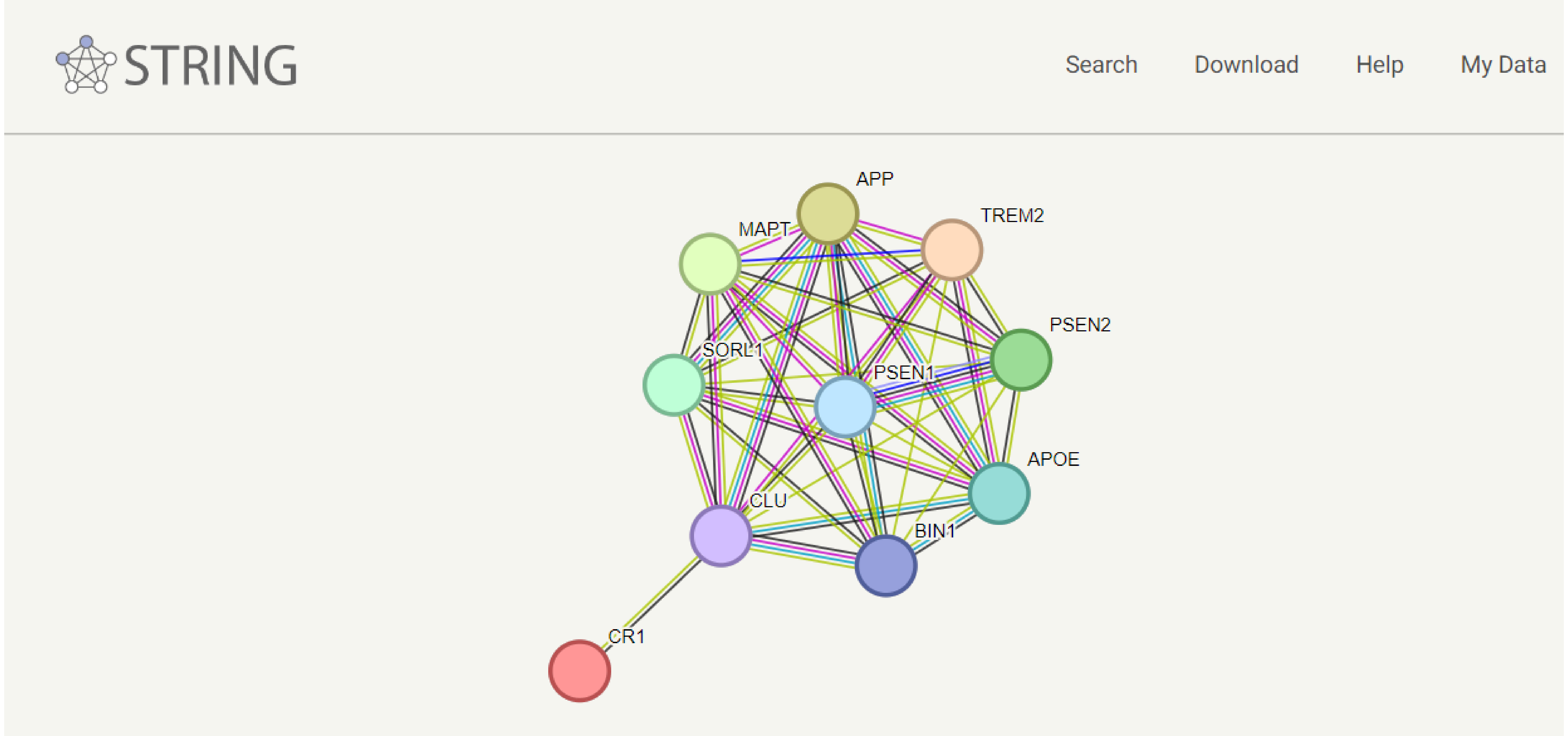
RSCUDO + CARET



Network representation demonstrates the separation and classification performance of the model, highlighting its ability to distinguish between the affected and control samples.

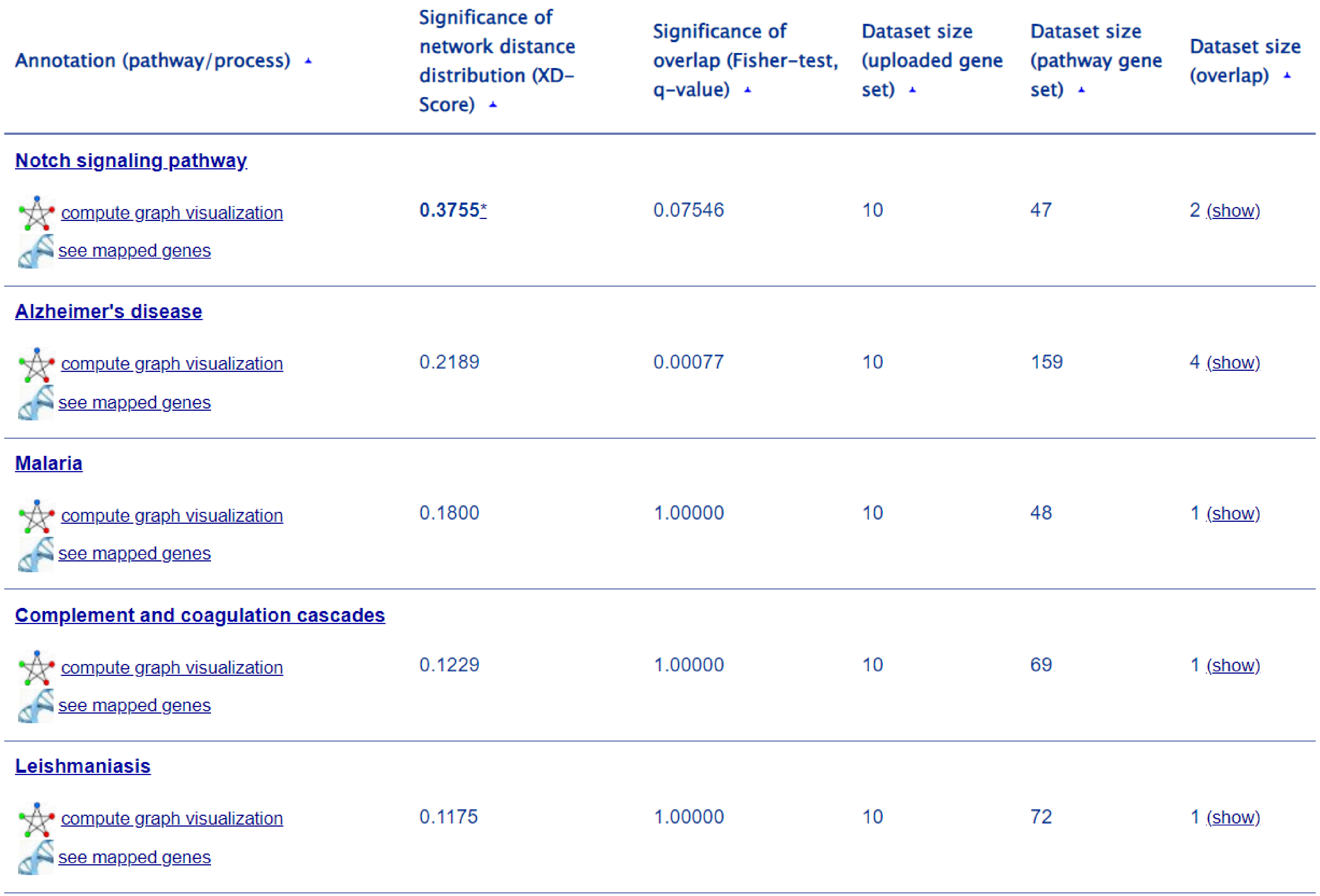
Enrichment Analysis

STRING network visualization shows the interactions among several key genes associated with Alzheimer's disease.



These interactions are crucial for understanding the complex molecular mechanisms underlying Alzheimer's disease.

The pathway enrichment analysis using **EnrichNet** identifies the Alzheimer's disease pathway as the most significantly enriched, with a q-value of 0.00077. This strong association highlights the relevance of the provided gene set to Alzheimer's disease.



Annotation (pathway/process) ▲	Significance of network distance distribution (XD-Score) ▲	Significance of overlap (Fisher-test, q-value) ▲	Dataset size (uploaded gene set) ▲	Dataset size (pathway gene set) ▲	Dataset size (overlap) ▲
Notch signaling pathway compute graph visualization see mapped genes	**0.3755***	0.07546	10	47	2 (show)
Alzheimer's disease compute graph visualization see mapped genes	0.2189	0.00077	10	159	4 (show)
Malaria compute graph visualization see mapped genes	0.1800	1.00000	10	48	1 (show)
Complement and coagulation cascades compute graph visualization see mapped genes	0.1229	1.00000	10	69	1 (show)
Leishmaniasis compute graph visualization see mapped genes	0.1175	1.00000	10	72	1 (show)

Conclusions

- The Lasso model is the most suitable for this dataset, effectively preventing overfitting and selecting important variables.

- Cross-validation techniques and the use of multiple analytical methods ensured the robustness and reliability of the results.

- This approach identified significant genes and pathways involved in Alzheimer’s disease, providing a strong foundation for further studies and investigations.

Thank you for the attention