Network Data Analysis Project

Virginia Leombruni

2024-06-13

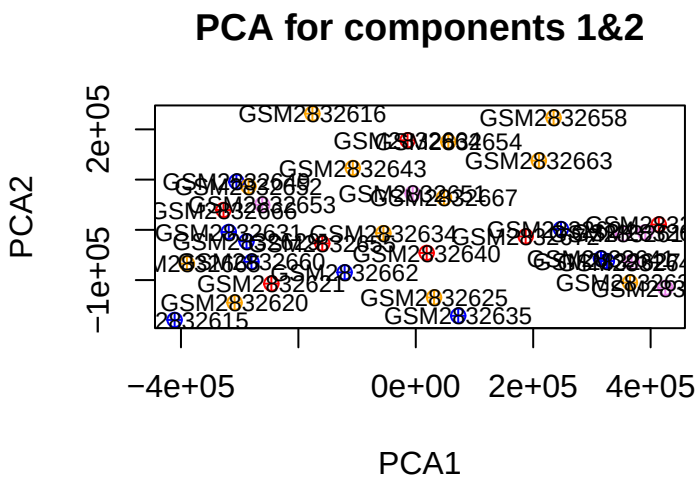
Introduction

The study analyzed the GSE106241 dataset, comprising 71 autopsied temporal cortical samples with varying degrees of Alzheimer's disease (AD)-related neurofibrillary pathology. The samples were categorized into seven groups based on Braak's staging, indicating the severity of the pathology. For statistical significance, stages 0 and 1 were combined into a control group, and stages 5 and 6 into an affected group, ensuring sufficient samples for a meaningful comparison of disease extremes. Subsequent analyses were performed on a total number of affected individuals of 19 and control individuals of 16.

The boxplot suggests that the data preprocessing steps, including normalization, were likely effective. The consistency and symmetry across channels indicate that the data is ready for further analysis, such as PCA.

PCA (Principal Component Analysis)

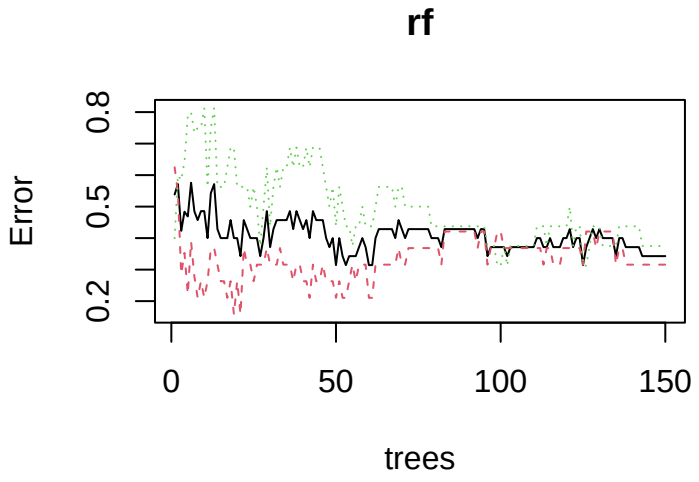
The PCA reduces the dimensionality of gene expression data and makes it possible to visualise relationships between samples. The PCA plot provided shows the distribution of samples based on the first two principal components (PCA1 and PCA2).



Different colours mixed without clear separation, suggest less distinction between the stages based on the data captured by PCA1 and PCA2. The dataset has a limited number of samples, clustering is not robust. For this reason, clustering of the data was not considered during the analysis.

Random Forest

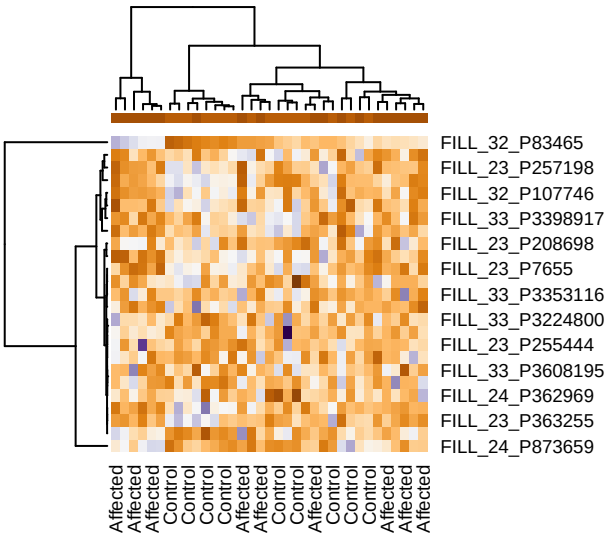
The Random Forest algorithm is used to analyse gene expression data and identify the most important genes distinguishing between the 'Affected' and 'Control' groups.



The performance analysis of the Random Forest model showed that 150 trees provided a stable error rate and good generalization, offering the best balance between computational efficiency and model reliability. To create a simpler, more interpretable model, the 200 most important genes were identified, thereby reducing the dimensionality of the dataset.

A random forest model identified and visualised the most important genes in the dataset. The selection of the 25 most important genes can be used for further analysis, such as pathway analysis or functional studies, to better understand the role of these genes in the biological processes under investigation.

Heatmap



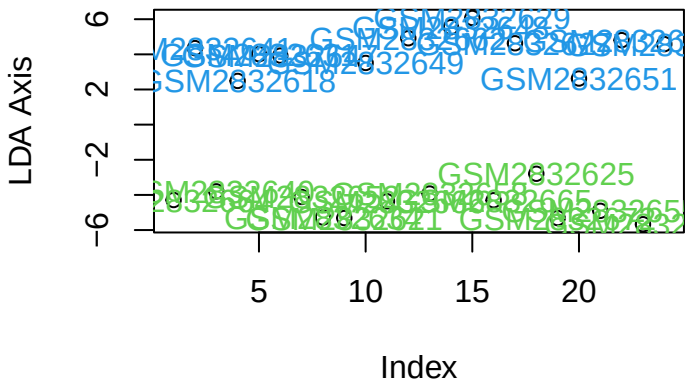
The heatmap displays differences in gene expression between the two groups, with certain genes being more highly expressed in one group compared to the other. The expression patterns are consistent within each group, suggesting distinct molecular profiles that could indicate potential biomarkers.

Identification of Genes with Significant Differences in Expression

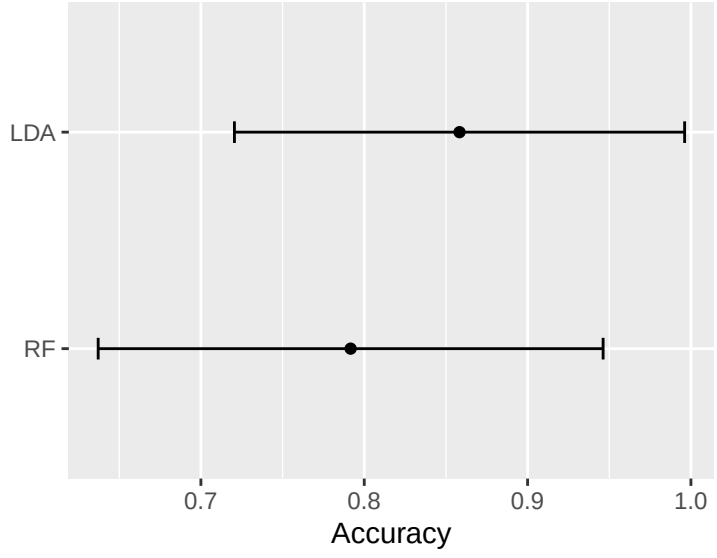
From the analyses performed, 16 genes exhibited significant differences in expression between the control and affected groups, as indicated by an adjusted p-value of less than 0.05. These genes show marked differences in expression, suggesting their potential involvement in disease-related pathological processes.

LDA (Linear Discriminant Analysis)

By running LDA, the variables are collinear. This means that two or more independent variables in the regression model are highly correlated with each other. To reduce the collinearity between the variables before running the LDA, PCA was applied, thus improving the stability and interpretability of the model.

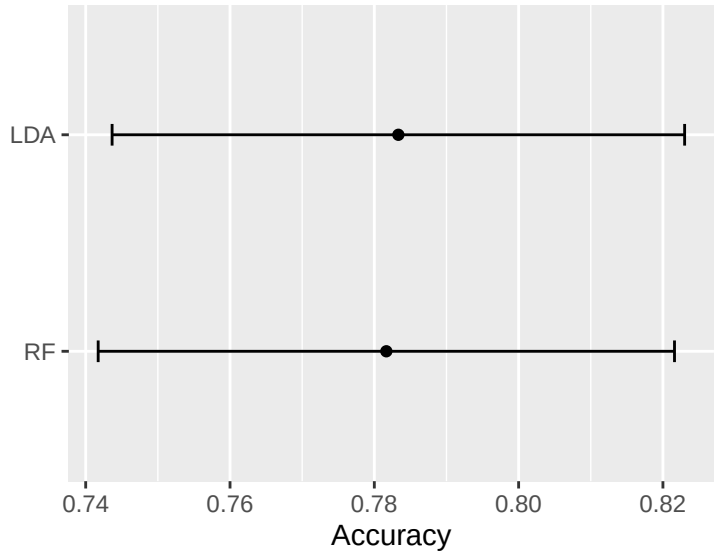


The LDA showed excellent discriminating ability, with all samples correctly classified in their respective groups. This is indicated by the perfect separation of the samples in the graph and the confusion matrix with no misclassification errors. However, to ensure the robustness of the model, cross-validation techniques must be used to exclude the possibility of overfitting."



Repeated Cross Validation

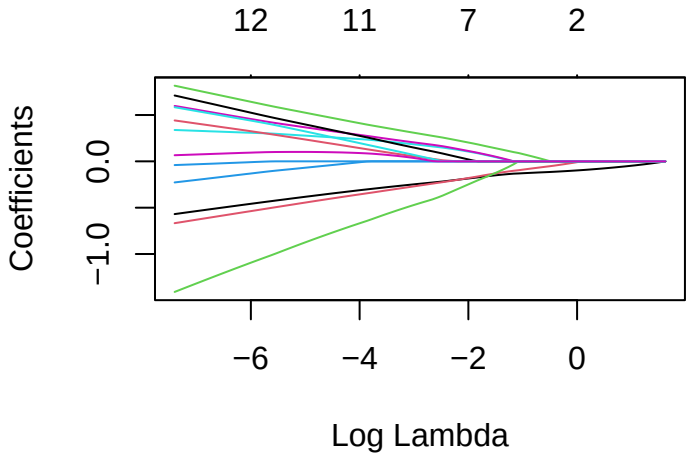
Repeated cross-validation is used to obtain more reliable estimates of model performance by reducing the variance associated with a single cross-validation run.

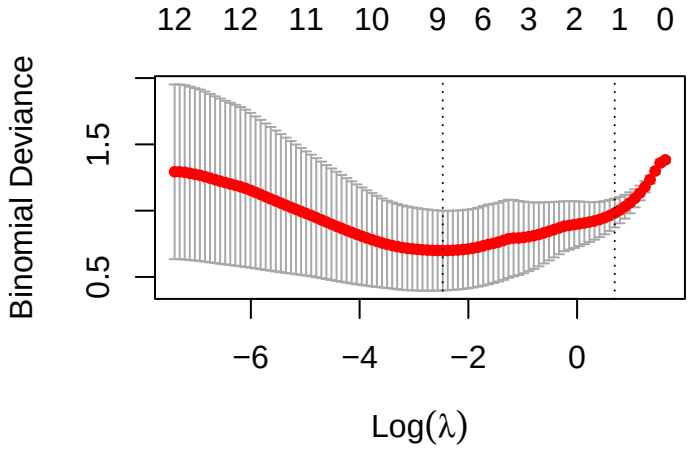


The results of the cross-validation indicate that both LDA and RF models exhibit strong performance with high accuracy and consistent results across folds.

Lasso and Ridge

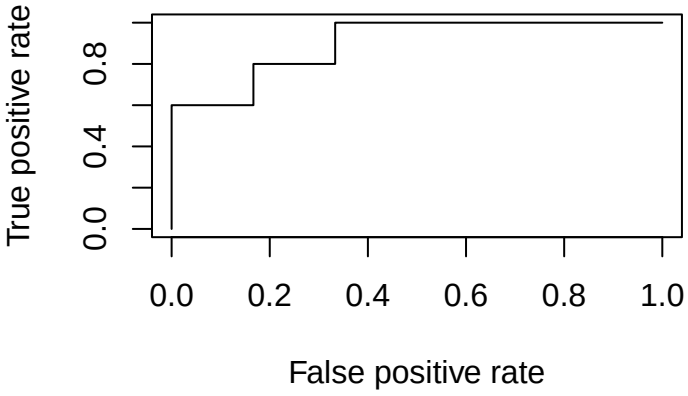
Lasso and Ridge are methods that help improve the model's generalization by adding a penalty that prevents overfitting.

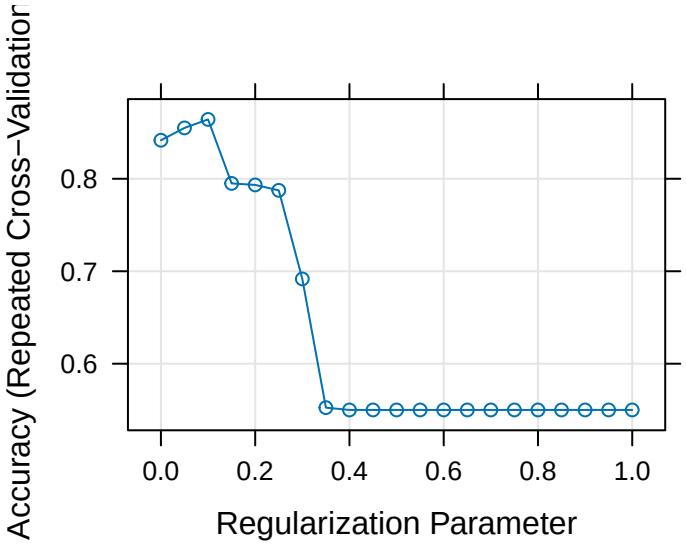


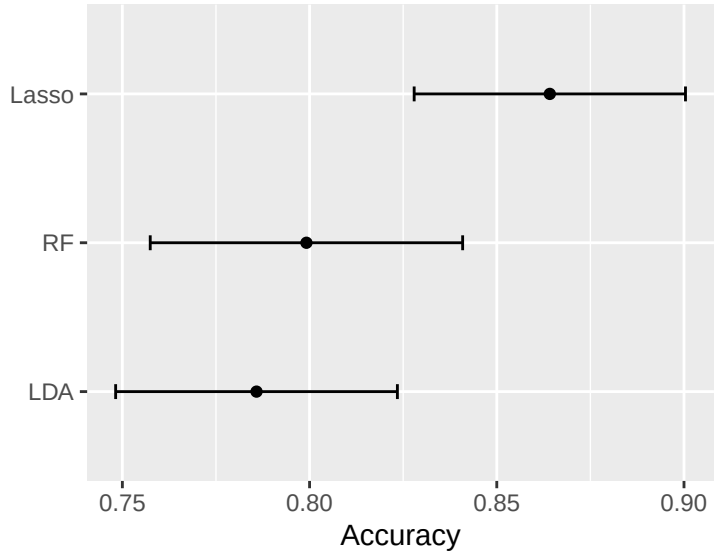


The Lasso model identified key principal components for classification with non-zero coefficients. Cross-validation found the optimal lambda, minimizing binomial deviance, indicating good generalization. Test predictions confirmed the model's effectiveness in selecting important variables and performing well on unseen data.

Plot ROCR


True positive rate
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

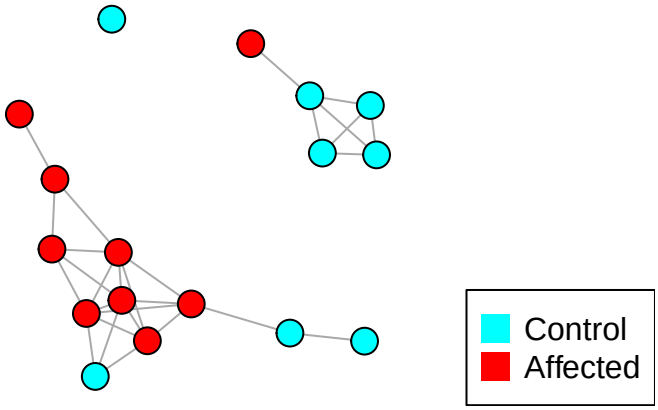

Accuracy (Repeated Cross−Validatior
0.8
0.7
0.6
0.0
0.2
0.4
0.6
0.8
1.0
Regularization Parameter



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

RSCUDO + CARET

```
library("igraph")
library("rScudo")
library("caret")
set.seed(123)
# Create training and test partitions
inTrain <- createDataPartition(f, list = FALSE)
trainData <- ex3[, inTrain]
testData <- ex3[, -inTrain]
# Perform training with rScudo
trainRes <- scudoTrain(trainData, groups = f[inTrain], nTop = 25, nBottom =
 ↪  25, alpha = 0.5)
model <- scudoModel(nTop = (2:6)*5, nBottom = (2:6)*5,N = 0.25)
control <- caret::trainControl(method = "cv", number = 5, summaryFunction =
 ↪  caret::multiClassSummary)
cvRes <- caret::train(x = t(trainData), y = f[inTrain], method = model,
 ↪  trControl = control)
testRes <- scudoTest(trainRes, testData, f[-inTrain], cvRes$bestTune$nTop,
 ↪  cvRes$bestTune$nBottom5)
testNet <- scudoNetwork(testRes, N = 0.2)
scudoPlot(testNet, vertex.label = NA) # perform classification of testing
 ↪  samples using best nTop & nBottom values
```




```
classRes <- scudoClassify(trainData, testData, 0.25, cvRes$bestTune$nTop,
 ↪  cvRes$bestTune$nBottom, f[inTrain], alpha = 0.05)
#caret::confusionMatrix(classRes$predicted, f[-inTrain])
```

The network visualization illustrates the clustering of samples into two distinct groups: the red nodes represent the affected group, and the blue nodes represent the control group. This visual representation effectively demonstrates the separation and classification performance of the model, highlighting its ability to distinguish between the affected and control samples.

Enrichment Analysis

```
library(gprofiler2)
gene_list <- c("APP", "PSEN1", "PSEN2", "APOE", "MAPT", "CLU", "BIN1", "CR1",
 ↪  "SORL1", "TREM2")
gostres <- gost(query = gene_list, organism = "hsapiens", ordered_query =
 ↪  FALSE, multi_query = FALSE, significant = TRUE, exclude_iea = FALSE,
 ↪  measure_underrepresentation = FALSE, evcodes = FALSE, user_threshold =
 ↪  0.05, correction_method = "g_SCS", domain_scope = "annotated", custom_bg
 ↪  = NULL, numeric_ns = "", sources = NULL, as_short_link = FALSE)
p <- gostplot(gostres, capped = TRUE, interactive = FALSE)
```

```
highlight_terms <- c("GO:0036477", "GO:0030425", "GO:0097447", "GO:0009986",
 ↪  "GO:0043005", "GO:0045202", "GO:0030054", "REAC:R-HSA-3928663")
highlight_term <- c("GO:0036477", "REAC:R-HSA-3928663")
```

From the enrichment analysis, it can be observed that the most relevant GO term is GO:0036477. This term refers to a specific component in the neuron involved in Alzheimer's disease.

ID translation in R

```
library(hgu133a.db)
library(AnnotationDbi)
genes_of_interest <- c("APP", "PSEN1", "PSEN2", "APOE", "MAPT", "CLU",
 ↪  "BIN1", "CR1", "SORL1", "TREM2")
probes <- AnnotationDbi::select(hgu133a.db, keys = genes_of_interest, columns
 ↪  = "PROBEID", keytype = "SYMBOL")
entrez_ids <- AnnotationDbi::select(hgu133a.db, keys = probes$PROBEID,
 ↪  columns = "ENTREZID", keytype = "PROBEID")
file_path <- "C:/Users/virgi/Downloads/enrichnet_ranking_table.txt"
enrichnet_data <- read.delim(file_path, header = TRUE, sep = "\t")
```

Network-Based Analysis

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, offering valuable insights into the molecular mechanisms underlying the condition.

Conclusions

The analysis indicated that 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.