Optimizing RNA-Seq-Based Prediction of Major Depressive Disorder Using Machine Learning

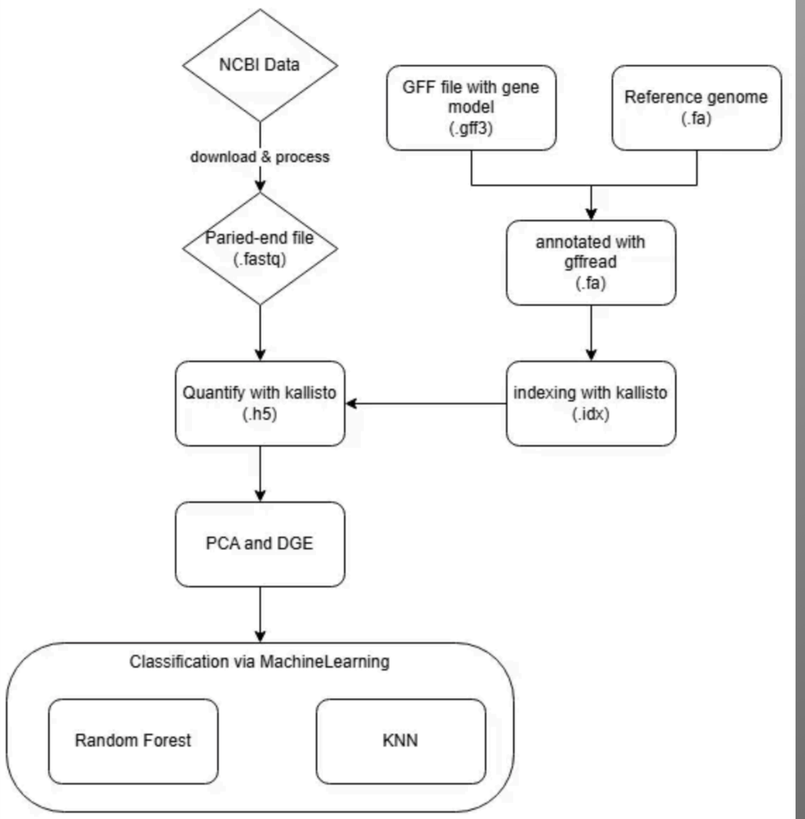
Umit Akirmak & Quoc Dung Van

University of Victoria

ABSTRACT

- RNA-Seq data for 59 participants from the NCBI database.
- Data processed with Kallisto; transcripts selected via PCA on raw and DEG-based data.
- Random forest and k-nearest neighbor (KNN) used to classify controls, MDD, and MDD-S groups.
- Key Findings:
  - Random forest outperformed KNN.
  - Gene expression features yielded better results than raw transcript features.
  - Differences in PCA and classifier performance

METHOD



RESULTS

Table 1. Comparison of classifier accuracy based on different approaches

	1197 raw transcripts		99 DGE transcripts	
ML Classifier	Train accuracy	Test accuracy	Train accuracy	Test accuracy
Random Forest	100%	48%	100%	96%
K-Nearest Neighbor	79.4%	44%	70.6%	48%

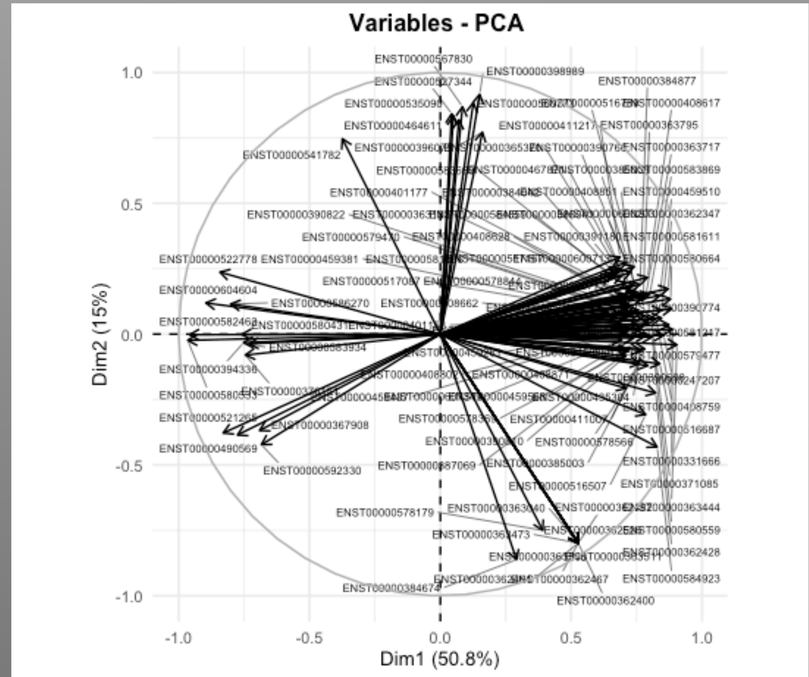


Figure 1. PCA plot of the 99 differentially expressed genes, with Ensembl transcript IDs displayed as data labels

REFERENCE

Verma, Pragya and Shakya, Madhvi (2022) Machine learning model for predicting Major Depressive Disorder using RNA-Seq data: optimization of classification approach. *Cognitive Neurodynamics*, **16**, 443–453.

Table 2. Confusion matrix for the classification problem of recognizing CON, MDD, and MDD-S using Random Forest

	Predicted class		
Actual Class	Control	MDD	MDD-S
Control	12	0	0
MDD	1	3	0
MDD-S	0	0	9

CONCLUSION

We replicated and extended the analysis of RNA-Seq data to predict Major Depressive Disorder (MDD) using machine learning. Transcripts were identified through PCA and gene expression analysis, using Kallisto for quantification, whereas the original study employed Cufflinks. Classification was performed using random forest and k-nearest neighbor algorithms. Consistent with the original study, random forest outperformed KNN, and gene expression features improved classification accuracy. Differences in PCA and classifier performance estimates underscore the impact of preprocessing and feature selection. Our results highlight the importance of optimizing analytical pipelines for reliable and reproducible RNA-Seq machine learning applications.