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

Umit Akirmak [1*], **Quoc Dung Van** [2*]

[1]Department of Computer Science, University of Victoria, 3800 Finnerty Road, Victoria BC, V8P 5C2, Canada
[2]Department of Electrical and Computer Engineering, University of Victoria, 3800 Finnerty Road, Victoria BC, V8P 5C2, Canada

ABSTRACT

We present our replication of the analyses conducted in the paper titled "*Machine learning model for predicting Major Depressive Disorder (MDD) using RNA-Seq data: optimization of classification approach*". RNA-Seq data for 59 participants were accessed from the NCBI database and processed using Kallisto. Relevant transcripts were identified based on PCA of raw transcripts and PCA of the results from differential gene expression (DGE) analysis. Two machine learning algorithms, random forest and k-nearest neighbor (KNN), were evaluated to classify participants into control, MDD, and MDD-S (suicidal) groups based on transcript data. Our results replicated the key findings of the original study, showing that (i) the random forest classifier achieved high accuracy and outperformed KNN, and (ii) using DGE transcripts yielded better results than raw transcripts. However, notable differences were observed in the specific estimates of PCA and machine learning performance, likely due to variations in processing pipelines, which are discussed further.

INTRODUCTION

The sequencing of an organism's transcriptome is referred to as RNA sequencing and is widely used to examine the expression levels of transcripts (1). Classification algorithms, such as machine learning, can leverage gene expression data to make predictions about diagnoses and disease classifications, which can be crucial for prevention or medication (2).

A prior study investigated how RNA-seq data can be utilized to accurately classify individuals with depression and healthy controls (1). The study employed machine learning algorithms, including random forest and k-nearest neighbor (KNN), to classify RNA-seq data into control and depressed groups. The authors reported that random forest outperformed KNN in classification accuracy, particularly when utilizing differentially expressed genes (DGEs) instead of raw transcript data. Building on these findings, our study aims to replicate the analyses, assess the reproducibility of results, and explore the impact of varying preprocessing pipelines on classification performance.

METHOD

In this project, we replicated the method of the original paper with modified in quantification process, illustrated in figure 2. The difference is instead of using Cufflink, we used Kallisto for quantification process. The result had lead to significant improvement of classification accuracy which we will present in the result section.

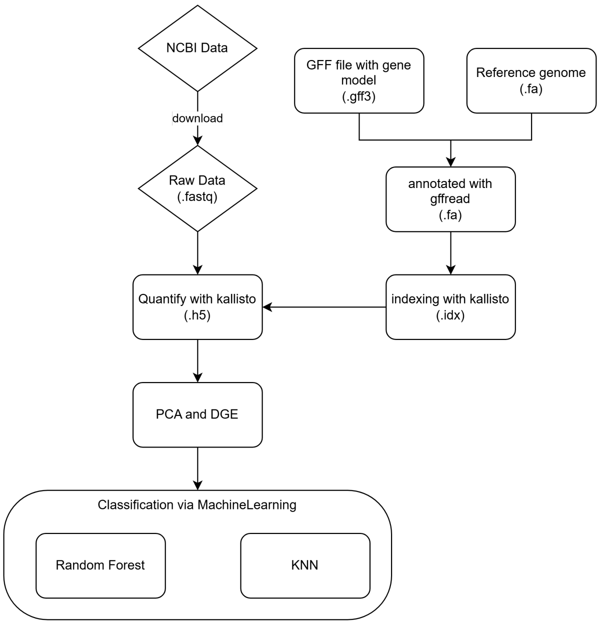

Figure 1. Our modified flow in this project which lead to improvement of classification accuracy

RNA sequencing and gene quantification expression

We analyzed the same RNA-seq data for 59 participants obtained from the NCBI database (3). This database contains 21 MDD-S (subjects with major depressive disorder and suicide), 9 MDD (subjects with MDD and no suicide)

[*]To whom correspondence should be addressed. Email: uakirmak@uvic.ca

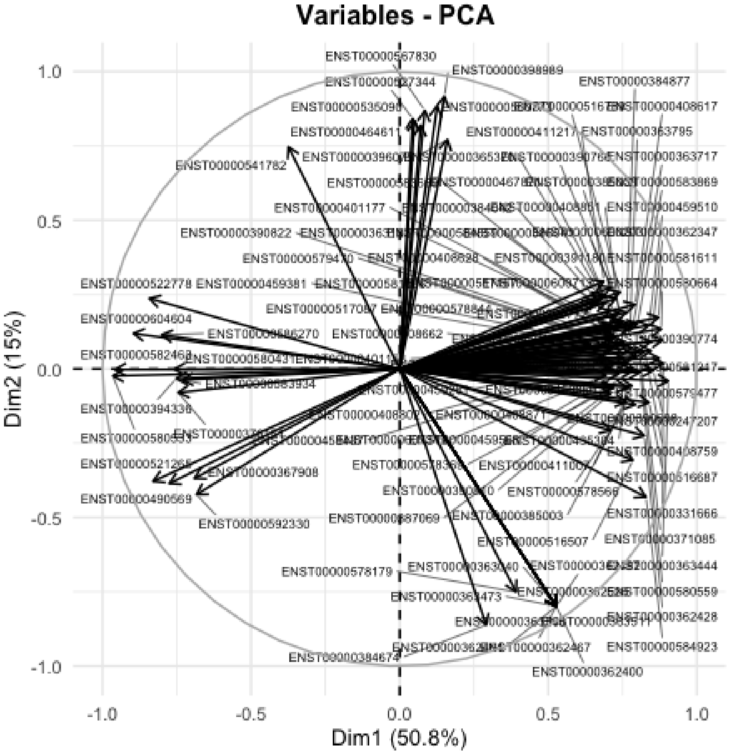

Figure 2. PCA plot of the 99 differentially expressed genes, with Ensembl transcript IDs displayed as data labels. PC1 and PC2 together explain 65.8% of the variability in the data.

and 29 CON (sudden death healthy control) samples. The downloaded data is processed into paried-end FASTQ files before passing to Kallisto software for transcript quantification.

Before quantifying with Kallisto, we also prepared an annotation file for the genomic sequence using Homo_sapiens.GRCh37.87.gff3 and using Homo_sapiens.GRCh37.dna.primary_assembly.fa from Ensemble as referenced genome. After indexing and quantifying gene expression, Kallisto generated .h5 abundance files for each sample, providing estimates of transcript expression levels. These abundance files were then imported into R using custom scripts for downstream analysis, including data normalization and feature extraction.

Feature selection PCA and differential gene expression DGE

Principal Component Analysis (PCA) was conducted separately on the raw transcript expression levels and on the results of differential gene expression (DGE) analysis. Differentially expressed genes were identified by applying thresholds for log2-transformed transcript counts and adjusted p-values, ensuring the inclusion of only statistically significant and biologically meaningful transcripts. The log transformation was applied to stabilize variance and handle outliers, enhancing the robustness of PCA and further machine learning analyses. Initial differential expression was determined using likelihood ratio testing, with a significance threshold set at $p < 0.5$ for transcript selection.

Using Sleuth, we identified a set of statistically significant differentially expressed transcripts. From this set, an initial subset of 1,000 transcripts was selected based on variance. This subset was further refined to 99 transcripts, chosen for their highest contributions to the first two principal components, in accordance with the methodology of the original study. PCA was subsequently performed on this refined set, focusing on 80 transcripts contributing to the first dimension and 19 transcripts contributing to the second dimension (see Figure 2). This targeted approach facilitated the analysis of the most informative transcripts within the differentially expressed gene set.

For the raw transcript analysis, PCA was similarly performed on the full dataset. Transcripts were ranked by variance, and the top 1,197 transcripts were selected for further analysis. This analysis was conducted to replicate the methodology and approach of the original article, enabling a consistent comparison between Sleuth-based differential expression results and raw transcript data to evaluate which would perform better in machine learning algorithms.

Classification with machine learning: Random forest and KNN

Two machine learning algorithms, random forest and k-nearest neighbor (KNN), were applied to classify the RNA-seq data into three categories: controls, individuals with Major Depressive Disorder (MDD), and those with suicidal tendencies, MDD-S. The performance of the classifiers was evaluated based on accuracy and other key metrics. The analyses were conducted on datasets derived from both

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%

raw transcript expression levels and DGE-transformed data to assess the impact of feature selection on classification performance.

RESULTS

To assess the benefit of feature selection and DGEs analysis, the classification is performed with 2 approaches. First approach is classifying transcripts after applying PCA analysis on the 76486 raw transcripts (gave 1197 important transcripts). Another is after applying PCA on 1000 DGEs transcripts (gave 99 most significant transcripts).

To test our classification, the 59 samples are divided into training data (60% of the data) and testing data (40% of the data). The metric for evaluating the classification is the accuracy derived from the confusion matrix. The diagonal of the matrix represents the number of samples correctly classified for each class, while off-diagonal elements indicate misclassifications. By dividing the total number of correctly classified samples (sum of the diagonal elements) by the total number of samples, we obtain the accuracy. This provides a simple way to evaluate the classifier's performance.

The confusion matrix for predicting the testing data using Random Forest is shown in Table 2. And Table 1 provides accuracy metric results for 2 approaches with 2 types of classifier (Random Forest and KNN). It is obvious that the classifier for DGE transcripts show superior performance than the raw transcripts, and in general Random Forest performs better than KNN in any cases.

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

In this study, we replicated and extended the methodology for classifying experimental conditions using transcript data and machine learning models. Our analysis utilized both raw transcript counts and differentially expressed genes identified through Sleuth. By applying dimensionality reduction with PCA and feature selection, we aimed to evaluate whether focusing on DGEs improved classification performance compared to raw transcript data.

Our results demonstrated that models trained on DGEs achieved higher classification accuracy, particularly for the Random Forest classifier, which reached 96% test accuracy with the 99 DGEs subset compared to 48% with 1197 raw transcripts. This aligns with the hypothesis that biologically meaningful feature subsets can enhance model performance. However, our findings differed from the original study in multiple aspects, including classification accuracy, confusion matrix outcomes, and PCA results. The original study reported a distinct PCA pattern, whereas our results showed variations in variance capture, likely due to differences in pre-processing and feature selection.

Several factors may explain these discrepancies. While we applied log2 transformation to stabilize variance and manage outliers, the original study did not specify this step. Furthermore, the methods used to select DGEs and extract features from PCA likely influenced the results. Finally, differences in the way classifiers were implemented could also explain the observed variations.

Despite these variations, our findings confirm the importance of dimensionality reduction and feature selection in improving classification accuracy in transcriptomic datasets. Future studies should aim to standardize pre-processing pipelines, refine feature selection methods, and validate findings across diverse datasets to enhance reproducibility and generalizability. Integrating alternative machine learning frameworks may further improve classification robustness.

SUPPLEMENTARY MATERIALS

The code, data, and supplementary materials for this project are available on our GitHub repository (https://github.com/uakirmak/bioinformatics-course-project). The repository includes detailed documentation on methods, data pre-processing, and machine learning implementation. These materials are provided for research and educational purposes.

REFERENCES

1. 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.
2. Zararsiz G, Goksuluk D, Korkmaz S, Eldem V, Duru IP, Unver T, & Ozturk A (2014). Classification of RNA-Seq data via bagging support vector machines. *b*ioRxiv, 007526.
3. NCBI Dataset `https://www.ncbi.nlm.nih.gov/biosample?LinkName=bioproject_biosample_all&from_uid=394722`