Project Proposal

Replication of Research paper entitled *"Molecular markers of early Parkinson's disease based on gene expression in blood"*

PREPARED FOR

Dr. Suhila Sawesi, Ph.D., MPharm, BPharm

sawesis@gvsu.edu

PREPARED BY

Rashmita Vaggu

Meher Nivedita Avdut

Muttaki I. Bismoy

October 3, 2023

Project Overview

1. Introduction:

A multifaceted etiology characterizes Parkinson's disease (PD), a complicated neurological condition. For a successful intervention and better patient outcomes, PD must be diagnosed early. A promising method for identifying potential molecular markers for early PD diagnosis is presented in the 2007 study "Molecular markers of early Parkinson's disease based on gene expression in blood" by Scherzer et al. With the aid of contemporary data analytic methods and datasets, this replication study seeks to confirm and expand the conclusions of this fundamental work.

2. Aims and Objectives:

To replicate the following from the paper:

a. Expression data matrix of eight marker genes
b. Validation of the risk marker on independent test samples
c. The ROC curve in the test set
d. Exploring additional insights and potential markers using advanced analytical techniques
e. Disseminating the findings through publication and contributing to the understanding of early PD diagnosis.
f. Dopamine replacement medication

The primary objective of this study is to replicate the methods and findings of the original paper. The focus will be on reproducing the original methodology and analyses to assess the robustness and generalizability of the reported results.

The key aspects of the study are:

- Effectiveness of the Molecular Marker: Investigating how well the molecular marker, characterized by risk scores and gene expression patterns, distinguishes between individuals with PD and healthy controls
- Robustness Across Independent Samples: Assessing the robustness and reliability of the identified molecular marker by validating its predictive value across an independent set of samples

- Influence of Dopamine Replacement Medication: Examining whether the use of dopamine replacement medication by PD patients affects the predictive accuracy of the molecular marker, making sure that the marker is effective no matter the status of the medication.

3. Research questions:

Our research questions are centered around identifying molecular markers for early Parkinson's disease. We aim to address the following specific aims:

1. To what extent does the identified molecular marker, determined through risk scores and gene expression patterns, effectively differentiate individuals with Parkinson's disease (PD) from healthy controls and other neurodegenerative disease controls?
2. Replicate the statistical analysis methods used in the paper to identify potential molecular markers.
3. Interpret the biological significance of the identified markers and their potential use in early PD diagnosis.
4. Additionally, how robust and reliable is the predictive value of the marker across independent test samples, and does the inclusion of individuals on dopamine replacement medication influence the predictive accuracy of the marker?

Methodology & Approach

1. Data Collection:

The acquisition of gene expression data will be conducted using a dataset that is pertinent and up-to-date, with a preference for utilizing publicly accessible sources or establishing cooperation with esteemed academic institutes. The dataset will be curated to encompass samples from both people diagnosed with Parkinson's disease (PD) and those without PD, in addition to incorporating pertinent clinical data.

2. **Dataset:**

   To create the dataset, neurology board-certified movement disorders specialists enrolled 50 PD patients and 55 age-matched healthy and neurodegenerative disease controls.
   Link :- https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2519

3. **Data Analysis:**

a. Replication of Statistical Analysis:
   The statistical analytic methods outlined in the original research, encompassing differential gene expression analysis and the identification of possible markers, will be replicated in our study.

b. Validation:
   The found markers will be validated by the utilization of cross-validation on the replication dataset and, if possible, on other datasets.

c. Machine Learning:
   In order to enhance the predicted accuracy, we want to employ machine learning algorithms, such as random forests and deep learning, to find supplementary markers.

4. **Interpretation:**

   The assessment of the identified indicators' biological significance and possible clinical utility will be conducted by means of pathway analysis and integration with pre-existing knowledge on the biology of Parkinson's disease (PD).

Statistical/Machine Learning Algorithms

SVM, or Logistic Regression: is suitable for predicting and classifying individuals based on gene expression data. SVM excels at capturing non-linear relationships, while logistic

regression provides interpretable results and is well-suited for binary classification tasks. The following are the steps:

1. **Data Preprocessing:**
   Standardize and preprocess the gene expression data to ensure consistency.

2. **Model Training:**
   Train SVM, Logistic Regression, and Random Forest models using the original study's training set.

3. **Parameter Tuning:**
   Optimize model parameters through cross-validation to enhance performance.

4. **Testing:**
   Apply the trained models to the independent test set and compare the results with the original study.

5. **Evaluate Performance:**
   Assess the accuracy, precision, recall, and other relevant metrics to ensure the reliability of the replication.

Expected Outcomes and Contributions

1. Successful replication of the molecular marker associated with Parkinson's disease (PD) risk
2. Discovery of potential novel markers using advanced analytical techniques
3. Evaluation of the accuracy of risk prediction models using SVM, or Logistic Regression
4. Identification of genes contributing significantly to the risk prediction
5. Contribution to the field of Parkinson's disease research and early diagnosis
6. Comparative analysis of results with the original study

Insight to uncover

1. Insights into the performance and generalizability of different algorithms for predicting PD risk based on gene expression data
2. Uncovering specific genes that play a crucial role in distinguishing PD patients from controls provides potential targets for further research.

Timeline

Check the enclosed Gantt chart for a comprehensive schedule delineating significant milestones and dates pertaining to various stages of the project.

Task Name	Start	End	Duration (days)
Project Team Contract	9/25/2023	10/2/2023	7
Project Proposal	9/29/2023	10/4/2023	5
Literature Review	10/9/2023	10/20/2023	11
Dataset Collection	10/16/2023	10/23/2023	7
Project Data Analysis	10/23/2023	11/1/2023	9
Replication Analysis	11/7/2023	11/13/2023	6
Validation	11/11/2023	11/15/2023	4
Documentation	10/16/2023	11/15/2023	30
Final Project Report	11/11/2023	11/22/2023	11
Project Presentation	11/20/2023	11/29/2023	9

Fig. Project Timeline

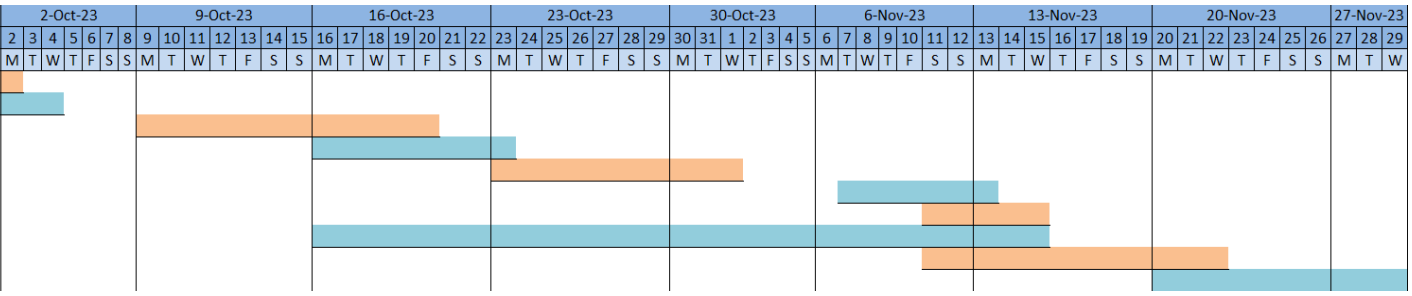

Fig. Gantt Chart

Conclusion

The objective of this study is to reproduce and extend the results presented in the research article titled "Molecular markers of early Parkinson's disease based on gene expression in blood," authored by Scherzer et al. Our methodology encompasses the gathering of data, meticulous examination, verification, and comprehension, employing contemporary analytical methodologies. The anticipated results will enhance our comprehension of early Parkinson's disease diagnosis and may have significant clinical implications.

Reference

1. Scherzer, C. R., Eklund, A. C., Morse, L. J., Liao, Z., Locascio, J. J., Fefer, D., Schwarzschild, M. A., Schlossmacher, M. G., Hauser, M. A., Vance, J. M., Sudarsky, L. R., Standaert, D. G., Growdon, J. H., Jensen, R. V., & Gullans, S. R. (2007) Molecular markers of early Parkinson's disease based on gene expression in the blood *Proceedings of the National Academy of Sciences of the United States of America*, *104*(3), 955–960. https://doi.org/10.1073/pnas.0610204104