Ben-Gurion University of the Negev
Industrial Engineering and Management

אוניברסיטת בן-גוריון בנגב
המחלקה להנדסת תעשיה וניהול

Project No.	**Project Title**	
2022-01-105	Clustering-based Prediction of ALS deterioration	
Academic Advisor	**Co-Advisor**	
Prof. Boaz Lerner	-	
Team Members		
Roy Wolf		
wolfr@post.bgu.ac.il		

Abstract

ALS is a devastating incurable disease that affects the patient's motor neurons. The disease causes progressive loss of structure or functionality of neurons that leads to muscle weakness, disability, and eventually death. ALS has no sufficient risk factors as only 10% of the cases can be attributed to genetic/familial reasons, and most cases are sporadic. In this project, I use the PRO-ACT database, created by the non-profit organization Prize4Life, which contains over 10,000 ALS patient records from multiple completed clinical trials.

The main goal of this project is to create a pipeline that will be used to predict a patient's future disease deterioration rate. The pipeline consists of two main phases. In the first phase, we are using unsupervised machine learning approach to divide the patient's population by their disease progression rates into homogenous clusters. In the second phase, prediction models are trained based on the information obtained in the first phase. I hope that utilizing information from the first phase will improve our ability to predict the patient's future deterioration rate. Using this pipeline, when a new patient arrives, he will be assigned to a cluster, and using the cluster specific prediction model, his future deterioration rate will be predicted.

To test the pipeline, I constructed four different experiments, designed to test each module. The first experiment is designed to determine the optimal number of clusters. We used different metrics for the evaluation of the clustering scheme and selected the best preforming number of clusters. The second experiment is designed to determine the most important features that should be used for prediction of future deterioration rates for each cluster. I created a prediction task and used feature importance measures from random forest algorithm to evaluate each variable. The third experiment is testing different assignment methods of a new (test) patient to a cluster. To evaluate the assignment methods, I compared the partition created by the assignments of the model and the true clustering scheme. The last experiment examined different approaches to use the clustering information for the prediction of future disease deterioration, comparing them to a baseline model, which is trained on the entire patient's population.

In this work, we demonstrated that using unsupervised machine learning to cluster patients, can be beneficial in both understanding the disease as well as predicting the future disease deterioration rates. This pipeline can be implemented on different databases.

Keywords: Machine Learning, Clustering, Prediction, Random Forest, ALS