Project No.	Project Title	
2022-01-108	Concept drift detection and re-learning to model ALS progression	
Academic Advisor		**Co-Advisor**
Prof. Boaz Lerner		
Team Members		
Moraz Finegold		
Morazf@post.bgu.ac.il		

Abstract

Most machine learning algorithms assume that the data it is using to generate a model come from a stable process. A major challenge in data stream applications is the changes in the target and other variables over time in unexpected ways, a phenomenon called concept drift (CD). Often, these changes make the model built on old data inconsistent with the new data, and regular updating of the model is necessary. By detecting concept drift in the medical field, the identification of changes can provide information about how the different variables can affect or be affected throughout disease progression.

Amyotrophic lateral sclerosis (ALS) is an incurable neurodegenerative disease. The risk factors are still not entirely known to clinicians, but it is known that only 5% to 10% are hereditary, and the rest are sporadic. Since so little is known about the disease, it takes a long time to diagnose it. Early detection may allow patients to adjust their lifestyle to cope with the changes expected and to avoid the stress of being left without options when they have suddenly lost an ability.

The purpose of this study is to model ALS progression using a Bayesian network (BN) that detects a concept drift due to changes occurring in the values of laboratory test results of a patient, before re-learning a new network using the current state data. The BN graphical model demonstrates relationships between variables representing the laboratory tests and ALS disease throughout time, allowing us to interpret the drift in the model.

The database we used included electronic medical records of 810 subjects from Meuhedet HMO, of which 162 are ALS patients and 648 are control subjects matched to patients by their age, gender, and area of residence in a 1:4 ratio. Since a patient medical record, regarded as time-series data, is sometimes missing, imputation was performed using different methods to reach a database that would model disease progression well. To test the methods, random values were deleted and filled using each method. The evaluation measure was root-mean-square error (RMSE) when the error is between the true and filled values. First, imputation of missing values between existing periods was examined. The results show that for most laboratory tests the method that yielded the lowest RMSE was linear interpolation according to time, but for some other laboratory tests, the mean of the K nearest neighbors, when neighbors are found in half a year time windows, was best. Extrapolation methods for filling missing values at periods not between existing periods will be examined in a similar way.

Keywords: ALS, Machine learning, Concept drift, Bayesian network.