Research Plan Proposal

**Title:** Disease Module Detection Using a Combined Machine Learning

approach

**Professor:** Jagath Rajapakase

**Supervisor:** Rama Kaalia

**Author:** Li Mengyang

**Background**

During the first semester, we had a deeper insight into the disease module detection problem history and observed from the several commonly used algorithm including hierarchical based agglomerative algorithm and divisive algorithm, greedy based Louvain algorithm, and random walk based Markov Chain Cluster algorithm (MCL). Inspired by the idea of using multi-relational association and converting the data used for each dimension into feature [1], we plan to consolidate the idea of combined machine learning approach. To fit the data more accurate into the model for future research, it is proposed that we build the machine learning model for a certain type of disease, where dementia would be a great choice.

Combined machine learning approach is to combine some brilliant ideas used detecting disease module in commonly used algorithms and the idea of model building by tuning the model data. Detailed methods are explained in Method section [2].

**Aim**

By the end of January, we should be able to obtain the following:

1. Three layers of feature data of mathematical feature, topological feature and functional feature respectively
2. Further algorithmic improvement on the convolutional neural network to fit the disease module model better

**Method**

In a general way, we aggregate nodes with the highest similarity with the known disease nodes together. The way we define "similarity" is specifically comprehensive for biological disease module, which is to combine most significant mathematical features, topological features, and functional features.

Mathematical feature: The covariance of each node to disease nodes of several types of dementia is calculated according to the equation i). As covariance ranges in , we are able to obtain the distinct value difference for sample nodes to each disease.

Cov(x,y)=∑(xi−¯x)(yi−¯y)n=⟨x−¯x, y−¯y⟩n (i)

Topological feature: To improve the method developed in [1], four topological properties are being used here, connectivity significance, global modularity, degree centrality, and articulation protein, where connectivity significance is calculated by the definition defined in [3].

Functional feature: GO term, biological pathway and protein domain are used here (we can add more here).

With the above featured data, we convert them into several layers of feature map with each row represents one type of dementia disease and each column represents one sample node. (Not sure if a long rectangle kernel is feasible).

At the same time, we will seek an algorithm-based way to improve convolutional neural network and decision tree. For example, by storing intermediate data into a max heap will greatly reduce the linear searching time to logarithmic searching time. By combining various algorithm and data-structures concepts, with the same code, building models for other types of disease is achievable in a shorter time as well as in a more disease-type-oriented accurate way.

**Validating**

Different from ordinary machine learning project, statistics including accuracy of disease detecting module is not measurable directly. Series of indirect measuring technique thus can be used:

1. By removing a certain fraction of seed nodes, observe the change in modularity

computed from the original model and computed from the new model.

2. Compare the number of hints hit (through text mining) computed from the newly trained model with the multi-relational association mining model, DIAMOnD model, Louvain algorithm, and MCL.

3. Performance comparison of fitting known, predictable nodes between the newly trained model with others.