NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

School of Computer Science and Engineering
College of Engineering

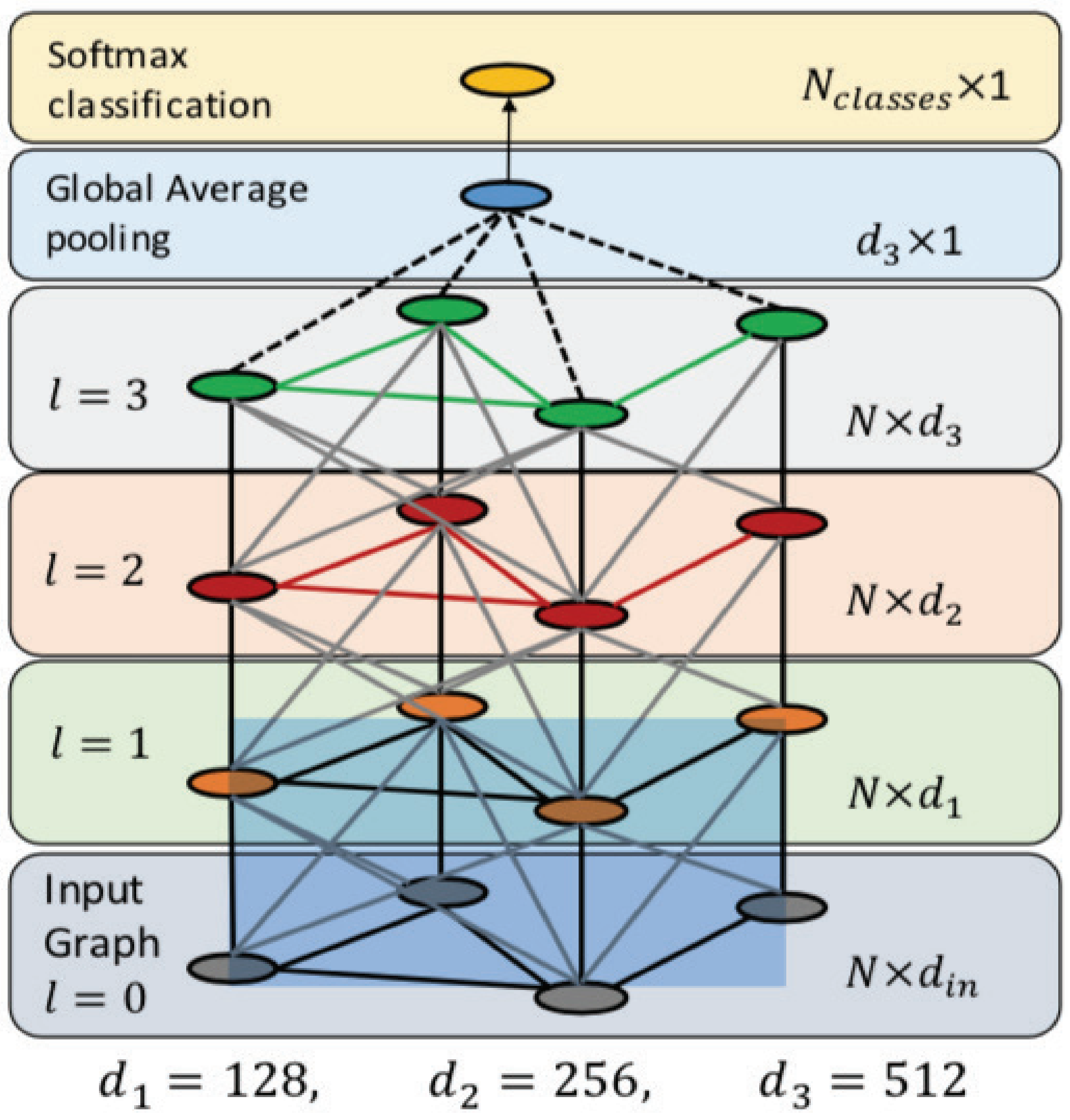
Interpretable Graph Classification

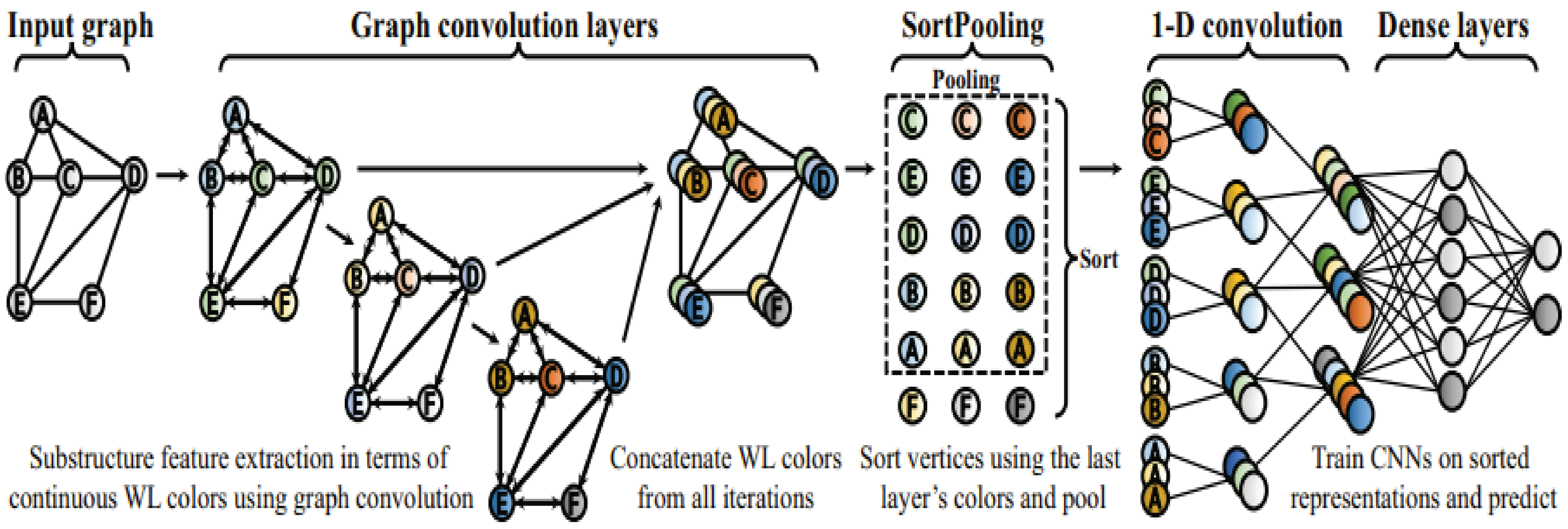
Evaluation and interpretation of graph classification using deep learning

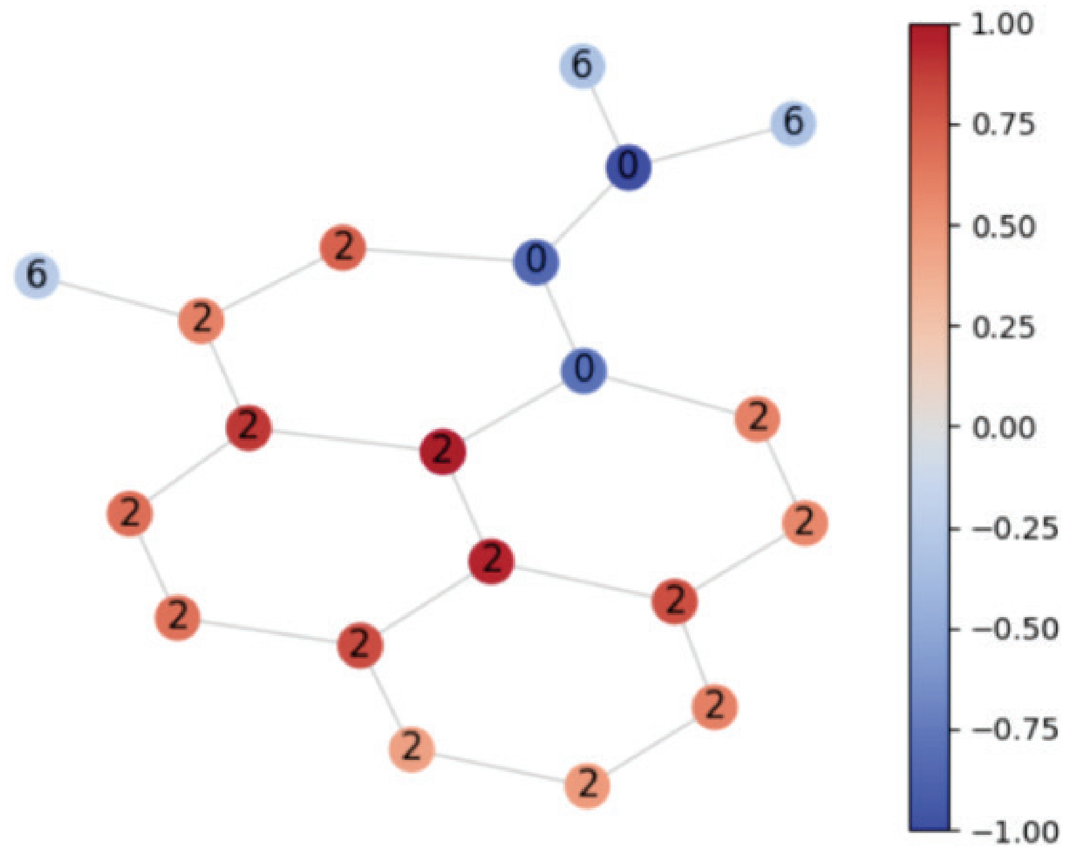
Student: Eko Edita Limanta **Supervisor:** Asst Prof Arijit Khan

Neural networks are black-box in nature. In many real-life applications, it is of vital importance to know why the neural networks make certain decisions. For example, if a protein is classified as toxic due to presence of certain molecules.

This project aims to evaluate several interpretability methods for graph classification using deep learning. Two deep learning architectures are used: Deep Graph Convolutional Neural Network (DGCNN) and Graph Convolutional Networks(GCN).




Input graph
Graph convolution layers
SortPooling
Pooling
Sort
1-D convolution
Dense layers
A
B
C
D
E
F
Substructure feature extraction in terms of continuous WL colors using graph convolution
Concatenate WL colors from all iterations
Sort vertices using the last layer's colors and pool
Train CNNs on sorted representations and predict



Two interpretability methods – DeepLIFT and Contrastive Gradient – are then applied to these architectures. The interpretability results are then evaluated using three quantitative metrics: fidelity, contrastivity, and sparsity.