# Graph Neural Networks for Tabular Data Prediction

*Granick, Joseph; Gupta, Sarthak*

## Abstract
While deep neural networks have revolutionized machine learning for structured data types such as images and text, their success has been less pronounced in tabular data settings. Tabular data lacks the inherent spatial or sequential structure that convolutional neural networks (CNNs) and recurrent neural networks (RNNs) can exploit. Recent advancements in Graph Neural Networks (GNNs) suggest that graph-based models may offer a promising solution for learning latent relationships within tabular data.

## Introduction
Since the resurgence of neural networks in 2011, driven by increased computational power and availability of large datasets, they have dominated machine learning tasks, particularly in image and text domains. However, their application to tabular data, which is common in many industry applications like finance and healthcare, remains limited. Traditional machine learning models like decision trees and gradient boosting tend to outperform more complex deep learning models when the structure of the data is more arbitrary, particularly for smaller datasets.[2] Graph Neural Networks (GNNs) show promise improving prediction performance for tabular data by learning and representing latent factors that other neural network architecture is not conducive to learning without the mathematical rigidity of the grid structure of images or the sequential structure of text.[1,2]

## Motivation
Difficulties applying deep learning models to tabular data is primarily driven by the often arbitrary relationships between features. Shallower ML methods and manual feature engineering are still typically the preferred approach for practical applications on tabular datasets. Unlike CNNs and RNNs, which struggle with tabular data since it lacks the rigid structure of representations like pixel grids for images, or sequential dependence for text data.[1]

GNN research indicates that embedding the graph structure can exploit feature interactions by constructing graph structures over tabular data columns.[2] This motivates the exploration of GNNs as a means to improve model performance on tabular datasets, where other approaches are lacking.[1,2]

## Objectives

1. **Constructing graph structures**: Various graph structures will be explored, including homogeneous and heterogeneous graphs.[1,2]
2. **Model comparison**: We will compare the performance of GNN-based models against CNNs, RNNs, and traditional machine learning models (e.g., random forests and gradient boosting) in both classification and regression tasks.[1,2]
3. **Evaluation of sample size impact**: The project will assess how the models' performance changes with different sample sizes, testing GNNs' ability to generalize better with relatively smaller datasets exploiting structure of connectivity in the graph representation structure.[1]
4. **Evaluation of feature complexity**: We will attempt to evaluate how the model performance changes with different dimensionality of complexity, both in terms of complexity of fit with respect to the size of the network as well as complexity of the representation with respect to feature vector dimensionality.

## Expected Result

The project aims to contribute insights into how GNNs can enhance the performance of machine learning models on tabular data. By systematically evaluating GNN-based approaches alongside conventional methods, the work seeks to provide a clearer understanding of the practical scenario where the nature of the application and the structure, complexity, and volume of data make GNNs advantageous to both traditional machine learning methods and more common neural networks like CNNs and GNNs

**References**

Cheng-Te Li, Yu-Che Tsai, Chih-Yao Chen, Jay Chiehen Liao. *Graph Neural Networks for Tabular Data Learning: A Survey with Taxonomy & Directions.* arXiv preprint, 2024. Available at: https://arxiv.org/abs/2401.02143.

Xiawei Guo, Yuhan Quan, Huan Zhao, Quanming Yao, Yong Li, Weiwei Tu. *TabGNN: Multiplex Graph Neural Network for Tabular Data Prediction.* Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2021. Available at: https://doi.org/10.1145/3447548.3467327.