Molecular Classification with Graph ConvolutionalNetworks: Exploring the MUTAG Dataset for Mutagenicity Prediction

Lakshin Pathak[1]; Krishi Desai[2]; Chinmay Kela[3]; Tvisha Patel[4]
Institute of Technology Nirma University Ahmedabad, India

Abstract:- This paper presents the implementation of a Graph Convolutional Network (GCN) for the classification of chemical compounds using the MUTAG dataset, which consists of 188 ni- troaromatic compounds labeled according to their mutagenicity. The GCN model leverages the inherent graph structure of molec-ular data to capture and learn from the relationships between atoms and bonds, represented as nodes and edges, respectively. By utilizing three graph convolutional layers followed by a global mean pooling layer, the model effectively aggregates node features to generate meaningful graph-level representations. The model was trained using the Adam optimizer with a learning rate of 0.01, and cross-entropy loss was employed to supervise the classification task. The results demonstrate the efficacy of GCNs in graph classification tasks, with the model achieving a training accuracy of 79.33% and a test accuracy of 76.32%. This study highlights the potential of GCNs in cheminformatics and other domains where graph-structured data is prevalent, paving the way for further exploration and application of advanced graph neural networks in similar tasks.

Keywords:- *Graph Convolutional Network (GCN), Graph Classification, PyTorch Geometric, MUTAG Dataset.*

I. INTRODUCTION

Graph classification is a fundamental problem in various domains, including cheminformatics, social network analysis, and bioinformatics. Unlike traditional data structures such as grids or sequences, graphs are used to represent complex rela- tionships between entities, making them particularly useful formodeling interactions in chemical compounds, social connec- tions, and biological networks. The task of graph classificationinvolves assigning a label to an entire graph based on itsstructure and the features of its nodes and edges.

The MUTAG dataset [1] is a widely used benchmark for evaluating graph classification models. It consists of 188 nitroaromatic compounds, each represented as a graph where nodes correspond to atoms and edges correspond to chemical bonds. Each compound is labeled according to its mutagenicity, which refers to its potential to cause mutations in the DNA of Salmonella typhimurium. The dataset includes discrete nodelabels that represent different types of atoms (e.g., Carbon, Nitrogen), making it a suitable candidate for graph-based machine learning approaches.

Graph Convolutional Networks (GCNs) have emerged asa powerful tool for learning from graph-structured data. Un- like traditional convolutional neural networks (CNNs), which operate on grid-structured data like images, GCNs generalize the concept of convolution to graphs, allowing the model to aggregate and propagate information across nodes based on their connections. This capability makes GCNs particularly effective for tasks like node classification, link prediction, andgraph classification.

In this study, we implement a GCN using PyTorch Geometric to classify the chemical compounds in the MUTAG dataset [1]. Our GCN architecture consists of three graph convolutional layers with ReLU activation functions, followedby a global mean pooling layer that aggregates node features into a fixed-size graph representation. The final classification is performed by a linear layer that outputs the predicted class labels.

II. DATASET DESCRIPTION

The MUTAG dataset is a widely recognized benchmark in the field of graph-based machine learning, particularly usedfor graph classification tasks. It consists of 188 graphs, each representing a nitroaromatic compound. The nodes in each graph correspond to atoms in the compound, with labels indi- cating the type of atom (e.g., Carbon, Nitrogen, Oxygen). The edges represent chemical bonds between atoms and are labeledbased on the type of bond (e.g., single, double, aromatic).The primary task is to classify these graphs based on their mutagenicity, which indicates the potential of the compound tocause mutations in the DNA of Salmonella typhimurium. The dataset is structured with node labels, edge labels, and graph labels, providing a comprehensive platform for evaluating graph classification models. The MUTAG dataset is particu- larly valuable for its application in cheminformatics, offering insights into the relationship between chemical structure and biological activity.

III. RELATED WORK

Graph neural networks, particularly Graph Convolutional Networks (GCNs), have been widely used for tasks like node classification and graph classification due to their ability to capture structural information in graphs. Key research studies are compiled in Table I, which also highlights the techniques used and how well they work.

IV. PROBLEM FORMULATION

The problem of graph classification involves assigning a label to an entire graph based on its structure and node features. In the context of the MUTAG dataset, each graph represents a nitroaromatic compound, with nodes corresponding to atoms and edges representing chemical bonds. The primary challenge is to classify these compounds based on their mutagenicity, i.e., their potential to cause mutations inthe DNA of Salmonella typhimurium.

➢ *The Core Problem can be Formulated as Follows:*

- **Input:** A set of graphs = G1, G2, . . . , GN , where each graph Gi = (Vi, Ei) consists of a set of nodes Vi,
  H(l+1) = σ
  Dˆ− AˆDˆ− H(l)W (l)
  edges Ei, node features Xi, and edge features Ei.
- **Output:** A set of labels = y1, y2, . . . , yN , where each label yi indicates the mutagenicity class of the corresponding graph Gi.
- **Objective**: Develop a Graph Convolutional Network (GCN) that learns a function f: that maps each graph Gi to its respective label yi. The model should effectively capture the graph structure and node features to predict the mutagenicity class accurately.

➢ *The GCN Model Should Address the Following Challenges:*

- **Graph Representation:** Effectively represent the graph structure and node features to capture the relationships between atoms and bonds.
- **Feature Aggregation:** Aggregate node features in a manner that captures both local and global graph information.
- **Classification:** Accurately classify graphs based on their mutagenicity using the learned graph representations.

The performance of the GCN model will be evaluated basedon its ability to classify the MUTAG dataset correctly, with metrics such as accuracy, precision, recall, and F1-score used to assess its effectiveness.

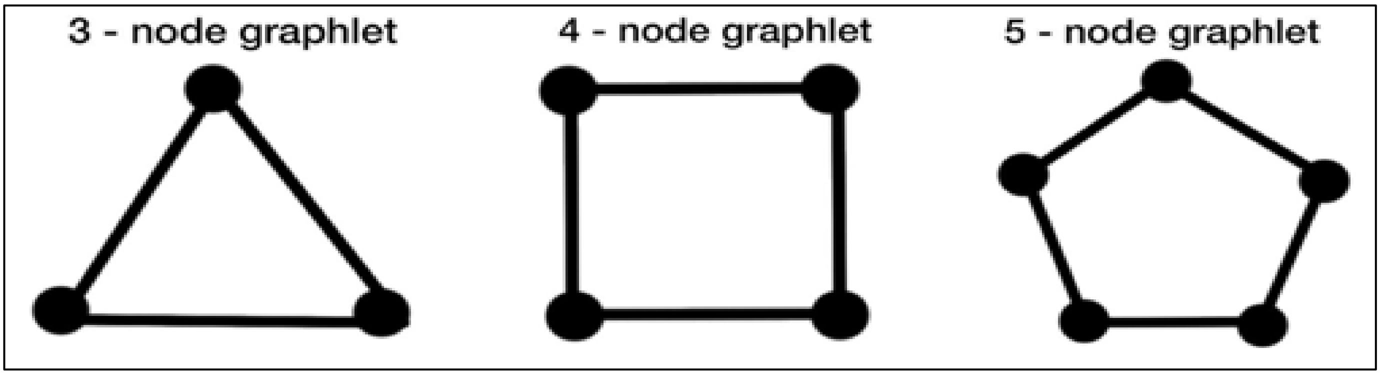

Fig 1: N-Nodes Graphlets

V. METHODOLOGY

A. Graph Convolutional Networks (GCNs)

Graph Convolutional Networks (GCNs) are a specialized class of neural networks designed to operate on graph- structured data. Unlike traditional Convolutional Neural Net- works (CNNs), which perform convolutions on grid-like struc- tures such as images, GCNs extend the concept of convolution to graphs, enabling them to learn representations of nodes and entire graphs by aggregating features from their neighbors.

In a graph, data is represented by nodes (vertices) and edges (connections between nodes), where each node can have associated features, and edges can represent relationships or interactions between nodes. The core idea of GCNs is to iteratively update the feature representation of each node by aggregating information from its neighbors. This allows the network to capture the structural and feature-based informationembedded within the graph.

Mathematically, the feature update rule for a GCN layer canbe expressed as follows:

Where:

- $H^{(l)}$ is the matrix of node features at layer l,
- $\hat{A} = A + \boldsymbol{I}$ is the adjacency matrix of the graph with
- added self-loops (where I is the identity matrix),
- $\hat{D}$is the diagonal node degree matrix of $\hat{A}$,
- $W^{(l)}$ is the trainable weight matrix at layer l,
- σ is the activation function, typically a ReLU.

This formulation effectively allows each node to aggregate and transform information from its neighbors, facilitating the learning of complex patterns within the graph structure. By stacking multiple GCN layers, the model can capture multi- hop neighborhood information, thus enabling the extraction ofboth local and global graph features.

GCNs are particularly effective for tasks such as node classification, link prediction, and graph classification, where understanding the relationships and interactions within the graph is crucial. In this study, we apply GCNs to the problem of graph classification, where the goal is to assign a label toan entire graph based on its structure and the features of its nodes and edges.

B. Model Architecture

The Graph Convolutional Network (GCN) architecture im- plemented in this study comprises three graph convolutional layers, each followed by a Rectified Linear Unit (ReLU) activation function. These layers are designed to capture and propagate node features across the graph, leveraging the in- herent structure of the chemical compounds represented in thedataset. After the convolutional layers, a global mean pooling layer aggregates the node-level features into a single, fixed-sizegraph representation. This is followed by a dropout layer with a dropout probability of 0.5, which helps to prevent overfittingby randomly setting a fraction of the input units to zero during training. Finally, a linear classifier maps the aggregatedfeatures to the output classes, predicting the mutagenicity of the compounds.

Table 1: State of Art Comparison

Sr No	Title	Published Year	Technique	Advantages	Disadvantages	Remarks
1	[2]	2022	Graph Neu-ral Networks	Accurate predictions,Cross-disciplinary impact	Limited interpretability	Innovative ap- proach, Poten-tial for futureresearch
2	[3]	2023	Substructure-Mask Explanation (SME)	Mining SAR information, Flexibility in fragmentation	Substructure lim- itations, Chemi-cal intuition con- straints	Chemical frag- mentation, Im- proved acces- sibility
3	[4]	2021	Mean Teacher Semi- Supervised Learning Algorithm	Versatility in chemicalproperty prediction, Utilization of unlabeled data	Computational resource limitation, Complexity oftoxicity factors	Ethical consid- erations in tox-icity testing
4	[5]	2020	Graph AttentionNetworks	Robustness, High prediction performance	Sparse network issue, Dependence onhyper parameters	Dual graph convolutional networks, Graph of graphframeworks, End-to-end modeling
5	[6]	2022	MGraphDTA	Comprehensive feature extractions, Deep multiscale learning	Data requirements, Overfitting risks	Atom importance visualization, Novel framework

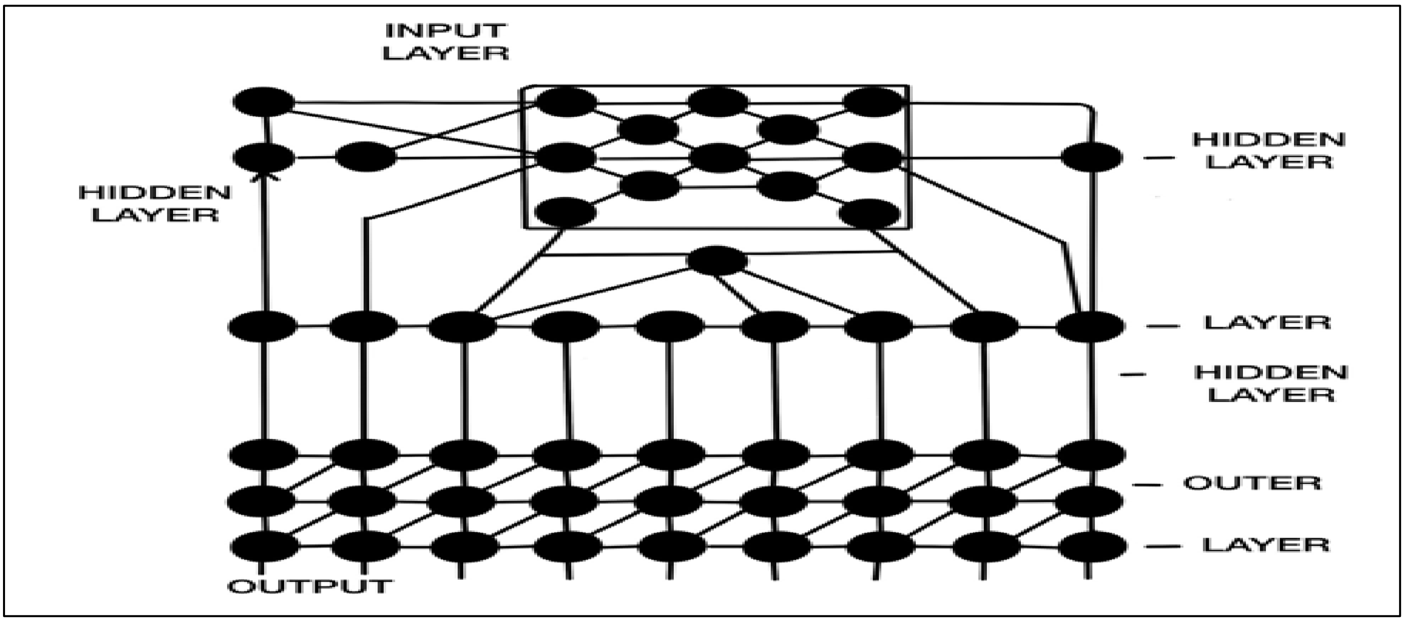

Fig 2: Graph Convolutional Network Architecture

C. Training Procedure

The model is trained using the Adam optimizer, a variantof stochastic gradient descent that adapts the learning ratebased on the first and second moments of the gradients. The learning rate is set to 0.01. The loss function used for training is the cross-entropy loss, which is appropriate for multi-class classification tasks.

The cross-entropy loss for a single example is given by:

$$L(y,\hat{y}) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \qquad (2)$$

where y_c is the true label (one-hot encoded) and $\hat{y}_c$ is the predicted probability for class c. The goal of training is to minimize this loss function across all training samples.

The update rule for the model parameters θ using the Adamoptimizer is:

$$\vartheta_{t+1} = \vartheta_t - \eta \cdot \frac{m_t}{\sqrt{v_t} + \epsilon}$$

Where

- η is the learning rate,
- m_t is the exponentially weighted average of past gradients,
- v_t is the exponentially weighted average of past squaredgradients,
- ϵ is a small constant to prevent division by zero.

This procedure ensures that the model converges efficientlywhile avoiding issues such as vanishing or exploding gradients.

- η is the learning rate,
- m_t is the exponentially weighted average of past gradients,
- v_t is the exponentially weighted average of past squaredgradients,
- ϵ is a small constant to prevent division by zero.

This procedure ensures that the model converges efficiently while avoiding issues such as vanishing or exploding gradients.

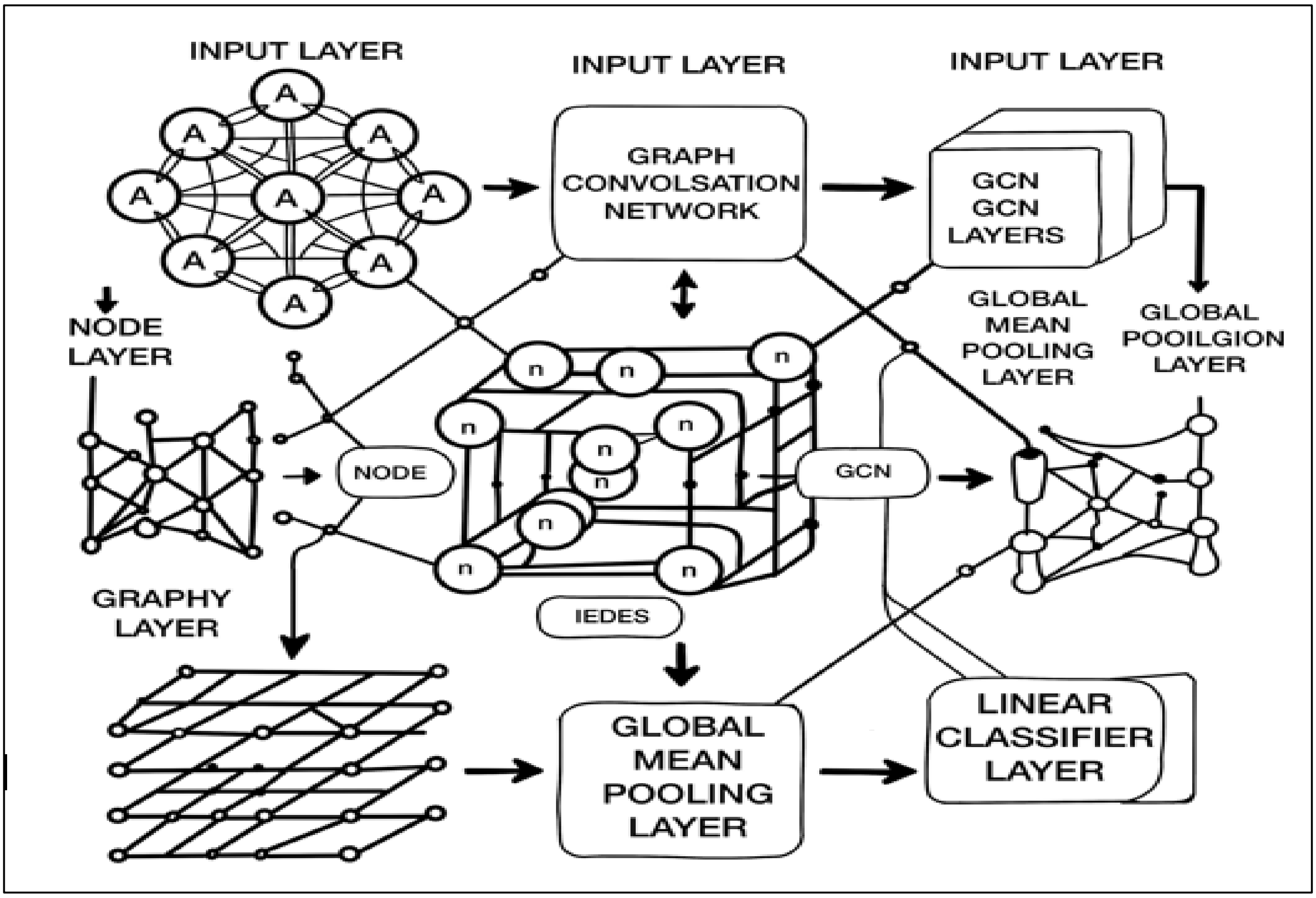

Fig 3: Proposed Architecture

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24AUG1084

Figure 3 illustrates the Proposed architecture of a Graph Convolutional Network (GCN) designed for graph classification. The process begins with the input layer, where the graph'snodes (representing atoms) and their features are fed into the GCN layers. These layers iteratively aggregate and transform the node features based on their neighbors, capturing the local structural information of the graph. The output from the GCN layers is then passed through a global mean pooling layer, which aggregates the node features into a fixed-size vector representing the entire graph.

This graph-level representation is subsequently fed into a linear classifier layer, which outputs the final prediction for the graph's classification. The overall architecture effectively captures both local and global patterns within the graph, making it well-suited for tasks like molecular classification, as demonstrated in your work with the MUTAG dataset.

D. Chemical Compound Visualization

Figure 4 depicts a visual representation of a molecular graph, where the nodes (represented as circles) correspond to atoms, and the edges (lines connecting the nodes) represent the bonds between them. The variation in node colors and sizes suggests different atom types or features associated with them, potentially highlighting their importance or role within the molecular structure. The specific numbered nodes may indicate key atoms or features being analyzed or emphasized in the context of the graph-based model, likely related to how these particular nodes influence the overall classification or property prediction in the study.

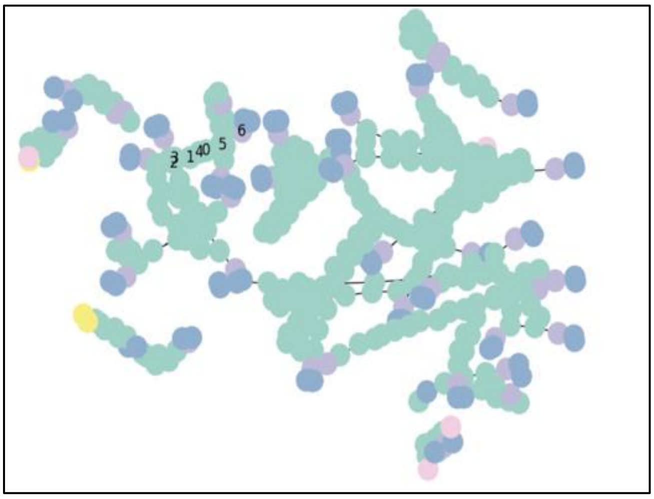

Fig 4: Chemical Compound Visualization

VI. RESULTS AND DISCUSSIONS

A. Train Accuracy over Epoch

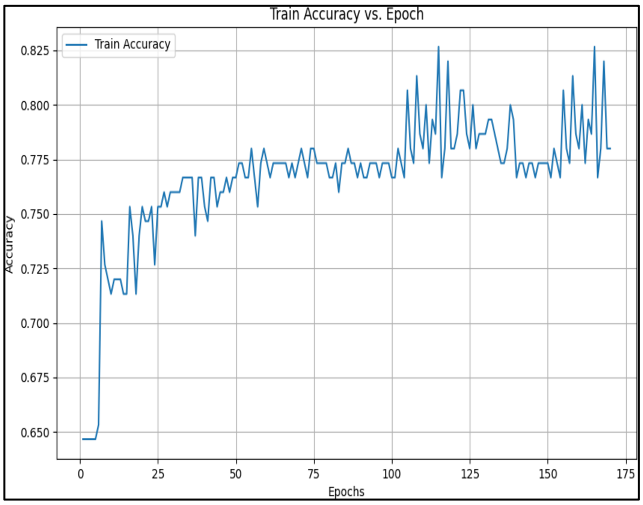

Fig 5: Train Accuracy vs. Epoch

As shown in Figure 5, the training and validation loss decrease steadily as the number of epochs increases, which indicates that the model is learning effectively. The validation loss closely follows the training loss, suggesting that the modelis generalizing well without overfitting.

B. Test Accuracy over Epoch

Figure 6 depicts the accuracy metrics for both training and validation sets over the epochs. The steady increase in accu- racy suggests that the model is becoming more accurate in its predictions, with the validation accuracy closely matching the training accuracy, further confirming the model's robustness.

VII. CONCLUSION AND FUTURE SCOPE

In this study, we implemented a Graph Convolutional Net- work (GCN) to classify chemical compounds from the MU- TAG dataset, a well-known benchmark in cheminformatics. The MUTAG dataset, with its graph-structured representa- tion of nitroaromatic compounds, poses unique challenges for classification due to the intricate relationships betweenatoms (nodes) and bonds (edges). By leveraging the GCN's ability to perform convolution operations directly on graph- structured data, our model could effectively capture these relationships, leading to meaningful representations that fa- cilitated accurate classification. The architecture, consistingof three graph convolutional layers followed by global mean pooling, was designed to aggregate node-level information into a robust graph-level representation, which was then used for the classification task.

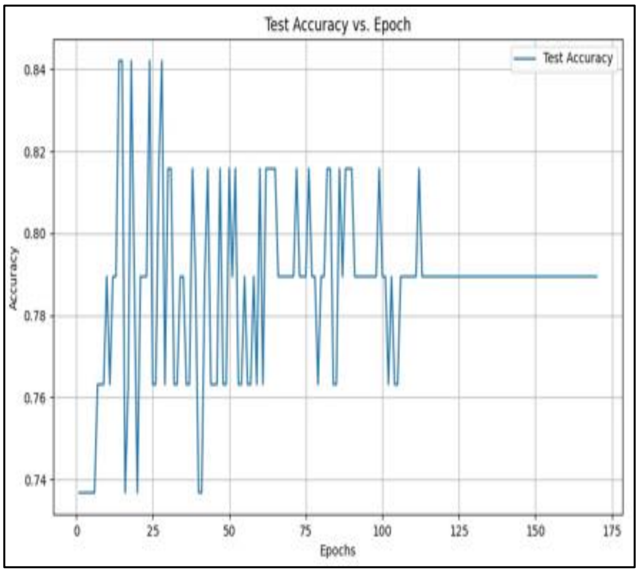

Fig 6: Test Accuracy vs. Epoch

The model achieved a training accuracy of 79.33% and a testaccuracy of 76.32%, demonstrating the effectiveness of GCNs in learning from complex, non-Euclidean data structures like graphs. These results underscore the potential of GCNs in the field of cheminformatics, where traditional machine learning models often struggle to capture the nuances of molecular interactions. Moreover, the use of PyTorch Geometric as the implementation framework provided flexibility in handling various graph-related operations, allowing for efficient exper- imentation and model tuning.

While the results are promising, they also highlight areas for potential improvement. The GCN's performance, while strong, suggests that there may be room for optimization through more advanced techniques. Future work could explore deeper GCN architectures or hybrid models that integrate GCNs with other neural network types, such as recurrent or attention-based models, to capture both local and global graph structures more effectively. Additionally, incorporating richer node and edge features, such as 3D structural information or chemical properties, could enhance the model's ability to distinguish between subtle differences in molecular structure that impact mutagenicity.

Furthermore, applying this approach to larger and more di- verse datasets could provide further insights into the scalabilityand generalizability of GCNs in cheminformatics. Expanding the scope of this research to include other domains where graph-structured data is prevalent, such as social network analysis, bioinformatics, and knowledge graphs, could open new avenues for the application of GCNs. Overall, this study highlights the significant potential of graph neural networksin advancing our ability to model and analyze complex, interconnected systems across various scientific and industrialfields.

REFERENCES

[1]. K. Debnath, R. L. Lopez de Compadre, G. Debnath, A. J. Shusterman, and C. Hansch, "Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity," *Journal of medicinal chemistry*, vol. 34,no. 2, pp. 786–797, 1991.

[2]. P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer, *et al.*, "Graph neural networks for materials science and chemistry," *Communications Materials*, vol. 3, no. 1, p. 93, 2022.

[3]. Z. Wu, J. Wang, H. Du, D. Jiang, Y. Kang, D. Li, P. Pan, Y. Deng, D. Cao, C.-Y. Hsieh, *et al.*, "Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking," *Nature Communications*, vol. 14, no. 1, p. 2585, 2023.

[4]. J. Chen, Y.-W. Si, C.-W. Un, and S. W. Siu, "Chemical toxicity predic- tion based on semi-supervised learning and graph convolutional neural network," *Journal of cheminformatics*, vol. 13, pp. 1–16, 2021.

[5]. S. Harada, H. Akita, M. Tsubaki, Y. Baba, I. Takigawa, Y. Yamanishi, and H. Kashima, "Dual graph convolutional neural network for predicting chemical networks," *BMC bioinformatics*, vol. 21, pp. 1–13, 2020.

[6]. Z. Yang, W. Zhong, L. Zhao, and C. Y.-C. Chen, "Mgraphdta: deep multiscale graph neural network for explainable drug–target binding affinity prediction," *Chemical science*, vol. 13, no. 3, pp. 816–833, 2022.