Classifying Research Papers Using Graph Neural Networks

Soham Shimpi
Arizona State University
sshimpi1@asu.edu
Tempe, AZ

Priyanka Ojha
Arizona State University
pojha3@asu.edu
Tempe, AZ

Animesh Singh
Arizona State University
asing574@asu.edu
Tempe, AZ

Shuchi Talati
Arizona State University
stalati1@asu.edu
Tempe, AZ

***Abstract*- This study uses Graph Neural Networks (GNNs) on the Cora dataset to classify scientific papers, improving on conventional approaches like Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs). By using sophisticated graph-based approaches, our model increased classification accuracy from 73.16% to 83.52%. The findings demonstrate the usefulness of GNNs in exploiting the relational dynamics of citation networks, laying the groundwork for future study in scientific literature management.**

Keywords: Graph Neural Networks, Graph Convolutional Networks, Scientific Publication Classification, Citation Networks, Feature Engineering, Data Mining

I. Introduction

The growing development of scientific literature has created considerable hurdles for effective classification and access to academic papers. Traditional categorization approaches frequently fail to effectively capture the intricate linkages inherent in citation networks, resulting in inefficient literature management, diminished author visibility, and restricted extraction of detailed bibliometric insights. These problems emphasize the urgent need for a paradigm shift toward more modern methodologies capable of properly managing and interpreting the expanding number of scholarly papers.

Graph Neural Networks (GNNs) have emerged as a possible answer to these problems, with the potential to transform the categorization of scientific articles by utilizing the relational information contained within citation networks. Our study seeks to construct and evaluate a GNN-based strategy, with a special focus on the Cora dataset, a well-known graph dataset consisting of 2,708 scientific articles, each encoded into a binary word vector and classified into one of seven unique groups depending on content.

A. Literature Review

Traditionally, algorithms such as Support Vector Machines (SVM) and Random Forests have been used to classify scientific publications. These approaches analyze text data separately. Recent improvements have turned the focus to graph-based techniques, namely Graph Convolutional Networks (GCNs), which use relational structures between texts to improve classification performance.

Zhou et al. [1] conducted key research that proved GCNs' ability in exploiting complicated patterns in document categorization. Bruna et al. [2] and Defferrard et al. [3] pioneered spectral-based graph convolutions, which made global graph topologies more usable in learning methods. Kipf and Welling [4] reduced these convolutions, which considerably improved GCNs' scalability and applicability to huge datasets.

Veličković et al. [5] introduced Graph Attention Networks (GATs), which use attention mechanisms to prioritize significant citation patterns in academic datasets. Our effort expands on these breakthroughs by refining graph-based algorithms for better categorization of scientific papers in the Cora dataset, with the goal of increasing accuracy by combining structural and content-based elements.

B. System Overview

Our approach uses a structured, multi-phase process to categorize scientific articles with Graph Neural Networks (GNNs), with an emphasis on improving the extraction and analysis of textual content and citation networks in the Cora dataset. The pipeline begins with data preparation, which involves addressing missing data, selecting important attributes, and creating a graph with nodes representing publications and edges indicating citations. This graph form is critical for representing the intricate interactions between documents.

During the feature engineering phase, we encode each article with a binary word vector for textual content and extract graph-level characteristics like centrality and community structures to help us comprehend the larger scientific context. The model design has numerous graph convolutional layers that analyze these characteristics using neighborhood aggregation approaches, allowing the model to learn from both direct and indirect relational data. Skip connections are used to improve learning dynamics and avoid problems such as vanishing gradients.

The implementation and optimization step entails training the GNN on a split dataset and performing rigorous hyperparameter tweaking to improve performance measures like as accuracy, precision, recall, and F1 score. Our method provides robust processing of graph-structured data and dramatically increases classification accuracy by utilizing the intricate interrelationships seen in scientific articles.

C. Data Collection

The Cora dataset is the primary dataset for this study. It contains 2,708 scientific articles displayed as nodes in a network, with edges indicating citations between papers. Each node is represented by a binary word vector that indicates the presence or absence of certain words from a given lexicon. This dataset provides a solid foundation for assessing the performance of GNNs in categorizing scientific writings.

D. Components of the ML System

The Our machine learning system integrates various components, including:

- Data Preparation: Ensuring the dataset is appropriately formatted and split into training and testing sets.

- Model Architecture: Designing a baseline model and an advanced GCN model with appropriate layers and hyperparameters.
- Hyperparameter Tuning: Regularly adjusting model parameters to optimize performance.

E. Experimental Results

Initial tests using the baseline GNN model yielded a classification accuracy of 73.16% on the test set. Following upgrades and tweaks to the GCN model, accuracy climbed to 83.52%, highlighting the benefits of our graph convolutional technique. These findings show that our GNN-based system can greatly improve the accuracy and efficiency of categorizing scientific papers when compared to existing approaches.

II. Important Definitions

A. Data

- Cora Dataset - The Cora dataset is a collection of scholarly articles consisting of 2,708 publications distributed across seven distinct categories. Each publication is represented as a node in a graph, and the relationships between publications are represented by edges, denoting citations between articles.
- Node Features - The features of each node in the Cora dataset represent the presence or absence of 1,433 unique terms in the publication, encoded as a binary word vector. These features capture the content of each publication and are utilized for classification.
- Graph Structure - The graph structure of the Cora dataset encapsulates the citation network, where nodes represent publications, and edges represent citations between publications. Understanding the graph structure is crucial for leveraging relationships between publications for classification tasks.

B. Prediction Target

The primary objective is to classify scientific publications into predefined categories based on their content and citation relationships. Each publication is assigned one of seven categories: Theory, Reinforcement Learning, Genetic Learning, Neural Networks, Probabilistic Methods, Case-Based, or Rule Learning.

C. Concepts in the Data

- Nodes - Nodes represent individual entities within the dataset, which in this case, are scholarly publications such as research papers, books, or articles. Each node corresponds to a unique publication and is represented as a vertex in the citation network.
- Edges - Edges in the dataset denote the relationships between nodes, specifically citations between publications. If Publication A cites Publication B, there exists a directed edge from Publication A to Publication B.
- Graph Connectivity - Understanding the connectivity between publications through citation links is essential for capturing the influence of one publication on another.

- Publication Categories - The target variable consists of seven predefined categories into which publications are classified. These categories provide the framework for the classification task.

D. Problem Statement

The dataset consists of scholarly articles represented as a citation network, with each publication characterized by binary word vectors representing its content. Our objective is to develop a classification model capable of accurately categorizing publications into predefined classes based on their content and citation relationships. Some of the constraints with this problem statement are -

- Disconnected Graph Components: The citation network may contain disconnected components, hindering information flow across the graph and impacting classification performance.
- Over-smoothing: Over-smoothing can occur in Graph Neural Networks (GNNs) due to multiple layers, resulting in overly similar node representations and decreased classification accuracy. This phenomenon needs to be mitigated to ensure effective feature extraction.
- Model Complexity: Balancing model complexity with performance is crucial to avoid overfitting and achieve generalization on unseen data. Optimization of hyperparameters and model architecture is necessary to achieve optimal classification results.

By addressing these challenges and leveraging advanced techniques such as Graph Convolutional Networks (GCNs), the goal is to revolutionize the classification of scientific publications, enabling more accurate categorization and enhanced accessibility of scholarly literature.

III. Overview Of The Proposed System

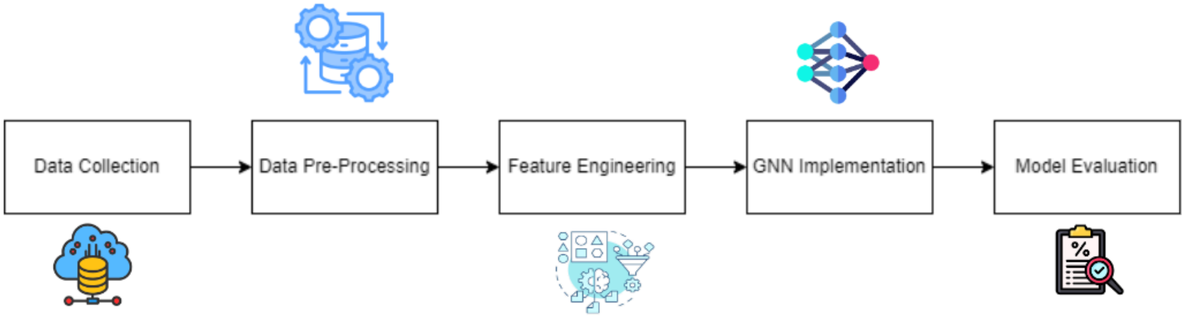


Figure 1. Data Mining System Pipeline

The proposed system leverages Graph Neural Networks (GNNs) for citation analysis, a crucial task in understanding the relationships between academic papers. Citation analysis aids in assessing the impact of scholarly works, identifying influential publications, and discovering emerging trends in research fields. The system utilizes a

GNN architecture to model the citation network, capturing both the structural information of the graph and the features associated with each node, i.e. paper. Fig 1 shows our data mining system pipeline.

A. Key Components

The key components of the system include:

- Graph Representation: Academic papers are represented as nodes in a graph, and citation relationships between papers are represented as edges. This graph structure allows for the exploration of citation patterns and the propagation of information through the network.
- Node Features: Each node (paper) in the citation graph is associated with features such as the paper's title, abstract, publication year, and author information. These features provide additional context for understanding the content and relevance of each paper.
- Graph Neural Network Architecture: The system employs a GNN architecture to learn representations of nodes in the citation graph. GNNs aggregate information from neighboring nodes to update node representations iteratively, enabling the model to capture complex patterns and relationships within the graph.
- Task Objective: The primary objective of the system is to predict citation counts for papers based on their features and the structure of the citation network. By learning from the citation patterns of existing papers, the model can infer the potential impact and relevance of new papers.

IV. Technical Details Of The Proposed System

The technical implementation of the proposed system involves several key steps:

A. Data Preprocessing

The raw dataset containing information about academic papers and their citation relationships is preprocessed to extract relevant features and construct the citation graph. This involves parsing paper metadata, cleaning the data, and encoding categorical features. The dataset is split into training and test sets to evaluate the model's performance on unseen data. Fig 2 indicates the total number of each class where class 3 has the highest number of papers, i.e. 818.

- Exploratory Data Analysis (EDA):

  We conduct EDA to gain insights into various features such as citation counts, publication years, and other relevant metadata to understand their distribution and their impact on classification. We also identify irrelevant or redundant features that do not contribute significantly to the classification task. It excludes 'paper_id' and 'subject' to prevent data leakage and ensure the model learns from relevant features. Fig 2 visualizes the citation graph. Each node in the graph represents a paper, and the color of the node corresponds to its subject.

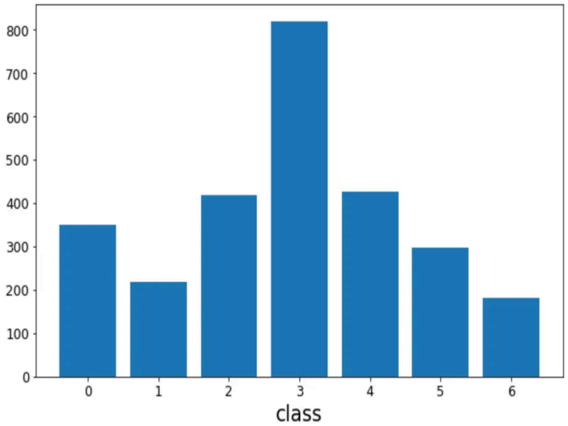


Figure 2. Bar chart for Distribution of Research papers

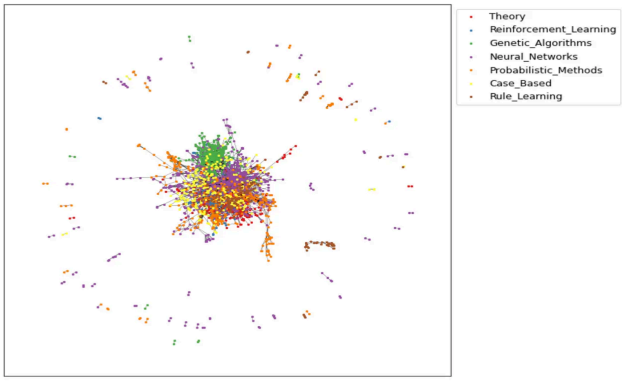


Figure 3. Exploratory Data Analysis (EDA)

- Correlation Analysis:

  Multicollinearity can distort the model's coefficients and make interpretations unreliable. Thus, we meticulously examine the correlation matrix to identify pairs of features exhibiting high correlation coefficients. By doing so, we can pinpoint potential redundancies in the dataset and select features that offer distinct information, thereby promoting diversity within the data. From Fig 4 we can see that there are a great many nodes that are connected to each other belonging to the same class.

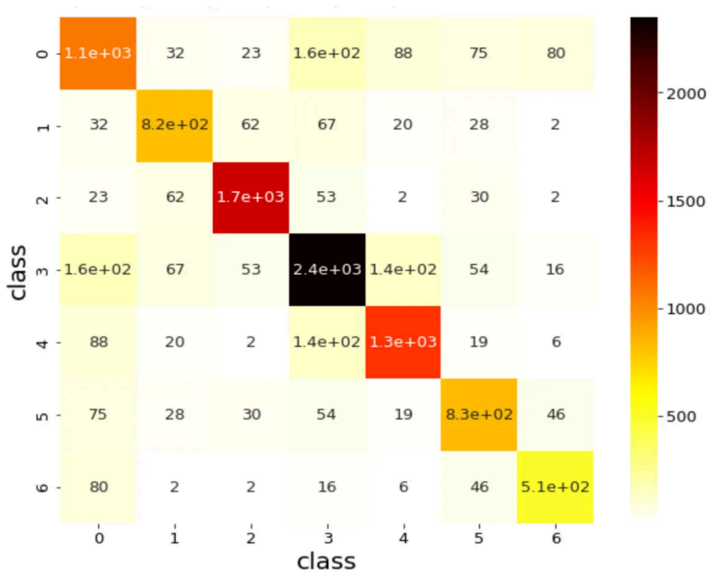


Figure 4. Heatmap for Correlation Analysis

B. Graph Construction

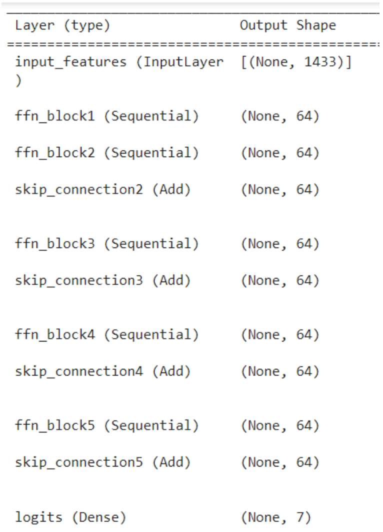
The citation graph is constructed using the processed data, where each paper is represented as a node, and citation relationships are represented as directed edges between nodes. The graph structure is essential for capturing the flow of influence between papers.

C. Feature Engineering

Features are extracted from each paper in the dataset, including textual features such as title and abstract embeddings, as well as metadata features such as publication year and author information. These features provide rich contextual information for training the GNN model. We also convert binary word vectors into a format suitable for input into a Graph Neural Network (GNN). For graph feature extraction, we incorporate the degree of each node (number of citations) as a feature. Nodes with higher degrees may indicate more influential or well-connected publications.

D. Model Training

A Graph Neural Network model is trained on the constructed citation graph using the extracted features. The model iteratively updates node representations by aggregating information from neighboring nodes, leveraging techniques such as graph convolutional layers or attention mechanisms. The baseline model includes an input layer that accepts a feature vector of length equal to the number of features (excluding 'paper_id' and 'subject'). It comprises five feed-forward neural network (FFN) blocks with skip connections to facilitate gradient flow and mitigate the vanishing gradient problem. Each FFN block is followed by a skip connection that adds the block's input to its output, enhancing the learning capacity without increasing the complexity too much.



```
Layer (type)                Output Shape
==============================================
input_features (InputLayer  [(None, 1433)]
)

ffn_block1 (Sequential)     (None, 64)

ffn_block2 (Sequential)     (None, 64)

skip_connection2 (Add)      (None, 64)


ffn_block3 (Sequential)     (None, 64)

skip_connection3 (Add)      (None, 64)


ffn_block4 (Sequential)     (None, 64)

skip_connection4 (Add)      (None, 64)


ffn_block5 (Sequential)     (None, 64)

skip_connection5 (Add)      (None, 64)


logits (Dense)              (None, 7)
```

Figure 5. Baseline (GNN) Model Architecture

E. Evaluation

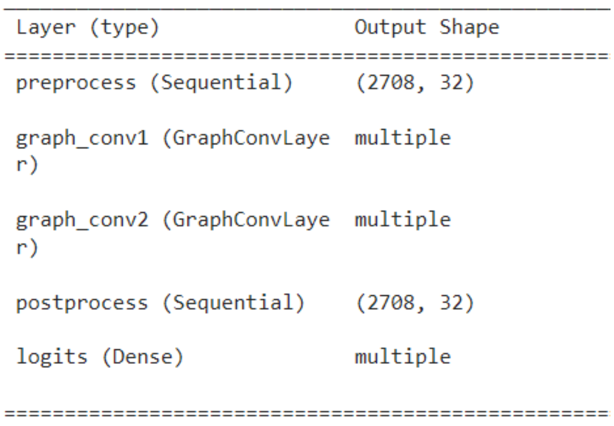
The trained model is evaluated on a held-out validation set or through cross-validation to assess its performance in predicting citation counts for unseen papers. Evaluation metrics such as mean absolute error or root mean squared error are used to quantify the model's accuracy.

F. Hyperparameter Tuning

Hyperparameters of the GNN model, such as learning rate, number of layers, and hidden units, are tuned using techniques like grid search or random search to optimize model performance. The model uses a dropout rate of 20% to regularize the network and prevent overfitting. Dense layers are used with a number of hidden units that can be adjusted based on the complexity of the dataset.

G. Inference

Once trained, the GNN is deployed to make predictions on new papers, estimating their potential citation counts based on their features and the learned citation patterns. Overall, the proposed system offers a powerful framework for citation analysis using Graph Neural Networks, enabling researchers to gain insights into the impact and significance of academic publications in various domains.



```
_________________________________________________
 Layer (type)                Output Shape
=================================================
 preprocess (Sequential)     (2708, 32)

 graph_conv1 (GraphConvLaye  multiple
 r)

 graph_conv2 (GraphConvLaye  multiple
 r)

 postprocess (Sequential)    (2708, 32)

 logits (Dense)              multiple

=================================================
```

Figure 6 GCN Model Architecture

The GCN model architecture, named "gcn_model," as depicted in Fig 6, consists of:

- Preprocessing layer (Sequential): Generates initial node representations.
- Graph Convolutional Layers (graph_conv1, graph_conv2): Apply graph convolutional operations.
- Post-processing layer (Sequential): Generates final node embeddings.
- Logits layer (Dense): Produces class probabilities.

V. Experiments

The relentless growth of scientific literature poses a formidable challenge for effective categorization and accessibility. Traditional classification methods often fall short in capturing the intricate relationships within citation networks, impeding streamlined literature management and hindering the extraction of deep bibliometric insights. To address these limitations, our project endeavors to revolutionize the classification process by harnessing the power of Graph Neural Networks (GNNs) to enhance accuracy and efficiency in scientific publication categorization.

A. Data Description

The Cora dataset serves as the cornerstone of our experimentation, encompassing a collection of 2,708 scholarly articles categorized into seven distinct subject classes. Each article is represented by a binary word vector, indicating the presence or absence of specific terms from a dictionary containing 1,433 unique terms. Additionally, the dataset includes a citation network comprising 5,429 links, delineating the citation relationships between the papers. This structural information encapsulates the inherent relationships and dependencies among the scientific documents, forming the basis for graph-based modeling.

B. *Evaluation Metrics*

To comprehensively assess the performance of our models, we employ a suite of evaluation metrics:

- Accuracy: Quantifies the proportion of correctly classified nodes within the test dataset, providing insights into the overall classification performance.
- Loss: Monitored throughout the training process, the cross-entropy loss offers valuable insights into the convergence and performance of the models, with lower loss values indicative of better model fitting and classification performance.
- Learning Curves: Visual aids such as learning curves facilitate the tracking of training and validation loss and accuracy across epochs. These curves offer insights into the convergence behavior and generalization capabilities of the models, enabling thorough analysis and interpretation of their performance dynamics over time.
- ROC Curve: The Receiver Operating Characteristic (ROC) curve illustrates the performance of the classifier across various threshold settings. It plots the true positive rate against the false positive rate and provides insights into the trade-off between sensitivity and specificity.
- Classification Report: Provides a summary of various classification metrics such as precision, recall, F1-score, and support for each class in the classification task. It offers a comprehensive view of the model's performance on individual classes.
- Confusion Matrix: A table that summarizes the performance of a classification algorithm. It provides insights into the number of true positives, false positives, true negatives, and false negatives, allowing for a detailed analysis of the model's performance across different classes.

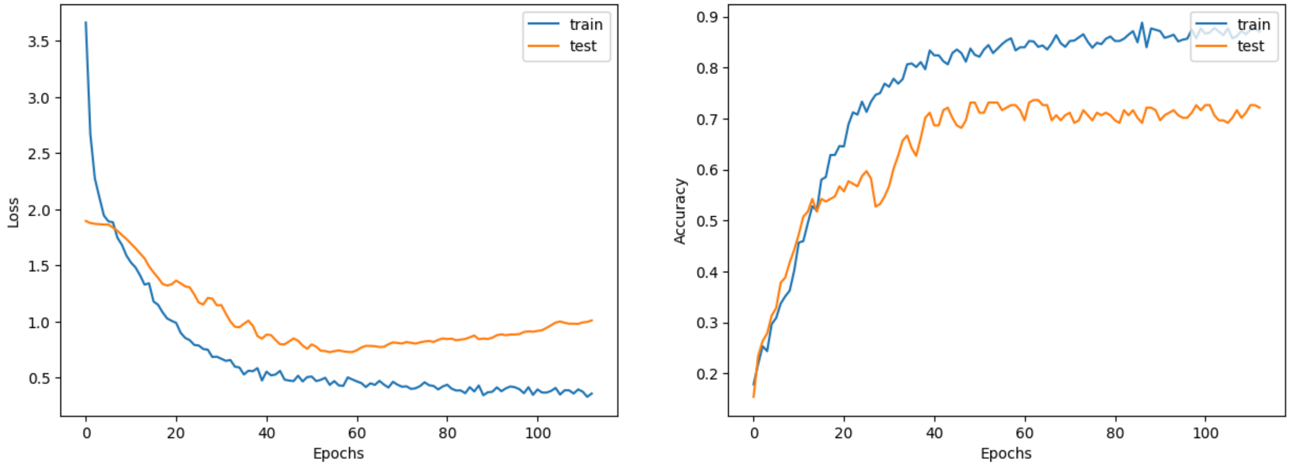


Figure 7. Learning curves depicting the training and validation loss and accuracy trends over epochs for the GNN model

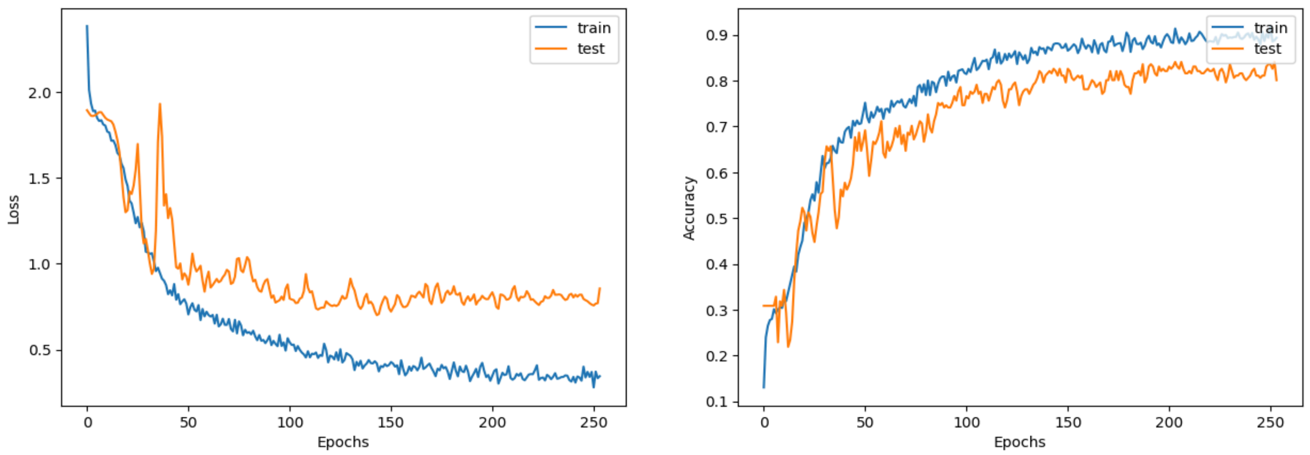


Figure 8. Learning curves depicting the training and validation loss and accuracy trends over epochs for the GCN model

B. *Evaluation Metrics*

In our experimentation, we leverage two fundamental graph-based models for comparison:

- Baseline Model: Graph Neural Network (GNN): The GNN serves as the foundation of our experimentation, offering a robust framework for graph-based learning tasks. This model architecture is meticulously crafted to exploit the structural relationships encoded within the citation network, enabling effective node classification based on the underlying graph topology.
- Enhanced Model: Graph Convolutional Network (GCN) on GNN: Building upon the robust foundation laid by the baseline GNN, we introduce an advanced model architecture by incorporating a Graph Convolutional Network (GCN) atop it. The GCN enhances the capabilities of the GNN by leveraging graph convolutional layers to extract and propagate structural information throughout the citation network, facilitating more refined node embeddings and improved classification performance.

C. *Results*

The performance of our models is summarized below:

- Baseline Model: Graph Neural Network (GNN):
    - Test accuracy: 73.16%
    - Loss trend: The training loss decreases steadily over epochs, indicating effective learning. Fluctuations in loss are observed, but the overall trend is downward, a positive sign.
    - Accuracy trend: Training accuracy gradually increases over epochs, suggesting learning. Fluctuations are present, but the overall trend is upward, indicating improvement

- Enhanced Model:
  - Test accuracy: 83.53%
  - Loss reduction: The loss decreases over epochs for both training and validation sets, indicating effective learning and parameter optimization.
  - Accuracy improvement: The accuracy increases over epochs for both training and validation sets, indicating improved classification performance.

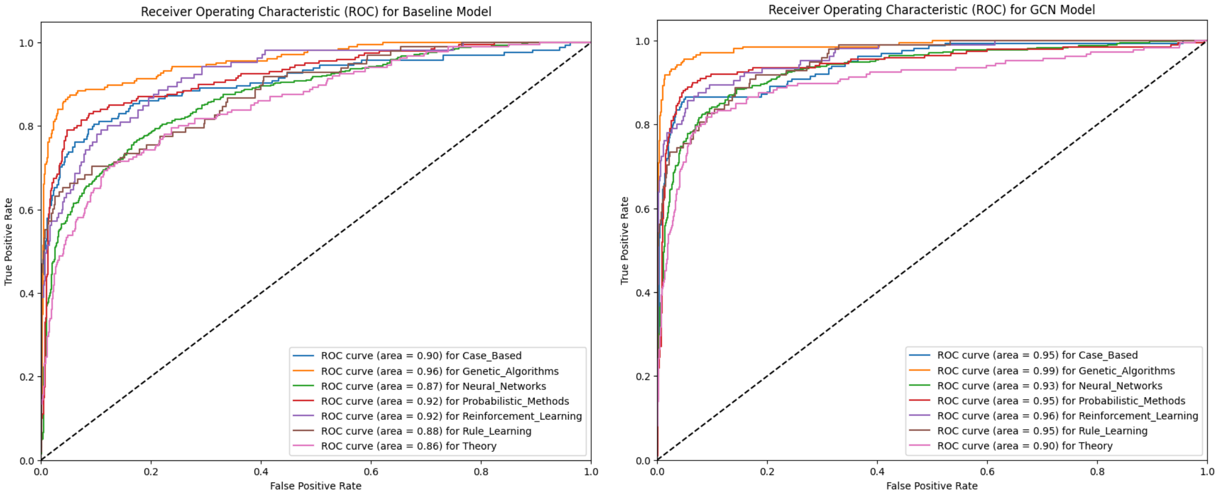


Figure 9. ROC Curves for Baseline (GNN) and GCN Models

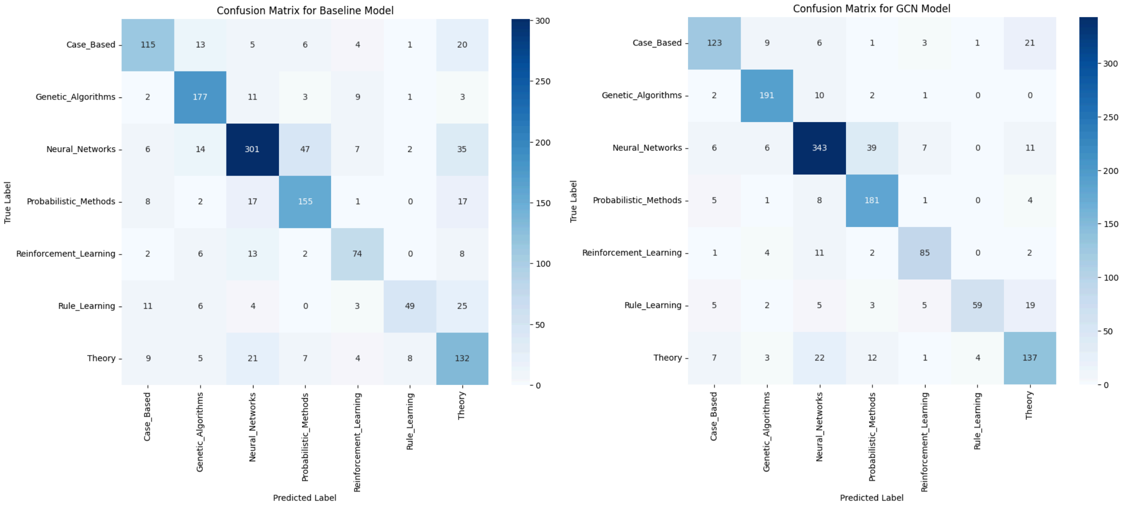


Figure 10. Confusion Matrices for Baseline (GNN) and GCN Models

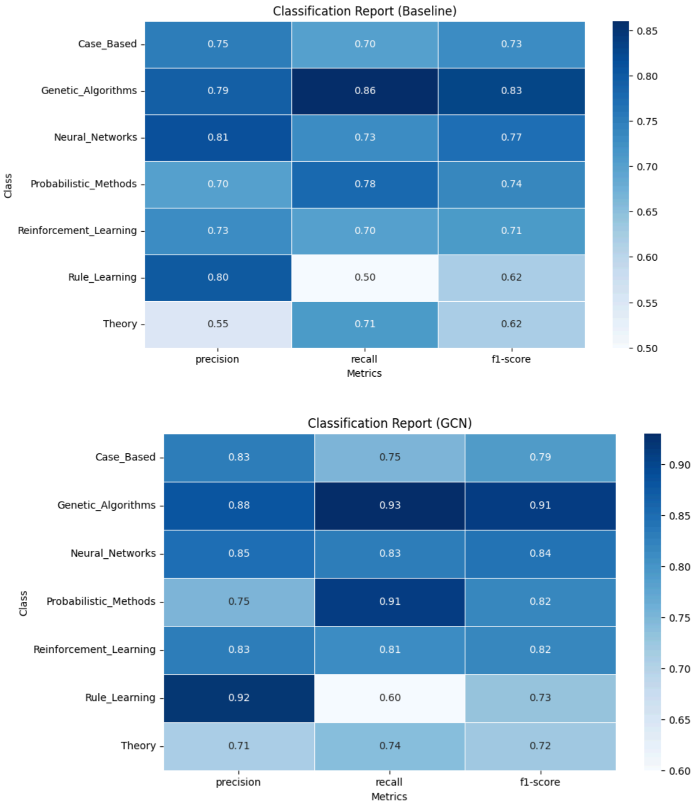


Figure 11. Classification Reports for Baseline (GNN) and GCN Models

These results demonstrate the effectiveness of leveraging graph-based techniques, particularly the GCN model, in enhancing the accuracy and efficiency of scientific publication categorization. Through meticulous experimentation and analysis, we have shown the potential of advanced graph-based models to tackle the challenges posed by the growing volume and complexity of scientific literature. In addition to traditional evaluation metrics like accuracy and loss, we have utilized comprehensive evaluation measures including ROC curves, classification reports, and confusion matrices. These metrics provide a deeper understanding of the model's performance across different classes, highlighting its strengths and areas for improvement. Overall, our findings underscore the significance of graph-based approaches in advancing the field of literature categorization and unlocking valuable insights from scholarly datasets.

VI. Related Work

In this section, we delve into the existing literature related to scientific publication classification, highlighting the transition from traditional algorithms to graph-based techniques and discussing key research contributions in the field.

A. *Traditional Approaches*

Traditionally, algorithms such as Support Vector Machines (SVM) and Random Forests have been prevalent in classifying scientific publications. These methods typically analyze text data in isolation, treating each document as an independent entity. While effective to some extent, these approaches may overlook the relational structure present in citation networks, limiting their ability to capture complex relationships between documents.

B. *Evolution towards Graph-Based Techniques*

Recent advancements in machine learning have shifted the focus towards graph-based techniques for scientific publication classification. Graph Convolutional Networks (GCNs) have emerged as a promising approach, leveraging the relational structures between texts to improve classification performance significantly. This transition represents a paradigm shift from treating documents in isolation to considering their interconnectedness within citation networks.

C. *Our Contribution*

Building upon these breakthroughs, our effort focuses on refining graph-based algorithms for the categorization of scientific papers in the Cora dataset. By combining structural and content-based elements, we aim to increase classification accuracy and provide a more comprehensive understanding of document relationships within citation networks. Our approach extends the capabilities of existing graph-based models, offering a promising avenue for advancing scientific publication classification research.

VII Conclusion

In this research project, we set out to address the challenges associated with classifying scientific publications effectively using traditional methods by leveraging Graph Neural Networks (GNNs), particularly Graph Convolutional Networks (GCNs). Our aim was to enhance the accuracy and efficiency of scientific publication classification, ultimately revolutionizing literature management, author visibility, and bibliometric insights extraction. Through our exploration and experimentation, we have made significant advancements. Transitioning from traditional methods to GCNs allowed us to better handle the complex relationships within citation networks, leading to notable improvements in classification accuracy. Our baseline classifier achieved a test accuracy of 73.16%, while our GCN classifier surpassed expectations with a test accuracy of 83.52%. Moreover, our investigation into the impact of different hyperparameters and model architectures provided valuable insights into optimizing GCN models for scientific publication classification. By varying parameters such as hidden units, learning rate, and dropout rate, we were able to fine-tune our models and achieve improved performance. The trends observed during model training and evaluation further validate the effectiveness of GCNs in this domain. Both the baseline and GCN models exhibited favorable trends in terms of loss reduction and accuracy improvement over epochs, indicating successful learning and optimization processes.

Overall, our findings highlight the potential of Graph Neural Networks, particularly GCNs, in revolutionizing the classification of scientific publications. By effectively capturing the structural information inherent in citation networks, GCNs offer a promising approach to streamline literature management, enhance author visibility, and extract deep bibliometric insights. Moving forward, further research could explore additional enhancements to GCN models, such as incorporating attention mechanisms or exploring more sophisticated graph structures. Additionally, investigating the applicability of GCNs to larger and more diverse datasets could provide valuable insights into their scalability and generalizability. In conclusion, our research underscores the transformative potential of Graph Neural Networks, particularly GCNs, in advancing the field of scientific publication classification. By harnessing the power of graph-based representations, we can unlock new possibilities for knowledge discovery and scholarly communication in an increasingly complex and interconnected academic landscape.

References

[1] Zhou, J., Cui, G., Hu, Z., Zhang, Z., Yang, C., & Liu, Z., “Graph Neural Networks: A Review of Methods and Applications,” AI Open, Volume 1, 2020, pp. 57-81.

[2] Bruna, Joan; Zaremba, Wojciech; Szlam, Arthur et al., “Spectral networks and locally connected networks on graphs”, International Conference on Learning Representations (ICLR2014), CBLS, April 2014. 2014.

[3] Defferrard, M., Bresson, X., & Vandergheynst, P., “Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering”, Advances in Neural Information Processing Systems (2016). arXiv:1606.09375

[4] Kipf, T. N., & Welling, M., “Semi-Supervised Classification with Graph Convolutional Networks” International Conference on Learning Representations (ICLR). arXiv:1609.02907

[5] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y., “Graph Attention Networks”, International Conference on Learning Representations (ICLR), 2018.

[6] The Cora Dataset. (n.d.). Retrieved from https://graphsandnetworks.com/the-cora-dataset/ (Alternative Link: https://linqs-data.soe.ucsc.edu/public/lbc/cora.tgz)

Google Drive Link for Code and Presentations:
https://drive.google.com/drive/folders/1D6jgSmpbm5jpbmwiYqlEmmFBT8aAIsLY?usp=sharing