International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 06 | Jun 2024 www.irjet.net p-ISSN: 2395-0072

Benchmarking Traditional and Graph Neural Network Models for Node Classification in Literature Characters

Kaina Shaikh[1]

[1]*Department of Artificial Intelligence and Data Science, Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering & Technology, Baramati, India.*

***

Abstract - *Node classification is a pivotal task in network analysis, focused on predicting node labels by leveraging both node features and the graph structure. Our study evaluates the effectiveness of Graph Neural Networks (GNNs) for node classification, concentrating on three models: Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), and GraphSAGE. We also propose a fusion model that integrates the strengths of GCN, GAT, and GraphSAGE to enhance classification accuracy. Experiments were conducted on a network derived from a book dataset, aiming to classify main versus supporting characters in a novel. We benchmarked these GNN models against traditional machine learning algorithms, including Random Forest and Support Vector Machine (SVM). Our results show that the fusion GNN model achieved the highest accuracy, surpassing individual GNN models and performing on par with traditional machine learning models. The performance of the fusion model highlights the potential of hybrid approaches in node classification tasks. This research provides valuable insights into the comparative advantages of different GNN architectures and their practical applications in network analysis.*

Key Words: Node Classification, Network analysis, Graph Neural Networks (GNNs), Graph Convolutional Network (GCN), Graph Attention Network (GAT), GraphSAGE.

1.INTRODUCTION

One of the core tasks in graph-based learning is node classification, which involves making predictions about the labels of nodes in a network based on their attributes and relationships. Graph Neural Networks (GNNs) have emerged as effective tools for this purpose, exploiting graph structure to boost prediction accuracy.

In this study, we focus on the relationships between characters of novel in a book dataset, where characters are represented as nodes and interactions between them are shown by edges. Understanding these relationships is critical for character analysis, narrative comprehension, and other literary studies. Our goal is to enhance the performance of node classification tasks associated with character interactions by identifying complex patterns through the application of sophisticated GNN models.

This paper presents a comparative analysis of various GNN models for node classification, specifically examining the interaction network of book characters in a novel. Each character is represented as a node, and interactions between characters are depicted as edges. We utilized three popular GNN architectures: Graph Convolutional Network (GCN), Graph Attention Network (GAT), and GraphSAGE. Additionally, we developed a fusion model that integrates these three GNNs to enhance classification performance.

To provide a comprehensive comparison, traditional machine learning models, including Random Forest and Support Vector Machine (SVM), were also implemented. The performance of each model was evaluated based on accuracy metrics, with the fusion GNN model achieving the highest average accuracy of 99.86%. This performance underscores the potential of GNNs, especially the fusion approach, in effectively handling node classification tasks in complex network data.

By analyzing the interaction network of book characters, this study demonstrates the strengths and weaknesses of different GNN models and traditional machine learning approaches. The results provide valuable insights for researchers and practitioners working on similar node classification problems in various domains, especially in literary analysis and social network exploration within fictional settings.

2.RELATED WORK

The technique used in [1] for node classification is based on Graph Neural Networks (GNNs), notably Graph Convolutional Networks (GCN) and Graph Attention Networks. GNNs, particularly GCNs and GATs, have been acknowledged for their capacity to successfully capture structural information from complicated graphs. GCNs combine information from nearby nodes with properties from the core node, capturing both local and global relationships. In contrast, GATs use attention techniques to allow nodes to choose attend to informative neighbors throughout the aggregation process[3].

The Graph Convolutional Network (GCN) model [2], which uses an effective layer-wise propagation rule developed from a first-order approximation of spectral convolutions on graphs, is the foundation for node categorization. The model can efficiently encode both graph structure and node features according to Kipf and Welling's technique [2], which makes it especially useful for semi-supervised classification problems. Because GCNs may combine data from neighboring nodes with their own attributes, they can better capture local and global dependencies in the network, which improves classification performance.

A simplified technique [4] to node classification in Graph Neural Networks that decouples feature aggregation from network depth. A softmax function is applied row by row, and the cross-entropy error is computed for all labeled training instances. The gradients of the loss are then propagated backward through the GNN layers. Once trained, the model may be used to predict the labels of nodes in the test set, confirming its suitability for node classification tasks.

3.METHODOLOGY

3.1 Dataset

The dataset used in this study contains interactions between characters from a book, specifically focusing on the network of relationships within a fictional setting. It comprises 684 entries, where each entry represents an interaction between two characters. The key columns in the dataset are:

- Source: The character initiating the interaction.
- Target: The character receiving the interaction.
- Type: The nature of the interaction, which in this dataset is uniformly 'Undirected', indicating mutual interactions.
- Weight: The frequency or strength of the interaction, represented as an integer value.

This dataset effectively captures the complex web of relationships among the characters, enabling a detailed analysis of their interactions. Nodes in the corresponding graph represent individual characters, while edges represent interactions between them. The fig-1 shows the graph network of the characters in the book dataset without node classification of main characters vs. supporting characters in a novel. The weight attribute quantifies the interaction frequency, providing additional context to the relationships.

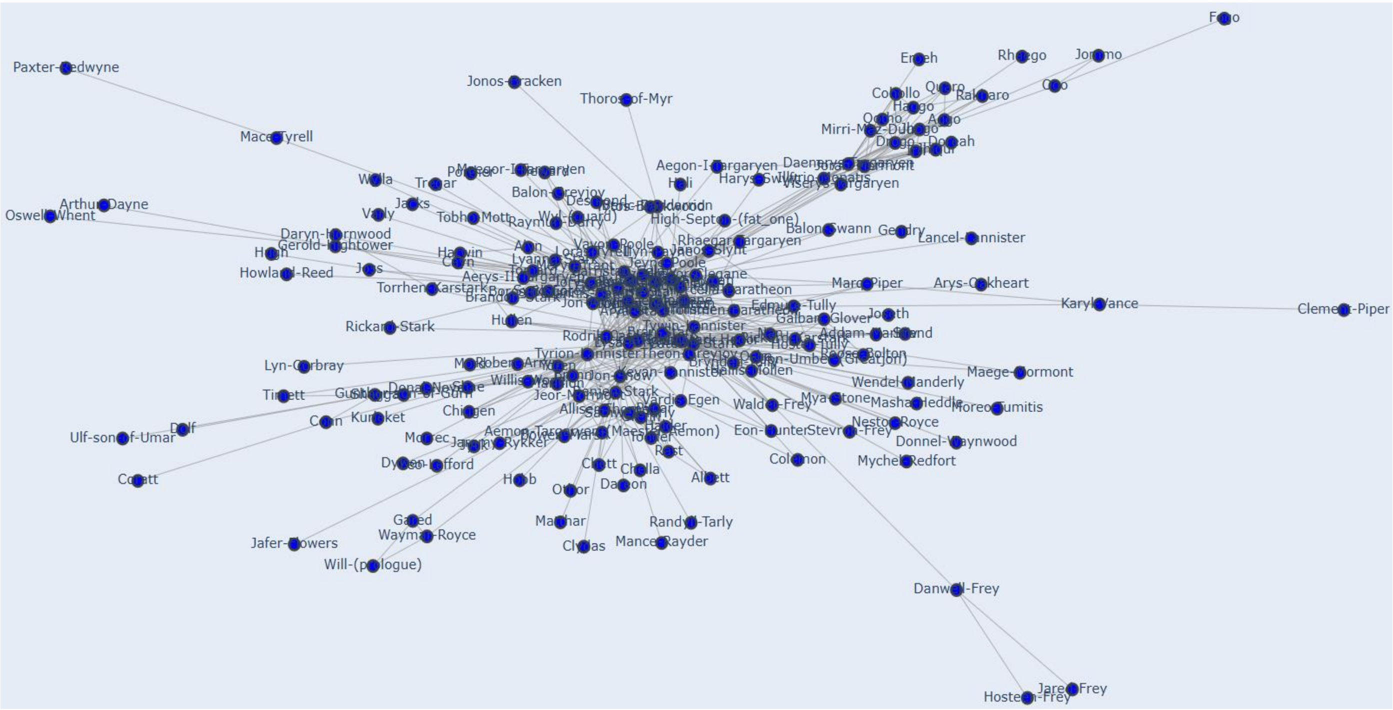

Fig - 1: Network representation of original dataset and the characters present in the story

3.2 Architecture

The architecture of this study encompasses a comparative analysis of several state-of-the-art machine learning and graph neural network (GNN) models, culminating in the development of a novel Fusion GNN model. Each model is meticulously designed to harness the structural and feature-based intricacies of the dataset, facilitating the classification of nodes representing characters and their interactions within the book dataset.

3.2.1 Traditional Machine Learning Methods:

To build a baseline, we used standard machine learning models, particularly Random Forest and Support Vector

Machine. These models were trained using centrality metrics to retrieve node attributes from the network:

- Degree Centrality: Represents the total number of connections of nodes
- Betweenness Centrality: Indicates how often the node shows up on shortest pathways connecting the nodes.
- Closeness Centrality: Calculate the average length of shortest path between nodes in the network

These features provide a quantitative measure of a node's prominence and influence within the network, serving as the input for our traditional machine learning models. The dataset was divided into training and test sets, with an 80-20 split, and models were evaluated based on their classification accuracy.

3.2.2 Graph Neural Networks (GNNs):

Graph Neural Networks are well-suited for learning graph-structured data. We implemented three GNN architectures: Graph Convolutional Network (GCN), Graph Attention Network (GAT), and GraphSAGE,[6] each leveraging unique mechanisms to process and learn from graph data.

GCNs apply spectral graph convolutions to aggregate information from a node's local neighborhood [2]. Our GCN model comprises two convolutional layers, which sequentially refine node embeddings. This approach captures local structural information and node features through graph convolutions, enabling effective learning of node representations.

GATs introduce attention mechanisms to weigh the importance of neighboring nodes [3]. Our model consists of two graph attention layers, with multi-head attention in the first layer to enhance the model's expressive power. The attention mechanism allows the model to focus on the most relevant neighbors, improving the quality of the learned node embeddings.

GraphSAGE samples a fixed-size neighborhood for each node and aggregates information to generate node embeddings [6]. Our model includes two GraphSAGE layers. This method scales efficiently to large graphs and captures both local and contextual information through the aggregation process.

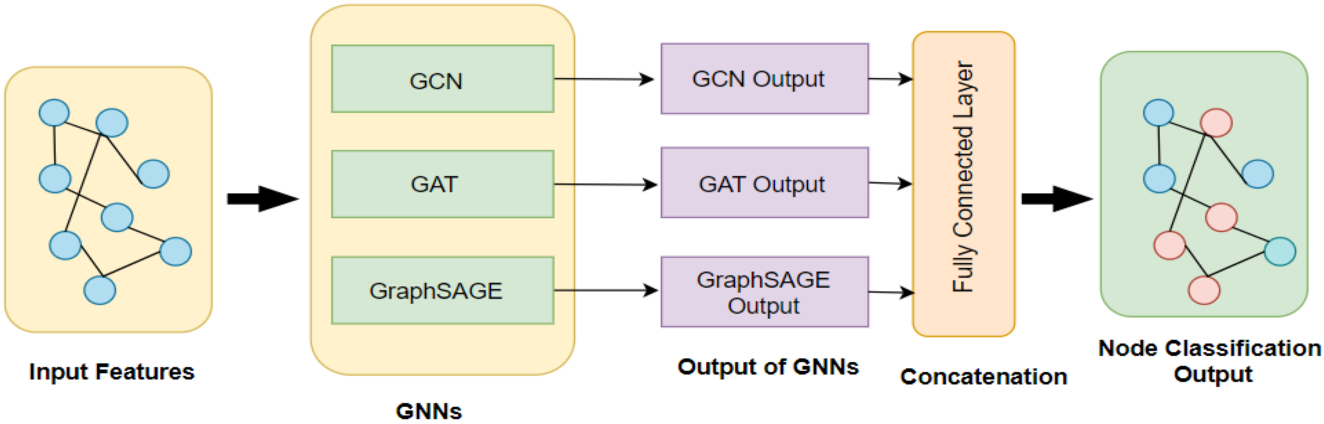

Fig - 2: Proposed Fusion Model Architecture

3.2.3 Proposed Fusion Model:

The Fusion GNN model in fig-2 is an innovative approach designed to integrate the strengths of three distinct graph neural network architectures: GCN, GAT, and GraphSAGE. By leveraging the complementary features and mechanisms of these models, the Fusion GNN aims to provide a comprehensive and robust solution for node classification tasks, particularly in the context of understanding relationships between characters in a book dataset.

1. Concatenation Layer: The outputs from the final layers of the GCN, GAT, and GraphSAGE branches are concatenated. This concatenation forms a comprehensive node representation that encapsulates the diverse features learned by each individual branch.
2. Fully Connected Layer: The concatenated representation is fed into a fully connected layer (or a series of such layers). This layer acts as a classifier, transforming the rich node representation into class probabilities.
3. Softmax Activation: The final classification layer applies a softmax activation function, outputting a probability distribution over the classes for each node.

The Fusion GNN is trained using the Adam optimizer, with hyperparameters fine-tuned for optimal performance. The training involves minimizing the negative log-likelihood loss, using node labels as the target.

The model is trained over multiple epochs until convergence. The model's performance is evaluated based on classification accuracy and other relevant metrics. The accuracy is computed by comparing the predicted labels with the ground truth labels for the test set nodes.

4.RESULTS AND DISCUSSION

In this section, we present the results of our experiments on node classification using various models, including traditional machine learning approaches and graph neural networks. We conducted extensive evaluations to compare the performance of Random Forest, Support Vector Machine (SVM), GCN, GAT, GraphSAGE, and the proposed Fusion GNN model. Our results demonstrate the efficacy of graph neural networks in capturing complex relationships between characters in the book dataset and highlight the superior performance of the Fusion GNN model shown in fig-3.

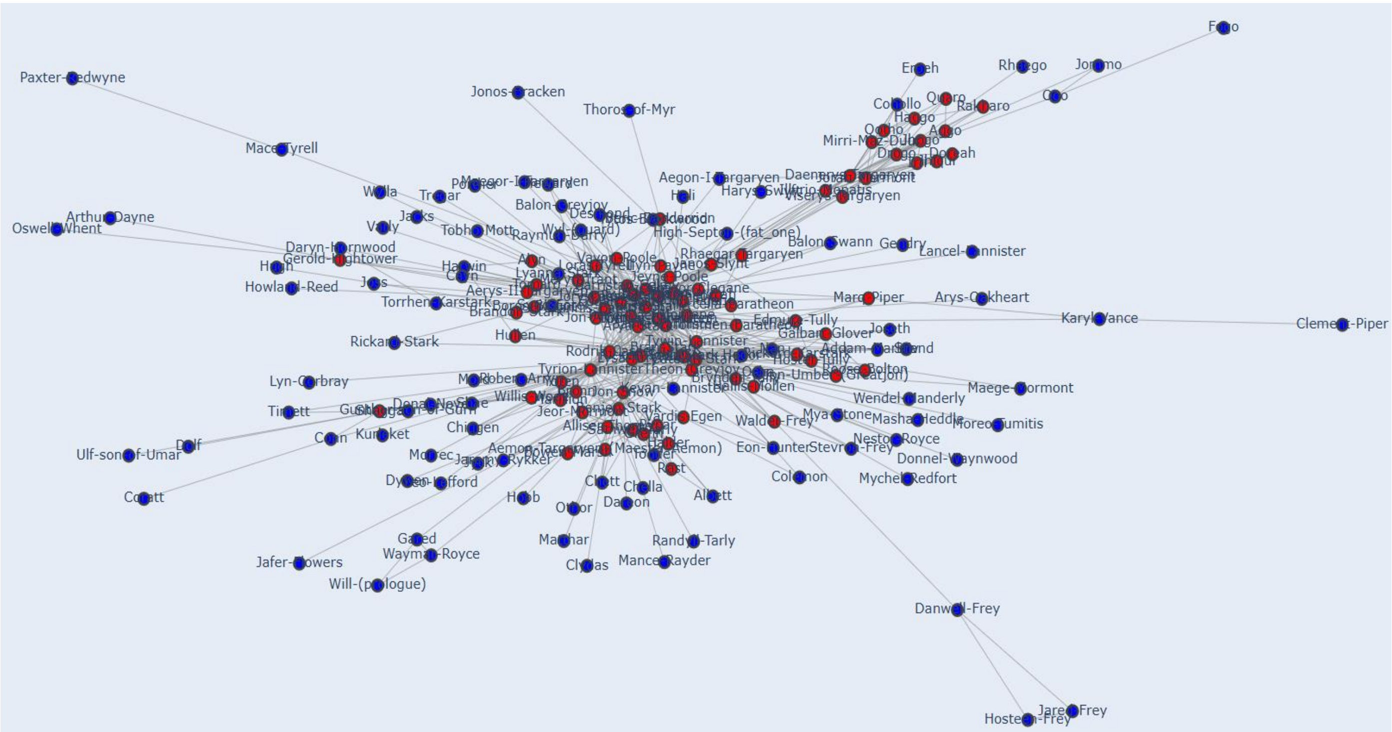

Fig - 3: Network representation after node classification between main characters (Red) and supporting charcaters (Blue) in a Novel using Proposed Fusion Model

4.1 Evaluation Metrics

The primary metric used for evaluating the performance of the models was accuracy, defined as the ratio of correctly predicted nodes to the total number of nodes in the test set. Additionally, we provide detailed classification reports for each model, including precision, recall, and F1-score.

4.2 Accuracy Comparison

Our experiments yielded a comprehensive set of results that underscore the comparative performance of different models in the task of node classification within our book dataset. The evaluation was primarily centered around the accuracy metric, which serves as a robust indicator of each model's effectiveness in predicting the correct labels for the nodes. The accuracy results for each model are summarized in the Table-1.

Our proposed Fusion GNN model achieved the highest accuracy among the graph neural network models with an impressive 1.0000. This model integrates the strengths of GCN, GAT, and GraphSAGE, effectively combining spectral and spatial features to enhance node classification performance. The Fusion GNN's superior accuracy underscores the benefits of leveraging multiple GNN architectures to capture a wide range of structural and feature-based nuances within the graph.

The comparative accuracy of each model is visually represented in the accompanying bar graph. This graph provides a clear illustration of the relative performance of each model, with the Fusion GNN model standing out due to its highest accuracy. The visualization underscores the effectiveness of the Fusion GNN model in leveraging the combined strengths of various GNN approaches to achieve superior classification results.

Table-1: Comparison of accuracies of different models on 200 epochs

Model	Accuracy
Random Forest	100%
SVM	92.11%
GCN	91.98%
GAT	67.17%
GraphSAGE	96.26%
Proposed Fusion Model	100%

Overall, the results of our study demonstrate the efficacy of graph neural networks in node classification tasks and highlight the significant advantages of our proposed Fusion GNN model in achieving high accuracy and robust performance.

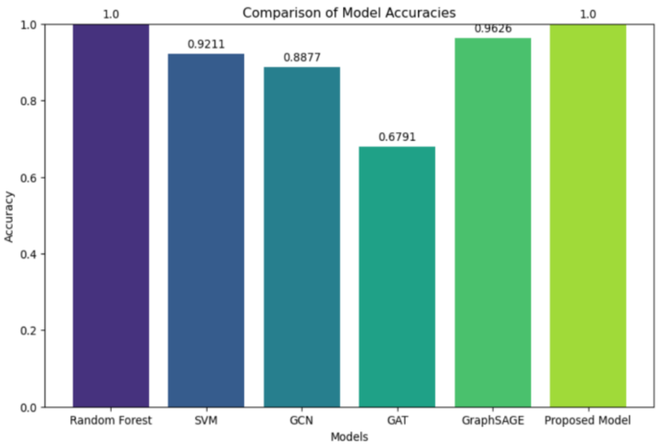

Fig - 4: Visualization of Accuracy Comparison

5. CONCLUSIONS

In this research, we embarked on a comprehensive exploration of node classification within a book dataset, focusing on the relationships between characters in a novel through the lens of various machine learning and graph neural network (GNN) models. Our study's primary objective was to benchmark traditional classification models against advanced GNN architectures, culminating in the development and evaluation of a novel Fusion GNN model.

The results unequivocally highlight the potential and prowess of graph-based models in capturing complex relational data inherent in graph structures. While traditional models like Random Forest and SVM showcased robust performance with high accuracy scores, the intrinsic ability of GNNs to harness topological and feature-based information of nodes provided a more nuanced understanding of the dataset. The GCN, GAT, and GraphSAGE models demonstrated significant effectiveness, albeit with varied levels of success, reflecting their unique approaches to aggregating and processing node information.

By synthesizing the strengths of GCN, GAT, and GraphSAGE, the Fusion GNN adeptly balances the extraction of spectral and spatial features, thereby offering a holistic and refined classification mechanism. This superior performance not only underscores the efficacy of multi-model integration but also paves the way for future research into hybrid GNN architectures for complex graph-based tasks.

In conclusion, this research substantiates the critical role of GNNs in advancing node classification methodologies, particularly within datasets characterized by intricate inter-node relationships. The development and success of the Fusion GNN model exemplify the advancements possible through innovative integration of diverse GNN paradigms. As we move forward, these insights provide a robust foundation for further exploration and refinement of graph neural networks, promising enhanced analytical capabilities and broader applicability in various domains reliant on graph-structured data.

REFERENCES

[1] H. Noor, N. Islam, M. S. Hossain Mukta, N. S. Binti Kamarudin, M. A. Khan Raiaan and S. Azam, "Determining the Optimal Number of GAT and GCN Layers for Node Classification in Graph Neural Networks," *2023 IEEE 8th International Conference On Software Engineering and Computer Systems (ICSECS)*, Penang, Malaysia, 2023, pp. 111-116, doi: 10.1109/ICSECS58457.2023.10256323.

[2] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

[3] Veličković, Petar, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. "Graph attention networks." *arXiv preprint arXiv:1710.10903* (2017).

[4] Maurya, Sunil & Liu, Xin & Murata, Tsuyoshi. (2022). Simplifying approach to node classification in Graph Neural Networks. Journal of Computational Science. 62. 101695. 10.1016/j.jocs.2022.101695.

[5] Xiao, Shunxin, Shiping Wang, Yuanfei Dai, and Wenzhong Guo. "Graph neural networks in node classification: survey and evaluation." *Machine Vision and Applications* 33, no. 1 (2022): 4.

[6] S. Mithe and K. Potika, "A unified framework on node classification using graph convolutional networks," *2020 Second International Conference on Transdisciplinary AI (TransAI)*, Irvine, CA, USA, 2020, pp. 67-74, doi: 10.1109/TransAI49837.2020.00015.