Creating Semantic Learner Groups in Distance Education Using the GraphSAGE approach

Ismail Chetoui[1,*], *Essaid* El Bachari[2], and *Yassine* Ait Lahcen[3]

[1,2,3]Department of Computer Science, Faculty of Sciences Semlalia, Cadi Ayyad University, Morocco

Abstract. In this article, we present a novel approach for creating semantic groups of learners in an educational platform using Graph Neural Networks (GNN) and GraphSAGE. The increasing availability of educational data necessitates advanced methodologies to enhance personalized learning experiences. Traditional techniques often fall short in capturing the complex relationships inherent in such data. To address this, we leverage GraphSAGE, an inductive framework, to generate meaningful embeddings that represent the diverse attributes and interactions of learners within the educational network. By sampling and aggregating information from the local neighborhoods of each learner, GraphSAGE effectively captures both individual and group-level learning patterns. These embeddings are then utilized to form semantic groups of learners, facilitating personalized recommendations, collaborative learning, and targeted interventions. Our approach demonstrates significant improvements in the ability to identify and cluster learners with similar learning behaviors and needs, thereby enhancing the overall educational experience. The results, evaluated on a comprehensive educational dataset, underscore the potential of GraphSAGE in transforming educational data into actionable insights for semantic group creation.

Keywords: Graph Neural Networks, GraphSAGE, Education, Semantic Groups.

1 Introduction

In the rapidly evolving field of education, the need for personalized learning experiences is becoming more critical. The expansion of online learning platforms and the massive influx of educational data present a unique opportunity to utilize advanced machine learning models to improve learning outcomes. Traditional approaches often fall short in capturing the complex, interconnected nature of educational data, resulting in less effective personalization. Graph Neural Networks (GNNs) have emerged as a promising solution for modeling relational data due to their ability to efficiently aggregate and distribute information across graph structures. Specifically, GraphSAGE (Graph Sample and AggregatE) [1] has gained recognition for its inductive learning capabilities and scalability, making it ideal for processing large, dynamic educational networks. This approach allows for the representation of intricate relationships among learners, courses, and interactions, paving the way for more meaningful and personalized educational experiences. This study aims to harness the potential of GraphSAGE to

*e-mail: ismail.chetoui@ced.uca.ma

create semantic groups of learners within an educational platform. By representing learners and their interactions as nodes and edges in a graph, GraphSAGE can generate meaningful embeddings that capture the nuanced relationships and attributes of each learner. These embeddings are used to form semantic groups, which are clusters of learners with similar learning patterns and needs. To create these groups, we start by constructing a graph where each node represents a learner and each edge represents an interaction or similarity between learners, such as shared courses or similar performance metrics. GraphSAGE then samples a fixed number of neighbors for each node and aggregates their features to generate a comprehensive embedding for each learner. These embeddings, which encapsulate both individual attributes and neighborhood information, are then fed into a clustering algorithm, such as k-means or hierarchical clustering, to form distinct groups of learners. These semantic groups enable more targeted and personalized educational experiences. For instance, learners within the same group can receive customized course recommendations, engage in collaborative learning activities, and benefit from interventions tailored to their specific needs. By leveraging the advanced graph-based methodology of GraphSAGE [1], we aim to demonstrate significant improvements in identifying and clustering learners, thereby enhancing the overall educational experience. Through the implementation of this GNN-based framework, we hope to contribute to more effective and engaging learning environments, showcasing the transformative potential of advanced machine learning techniques in education.

2 Related Work

2.1 Background

Graph Neural Networks (GNNs) have garnered significant attention in recent years due to their effectiveness in handling graph-structured data. Several studies have explored the application of GNNs in various domains, including social network analysis [2], recommendation systems [3] [4], and bioinformatics [5] . For instance, a study introduced Graph Convolutional Networks (GCNs) [6], which have been widely adopted for tasks such as node classification, link prediction [7], and graph clustering [8]. GCNs operate by performing convolutions directly on the graph, aggregating feature information from a node's local neighborhood to learn a more comprehensive representation. Building on the success of GCNs, Another model proposed called Graph Attention Networks (GATs) [9], which incorporate attention mechanisms to dynamically assign different weights to different neighbors. This allows the model to focus more on the most relevant parts of a node's neighborhood, improving performance in various tasks such as node classification and graph representation learning.

In the context of education, GNNs have been employed to address challenges such as predicting student performance [10], modeling learner interactions [11] , and recommending learning resources [3]. For example, [12] demonstrated the potential of GNNs in improving learner modeling and personalized learning by capturing complex relationships in educational data. Their study highlighted how GNNs could effectively model the intricate dependencies between students, courses, and educational materials, leading to more accurate predictions and recommendations. Despite these advancements, traditional GNN approaches often face scalability issues when applied to large and dynamic educational datasets. These limitations stem from the need to process the entire graph structure during training, which can be computationally intensive and impractical for large-scale graphs. To address these limitations, Hamilton [1] introduced GraphSAGE (Graph Sample and AggregatE), a scalable and inductive framework for generating node embeddings.

GraphSAGE operates by sampling and aggregating features from a fixed number of neighbors, rather than considering all neighbors in the graph. This approach allows GraphSAGE to generate embeddings for nodes that were not present during training, making it particularly suitable for dynamic environments where new nodes (such as new students or courses) are continuously added. GraphSAGE's framework includes various aggregation functions, such as mean, LSTM, and pooling, to capture different aspects of neighborhood information, enhancing its flexibility and effectiveness. Several recent works have applied GraphSAGE to diverse fields. For example, researchers [13] applied GraphSAGE to the problem of large-scale recommender systems, demonstrating its ability to handle vast datasets with complex user-item interactions. In healthcare, a recent study [14] showcases a novel application of GNNs in emergency departments. This work addresses the limitations of traditional triage methods by employing GNNs to process complex patient data, including vital signs, symptoms, and medical history. By representing patient information as a graph, the authors leverage the power of GNNs to capture intricate relationships between different health indicators and patient characteristics. This approach bears strong similarities to our use of GNNs in educational recommender systems, as both leverage the graph structure to model complex interactions. Their reported high accuracy in predicting patient triage categories underscores the potential of GNNs to enhance decision-making processes in high-stakes environments. This study not only reinforces the effectiveness of GNNs in handling complex, interconnected data but also suggests that similar graph-based approaches could be valuable in educational contexts, potentially improving the accuracy and efficiency of course recommendations and student support systems. In this study, we leverage GraphSAGE [1] to create semantic groups of learners within an educational platform. By representing learners and their interactions as nodes and edges in a graph, GraphSAGE can generate meaningful embeddings that capture the nuanced relationships and attributes of each learner. These embeddings are then used to form semantic groups, which are clusters of learners with similar learning patterns and needs. To create these groups, we start by constructing a graph where each node represents a learner and each edge represents an interaction or similarity between learners, such as shared courses or similar performance metrics. GraphSAGE then samples a fixed number of neighbors for each node and aggregates their features to generate a comprehensive embedding for each learner. These embeddings, which encapsulate both individual attributes and neighborhood information, are then fed into a clustering algorithm, such as k-means or hierarchical clustering, to form distinct groups of learners. These semantic groups enable more targeted and personalized educational experiences. For instance, learners within the same group can receive customized course recommendations, engage in collaborative learning activities, and benefit from interventions tailored to their specific needs. The ability to form such groups allows educational platforms to provide more nuanced support, addressing the diverse needs of students more effectively than traditional methods. Furthermore, the use of GraphSAGE ensures that the system remains scalable and adaptable to the continuously evolving landscape of online education. As new students enroll and new courses are added, GraphSAGE's inductive framework allows for the seamless incorporation of these new nodes into the existing graph, maintaining the relevance and accuracy of the recommendations and groupings. This study contributes to the growing body of research on GNNs in education, highlighting their potential to transform educational data into actionable insights for personalized learning [15]. By leveraging the advanced graph-based methodology of GraphSAGE, we aim to demonstrate significant improvements in identifying and clustering learners, thereby enhancing the overall educational experience. Our approach not only underscores the scalability and flexibility of GraphSAGE but also paves the way for more sophisticated and dynamic educational recommendation systems. In summary, this work builds on the strengths of GNNs and GraphSAGE to address the challenges of scalability and personalization in educational

data. By creating semantic groups of learners, we can better understand and cater to the unique learning journeys of individual students, ultimately contributing to more effective and engaging learning environments.

2.2 Gaps in the existing literature

Despite the substantial progress made in applying Graph Neural Networks (GNNs) to educational data, several gaps remain in the existing literature. One of the primary limitations is the scalability of traditional GNN methods. While Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs) have demonstrated significant improvements in modeling relational data, their applicability to large-scale and dynamic educational datasets is constrained by computational inefficiencies [16]. These methods often require the entire graph to be present during training, which becomes impractical as the size of educational networks grows. Another gap lies in the adaptability of existing models to new and unseen data. Many traditional GNN-based approaches are transductive, meaning they can only make predictions for nodes that were part of the training dataset [17]. This limitation is particularly problematic in educational settings where new students and courses are continuously introduced. The inability to generalize to unseen nodes restricts the usefulness of these models in real-world, dynamic educational environments. Furthermore, while there has been significant research on using GNNs for student performance prediction and course recommendation, there is a relative paucity of studies focused on leveraging these technologies for creating semantic groups of learners. Existing literature primarily addresses direct prediction tasks [18], often overlooking the potential benefits of clustering and group-based recommendations. Creating semantic groups can provide a more holistic understanding of student behavior and learning patterns, enabling more tailored and effective educational interventions. Additionally, the interpretability of GNN models remains a challenge. Educators and administrators often require clear and interpretable insights to make informed decisions. However, the complexity of GNN architectures can make it difficult to understand how specific predictions or recommendations are derived. This lack of transparency can hinder the adoption of GNN-based systems in educational contexts where interpretability is crucial. Finally, there is a need for more comprehensive evaluations of GNN models in educational settings. Many existing studies [19] focus on a narrow set of metrics or datasets, which may not fully capture the diverse and multifaceted nature of educational data. Evaluating GNN models across different types of educational platforms, learner demographics, and learning contexts is essential to validate their effectiveness and generalizability. In light of these gaps, this study aims to address several key challenges. By leveraging GraphSAGE, an inductive framework, we aim to overcome scalability and adaptability issues, allowing the model to handle large-scale and evolving educational datasets. Our focus on creating semantic groups of learners aims to fill the gap in the literature regarding group-based educational recommendations. Additionally, we emphasize the importance of interpretability and comprehensive evaluation to ensure that the proposed model can be effectively utilized in real-world educational settings.

3 Proposed Model

In our model, we employ a multi-step approach to leverage GraphSAGE for generating node embeddings and clustering learners based on their learning behaviors. The process starts with Data Preparation, where we identify nodes, define edges, extract relevant features, and label the data. In this step, nodes represent learners, and edges capture interactions or similarities between them, while features and labels are derived from learner profiles, historical performance, and other attributes.

Next, we construct a graph (G = V, E), where V represents the set of nodes (learners) and E represents the set of edges (interactions). Once the graph is defined, we implement the GraphSAGE algorithm to generate node embeddings. GraphSAGE performs neighbor sampling to select a subset of the node's neighborhood and applies aggregation functions (such as Mean, LSTM, and Pooling) to iteratively aggregate the features of neighboring nodes at each layer, capturing local graph structure and attribute information.

The embeddings produced by GraphSAGE encapsulate both relational and attribute information of each learner. These embeddings are then fed into the k-means clustering algorithm to group learners into semantic clusters based on similar learning patterns. The result is a set of clusters representing learners with similar educational needs, enabling personalized recommendations and interventions.

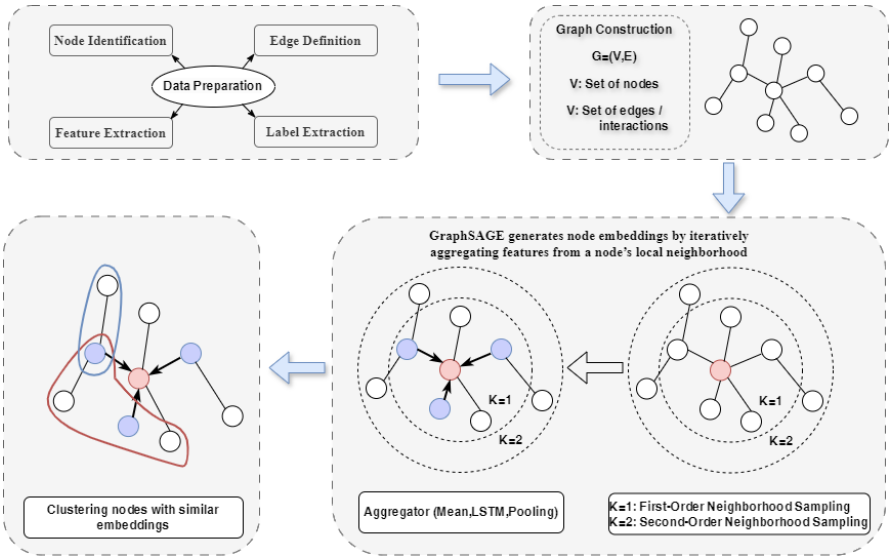

Figure 1. GraphSAGE-Based Learner Clustering Pipeline: From Data Preparation to Embedding Generation and Semantic Grouping.

3.1 Data Preparation

Node Identification: In the context of our educational platform, nodes represent learners. Each learner node is associated with various attributes, such as demographic information, past course enrollments, grades, and interaction data.

Edge Definition: Edges represent interactions or similarities between learners. These can include shared courses, similar performance metrics, peer discussions, or collaborative activities. Each edge can also have associated weights representing the strength of the connection or interaction frequency.

Feature Extraction: Node features are extracted and represented in a matrix form where each row corresponds to a learner and each column represents a specific feature or attribute. These features include both static attributes (e.g., age, major) and dynamic attributes (e.g.,

course completion status, engagement metrics).

Label Extraction: If available, labels can be used to supervise the training process. These could include learner proficiency levels, engagement scores, or other performance metrics.

3.2 Graph Construction

Using the prepared data, we construct a graph G=(V,E), where V represents the set of learner nodes and E represents the set of edges denoting interactions or similarities between learners. This graph structure captures the relational data necessary for effective learning representation.

3.3 GraphSAGE Implementationn

(a) **Neighbor Sampling**: For each node (learner), GraphSAGE samples a fixed number of neighbors to ensure scalability. This sampling reduces the computational complexity by limiting the number of nodes considered during the aggregation process.

(b) **Aggregation Functions**: GraphSAGE employs various aggregation functions to combine feature information from sampled neighbors. Common aggregation methods include:

- Mean Aggregator: Averages the features of neighboring nodes.
- LSTM Aggregator: Uses an LSTM (Long Short-Term Memory) network to aggregate neighbor features, capturing sequential dependencies.
- Pooling Aggregator: Applies a pooling operation (e.g., max pooling) to the neighbors' features.

For this study, we experiment with different aggregation methods to determine the most effective approach for our dataset. c. Node Embedding Generation: GraphSAGE generates node embeddings by iteratively aggregating features from a node's local neighborhood. The embedding for each node v at layer k is computed as:

$$h_v^{(k)} = \sigma\left(w^k \cdot \text{AGGREGATE}\left(\left\{h_u^{(k-1)}, \forall u \in N(v)\right\}\right)\right) \tag{1}$$

where $h_v^{(k)}$ is the embedding of node v at layer k, $N(v)$ is the set of neighbors of v, σ is a non-linear activation function, and w^k is a learnable weight matrix.

3.4 Clustering Learners

The embeddings generated by GraphSAGE encapsulate the relational and attribute information of each learner. These embeddings are then fed into a clustering algorithm to form semantic groups. We use the k-means clustering algorithm for this purpose, where learners with similar embeddings are grouped together, forming clusters that represent semantic groups of learners with similar learning patterns and needs.

3.5 Personalized Recommendations and Interventions

Once the semantic groups are formed, we can provide personalized educational experiences:

- **Course Recommendations**: Learners within the same group receive customized course recommendations based on the aggregated preferences and performance of the group.
- **Collaborative Learning**: Group-based activities and peer interactions are facilitated within the semantic groups, promoting collaborative learning.
- **Targeted Interventions**: Specific interventions and support are provided to learners based on the characteristics and needs of their respective groups.

Despite the promising capabilities of our proposed GraphSAGE-based framework for creating semantic groups of learners, there are several limitations that need to be addressed. First, scalability remains a challenge; processing large and complex educational datasets can be computationally intensive, requiring significant memory and processing power. Second, the effectiveness of our model heavily relies on the quality and completeness of the input data. In educational contexts, data can often be noisy, incomplete, or biased, which may lead to suboptimal groupings and recommendations. Third, the dynamic nature of educational data necessitates continuous updates to the graph and model, which can be resource-intensive and technically challenging. Fourth, the interpretability of GNN models is limited, making it difficult to provide clear explanations for the groupings and recommendations to educators and learners. Additionally, the generalizability of our model across different educational contexts and datasets can vary, necessitating customization and retraining for different environments. Ethical considerations, such as ensuring data privacy and mitigating biases, are also crucial and require ongoing attention. Lastly, the similarity metrics used to define edges in the graph may not capture all the nuances of learner behaviors, potentially affecting the accuracy of the semantic groups. Addressing these limitations will be key to further refining and enhancing the framework.

4 Experiments

To evaluate the effectiveness of our GraphSAGE-based framework for creating semantic groups of learners, we conducted a series of experiments using an educational dataset [20]. The dataset was divided into training, validation, and test sets to ensure a robust evaluation process. Here, we present key performance metrics and the configuration used for training our model.

Table 1. Training Configuration and Experimental Setup for GraphSAGE Model

Component	Details
Dataset	EdNet dataset including learner profiles, demographic information, past course enrollments, grades, and interaction data.
Graph Construction	Nodes represent learners; edges represent interactions or similarities (e.g., shared courses, similar performance metrics).
Model Architecture	GraphSAGE with three layers used to generate learner embeddings.
Aggregation Functions	Experimented with mean, LSTM, and pooling aggregators.
Optimizer	Adam optimizer with a learning rate of 0.001.
Batch Size	128
Epochs	200

The performance comparison of different aggregation functions in the GraphSAGE model is summarized in Table 2. We evaluated the mean, LSTM, and pooling aggregators using two key metrics: Clustering Quality and Silhouette Score.

Table 2. Performance Comparison of Different Aggregators in GraphSAGE Model

Aggregator	Clustering Quality	Silhouette Score
Mean	0.742	0.638
LSTM	0.768	0.659
Pooling	0.754	0.648

Clustering Quality measures how well learners are grouped into clusters based on their similarities. A higher value indicates that learners within the same group are more similar to each other, while those in different groups are more distinct. The Silhouette Score quantifies how well each learner is matched to its assigned cluster compared to other clusters. It ranges from -1 to 1, where a score closer to 1 indicates well-defined, coherent clusters, and a score closer to -1 suggests misclassification.

In our experiments, the LSTM aggregator outperformed the others, achieving the highest clustering quality (0.768) and silhouette score (0.659), indicating that it effectively captures complex sequential patterns in learner interactions and forms well-separated clusters. The pooling aggregator followed closely with a clustering quality of 0.754 and a silhouette score of 0.648, demonstrating its ability to aggregate rich neighborhood information. In contrast, the mean aggregator exhibited the lowest performance, with a clustering quality of 0.742 and a silhouette score of 0.638, though it still provided reasonable results. These findings highlight the importance of selecting the appropriate aggregation function, as it directly affects the quality of learner embeddings and, ultimately, the effectiveness of personalized learning interventions.

Table 3. Normalized Mutual Information (NMI) and F1-Score for Recommendations

Aggregator	NMI	Precision	Recall	F1-Score
Mean	0.815	0.77	0.72	0.745
LSTM	0.834	0.81	0.76	0.784
Pooling	0.821	0.79	0.74	0.764

NMI indicates how well the semantic groups correspond to the true group labels. Precision, Recall, and F1-Score evaluate the relevance and accuracy of personalized course recommendations made based on the learner groups. The results from Table 2 indicate that the LSTM aggregator achieves the highest clustering quality and silhouette score, suggesting that it captures the most effective representation of the learner data. In Table 3, the LSTM aggregator also demonstrates superior performance in terms of NMI and recommendation metrics, highlighting its capability to generate meaningful embeddings that align well with learner characteristics and preferences. Overall, the experimental results demonstrate that our GraphSAGE-based framework effectively creates semantic groups of learners and enhances the personalization of educational experiences. The use of advanced aggregation methods within GraphSAGE provides a significant improvement in capturing the complex relationships among learners, thereby facilitating more tailored and effective learning interventions. Moving forward, we plan to enhance the scalability of our GraphSAGE-based framework to handle larger datasets and more complex graphs efficiently. Implementing real-time updates will be crucial for maintaining the model's relevance as new data comes in. We also aim to

improve interpretability to build trust and transparency among educators and learners. Investigating cross-domain generalization will help ensure that our framework is applicable in various educational contexts. Addressing ethical and fairness considerations will be a priority to prevent biases and comply with privacy regulations. Integrating our framework with Learning Management Systems (LMS) will make it more practical and widely usable. Comprehensive evaluations across diverse educational settings will help validate its effectiveness, and exploring other advanced GNN models like GIN, GAT, and Graph Transformer Networks will allow us to identify the most effective approaches for different educational data and tasks.

5 Conclusion

In conclusion, this article presents a GraphSAGE-based framework for creating semantic groups of learners within an educational platform. By leveraging the power of Graph Neural Networks, specifically GraphSAGE, our model effectively captures and utilizes the complex relationships between learners and their interactions with educational content. This approach facilitates the formation of meaningful semantic groups, significantly enhancing personalized learning experiences. While our framework shows great promise, several limitations need to be addressed, including scalability, data quality, interpretability, and ethical considerations. Future work will focus on improving these aspects to ensure the model's robustness, transparency, and applicability across diverse educational settings. Overall, our framework represents a significant step forward in applying advanced machine learning techniques to personalized education. By continuously refining our approach and addressing the identified limitations, we aim to contribute to the development of more effective and equitable educational tools that can adapt to the evolving needs of learners.

References

[1] W. Hamilton, "Inductive Representation Learning on Large Graphs," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, 2017.

[2] Xiao Li, Li Sun, Mengjie Ling, Yan Peng, "A survey of graph neural network based recommendation in social networks", Neurocomputing, Volume 549, 2023.

[3] X. He, "LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2020.

[4] J. Wang, H. Xie, F. L. Wang, L.-K. Lee, and O. T. S. Au, "Top-N Personalized Recommendation with Graph Neural Networks in MOOCs," *Computers and Education: Artificial Intelligence*, vol. 2, 2021.

[5] Zhang X-M, Liang L, Liu L and Tang M-J (2021) "Graph Neural Networks and Their Current Applications in Bioinformatics". Front. Genet. 12:690049

[6] T. N. Kipf, M. Welling "Semi-Supervised Classification with Graph Convolutional Networks," 2017.

[7] M. Zhang, "Link Prediction Based on Graph Neural Networks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS)*, 2018.

[8] A. Tsitsulin, J. Palowitch, B. Perozzi, and E. Müller, "Graph Clustering with Graph Neural Networks," arXiv:2006.16904, 2023. [Online]. Available: https://arxiv.org/abs/2006.16904

[9] P. Velickovic et al., "Graph Attention Networks," in *International Conference on Learning Representations (ICLR)*, 2018.

[10] Huang, Q., Zeng, Y. "Improving academic performance predictions with dual graph neural networks". Complex Intell. Syst. 10, 3557–3575 (2024).

[11] Y. Zuo, H. Luo, and L. Xu, "Enhancing MOOCs Personalized Recommendation with Graph Neural Networks and Attention Mechanisms,", 2023.

[12] J. Chen, "Graph Neural Networks for Enhanced E-Learning Systems," *Journal of Educational Technology Systems*, vol. 50, no. 1, pp. 20-35, 2021.

[13] R. Ying, "Graph Convolutional Neural Networks for Web-Scale Recommender Systems," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2018.

[14] A. Defilippo et al., "Leveraging graph neural networks for supporting Automatic Triage of Patients," *Nature*, 2024.

[15] J. Klicpera, A. Bojchevski and S. Günnemann, "Predict then Propagate: Graph Neural Networks meet Personalized PageRank," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[16] X. Wang, "Heterogeneous Graph Attention Network," in *Proceedings of the 12th ACM International Conference on Web Search and Data Mining (WSDM)*, 2019.

[17] L. Zhao, "PairNorm: Tackling Oversmoothing in GNNs," in *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2019.

[18] M. Bronstein, "Geometric Deep Learning: Going beyond Euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18-42, 2017.

[19] M. Defferrard, "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering," in *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, 2016.

[20] K. Wang, "EdNet-KT1," *Kaggle*, 2020. [Online]. Available: https://www.kaggle.com/datasets/gmhost/ednetkt1.