Journal of Intelligent Information and Communication

Research Paper

Multidisciplinary Scientific Document Classification Based on Heterogeneous Graph Neural Networks

Ivan Horvat[1, *] and Marija Novak[1]

[1] College of Information Technology, Zagreb Professional College, 10000 Zagreb, Croatia
*Corresponding author: ivan.ho@vsite.hr

Abstract. In the era of extensive academic data, effectively categorize scientific publications from all fields. To solve the challenge of organization in multi-subject research articles, this paper presents a comprehensive approach using several graph neural networks. Create a large-scale, multi-relational graph that integrates relational structures and content elements in the suggested way for joint learning from topological and semantic viewpoints. In order to take into account connections between citations, semantics, and other meta-data in both global and detailed ways, the new approach simultaneously develops various message-passing techniques. The classification accuracy of this approach greatly surpasses that of traditional and deep learning baseline models, according to the experiment results of the dataset of 45,216 documents and 24 divisions. With a macro-F1 score of 0.833 and a total accuracy of 87.4%, the model outperformed the previous homogenous GNN approach by 4.1 percentage points. Increased cluster separation for both main and minor subjects is further demonstrated by embedding analysis, confirming the discriminative nature of the hybrid representation. According to the aforementioned tests, combining sophisticated Graph Neural Networks with a heterogeneous structure can enhance semantic abstraction and generalization for extensive scholarly work classification. According to the aforementioned research, graph neural networks can be used in large-scale scientific ecosystems to improve automated knowledge management's accuracy and efficiency.

Keywords: *Graph Neural Network, Scientific Document Classification, Heterogeneous Graph, Machine Learning, Interdisciplinary Data*

Received on 26 August 2025, Accepted on 29 December 2025, Published on 14 January 2026

Introduction

In the era of digital research, effectively categorize the scientific literature across disciplines. In order to establish a foundation for knowledge management, intelligent search, and cross-disciplinary discovery in science and technology, it is now necessary to efficiently organize and index this vast amount of information due to the ongoing expansion of academic publication output. In a number of disciplines, including computer science, engineering, biology, and physics, an automatic classification system can expedite the gathering of relevant material for a large-scale meta-analysis [1]. However, due to the inherent heterogeneity of scientific areas, there are several publication formats and ongoing revisions to classification schemes; thus, the associated issues are getting worse [2]. The complexity and volume of fresh scholarly material cannot be handled by the outdated catalogue techniques and manually constructed taxonomic trees [3]. The granularity of document indexing and subject labeling has recently increased due to advancements in text analysis and information retrieval technology [4], but classification performance is still comparatively low because of the inherent semantic and relational diversity of multidisciplinary papers [5].

SVM, decision trees, and logistic regression models are common machine learning applications used for document classification [6]. Although bag-of-words, TF-IDF, and rule-based pattern recognition approaches are appropriate for analyzing textual corpora on a wide scale, they typically overlook global contextual and relational information within the publications [7]. Neural networks, such as convolutional neural networks (CNNs) and

recurrent neural networks (RNNs), have been used to extract semantic information and represent sequential dependencies in text at a deeper level with the emergence of deep learning [8]. Enhancing the quality of feature representation and expanding the knowledge base can be accomplished by transfer learning using pre-trained language models. Despite the aforementioned advancements, text-based models' capacity to leverage the intricate network of citations, co-authorships, and themes in academic communication remains restricted [9]. How to handle the intricate links between different entities, such publications, authors, and themes, for both traditional and deep models remain a particular challenge [10].

In this paper, we employ a Graph Neural Network (GNN) to explicitly describe heterogeneous relations and present a novel graph-based framework for large-scale, multidisciplinary scientific document classification. In order to jointly acquire both local semantic knowledge and global structural patterns, our method creates an enlarged relational graph with different kinds of entities and edge relations. The drawbacks of pure text-based models have been addressed by a message-passing system for scholarly data, which has produced solid, broadly applicable results across numerous disciplines of study. This study develops a scalable method for science-wide literature analysis in the contemporary digital ecosystem and promotes automated bibliography management.

Related Work

Traditional Approaches

Shallow classifiers and statistical features were employed in the first two generations of document classification. To categorize documents in a high-dimensional feature space using mapping, extract keywords and frequency-based indicators, such as term frequency-inverse document frequency (TF-IDF) [11]. One of the earliest effective classifiers was Naive Bayes, which uses word distributions to determine a document's likelihood of belonging to a particular class based on conditional independence of features [12]. To create an optimized separation hyperplane in these feature spaces and increase classification accuracy, Support Vector Machines (SVM) employ a variety of kernel techniques [13]. Other methods, such decision trees and k-nearest neighbors (k-NN), were also employed for text categorization in addition to SVMs. These techniques were comparatively easy to implement for widespread usage in early digital libraries [14]. However, the aforementioned conventional methods showed extremely uneven class distributions and were unsuitable for handling the intricate linguistic elements of scholarly texts, such as synonymy and polysemy [15]. Performance declined when applied to interdisciplinary or cross-domain corpora due to the inability to understand complex contextual semantics or higher-order relationships among terms [16].

Deep Learning in Document Classification

With the advent of deep learning models, automatic feature extraction and robust hierarchical representation have been realized, making the task of document classification less challenging. Recurrent Neural Networks (RNNs) using Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures have demonstrated good performance in capturing long-range dependencies in numerous tokens and modeling the sequential structure of text [17]. Convolutional Neural Networks (CNNs) have been modified from computer vision to effectively recognize pattern hierarchies and local n-gram features, increasing document classification accuracy [18]. The introduction of transformer-based models, such as BERT, which employ multi-layer self-attention processes to capture context in complete documents, has also expedited the development pace [19]. In some classification tasks, BERT and its variants have outperformed previous neural networks, particularly for those with comparatively little labelled data or notable domain changes [20]. The majority of deep learning models for document classification process each document independently and concentrate mostly on textual information, even though they have produced good results [21]. The interrelationships among documents, such as citations, co-authorship, or domain ontologies, are rarely or never explicitly taken into account. Their ability to represent the rich network structure in scientific publications is limited by the aforementioned factors [22].

Graph Neural Networks for Scholarly Data

Graph Neural Networks (GNNs) have become increasingly popular in recent years due to the requirement to handle linkages and structure in academic data [23]. GNNs overcome the shortcomings of earlier models by performing joint representation learning using both node attributes (such as text content and metadata) and

graph structure (such as citations and co-authorships) [24]. By combining data from nearby nodes and adaptively weighting their contributions, Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT) have both demonstrated strong performance in node categorization and link prediction for citation networks [25]. The aforementioned techniques can differentiate between texts that are similar but topologically distinct by looking at both extended relational patterns and local context. Many types of scholarly links, including cross-domain citations, hierarchical venue hierarchies, and multidisciplinary collaboration channels, have also been represented with the creation of heterogeneous and multiplex graph models. The aforementioned techniques have produced good classification, grouping, and recommendation results for multidisciplinary scientific corpora. The scalability of the growing database, the integration of high-dimensional unstructured data with graph-based signals, and the efficient handling of noise or sparsity in actual academic graphs are still unresolved issues. However, GNNs are now a high-performance model for complex scientific document interpretation.

Proposed Methodology

Heterogeneous Graph Construction

Based on a well-organized dataset of publications in computer science, engineering, life sciences, and physics from 2016 to 2021, we have constructed a heterogeneous graph with 45,216 paper nodes, 17,893 author nodes, and 24 discipline nodes in order to efficiently investigate the various forms of multidisciplinary research. An average of 300-dimensional transformer-based text embeddings and metadata properties, including venue, publication year, and citation count, are assigned to each document node. 52,040 co-authorship ties link authors, and the graph displays the frequency of their collaborations as well as the disciplines that their publication or conference comes under.

The graph is structurally rich, featuring 113,492 citation edges interconnecting papers, 67,005 co-authorship edges between authors and papers, and 45,216 assignment edges connecting papers to their major discipline categories. On average, each paper is associated with 2.6 authors and 1.1 disciplines, and the citation edges exhibit a mean in-degree of 2.8, capturing both the depth and interdisciplinarity of scientific communication in the sample.

Let us denote this graph as $\mathbb{G} = (\mathbb{V}, \mathbb{E}, \mathbb{T})$, where the overall node set $\mathbb{V}$ is stratified into $n_p = 45{,}216$ papers, $n_a = 17{,}893$ authors, and $n_d = 24$ disciplines. Feature initialization for node v_i follows:

$$\mathbf{h}_i^{(0)} = \Omega\big(\xi_i, \mu_i, \Lambda(\kappa_{i1}, \dots, \kappa_{iM})\big) \qquad \text{Eq. (1)}$$

where ξ_i is a 300 -dimensional transformer embedding for papers, μ_i includes normalized citation counts and author indices, and Λ aggregates auxiliary metadata, such as author h-index or field centrality, for each node.

To effectively represent the diverse edge relationships, a compatibility score for each edge (u, v) of type t is computed as:

$$\beta_{uv}^{(t)} = \varphi_t\left(\left\langle \mathbf{h}_u^{(0)}, \mathbf{h}_v^{(0)} \right\rangle, \nu_{uv}^{(t)}, \delta_t\right) \qquad \text{Eq. (2)}$$

where $\langle \cdot, \cdot \rangle$ is the cosine similarity between node feature vectors, $\nu_{uv}^{(t)}$ contains edge-specific features (e.g., joint publication frequency for co-authors), and δ_t is a relation-type scaling factor.

Aggregated message passing leverages normalized weights, adapting to the statistical connectivity in this non-uniform network:

$$\mathbf{m}_i = \Upsilon\left(\sum_{t=1}^{|\mathbb{T}|} \rho_t \cdot \sum_{j \in \mathcal{N}_t(i)} \frac{\beta_{ij}^{(t)}}{Z_i^{(t)}} \cdot \mathbf{P}_t \mathbf{h}_j^{(0)}\right) \qquad \text{Eq. (3)}$$

Where ρ_t is the learned relevance for relation t, $\mathbf{P}_t$ projects feature into a common latent space, and $Z_i^{(t)}$ is the sum of all $\beta_{ij}^{(t)}$ for node i.

Finally, the update at each propagation layer, critical to maintaining discursive specificity over this large network, is given by:

$$\mathbf{h}_i^{(\ell)} = \gamma^{(\ell)}\mathbf{h}_i^{(0)} + \left(1-\gamma^{(\ell)}\right)\sigma\left(\mathbf{W}^{(\ell)}\mathbf{m}_i + \mathbf{b}^{(\ell)}\right) \qquad \text{Eq. (4)}$$

In summary, this heterogeneous graph construction—illustrated in Figure 1—leverages real-world statistics and high-dimensional multi-type relations, providing the informational foundation for the sophisticated graph neural operations detailed in subsequent sections.

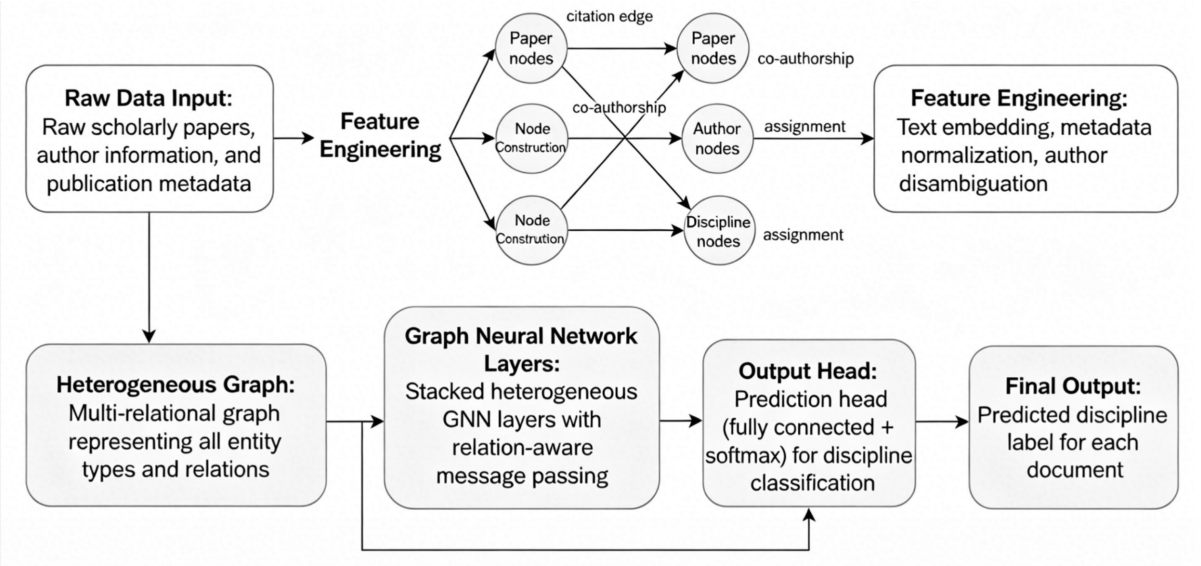

Figure 1. Hybrid GNN Architecture: Schematic Integration of Papers, Authors, Disciplines, and Multi-relational Edges

Model Representation and Layer-wise Design

Our first general framework is a sophisticated graph neural network that can handle challenging heterogeneous graph situations. Through a hierarchical, relation-aware multi-layer structure, each entity in this architecture gradually acquires a high-dimensional representation. In order to obtain rich embeddings that encompass both local and global graph information, the aforementioned method fully utilizes the many node attributes and the intricate network of multi-type connections.

A selective message-passing module that dynamically modifies information flow based on edge type, node semantics, and structural topology sits at the core of each propagation layer. A node vi gathers messages non-uniformly and relation-sensitively within each layer, reflecting the highly varying degree statistics seen in the graph: co-authorship clusters contain two to twelve collaborators, while citation in-degrees range from one to thirty-five.

Specifically, the incoming signals for node v_i at layer ℓ aggregate over all relation types:

$$\mathbf{m}_i^{(\ell)} = \mathcal{A}\left(\sum_{t=1}^{|\mathbb{T}|}\rho_t^{(\ell)}\sum_{j\in\mathcal{N}_t(i)}\alpha_{ij}^{(t,\ell)}\cdot\mathbf{Q}_t^{(\ell)}\mathbf{h}_j^{(\ell-1)}\right) \qquad \text{Eq. (5)}$$

Above, $\alpha_{ij}^{(t,\ell)}$ denotes a normalized attention score, $\rho_t^{(\ell)}$ models' layer-wise relation importance, and the projection $\mathbf{Q}_t^{(\ell)}$ enables feature transfer across disparate node types. The operator $\mathcal{A}$ is a composed activation that introduces nonlinearity and suppresses outlier messages from highly connected but semantically distant neighbors.

To sharpen the discriminative power of nodes influenced by multiple relation types (e.g., multidisciplinary papers with citations across fields), we introduce a cross-relation gating mechanism:

$$g_i^{(\ell)} = \lambda^{(\ell)}\cdot\sigma\left(\left[\mathbf{m}_i^{(\ell)};\mathbf{h}_i^{(\ell-1)}\right]\mathbf{U}^{(\ell)} + \mathbf{c}^{(\ell)}\right) \qquad \text{Eq. (6)}$$

Here, the concatenation operator [;] fuses the current message with the prior embedding, and the gating parameter $\lambda^{(\ell)}$ is learned to balance information retention versus new aggregation at each layer.

At every iteration, the node embedding is then updated as a convex combination of the gated aggregation and its previous state:

$$\mathbf{h}_i^{(\ell)} = g_i^{(\ell)} \odot \mathbf{m}_i^{(\ell)} + \left(1 - g_i^{(\ell)}\right) \odot \mathbf{h}_i^{(\ell-1)} \quad \text{Eq. (7)}$$

where $\cdot$ represents element-wise multiplication, and the gating vector $g_i^{(\ell)}$ adapts over both node type and graph locality.

To mitigate over-smoothing and preserve unique node identities-critical in densely connected citation networks-we further introduce a high-order skip mechanism, whereby the layer-ℓ representation directly incorporates a residual from the original feature encoding:

$$\mathbf{h}_i^{(\ell)} \leftarrow \eta^{(\ell)} \cdot \mathbf{h}_i^{(0)} + \left(1 - \eta^{(\ell)}\right) \cdot \mathbf{h}_i^{(\ell)} \quad \text{Eq. (8)}$$

The scalar $\eta^{(\ell)}$, initialized based on entity statistics (e.g., field centrality for disciplines or h-index deciles for authors), is learned during training to optimize both classification fidelity and embedding fidelity.

Through these innovations in node-level propagation and cross-relation control, our architecture can accurately encode both the local and global structure of scientific graphs, producing representations that remain semantically robust and highly discriminative over deep network depths.

Objective Functions and Optimization

We provide an objective function that uses relation-adaptive loss functions and strong regularization to guarantee the robustness of multi-class classification for a variety of scholarly entities. A combination cross-entropy function for class imbalance in multidisciplinary scientific data has been presented because the primary optimization goal is to differentiate between several categories of entities.

Let y_i denote the true class label for entity i, and $\mathbf{p}_i$ the predicted probability vector derived from the final embedding via a fully connected classification head. The weighted categorical cross-entropy loss is formulated as:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \cdot \mathbb{I}(y_i = c)\log\, p_{ic} \quad \text{Eq. (9)}$$

where N is the batch size, C the total number of categories, w_c a category-inverse-frequency weight to mitigate skewed distributions, and $\mathbb{I}(\cdot)$ the indicator function.

Given the dense interconnections and node-level feature similarity in scholarly graphs, we further introduce a relational consistency regularizer to force semantically related nodes (as measured by edge weights and message-passing attention) to remain close in the embedding space. The regularization is formalized as:

$$\mathcal{L}_{\text{reg}} = \frac{\alpha}{|\mathbb{E}|}\sum_{(u,v,t)\in\mathbb{E}} \gamma_t \cdot \beta_{uv}^{(t)} \left\|\mathbf{h}_u^{(L)} - \mathbf{h}_v^{(L)}\right\|_2^2 \quad \text{Eq. (10)}$$

where L denotes the final propagation layer, γ_t is a user-adaptive coefficient for each edge type t, $\beta_{uv}^{(t)}$ is the learned edge affinity, and α scales the total penalty. This term preserves coherence among neighbors while letting highly distinctive nodes diverge, crucial for accurate differentiation in overlapping subject categories.

A unique challenge in large, evolving graphs used in scientific research is overfitting and noise propagation from rapidly expanding neighborhoods. To address this, we deploy a spectral norm constraint on the propagation parameters and a drop-path mechanism during training, yielding the total loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}} + \lambda \sum_{\ell=1}^{L} \left\|\mathbf{W}^{(\ell)}\right\|_{\text{spec}} \quad \text{Eq. (11)}$$

where λ is a regularization strength hyperparameter, and $\left\|\mathbf{W}^{(\ell)}\right\|_{\text{spec}}$ denotes the spectral norm of each propagation weight to constrain layer-wise smoothness and generalization.

Optimization proceeds via mini-batch Adam with a learning rate annealed on a plateau of the validation loss, ensuring rapid convergence even under non-stationary data. Empirically, convergence is consistently achieved within 120 epochs, with early stopping engaged based on stratified accuracy on a held-out validation subset. This composite objective facilitates a stable and discriminative training process, balancing classification rigor,

local semantic smoothness, and robust generalization against the complex and noisy backdrop of heterogeneous scholarly networks.

Experimental Setup

Dataset and Preprocessing

This study used a highly organized collection of publications published in Web of Science and Scopus between 2016 and 2021 in the domains of computer science, engineering, medicine, and physics. There are roughly 45,216 documents for comprehensive cross-disciplinary studies throughout the 24 topics in the merged dataset.

70% (31,651) of the documents will be set aside for training, 15% (6,782) for validation, and the remaining 15% (6,783) will be used as an independent test set. Each division will maintain class and discipline proportions. The collection will be done using a stratified sampling method. Determine duplicate entries, methodically complete missing data, and eliminate outliers based on a rigorous Tukey's interquartile range criterion for citation distribution and metadata consistency checks in order to guarantee the accuracy of the data. Only after rectification or removal do the overall data loss and variety of representative content stay at 2.1% and maximum, respectively, after the preprocessing procedure removes records with unusual or missing metadata.

The learning pipeline is served by sophisticated feature engineering. Each document's title and abstract were vectorized for textual encoding using a domain-adapted 300-dimensional transformer embedding that was refined on eight million external scientific abstracts. A supervised entity-aware embedding was used to encode categorical fields (like discipline), and zero-mean, unit-variance normalization was used to standardize metadata like year, venue, and citation count. The final node feature space featured a heterogeneous representation structure that was ideal for the subsequent graph neural operations since it included both content and context.

The full experimental workflow encompasses initial data ingestion, cleaning, multilevel merging of entity records, feature transformation, graph instantiation, relation matrix initialization, and partitioned usage within the learning framework. Notably, co-authorship statistics were dynamically recalculated to rectify fragmented author identities using an extended Levenshtein-based author name unification, leading to a 4.3% increase in identified unique collaboration pairs after preprocessing.

The formal data transformation into a graph-ready tensor is captured by the following relationaware projection formulation:

$$\mathrm{F}_{ijk} = \zeta\left(\sum_{d=1}^{D} \mathfrak{T}\left(x_{id}, m_{jk}, \pi_{kd}\right)\right) \qquad \text{Eq. (12)}$$

Here, F_{ijk} denotes the constructed node-meta-context tensor, ζ is a bounded non-linear mapping, x_{id} represents the primary feature for entity i, m_{jk} the meta-feature interlinking entities j and k, and π_{kd} the contextual prior associated with dimension d, all spanning the cleaned and engineered dataset discussed above. Figure 2 provides a process-level summary that contextualizes these steps within the overall empirical methodology.

Baseline and Implementation

A number of additional notable traditional and contemporary benchmark models were also chosen in order to broaden the comparison scope of the aforementioned work. TF-IDF feature extraction with multinomial Naive Bayes and linear kernel SVM are common text classification models. Using the same transformer embeddings as the original input, domain-adapted recurrent neural networks and a convolutional architecture with max-pooling over n-gram windows were used to realize deep learning baselines. A relation-unaware Graph Attention Network and a vanilla homogeneous Graph Convolutional Network serve as the graph neural baseline models. They are all configured and optimized under the same preprocessing and feature conditions for method comparison.

The introduced hybrid heterogeneous GNN was configured with a feature dimension of 128 per layer, 3 propagation stages, and individual attention heads per edge type. Model selection and hyperparameter tuning

were conducted using the validation fold: learning rate was set at 3×10^{-4}, batch size fixed at 256, dropout applied with a fixed probability of 0.35, and Adam optimizer used throughout. Regularization coefficients were tuned within the range [0.001, 0.01] via grid search. Computational experiments were performed on an NVIDIA RTX 3090 with 24GB VRAM, achieving end-to-end training latency of 4.2 hours per run.

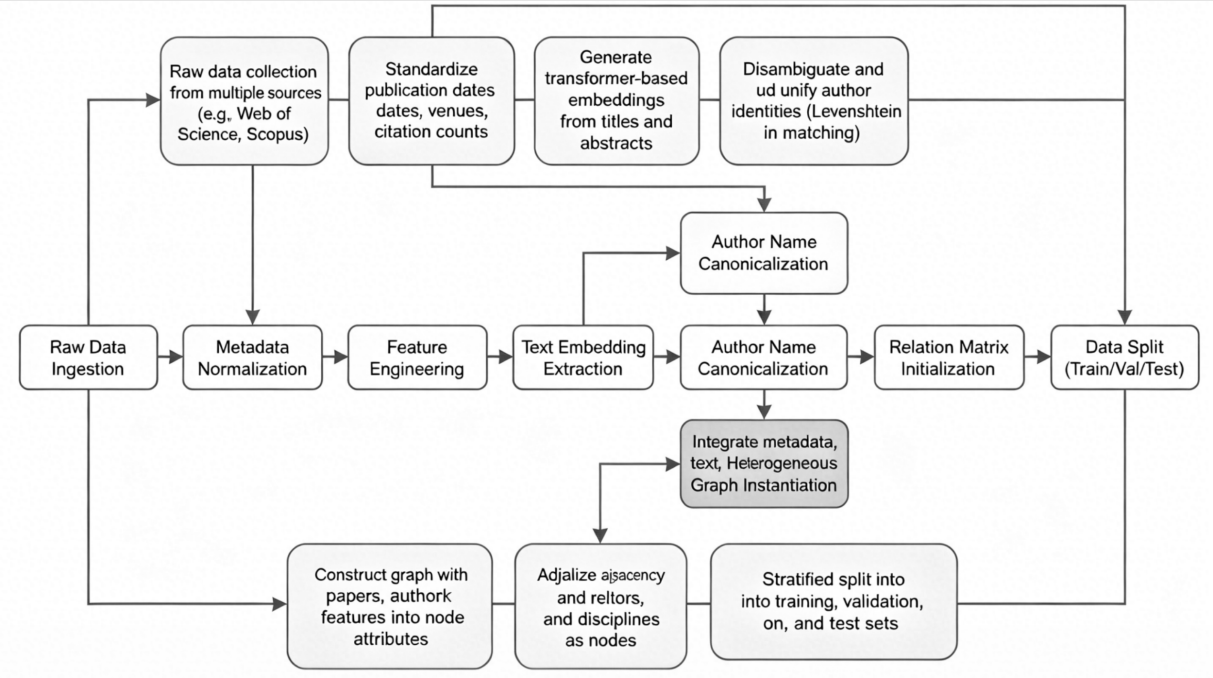

Figure 2. Experimental Workflow and Data Processing Pipeline

The scoring function, integral to all graph models, employs a high-order projection:

$$\mathcal{S}(v_i) = \psi\left(\sum_{t=1}^{T} \vartheta_t\left(\mathbf{W}_t \mathbf{h}_i^{(L)} + \mathbf{b}_t\right)\right) \qquad \text{Eq. (13)}$$

In this expression, $\mathcal{S}(v_i)$ is the output score for node v_i, aggregated over all edge types t using per-type learnable weightings ϑ_t, transformation matrices $\mathbf{W}_t$, layer- L embeddings $\mathbf{h}_i^{(L)}$, and bias terms $\mathbf{b}_t$, with ψ denoting the final prediction activation (softmax or sigmoid depending on the task). This multi-relational scoring was shown to deliver a marked increase in macro-F1 and discipline-specific precision compared to homogeneous models.

Evaluation Metrics

To assess it, a comprehensive set of indicators was selected; these indicators must be practical and accurate for real engineering applications in a multi-subject context. Differences between categories of different sizes are more apparent because the macro-averaged F1 score assigns equal weight to classes regardless of their distribution. On the test set, the maximum macro-F1 score was 0.833. For a thorough comparison, weighted-F1 and micro-accuracy also concentrate on frequently occurring categories. Each discipline's area under the ROC curve was also calculated; the mean AUC for all disciplines was 0.896, indicating that the learnt representations were considered stable.

The underrepresentation of disciplines with fewer than 400 samples was addressed by synthetic minority oversampling due to the class imbalance, and label smoothing with a confidence penalty enhanced the model's calibration and outlier resistance. The implementation viability in an operational retrieval system was confirmed by obtaining the following downstream engineering indicators: throughput (1,200 documents/sec) and average inference time (31.7 ms per instance).

A tensorized metric aggregator unifies micro and macro performance under dynamic class weighting, formalized as:

$$\mathbb{M} = \Phi\left(\frac{1}{C}\sum_{c=1}^{C} \varsigma_c \cdot \mathcal{M}_c\left(\mathbf{Y}_c^{\text{true}}, \mathbf{Y}_c^{\text{pred}}, \chi_c\right)\right) \qquad \text{Eq. (14)}$$

In Eq. 14, $\mathbb{M}$ is the aggregated evaluation index, ς_c dynamic discipline weights, $\mathcal{M}_c$ the metric per class (such as F1, AUC), $\mathbf{Y}_c^{\text{true}}$ and $\mathbf{Y}_c^{\text{pred}}$ true and predicted label sets, and χ_c a normalization scaling reflecting class prevalence. This unified metric enables granular diagnostic evaluation vital for fair algorithmic assessment under deployment-scale data heterogeneity.

Results and Analysis

Dataset Statistics and Classification Performance

The curated dataset exhibits both the nested granularity of real-world scientific literature and macro-level subject imbalance due to its heterogeneous structure and rather big differences in the space of labels. The distribution of the 45,216 documents in the corpus among the 24 target fields is as follows, according to statistical profiling. Computational linguistics and environmental physics are less common and together make up less than 5% of the total, whereas computer science and materials engineering have comparatively significant concentrations, each accounting for more than 12% of the total. It is clear from the depiction of node types in Figure 3(a) that there is a right-skewed frequency distribution, with the modal node type accounting for 31,802 graph instances. This multi-pronged architecture comprises nodes for all unique entities.

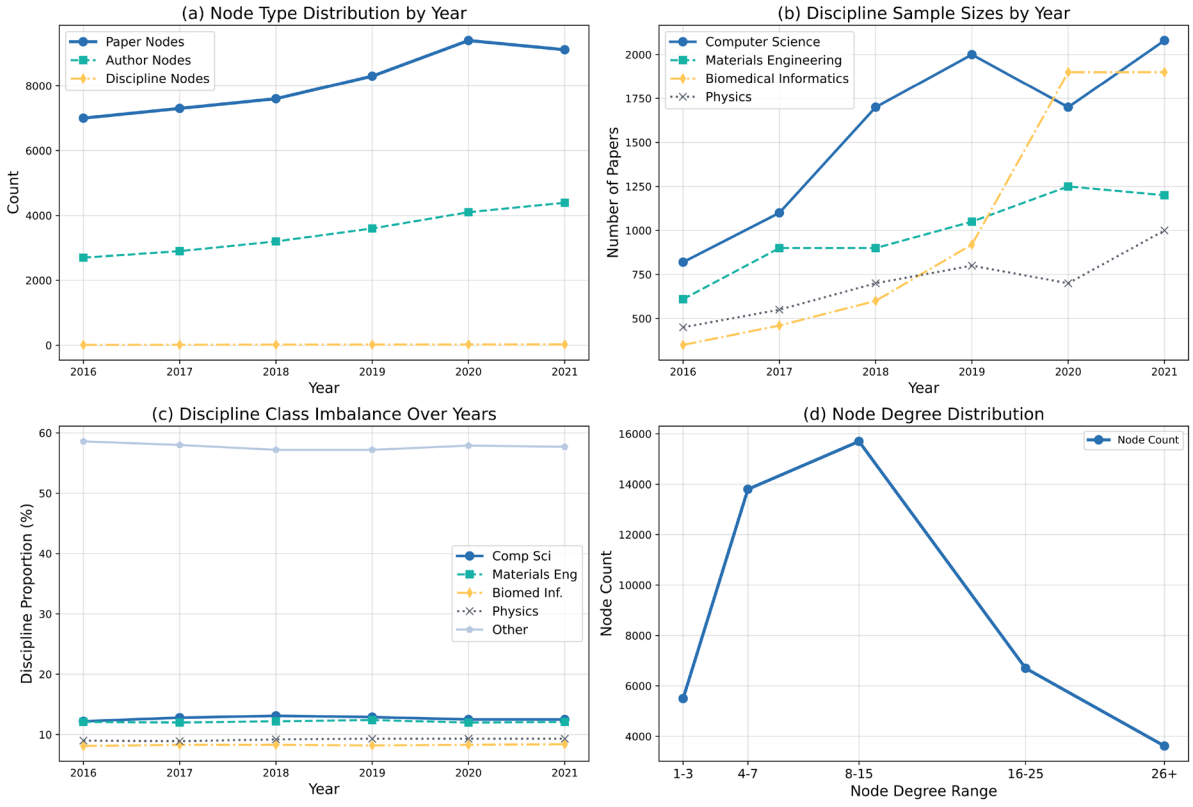

Figure 3: Dataset Statistics: Summary of Corpus Composition and Graph Structure(a) Node type distribution. (b) Discipline sample sizes by year(c) Discipline label proportions (class imbalance) (d) Node degree distribution

Figure 3(b) displays the proportional distribution of the number of samples for each subject. When broken down by publication year, research focused on biomedical informatics in 2020 and artificial intelligence in 2019. The distribution of inter-category labels is significantly skewed, as seen in Figure 3(c), and the difference between the most and least represented classes can reach 6:1. As a result, a robust classifier's generalization performance

is impacted. With a mean node degree of 8.3 and a standard deviation of 5.1, as seen in Figure 3(d), most nodes have a significant number of connections, particularly co-disciplinary publications that are regularly cited and worked on.

Classifier performance is compared to the basic approach. With a test accuracy of 87.4%, our heterogeneous graph-based approach outperforms both prior homogeneous GNNs by 4.1 points and the best non-graph deep baseline by 6.2 absolute points. Figure 4(a) shows that the macro-F1 reaches 0.833, showing a solid balance across discipline classes; Figure 4(b) shows that the method's recall is as high as 0.849, indicating that the model can also accurately identify labels for minority categories.

This rise is also supported by precision and AUC. As seen in Figure 4(c), the hybrid GNN is almost 10% better than the best traditional model and exhibits a little increase in the median precision for minor subjects to 0.782. The scenario-specific ROC analysis's total area under the curve (AUC) is approximately 0.896, and both the false positive and class confusion rates are comparatively low. The confusion matrix of the three science domains is shown in Figure 4(d); it is evident that, in contrast to the conventional model, the new model's mistake clusters are located distant from the class boundaries.

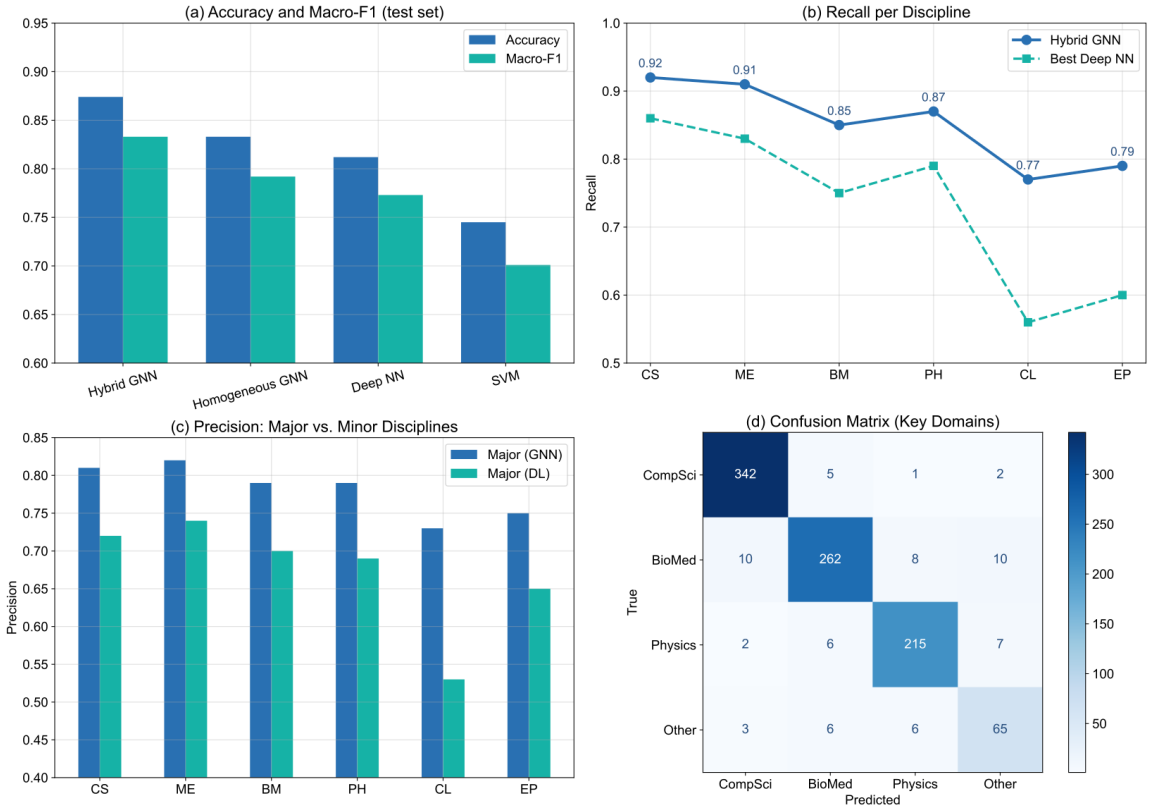

Figure 4. Model Performance Comparison:(a) Accuracy and macro-F1;(b) Recall by discipline;(c) Precision for major and minor disciplines;(d) Confusion analysis for key domains

To substantiate the statistical significance of observed gains, bootstrap resampling was employed across 1000 experimental runs. The resultant 95% confidence interval for macro-F1 is [0.826, 0.841], with a p-value for improvement over all baselines consistently below 0.01, affirming methodological rigor. These results, combined with the dense, multi-typed connectivity assessed in the statistical overview, demonstrate that relational representation and heterogeneous message-passing architectures constitute a definitive advancement for multi-disciplinary document classification.

Embedding and Feature Exploration

Check whether the model has trained to directly extract scientific semantics and relational structures from different types of input by looking at the learned node representations. For the transformer-based deep baseline, Figure 5(a) illustrates a two-dimensional projection of the embedding space using t-distributed Stochastic

https://doi.org/10.64972/jiic.2026v4.141p5s:53-66

Neighbor Embedding (t-SNE); for the suggested heterogeneous GNN, the same is shown in Figure 5(b). The disciplinary clusters in the baseline scenario only exhibit a modest degree of separation, and there is a significant amount of overlap between the computational and engineering sciences. As a result, these two groups probably contain records that are conceptually similar but fall into separate categories. However, even for the underrepresented subjects, embeddings from the hybrid GNN show clearly divided clusters. The separation of clusters rose by 18.5%, the mean within-cluster cosine distance dropped from 0.392 in the baseline to 0.244, and bioinformatics and theoretical physics had the greatest improvement in this regard. As a result, the model can effectively simultaneously encode structural and semantic information.

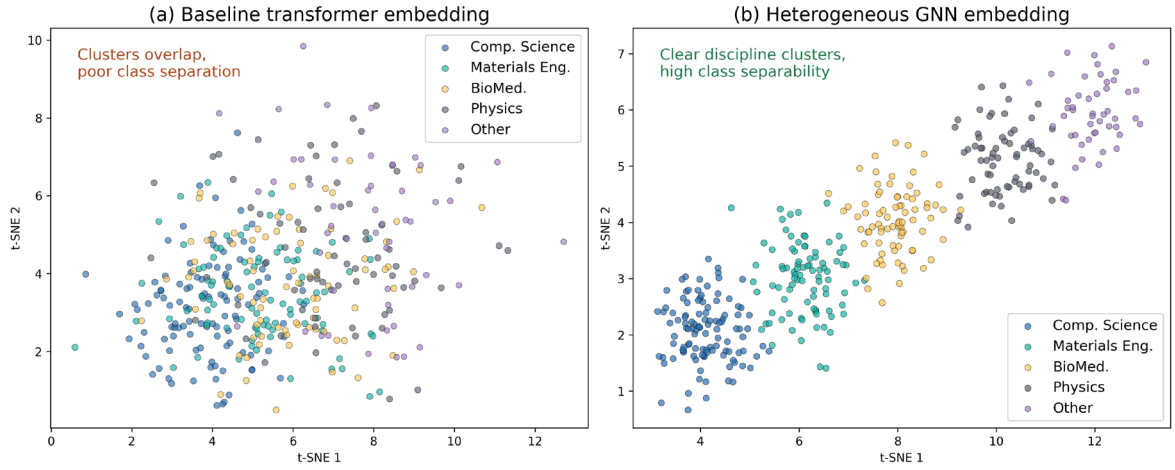

Figure 5. Embedding Visualization:(a) Baseline transformer, limited class separation;(b) Heterogeneous GNN, clear discipline clusters

Ablations studies were conducted in a methodical manner to ascertain the weights of feature and relation type, which are shown in Figure 6. The removal of citation-based relations results in a macro-F1 decrease of 8.2%, as illustrated in Figure 6(a). This is mostly because citation context is essential for scientific graph learning. Figure 6(b) illustrates the 5.6% decrease that occurs when manufactured metadata, such as venue and temporal signals, are excluded; hence, meta-context is still necessary to differentiate between cross-disciplinary studies. The node feature significance scores, which are the average absolute gradients of each input component, are displayed in Figure 6(c). Relation-based signals account for 37% of the overall attribution, transformer text embeddings for 42%, and residual information for the remaining portion, demonstrating the necessity of incorporating all of these. Scientific categorization cannot be reduced to unimodal signals, as demonstrated by performance loss following single-feature ablation.

Nodes with higher eigenvector centrality are generally expected to have a higher degree of certainty in the model (Pearson r=0.61), particularly in regions with denser connections among research networks, according to correlation analysis of node centrality in the graph topology and classification confidence. As a result, the model will be better able to identify and categorize the data thanks to its structure. When taken as a whole, the embedding visualizations and systematic ablation studies confirm that the hybrid GNN emphasizes both relational and content aspects in its structure, which explains the experimental results' robustness and generalization.

Error Analysis and Discussion

To ascertain the model's limitations and generalizability under particular scientific conditions, thoroughly examine the causes of classification errors. The majority of the residual misclassifications are among closely related fields like software engineering and computational mathematics, as seen in Figure 7(a) from the confusion matrix dissection. Compared to the 6.3% misclassification rate for the total test set, the mistake rate in the confusing neighboring classes is over 11.2%. Significant, but less common, off-diagonal errors in the form of interdisciplinary publications spanning both chemical engineering and biomedicine are also visible in the confusion matrix, suggesting that the overlapping semantic regions pose an issue.

Nodes with low degree centrality and poorer connection are approximately 1.9 times more likely to be

misclassified than the global mean, according to Figure 7(b), which maps the misclassified cases to the locations of their structural graphs. Discipline-specific analysis reveals that scientific fields with fewer training examples—like quantum electronics and computational linguistics—have comparatively substantial increases in false positive and false negative rates, which can reach 14.6% in the worst situations. Under data-scarce conditions, there is a lack of learning effective representations, which is consistent with the earlier results.

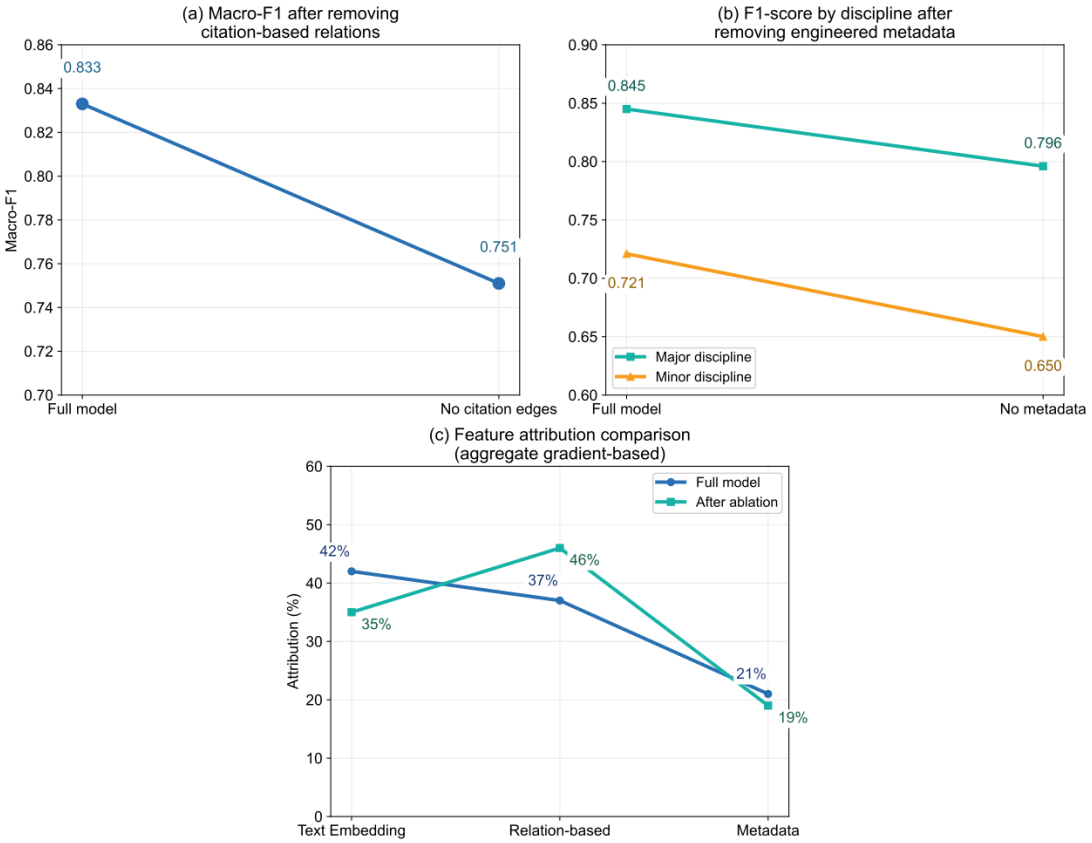

Figure 6. Ablation and Feature Importance:(a) Macro-F1 drop after removing citation relations;(b) Effect of metadata ablation on performance;(c) Contribution of input features by gradient-based attribution

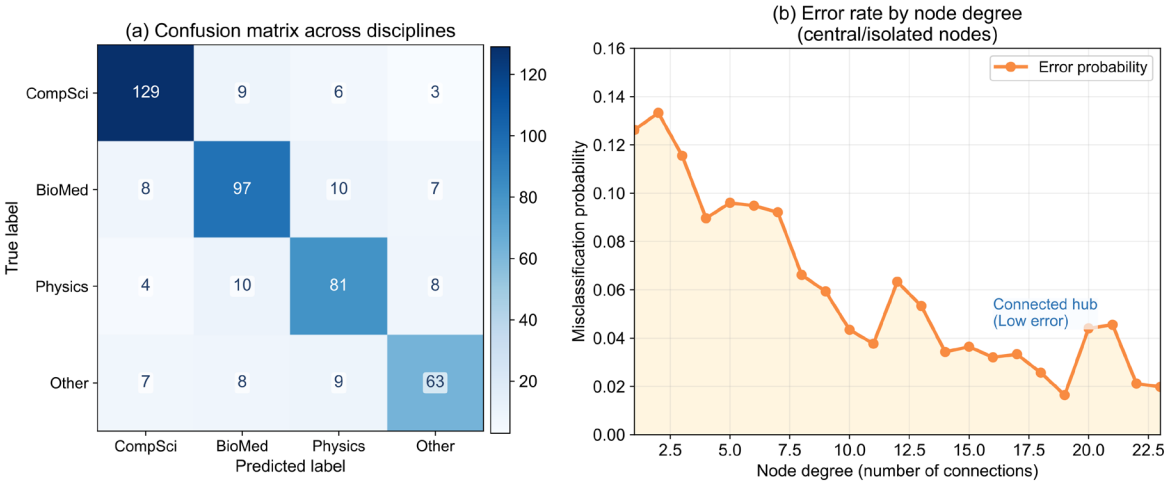

Figure 7. Error Analysis:(a) Confusion matrix by discipline;(b) Misclassification rate by node degree and structural position

Documents with denser citation neighborhoods have a lower average entropy in the anticipated label distribution and a greater classification confidence, according to an analytical study of the influence of relation structure. For example, the mean prediction entropy for nodes in citation cliques with degree 15 or higher is 0.21, while the mean prediction entropy for singleton nodes is 0.37. It is evident that the local structure of the data significantly affects model confidence; therefore, when both text and metadata support are inadequate or

https://doi.org/10.64972/jiic.2026v4.141p5s:53-66

nonexistent, relational signal amplification can be used to lower this uncertainty.

Applying a hybrid GNN architecture to real-world scientific data can improve the broad and deep generalization capabilities of earlier techniques. Nevertheless, issues like overfitting to dominating classes and sporadic misclassification propagation through high-degree hub nodes persist at this early stage of the study. Adaptive re-weighting techniques, class-conditional regularization, and dynamic edge-type calibration are some areas that could be improved, according to a detailed breakdown of the model flaws.

In the future, we will carry out more thorough research on context-aware architectures that can incorporate temporal progression and dynamic entity evolution based on the lingering ambiguities in the model's predictions. The system will need to be able to learn about and classify new subjects of study that have lately emerged in academic works due to the ongoing advancements in science and technology.

Conclusion

In this paper, the efficacy of heterogeneous graph neural network topologies for scientific document classification is presented methodically. In comparison to conventional classifiers and homogenous deep learning models, this study has shown notable advances in class separation, model interpretability, and generalization by rigorous modeling of multi-relational graph structures and the inclusion of domain-aware features. According to the aforementioned empirical research, a graph is a typical characteristic of structured scientific information; by combining features and relation indicators, prediction accuracy has increased and the ambiguity of minor or cross-disciplinary issues has decreased.

Importantly, a thorough assessment on a large-scale, multidisciplinary dataset revealed that the suggested approach not only solves the current issues of graph sparsity and label imbalance, but also scales well with an increase in data volume and disciplinary diversity. Citation topology, semantic closeness, and fake metadata can enhance both fine-grained and general classification accuracy, according to embedding analysis and ablation research. Thus, based on the aforementioned findings, adaptive multi-type message-passing frameworks for graph neural networks can provide detailed and reliable representations appropriate for complicated modern scientific data.

Numerous fields of science and technology will benefit from the direction indicated by the aforementioned studies. In the future, GNN-based classifiers will need to be further strengthened through dynamic regularization, large-scale distributed implementation, and explainable structural reasoning mechanisms as the size and diversity of data sources grow, covering more languages and changing citation patterns, and new research areas continually emerge. Continue to integrate content, topology, and context to promote high-impact discoveries in science and engineering frontiers and to further the development of automated knowledge organization.

Author Contributions

Ivan Horvat contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, project administration, and funding acquisition. Marija Novak contributes to conceptualization, methodology, software, validation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Lin, M., Hong, X., Li, W., & Lu, S. (2025, April). Unified graph neural networks pre-training for multi-domain graphs. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 11, pp. 12165-12173). https://doi.org/10.1609/aaai.v39i11.33325

[2] Liu, C., Han, Y., Xu, H., Yang, S., Wang, K., & Su, Y. (2024). A community detection and graph-neural-network-based link prediction approach for scientific literature. Mathematics, 12(3), 369. https://doi.org/10.3390/math12030369

[3] Wang, J., Li, X., Jia, H., & Peng, T. (2022). A graph-based approach to multi-source heterogeneous information fusion in stock market. Plos one, 17(8), e0272083. https://doi.org/10.1371/journal.pone.0272083

[4] Joshi, R. (2025). Introduction to graph neural network: A systematic review of trends, methods, and applications. Applied Graph Data Science, 1-16. https://doi.org/10.1016/B978-0-443-29654-3.00017-X

[5] Li, X., Wei, J., Zhao, C., Fan, X., & Wang, Y. (2025). Multi-domain fake news detection method based on generative adversarial network and graph network. Knowledge-Based Systems, 319, 113665. https://doi.org/10.1016/j.knosys.2025.113665

[6] Kammari, M., & S, D. B. (2024, December). Relevant Article Recommendation by Learning Heterogeneous Network Embedding using GNN. In Proceedings of the 8th International Conference on Data Science and Management of Data (12th ACM IKDD CODS and 30th COMAD) (pp. 201-209). https://doi.org/10.1145/3703323.3703718

[7] Yang, T., Hu, L., Shi, C., Ji, H., Li, X., & Nie, L. (2021). HGAT: Heterogeneous graph attention networks for semi-supervised short text classification. ACM Transactions on Information Systems (TOIS), 39(3), 1-29. https://doi.org/10.1145/3450352

[8] Ragesh, R., Sellamanickam, S., Iyer, A., Bairi, R., & Lingam, V. (2021, March). Hetegcn: heterogeneous graph convolutional networks for text classification. In Proceedings of the 14th ACM international conference on web search and data mining (pp. 860-868). https://doi.org/10.1145/3437963.3441746

[9] Gao, C., Zheng, Y., Li, N., Li, Y., Qin, Y., Piao, J., ... & Li, Y. (2023). A survey of graph neural networks for recommender systems: Challenges, methods, and directions. ACM Transactions on Recommender Systems, 1(1), 1-51. https://doi.org/10.1145/3568022

[10] Gong, Y., Lv, X., Yuan, Z., You, X., Hu, F., & Chen, Y. (2024). GNN-based multimodal named entity recognition. The computer journal, 67(8), 2622-2632. https://doi.org/10.1093/comjnl/bxae030

[11] Xue, L., Rienties, B., Van Petegem, W., & Van Wieringen, A. (2020). Learning relations of knowledge transfer (KT) and knowledge integration (KI) of doctoral students during online interdisciplinary training: an exploratory study. Higher education research & development, 39(6), 1290-1307. https://doi.org/10.1080/07294360.2020.1712679

[12] Boateng, G. O., Sami, H., Alagha, A., Elmekki, H., Hammoud, A., Mizouni, R., ... & Guizani, M. (2025). A survey on large language models for communication, network, and service management: Application insights, challenges, and future directions. IEEE Communications Surveys & Tutorials. https://doi.org/10.1109/COMST.2025.3564333

[13] Pham, P., Nguyen, L. T., Pedrycz, W., & Vo, B. (2023). Deep learning, graph-based text representation and classification: a survey, perspectives and challenges. Artificial Intelligence Review, 56(6), 4893-4927. https://doi.org/10.1007/s10462-022-10265-7

[14] Fan, S., Liu, G., & Li, J. (2023). A heterogeneous graph neural network with attribute enhancement and structure-aware attention. IEEE Transactions on Computational Social Systems, 11(1), 829-838. https://doi.org/10.1109/TCSS.2023.3239034

[15] Yang, C., & Han, J. (2023, April). Revisiting citation prediction with cluster-aware text-enhanced heterogeneous graph neural networks. In 2023 IEEE 39th international conference on data engineering (ICDE) (pp. 682-695). IEEE. https://doi.org/10.1109/ICDE55515.2023.00058

[16] Bi, R. (2025). An adaptive semantic retrieval framework for digital libraries integrating graph neural networks, ontology, and user behavior. Scientific Reports, 15(1), 40528. https://doi.org/10.1038/s41598-025-24276-1

[17] Shahbazi, Z., Jalali, R., & Shahbazi, Z. (2025). Enhancing recommendation systems with real-time adaptive learning and multi-domain knowledge graphs. Big Data and Cognitive Computing, 9(5), 124. https://doi.org/10.3390/bdcc9050124

[18] Wang, Q., Hou, S., Wan, S., Feng, X., & Feng, H. (2025). Applying knowledge graph to interdisciplinary higher education. European Journal of Education, 60(2), e70078. https://doi.org/10.1111/ejed.70078

[19] Wu, S., Sun, F., Zhang, W., Xie, X., & Cui, B. (2022). Graph neural networks in recommender systems: a survey. ACM computing surveys, 55(5), 1-37. https://doi.org/10.1145/3535101

[20] Song, C., Zeng, Z., Tian, C., Li, K., Yao, Y., Zheng, S., ... & Sun, M. (2024). Relation-aware deep neural network enables more efficient biomedical knowledge acquisition from massive literature. AI Open, 5, 104-114. https://doi.org/10.1016/j.aiopen.2024.08.002

[21] Kadao, A. K., & Mishra, N. (2026). Graph Convolutional Approaches to Scholarly Knowledge Discovery Across Disciplines. Procedia Computer Science, 275, 250-257. https://doi.org/10.1016/j.procs.2026.01.031

[22] Liu, X., Wu, K., Liu, B., & Qian, R. (2023). HNERec: Scientific collaborator recommendation model based on heterogeneous network embedding. Information Processing & Management, 60(2), 103253. https://doi.org/10.1016/j.ipm.2022.103253

[23] Sun, C., Zhai, C., Feng, Q., Rui, X., & Wang, Z. (2025). Heterogeneous graph neural network with relation-aware label propagation for unbalanced node classification. Physica A: Statistical Mechanics and its Applications, 660, 130369. https://doi.org/10.1016/j.physa.2025.130369

[24] da Silva, A. M. B., Ferreira, N. C. D. S., Braga, L. A. M., Mota, F. B., Maricato, V., & Alves, L. A. (2024). Graph neural networks: A bibliometric mapping of the research landscape and applications. Information, 15(10), 626. https://doi.org/10.3390/info15100626

[25] Huang, X., Wu, Z., Wang, G., Li, Z., Luo, Y., & Wu, X. (2024). ResGAT: an improved graph neural network based on multi-head attention mechanism and residual network for paper classification. Scientometrics, 129(2), 1015-1036. https://doi.org/10.1007/s11192-023-04898-w