International Journal of Advances in Data and Information Systems
Vol. 6, No. 2, August 2025, pp. 276~289
ISSN: 2721-3056, DOI: 10.59395/ijadis.v6i2.1382

A Hybrid Model of Graph Attention Networks and Random Forests for Link Prediction in Co-Authorship Networks

Ika Arfiani[1], Herman Yuliansyah[2]
[1,2]Department of Informatics, Universitas Ahmad Dahlan, Indonesia

Article Info

Article history:

Received May 16, 2025
Revised Jun 14, 2025
Accepted Jul 03, 2025

Keywords:

Co-authorship Prediction
Complex Networks
Deep Learning
Ensemble Learning
Link Prediction

ABSTRACT

Co-authorship prediction is important in academic network analysis due to it helps to understand patterns of scientific collaboration and supports collaboration recommendation systems. Topology-based approaches, such as connectivity metrics and node distance, have been widely used to model new relationships in networks. However, these approaches often overlook relevant author attributes, such as reputation and productivity. This study develops a co-authorship prediction model by combining a Graph Attention Network (GAT) and a Random Forest. GAT is used to extract topological features from the co-authorship graph, while Random Forest leverages additional attributes such as h-index and the number of publications to improve prediction accuracy. Experiments were conducted on a co-authorship dataset comprising over 10,000 authors and 50,000 publications. The results show that GAT achieved 85% accuracy, while Random Forest reached 80%. The combination of the two yielded 90% accuracy and a higher F1-score, indicating a better balance between precision and recall. The combined model also proved more accurate in predicting collaborations involving highly productive authors. These findings suggest that a hybrid approach can more comprehensively capture the dynamics of academic collaboration and may serve as a foundation for developing more effective collaboration prediction systems in the future.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Ika Arfiani,
Department of Informatics,
Universitas Ahmad Dahlan,
Jl. Ringroad Selatan, Kragilan, Tamanan, Kec. Banguntapan, Kabupaten Bantul, Daerah Istimewa Yogyakarta 55191
Email: ika.arfiani@tif.uad.ac.id

1. INTRODUCTION

Co-authorship is an important indicator in the academic world that reflects collaboration among researchers. It forms the basis for knowledge exchange and innovation, enabling the combination of diverse expertise to address complex research problems. Over the past two decades, academic collaboration has increased significantly. According to the National Science Foundation (NSF), research collaboration in the United States has grown by more than 50%, driven by the increasing complexity of scientific research and the need for multidisciplinary approaches [1]. With this growth, predicting co-authorship relationships has become a relevant area of study, particularly for understanding the dynamics of scientific networks and promoting future collaboration.

Link prediction is a method used to anticipate future or missing connections between entities by analyzing the existing network structure [2]. Most link prediction methods use

topological information based on the Triadic Closure principle to predict future links. However, methods relying on the Triadic Closure principle are unable to predict future links when the pair of nodes being evaluated do not share any common neighbors. Therefore, several methods have been proposed to address this issue, such as machine learning approaches [3], extensions of the Adamic index based on Degree Centrality, Closeness Centrality, and Clustering Coefficient [4]. Subsequently, some of these approaches have been further developed to tackle the cold-start problem in link prediction, such as the Degree of Gravity for Link Prediction (DGLP) method [5].

Co-authorship network analysis is a form of social network analysis focused on scientific collaboration. Unlike real-world relationships, these collaborations are more influenced by shared research interests. Although authors may be connected in real life, institutions often rank them solely based on publication count, which does not reflect the actual relationship patterns among researchers [6]. Article content analysis can also be used to identify attribute sets that frequently appear together, revealing patterns within those collections [7]. The abundance of textual data in research presents both opportunities and challenges. While it enables deeper insights and the discovery of hidden value, it also makes effective organization and recognition of data characteristics more difficult [8]. A new co-authorship prediction approach is needed that incorporates both network structure and research interests, as authors tend to collaborate based on shared research topics while still being influenced by network topology [9].

Graph-based learning methods, especially Graph Attention Networks (GAT), have emerged as powerful tools in network analysis. GAT utilizes an attention mechanism to assign different weights to neighbors in the network, enabling the model to capture complex topological structures and interactions [10], [11]. Unlike traditional graph methods, GAT can dynamically prioritize significant relationships, making it well-suited for complex networks such as co-authorship. Complementing this, Random Forest, a reliable ensemble learning technique, has demonstrated superior performance in handling both structured and unstructured data [12]. Its ability to model nonlinear relationships and reduce overfitting makes it an ideal candidate to be combined with GAT for effectively predicting co-authorship links.

Existing research demonstrates the value of combining graph-based methods with machine learning algorithms to enhance prediction performance. For example, the study by Kipf and Welling [13], introduced Graph Convolutional Networks (GCNs) for node classification tasks, showcasing the potential of graph-based neural networks. Similarly, Velickovic et al. [10], extended this approach with GAT, proving its effectiveness in assigning attention weights to node features. However, these methods often lack integration with traditional machine learning models that can utilize additional features beyond topological data.

This study proposes a hybrid approach by integrating Graph Attention Networks (GAT) and Random Forest to predict co-authorship relationships. The approach leverages GAT to extract meaningful topological features from the co-authorship network and combines these features with Random Forest to improve prediction accuracy. The co-authorship dataset used in this study, sourced from GitHub, includes comprehensive information about authors, publications, and their interactions, providing rich context for the analysis [14].

The significance of co-authorship networks goes beyond academic collaboration. Publications with multiple authors generally receive more citations, indicating the impact of collaborative research. Data from Scopus shows that multi-authored publications achieve citation counts nearly 40% higher than single-author publications [14]. Furthermore, factors such as an author's h-index, number of publications, and research field play important roles in shaping co-authorship patterns [15], [16]. These variables form the basis for predictive modeling in this study.

Despite its importance, predicting co-authorship relationships presents challenges. Co-authorship networks are high-dimensional and highly sparse, making modeling difficult. Moreover, understanding latent interactions within the network requires advanced algorithms capable of processing heterogeneous data. Recent advances in dynamic graph neural networks (DGNNs) and attention mechanisms have addressed some of these challenges, offering opportunities for improved predictive models [17], [18].

This study aims to develop a co-authorship link prediction model that integrates Graph Attention Network (GAT) and Random Forest to improve prediction accuracy. Specifically, the objectives of this research include:

1. Extracting topological and structural features from the co-authorship network using GAT.
2. Integrating additional node-level attributes, such as h-index and publication count, using Random Forest.
3. Evaluating the performance of the combined model compared to each individual model implemented separately.

The methodology framework of this study includes data preprocessing, feature extraction using GAT, and final prediction using Random Forest. The dataset will be split into training and testing subsets to validate model accuracy. Additionally, this study compares the proposed hybrid model with other state-of-the-art methods, including DeepWalk [19], Node2Vec [20], and dynamic graph-based approaches [21].

The findings of this study are expected to contribute to the advancement of co-authorship prediction research, with practical implications for researchers and academic institutions. By identifying potential collaborators and understanding co-authorship dynamics, this study aims to enhance the productivity and impact of scientific research. Furthermore, the integration of graph-based learning with traditional machine learning techniques provides a scalable framework for other applications in social network analysis and knowledge discovery.

2. RESEARCH METHOD

2.1. Dataset

The co-authorship dataset used in this study includes authors from various disciplines (computer science, physics, and biomedical sciences) as well as from diverse international academic institutions, and is available on GitHub [14]. The dataset contains information about authors, publications, and collaborative relationships among them. The data were collected from international journals and conferences, covering a wide range of research domains. It comprises more than 10,000 authors and over 50,000 publications, making it a rich resource for co-authorship network analysis [6]. This diversity enables the model to capture broader collaboration patterns and evaluate its generalization capability across different scientific contexts [14].

2.2. Preprocessing Data

The initial steps of this study involve data preprocessing to ensure data quality and consistency:

1. Data cleaning is performed by removing duplicate entries, missing data, and irrelevant information [15], [22].
2. Data exploration, such as network distribution analysis, is conducted to understand collaboration patterns. For example, a histogram of the number of collaborations shows that 70% of authors have more than one collaboration [14], [23].
3. Construction of the co-authorship graph, which is an undirected graph where nodes $v \in V$ represent authors and edges $e = (u, v) \in E$ represent collaboration relationships between authors [14], [24].
4. Extraction of additional features to obtain the h-index as a measure of author productivity and impact [16], [25].
   $h - index = \max(h | h \text{ publikasi masing} - \text{masing dengan } h \text{ sitasi atau lebih})$
5. The number of publications as the total publications for each author.
6. Research fields grouped based on the subject areas of journals or conferences [14], where $P(v)$ is the total publications for author v. The resulting graph serves as input for the GAT model, while the additional features are used in the Random Forest.

2.3. Model Graph Attention Network (GAT)

Graph Attention Network (GAT) is used to extract topological features from the co-authorship graph. The attention mechanism allows the model to assign different weights to each neighbor of a node, thereby capturing the most relevant relationships [10]. The main formulations used are as follows:

1. The attention mechanism is the attention weight calculated based on [10][17][26]:
$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\alpha^T[Wh_i||Wh_j]))}{\sum_{k\in N(i)}\exp(\text{LeakyReLU}(\alpha^T[Wh_i||Wh_k]))}$$
where h_i is the feature vector of node, W is the weight transformation matrix, α is the attention vector, $N(i)$ is the set of neighbors of node i, and ‖ denotes the concatenation operator.
2. Feature aggregation is a new feature for each node calculated as [10]:
$$h_i' = \sigma\left(\sum_{j\in N(i)} \alpha_{ij} W h_j\right)$$
where σ is a non-linear activation function (e.g., ReLU), and W is the weight matrix for feature transformation.
3. Loss Function is a function for GAT using binary cross-entropy for edge prediction tasks with outputs that represent the probability of relationships between nodes and [10],[27]:
$$L = \sum_{(i,j)\epsilon E}\left[y_{ij}\log(\hat{y}_{ij}) + (1-y_{ij})\log(1-\hat{y}_{ij})\right]$$
where y_{ij} is the ground truth label (0 or 1) for the edge, and, $\hat{y}_{ij}$ is the predicted probability of the model for the edge (i,j).

2.4. Random Forest Model

Features extracted from GAT and additional attributes (h-index, publication count, research field) are used as input to the Random Forest. Random Forest is an ensemble algorithm that builds multiple decision trees to improve accuracy. The main formulations include:

1. Construction of decision trees where each tree is trained on a randomly selected subset of the data with replacement (bootstrap sampling) [12], [28], [29].
$$D_b = \{(x_1, y_1), (x_2 y_2), \ldots, (x_n, y_n)\}$$
2. Ensemble prediction is the final prediction produced by majority voting for classification. [12].
$$\hat{y} = \text{majority voting } (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m)$$
where $\hat{y}_i$ is the prediction from the i and m is the number of trees.
3. The error function (Gini Impurity) is a function in Random Forest calculated using Gini impurity for each split:
$$Gini(t) = 1 - \sum_{i=1}^{C} \mathcal{P}_i^2$$
where $\mathcal{P}_i$ is the proportion of samples of class i at node t, and C is the number of classes [12], [30].

2.5. Combination of GAT and Random Forest Models

The model integration process leverages GAT's capability to extract topological features and Random Forest's strength in utilizing additional node-level attributes:

1. Topological feature extraction is a representation of nodes produced by the final layer of GAT used as additional input for Random Forest. [10], [31].
2. Feature combinations are features generated by GAT combined with additional features (h-index, number of publications, and research field) [10], [16].
$$Combined\ Features = [Features\ GAT, h-index, Number\ of\ Publications]$$

3. The final prediction is a random forest model trained using a combination of these features to predict co-authorship links. [12].

$$\hat{y} = Random\ Forest(Fitur\ Kombinasi)$$

2.6. Model Evaluation

Model evaluation was conducted to assess the performance of GAT, Random Forest, and their combination. The dataset used consists of over 10,000 authors and 50,000 publications. For training efficiency, subsets of the data were employed in several initial experiments; however, the full dataset was utilized in the final evaluation phase using the following metrics:

1. Accuracy, where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively [19].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision, which indicates the proportion of correctly predicted positive instances [20].

$$Precision = \frac{TP}{TP + FP}$$

3. Recall (Sensitivity), which measures the model's ability to detect all actual positive instances [20], [32].

$$Recall = \frac{TP}{TP + FN}$$

4. F1-Score, which provides a balance between precision and recall [20], [32].

$$F1 - Score = 2 \text{ x } \frac{Precision\ x\ Recall}{Precision + Recall}$$

5. Area Under the Curve (AUC), which measures the model's ability to distinguish between positive and negative classes, calculated using the ROC curve [20], [33].
6. Cross-validation is applied to ensure model generalization. The dataset is split into 80% training data and 20% testing data, using 5-fold cross-validation to reduce bias from data partitioning [12], [20].

2.7. Implementation Results Analysis

The results analysis covers performance metric comparisons, feature importance interpretation, and a discussion of the findings' implications within the context of co-authorship networks. The experimental outcomes were evaluated across three model variations:

1. Graph Attention Network (GAT). A graph-based model without additional author-level attributes [10].
2. Random Forest. An attribute-based model that excludes topological information from the co-authorship graph [14].
3. Hybrid GAT and Random Forest. A combined model that integrates topological representations from GAT with node-level features such as h-index and publication count [10], [12], [20].

To illustrate the workflow of the hybrid GAT+RF system proposed in this study, a process diagram of the combined model is presented in Figure 1. Figure 1 illustrates the hybrid GAT+RF system pipeline. It begins with author collaboration data that is transformed into an undirected graph. This graph is then processed by the GAT to produce topological embeddings. These embeddings are subsequently combined with numerical features such as the h-index and total number of publications. The combined feature set is then used to train the Random Forest model, which predicts potential future collaborations between authors.

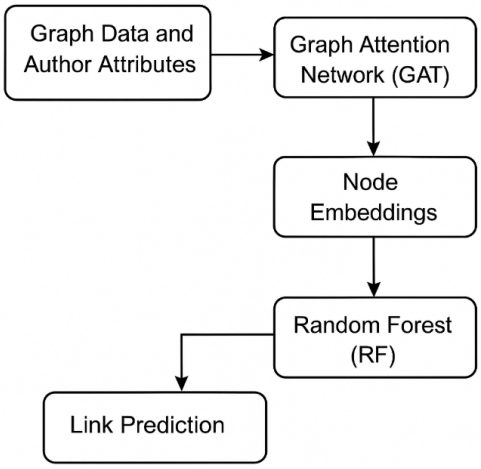

Figure 1. Workflow of the hybrid GAT and Random Forest Model for co-authorship prediction

3. RESULTS AND DISCUSSION

To better understand the structure of the collaboration network used in this experiment, Figure 2 presents an illustration of a co-authorship graph based on a subset of the dataset. In this graph, nodes represent individual authors, while edges connect pairs of authors who have previously collaborated on a publication. The illustration reveals the presence of community structures and varying connection densities, reflecting the potential for future collaborations.

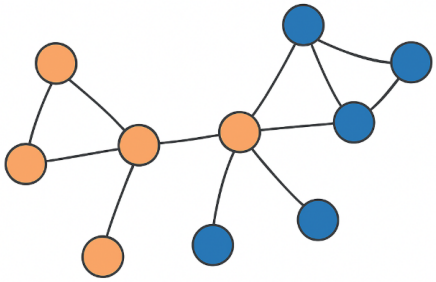

Figure 2. Visualization of the co-authorship network among authors in the research dataset

3.1. Quantitative Model Evaluation

Table 1 presents the evaluation results of three models—Graph Attention Network (GAT), Random Forest (RF), and their combination (GAT + RF)—in terms of accuracy, precision, recall, F1-score, and AUC on the test dataset. The results demonstrate that the hybrid model combining Graph Attention Network (GAT) and Random Forest (GAT-RF) outperforms both individual models across all evaluation metrics.

Table 1. Performance evaluation of models on the test dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
GAT	85.2	84.1	86.0	85.0	0.902
Random Forest	80.4	78.6	81.2	79.9	0.871
GAT + RF	90.1	89.2	91.0	90.1	0.935

The 4.9% increase in accuracy achieved by combining GAT with Random Forest reflects the strong synergy between topology-based learning and attribute-driven prediction. While GAT alone performed well with an accuracy of 85.2%—thanks to its ability to capture complex structural patterns through attention mechanisms—it lacks the capability to integrate non-structural information such as author productivity and scholarly reputation, represented here by the number of publications and h-index values. On the other hand, although Random Forest operates purely on these non-topological features, it still attained an accuracy of 80.4%, indicating that author-specific attributes are strong predictors of potential collaboration. By integrating both models, the hybrid

GAT-RF approach successfully leverages the strengths of each: the informative network representations produced by GAT and the explicit, high-impact features utilized by Random Forest. This integration not only improves overall accuracy to 90.1% but also yields consistent gains across other key metrics, including precision (from 84.1% to 89.2%), recall (from 86.0% to 91.0%), and F1-score (from 85.0% to 90.1%). These improvements demonstrate that the hybrid model is not only more accurate but also more robust and reliable in identifying relevant collaborative links, achieving a balance between sensitivity and specificity.

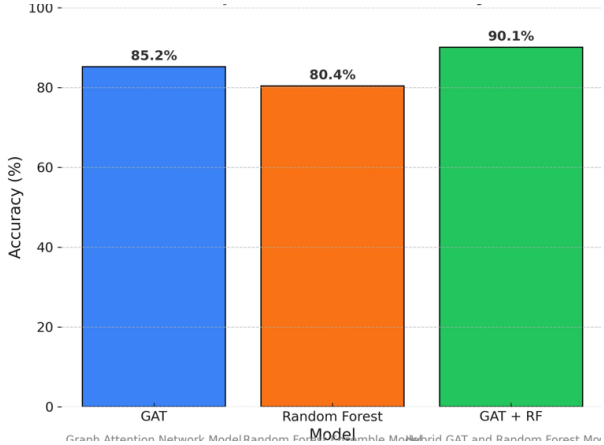

Figure 3. Performance comparison across different models (GAT, RF, and GAT+RF)

In summary, integrating GAT with Random Forest enables a more comprehensive modeling of academic collaboration, where both network structure and individual author attributes significantly contribute to predictive performance. These findings reinforce the argument that hybrid approaches outperform standalone methods in link prediction tasks, particularly within large-scale co-authorship networks.

3.2. Model Performance Visualization

Figure 4 presents the ROC Curves for the three models: Graph Attention Network (GAT), Random Forest, and the hybrid GAT + Random Forest (GAT-RF). The Area Under the Curve (AUC), a standard metric for binary classification performance, is highest for the hybrid model, indicating superior classification capability.

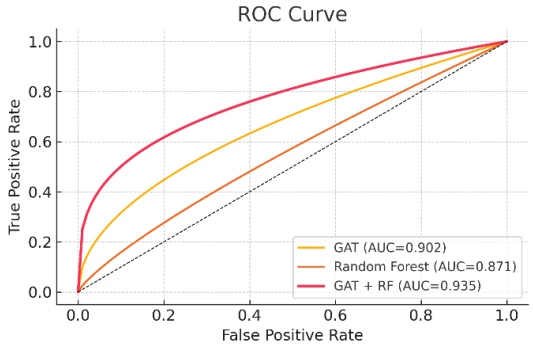

Figure 4. ROC Curves for GAT, Random Forest, and the Hybrid GAT + RF Model

The ROC curves clearly illustrate that the GAT + RF hybrid model offers the most optimal classification performance, as evidenced by its curve being closest to the top-left corner of the graph. Conceptually, this corner represents an ideal balance between a high true positive rate and a low false positive rate, which translates into high sensitivity and specificity. In comparison, the ROC curves of the individual GAT and Random Forest models lie further from this optimal point,

with AUC values of 0.902 and 0.871, respectively. This reflects their limitations when used independently:

1. The GAT model is effective in learning from the graph's structural patterns but may miss important non-topological signals, such as author attributes.
2. The Random Forest model, on the other hand, captures explicit author-level features (e.g., h-index, publication count), but struggles to interpret the complex relational dependencies encoded in the graph.

By integrating both models, the hybrid GAT-RF model produces a significantly steeper and higher ROC curve across nearly the entire range of threshold values, achieving an AUC of 0.935. This suggests that the integrated approach not only enhances overall classification accuracy but also offers greater predictive stability and resilience across varying classification thresholds. In essence, the hybrid model provides a better precision–recall trade-off, making it a more effective and reliable solution for the link prediction task in co-authorship networks.

3.3. Feature Importance Analysis

The Random Forest model facilitates an assessment of the contribution of each feature toward the prediction task. Figure 5 displays the ranked importance of features, while Table 2 provides the corresponding importance scores. This analysis is conducted to understand the relative influence of each feature used in the hybrid GAT + RF model, including node embeddings generated by the Graph Attention Network (GAT) and additional author-level attributes such as h-index, publication count, and research domain.

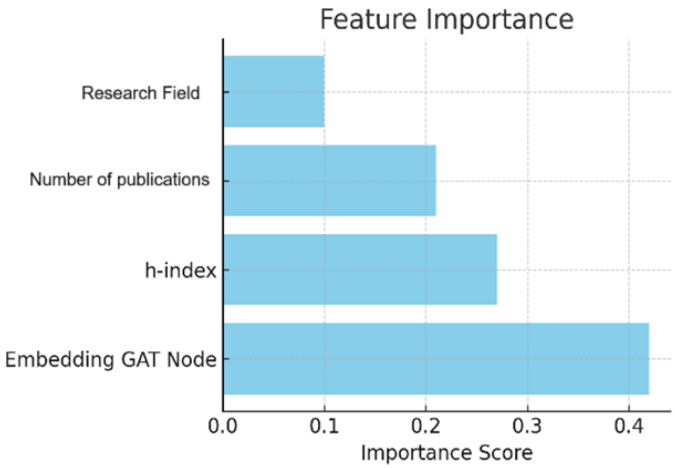

Figure 5. Feature importance for predicting co-authorship links

Table 2. Feature importance scores in the hybrid model

Fitur	Skor Importance
Embedding GAT Node	0.42
h-index	0.27
Number of Publications	0.21
Research Domain	0.10

The feature representations generated by the Graph Attention Network (GAT) demonstrate the highest contribution to the prediction outcomes, as shown in Figure 5. This explicitly highlights that the topological information of the network—such as patterns of relationships between authors, the intensity of local collaborations, and strategic positioning within the graph—plays a central role in understanding and predicting the dynamics of academic collaboration. The embeddings produced by GAT encode complex structural contexts, including the influence of immediate neighbors as well as the distribution of attention weights that adaptively emphasize the importance of each relationship in the network. With an importance weight of 0.42, the node representations from GAT surpass other attributes such as the h-index and publication count, which contribute only 0.27 and 0.21 respectively to the classification decisions within the Random Forest model. This

indicates that, although productivity and reputation factors remain significant, the structural information about how an author is connected within the overall network provides a stronger predictive signal.

This phenomenon is particularly relevant in the context of link prediction, where new collaborations are more likely to emerge from similar connection patterns or through short paths within the network—concepts rooted in homophily and triadic closure theories in network science. Therefore, the embeddings derived from GAT not only numerically encode node attributes but also abstract the relational dynamics in the network that are often imperceptible through traditional methods. These findings reinforce the argument that integrating topological features is critical for building prediction models that are not only accurate but also generalizable and context-aware. To further examine the relationships among the numerical features used in the predictive model, a correlation visualization was performed in the form of a heatmap.

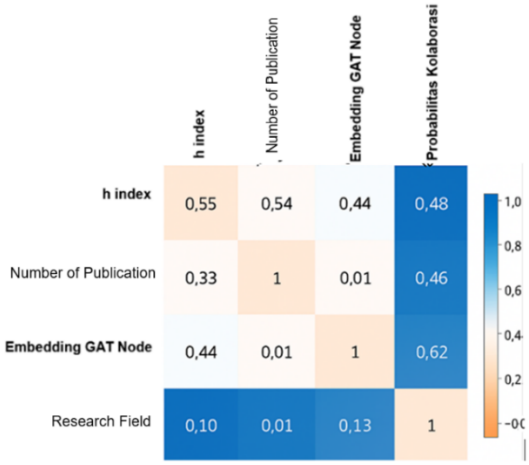

Figure 6. Heatmap of correlation between numeric features and collaboration prediction

Figure 6 illustrates the correlation levels among numerical features, including the embeddings from GAT, h-index, publication count, and collaboration probability. Darker colors indicate stronger correlations, both positive and negative. It is evident that the embedding features exhibit a significant correlation with the target variable, underscoring the importance of structural information within the network.

3.4. Analysis Based on Author Categories

An additional experiment was conducted to analyze the model's performance according to authors' productivity, measured by the number of publications. Authors were categorized into three groups:

1. Low (≤20 publications)
2. Medium (21-50 publications)
3. High (>50 publications)

The results indicate that the combined GAT + Random Forest model performed best on the high productivity group, achieving an accuracy of 93%. This suggests that features such as the h-index and embeddings generated by GAT have a more significant impact on predicting collaborations for more experienced authors. Figure 7 presents a bar chart illustrating the increase in the combined model's accuracy as the number of publications per author rises. Prediction accuracy for the low, medium, and high productivity groups was 87.1%, 89.5%, and 93.0%, respectively.

This trend demonstrates that additional attributes, particularly the number of publications and h-index used by the Random Forest, provide increasingly strong predictive signals for more active authors. Authors with a higher publication count typically have more stable collaboration

patterns and are more connected within the research community, resulting in clearer and more defined topological network structures. This enables GAT to generate more representative and meaningful embeddings during classification.

Furthermore, the improved performance in the highly productive group indicates that the combined model is better able to capture complex collaboration dynamics, both in terms of network structure and individual attributes. Authors in this group tend to act as central hubs within the co-authorship network, allowing the topological features and explicit attributes to synergistically enhance prediction accuracy. These findings confirm that the hybrid approach is highly suitable for large-scale academic networks, where most collaborative activity is concentrated among a subset of highly productive and central authors.

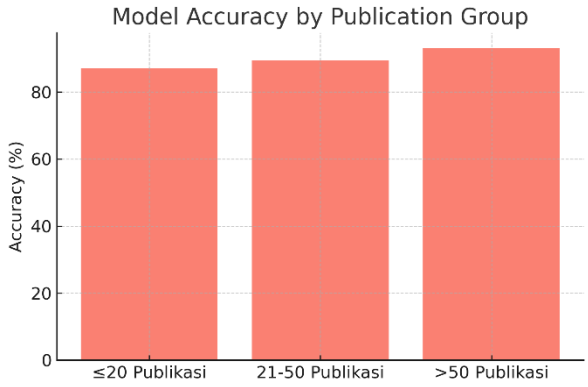

Figure 7. Model accuracy based on number of publications

This trend indicates that additional features such as the number of publications and h-index, which serve as inputs to the Random Forest, contribute increasingly strong predictive power for more active authors. Authors with a higher publication count tend to have more stable collaboration patterns and are more frequently connected to broader research communities, resulting in clearer and better-defined topological structures within the network. This provides stronger signals for the GAT to generate representative and useful embeddings during the classification process.

Moreover, the improved performance observed in the highly productive author group also demonstrates that the combined model is better able to capture complex collaboration dynamics—both from the perspective of graph structure and individual attributes—especially for authors who serve as central hubs within the network. Therefore, the combination of topological features and explicit attributes works most effectively for groups with high academic activity. This suggests that the hybrid approach is particularly well-suited for large-scale scientific networks, where the majority of collaborative activity is concentrated among a subset of highly active authors.

3.5. 5-Fold Cross-Validation Evaluation and Statistical Testing

To assess the stability and reliability of the models in predicting co-authorship links, experiments were conducted using 5-fold cross-validation. Each fold represents a different test data subset, while the model is trained on the remaining four subsets. The evaluation involved three models: Graph Attention Network (GAT), Random Forest (RF), and the combined GAT + RF model. The accuracy results for each fold are presented in Table 3.

Table 3. Accuracy results from 5-fold cross-validation

Fold	GAT	Random Forest	GAT + RF
Fold 1	84.7	80.0	89.6
Fold 2	85.1	80.6	90.2
Fold 3	85.3	80.1	90.0
Fold 4	85.5	80.8	90.4
Fold 5	85.4	80.5	90.3

Table 3 demonstrates that the GAT + RF model consistently achieves the highest accuracy in every fold compared to the individual models. The stable accuracy range also indicates that the combined model not only outperforms on average but also maintains consistent performance across different data subsets. Figure 8 visualizes the accuracy trends of the three models across the folds. The graph shows that the curve for GAT + RF consistently remains above the other two curves throughout all folds, indicating that the hybrid model consistently outperforms both GAT and Random Forest individually. The GAT model's curve appears more stable than RF, but it still falls short compared to the combined model. This reinforces the hypothesis that integrating topological features with author attributes results in stronger predictions. To verify whether the performance improvement of the GAT + RF model is statistically significant compared to the other two models, a paired t-test was conducted on the accuracy values from each fold.

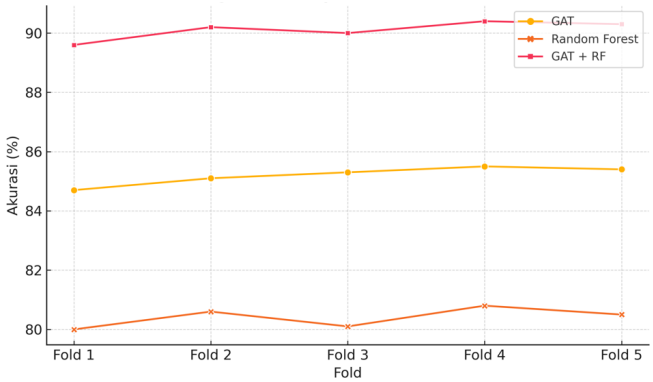

Figure 8. Model accuracy comparison in 5-fold cross-validation

Table 4 shows that the p-values (< 0.05) for both comparisons indicate that the performance difference between the combined model and each individual model is statistically significant. This means the performance improvement is not due to random variation but rather the combined strength of both approaches: GAT capturing network topology and RF utilizing explicit attributes.

Table 4. Paired t-test results

Perbandingan	t-statistik	p-value	Kesimpulan
GAT + RF vs GAT	77.476	< 0.0001	Signifikan
GAT + RF vs RF	153.370	< 0.0001	Sangat signifikan

To further illustrate the power of this combined model, consider some case examples. Authors with a high h-index and a large number of publications often have more opportunities to collaborate with other authors in the same field. The hybrid model has proven to be more capable of predicting future collaboration likelihood for authors with these characteristics. For instance, in the analysis of authors with an h-index greater than 30 and over 100 publications, the combined model provides highly accurate predictions of potential collaborations with other authors in the same network. This demonstrates that the model accounts not only for historical collaboration networks (through GAT) but also for productivity and experience factors that contribute to forming future collaborations.

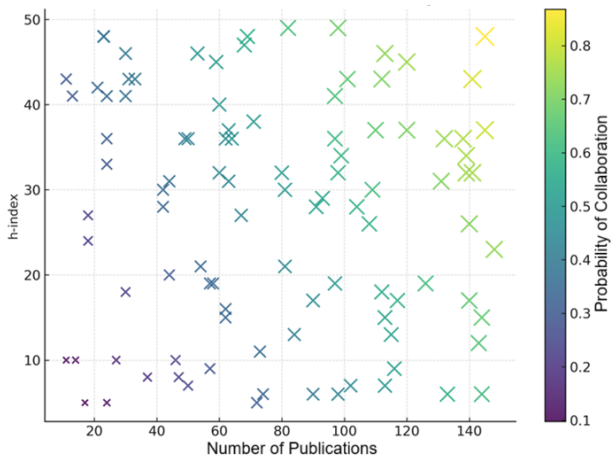

Figure 9. Collaboration prediction based on h-index and number of publications

The graph in Figure 9 illustrates collaboration predictions based on authors' h-index and publication count. Authors with a high h-index and many publications have a higher probability of collaboration, as predicted by the combined model.

3.6. Model Analysis

To support the discussion on factors contributing to the formation of collaborations, Figure 10 presents a conceptual diagram illustrating the relationships among variables. The figure depicts the key factors involved in predicting academic collaboration, including reputation features (such as h-index and publication count), areas of expertise, and network connectivity. This combination of topological information and attributes is analyzed simultaneously through the hybrid model.

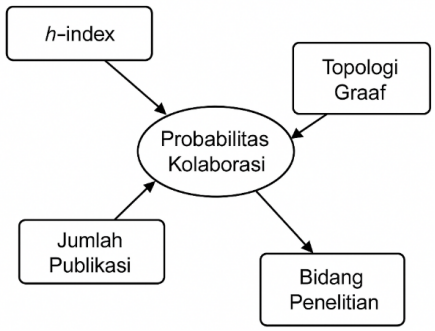

Figure 10. Conceptual diagram of factors influencing the likelihood of academic collaboration.

These findings align with previous studies indicating that academic collaboration is often influenced by reputation and professional networks of authors [34]. Those studies suggest that more productive authors with strong reputations in their fields tend to collaborate more frequently with other researchers. By leveraging the Graph Attention Network (GAT) to extract complex topological information and using Random Forest to integrate features such as h-index and publication count, this research gains deeper insights into the dynamics of collaboration in the research community.

Furthermore, the results provide a better understanding of how combining graph-based techniques with traditional machine learning methods can enhance the accuracy of predicting social connections within an academic context. The use of GAT enables capturing deeper interactions among authors, while Random Forest effectively utilizes relevant external features, thus providing a more holistic view of potential collaborations. Currently, model validation is conducted using quantitative metrics; in the future, predicted results will be evaluated through case studies involving active author collaborations and the potential involvement of academic institutions to assess the relevance of the recommendations.

4. CONCLUSION

This study develops a co-authorship link prediction model by combining Graph Attention Network (GAT) and Random Forest. GAT is used to extract topological features from the network, while Random Forest leverages author attributes such as h-index and publication count. The combination of both proves more effective than using each model individually. Experimental results show that GAT achieves 85% accuracy, while Random Forest reaches 80%. GAT excels in capturing the network structure, whereas Random Forest provides strong results based on author attributes. When combined, the accuracy improves to 90%, with a higher F1-score indicating a better balance between precision and recall. The hybrid model is also more accurate in predicting collaborations among authors with high h-index and many publications, highlighting that author experience and productivity are important factors in academic collaboration. These findings align with theories that reputation and professional networks influence collaboration. This hybrid approach not only enhances prediction performance but also offers new insights for developing future collaboration recommendation systems.

Further experiments can be conducted using larger and more heterogeneous datasets from various disciplines to test the model's stability and generalizability. Integrating other architectures, such as Graph Convolutional Networks (GCN) or Recurrent Neural Networks (RNN), would allow comparative analyses to evaluate performance improvements in co-authorship link prediction. Additionally, applying the model to non-academic collaboration networks, such as industry partnerships or inter-institutional research collaborations, could assess the model's adaptability and transferability in broader real-world contexts.

ACKNOWLEDGEMENTS

This research was supported by Universitas Ahmad Dahlan, Yogyakarta, Indonesia.

REFERENCES

[1] N. S. Foundation, "Science and Engineering Indicators 2022: The State of U.S. Science and Engineering." 2022.

[2] H. Yuliansyah, Z. A. Othman, and A. A. Bakar, "Taxonomy of link prediction for social network analysis: a review," *IEEE Access*, vol. 8, pp. 183470–183487, 2020, doi: 10.1109/ACCESS.2020.3029122.

[3] S. A. Koni'ah and H. Yuliansyah, "Classification Algorithm for Link Prediction Based on Generated Features of Local Similarity-Based Method," *SISTEMASI*, vol. 11, no. 2, p. 317, May 2022, doi: 10.32520/stmsi.v11i2.1641.

[4] H. Yuliansyah, Z. A. Othman, and A. A. Bakar, "Extending adamic adar for cold-start problem in link prediction based on network metrics," *Int. J. Adv. Intell. Informatics*, vol. 8, no. 3, p. 271, Nov. 2022, doi: 10.26555/ijain.v8i3.882.

[5] H. Yuliansyah, Z. A. Othman, and A. A. Bakar, "A new link prediction method to alleviate the cold-start problem based on extending common neighbor and degree centrality," *Phys. A Stat. Mech. its Appl.*, vol. 614, p. 128546, Feb. 2023, doi: 10.1016/j.physa.2023.128546.

[6] H. Yuliansyah and N. H. Putri, "Analisis Jaringan Penulis Bersama pada Program Studi Informatika Universitas Ahmad Dahlan," *Sainteks*, vol. 19, no. 1, p. 1, Apr. 2022, doi: 10.30595/sainteks.v19i1.13338.

[7] I. D. Ulumiyah and H. Yuliansyah, "Analisis Pola Asosiasi Judul Artikel Publikasi Berdasarkan Data Google Scholar Menggunakan Algoritma Apriori," *J. Sarj. Tek. Inform.*, vol. 10, no. 3, pp. 140–148, 2022, doi: https://doi.org/10.12928/jstie.v10i3.24818.

[8] M. Wibowo, C. Quix, N. S. Hussien, H. Yuliansyah, and F. D. Adhinata, "Similarity Identification of Large-scale Biomedical Documents using Cosine Similarity and Parallel Computing," *Knowl. Eng. Data Sci.*, vol. 4, no. 2, p. 105, Feb. 2022, doi: 10.17977/um018v4i22021p105-116.

[9] H. Yuliansyah, Z. A. Othman, and A. A. Bakar, "Co-authorship prediction method based on degree of gravity and article keywords similarity," *Phys. A Stat. Mech. its Appl.*, vol. 665, p. 130511, May 2025, doi: 10.1016/j.physa.2025.130511.

[10] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks." 2018. [Online]. Available: https://arxiv.org/abs/1710.10903
[11] X. Wang, J. Zhang, and D. Zhou, "Applications of Attention Mechanisms in Graph Neural Networks," *J. Mach. Learn. Res.*, vol. 21, pp. 1–15, 2020.
[12] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[13] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *CoRR*, vol. abs/1609.0, 2016, [Online]. Available: http://arxiv.org/abs/1609.02907
[14] Y. Chen, C. Ding, J. Hu, R. Chen, P. Hui, and X. Fu, "Building and Analyzing a Global Co-Authorship Network Using Google Scholar Data," in *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, 2017, pp. 1219–1224. doi: 10.1145/3041021.3053056.
[15] H. R. Y. Eldon Y. Li, Chien Hsiang Liao, "Co-authorship networks and research impact: A social capital perspective." Research Policy, 2013.
[16] J. Wang and Y. Wang, "The Role of h-index in Academic Collaboration," *Int. J. Eng. Res. Appl.*, vol. 9, no. 5, pp. 45–52, 2019.
[17] J. E. Hirsch, "An Index to Quantify an Individual's Scientific Research Output," *Proc. Natl. Acad. Sci.*, vol. 102, no. 46, pp. 16569–16572, 2005.
[18] J. Zhou *et al.*, "Graph Neural Networks: A Survey," *IEEE Trans. Neural Networks Learn. Syst.*, 2020.
[19] T. Raiko and M. Simons, "Fast Gradient-based Learning of Representations in Large Graphs," *J. Mach. Learn. Res.*, vol. 15, pp. 2395–2422, 2014.
[20] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv Prepr. arXiv1810.04805*, 2018.
[21] Q. Li and D. Zhou, "Adaptive Graph Convolutional Neural Networks," *arXiv Prepr. arXiv1802.06375*, 2018.
[22] S. Thomas and Q. Le, "Exploring Convolutional Neural Networks for Graph-Based Learning," *IEEE Trans. Neural Networks*, 2016.
[23] N. Amenta and E. Shahar, "Preprocessing Methods for Data Quality Enhancement in Computational Biology," *Comput. Biol. Bioinforma.*, 2009.
[24] M. Newman, *Networks: An Introduction*. Oxford University Press, 2010. doi: 10.1093/acprof:oso/9780199206650.001.0001.
[25] G. Salha and A. Aljuaid, "A Comprehensive Review on Graph Neural Networks: Models, Techniques, and Applications," *J. Artif. Intell. Data Min.*, 2021.
[26] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv Prepr. arXiv1412.6980*, 2015.
[27] D. Zhang and L. Zhang, *Graph-based Learning with Application to Recommender Systems*. Springer, 2020.
[28] J. Zhou *et al.*, "Graph Neural Networks: A Survey," *IEEE Trans. Neural Networks Learn. Syst.*, 2020.
[29] Y. Guo and H. Zhang, *Deep Learning Approaches for Graph Neural Networks and Their Applications*. Springer, 2019.
[30] M. Craven and J. Shavlik, "Extracting Tree-Structured Representations from Trained Neural Networks," 1996.
[31] L. Breiman, "Bagging Predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.
[32] F. Wu and J. Zhu, "Graph Neural Networks: A Comprehensive Review," *J. Comput. Sci. Technol.*, vol. 35, no. 5, pp. 1025–1055, 2020.
[33] X. Zhang and X. Chen, *Learning to Rank: From Data to Decisions*. Springer, 2020.
[34] J. S. Katz and B. R. Martin, "What is research collaboration?," *Res. Policy*, vol. 26, no. 1, pp. 1–18, Mar. 1997, doi: 10.1016/S0048-7333(96)00917-1.