ZANCO Journal of Pure and Applied Sciences

TYPE Original article
DOI http://dx.doi.org/10.21271/ZJPAS.37.5.13

OPEN ACCESS

*Corresponding author

Hajar Ali Hasin
hajar.hasin@uod.edu.krd

RECEIVED :19 /11 /2024
ACCEPTED :02/06/ 2025
PUBLISHED :31/ 10/ 2025

KEYWORDS:
Link Prediction; Co-authorship Networks; Social Network Analysis; Graph Mining; Text Mining

LINK PREDICTION BASED ON TOPOLOGICAL AND CONTENT ANALYSIS IN CO-AUTHORSHIP NETWORKS

Hajar A. Hasin* ,Diman Hassanand Ismael A. Ali

Department of Computer Science, Faculty of Science, University of Zakho, Zakho, Iraq

*

ABSTRACT

In network analysis, the prediction of the connections or associations between entities or nodes within the network becomes important. Link Prediction is the problem of predicting or identifying the existence of a link between two entities in a network. However, it still the main issue in the complex network data application field, particularly in the type of analysis related to co-authorship networks despite its wide usage. Topological methods and content-based methods are the two different approaches that have been proposed for the link prediction in collaboration networks. However, topological methods are based on the structural analysis of the network, and content-based approaches rely on textual information from academic papers in the network. In this paper, we introduce the Content and Graph-Based Link Prediction (CGLP) approach, which integrates topological and content-based features from networks in a hybrid manner for predicting links in co-authorship networks. The efficacy of the proposed approach was already tested using three academic datasets: Hep-th, Hep-lat, and AMC by applying various machine learning models. Results indicated that all models showed almost the same efficiency on all three datasets and outperformed the state-of-the-art approach with a maximum F1 score of 98.05% and ROC AUC of 98.74%.

1.Introduction

A social network (SN) may be defined as a finite number of nodes with several kinds of relations between them. The interactions in a social network represent the particular characteristics of the nodes and the interaction properties of their respective links, i.e., engaging in cooperative activities, working on projects, or coordinating scientific efforts (Bergmeir et al., 2018). A good example is a community of researchers collaborating to transcend multidisciplinary, intricate issues. Research quality and novelty can be improved through reciprocal exchange of innovative ideas, exchange of knowledge, and interaction among academics (Kong et al., 2019). An extremely good example of a social network is a co-authorship network, in which two authors get together with the objective of composing a scientific paper. For co-authorship networks, the authors of a manuscript are the nodes, and co-authorship or publication order between two different nodes that have collaborated to produce a co-authored manuscript is the edge.

A SN is composed of several points, known as nodes, and various kinds of links between them. The links are usually caused by specific attributes of the nodes and their interactions, e.g., cooperating, collaborating on projects, or attending scientific conferences (Wang et al., 2014). For example, studying a community of scientists collaborating on problems that cover many various areas. Research quality and innovation rise when there is exchange of knowledge, sharing of new ideas, and collaboration among researchers through the study of these collaboration (Kong et al., 2019).

A further compelling illustration of a social network is the co-authorship network. This network is an excellent example of the relationship between authors since they work together to come up with a research paper. An author is denoted by a node, and the line denotes the relationship between two authors who have co-authored a paper. Various symbols can be used to denote information regarding the frequency of their co-authorship or the year when their papers were accepted. A feature vector can be constructed for all research papers in order to discover similarities among titles and abstracts. By doing this, it is easy to predict potential future collaborations among authors. It is a challenging problem to predict the co-authorship relationship among authors and thus such a prediction task is referred to as LP.

To view the LP problem, we need to look at a provided graph or network referred to as $G=(V,E)$ at time t. G here is a graph or network made up of authors connected together. Here, V is a collection of points or nodes that are tantamount to scholarly authors, and E is a collection of lines or links at time t that represent relationships or links between the authors in network G. The goal of the link prediction (LP) task is to predict potential links that can exist between time steps t and $t+1$, with the condition that t must be smaller than $t+1$ (Yuliansyah et al., 2020, Zhang and Chen, 2018).

Based on that, pairs of nodes connections will be generated at time t to $t+1$ for the prediction of missing links or future links that might represent the future collaboration in the network G (Newman, 2004). **Fig**.1 theoretically shows the problem of LP concerning the parameter of time. Most of the LP literature has relied on basic and simple analysis metrics (Affonso et al., 2022, Daud et al., 2020, Hasin and Hassan, 2022, Razzaq et al., 2022, Resce et al., 2022). The literature has successfully proposed LP approaches based on the topological methods and content-based methods to predict the future links in the co-authorship networks.

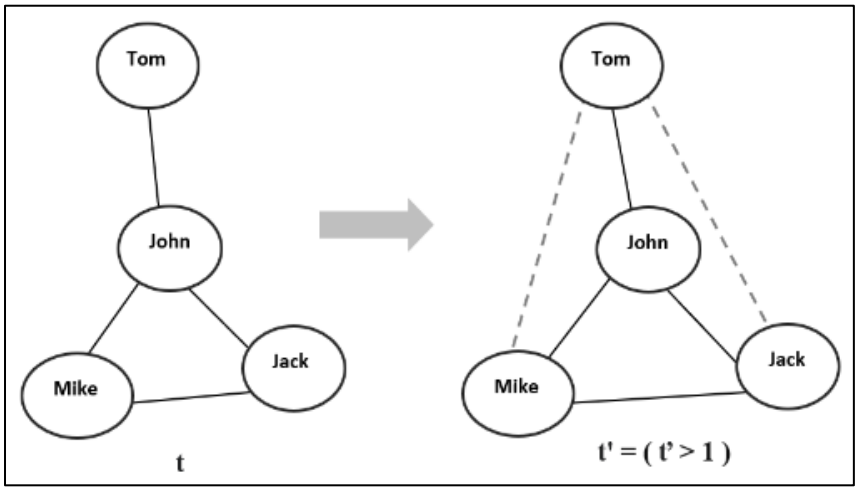

Fig.1 Graphical representation of the LP problem; the dotted lines represent the predicted links at time (t').

Topological methods mean the understanding of the topological or the structural properties of the graph and the relationship between the connected nodes using mathematical concepts from the network structure (Chen et al., 2021, Nasiri et al., 2021). On the other hand, the content-based methods rely on attributes, the content, and the textual information about the nodes in the network (Do et al., 2019, Lande et al., 2020).

The main property of the topological-based approaches is the simplicity and the ease of achieving score ranking for each unobserved pair of nodes with a high computational time (Kumar et al., 2020b). Whereas, the content-based approaches are considered effective only when rich contextual information is available otherwise it considered less accurate because content might contain noisy and irrelevant data (Lü and Zhou, 2011, Kumari et al., 2022). Moreover, the content-based approaches require deep understanding of the domain specified for the problem.

Various content-based methods have been proposed in the literature specifically by (Hassan, 2019, Quercia et al., 2012, Sachan and Ichise, 2010). The shortcomings in both topological and content-based methods have inspired this study, along with other such research efforts in the domain, to explore hybrid methods that can exploit the merits of both methods to further enhance link prediction in co-authorship networks. In (Antunes et al.), Antunes et al. proposed a hybrid linear programming approach named ConPredict that integrates structural network patterns and contextual information derived from nodes, i.e., the titles and abstracts of scientific papers authored by scientists. To implement this approach, the authors employed a collection of metrics such as the Shortest Path (SP) distance between nodes, the Jaccard distance, and node similarity according to the frequency of word use within the nodes. The authors asserted that ConPredict performed better than the techniques they had previously employed with an F1 score of 86.02%. (Chuan et al., 2018) proposed a novel hybrid approach, namely LDAcosin, to predict author-author relationships among three different co-authorship networks: AMC, Hep-th, and Hep-lat. It was based on analyzing structural and content-based similarities of pairs of research articles authored by the same two authors in the specified individual networks. The LDAcosin approach employs a content similarity measuring method called Latent Dirichlet Allocation (LDA) that is reinforced by various weighted topological indices such as weighted common neighbors, weighted Jaccard coefficients, and weighted Adamic-Adar indices. To achieve its predictive functionality, LDAcosin engagements a weighted binary classification technique, specifically a weighted Support Vector Machine (SVM). The LDAcosin's performance was evaluated through experimentation on the three real-world co-authorship networks presented above. The results confirmed that LDAcosin outperformed content- or topology-based baseline methods

with an accuracy of 32.50% in F1 score and 66.26% in the area under the ROC curve.

In (Wu et al., 2021), Wu et al. presented an LP approach named LP-UIT (link prediction based on user information and topology). The approach used textual information from each node representing an author. The proposed approach's information representing the short-term and long-term interests of each user, graph information, social influence, and numerical information for the designing a multimodal framework for link prediction. A graph convolutional network (GCN) was used to represent the graph structure and a multilayer perceptron (MLP) model is used to tune the three types of features (textual, topological, and numerical) for link prediction. Two real world social networks were used to evaluate the effectiveness of the proposed method named as Zhihu and Epinions datasets. The results demonstrated that the LP-UIT achieved an AUC of 95.16%.

In this study and based on the reviewing of the literature and their results, the goal of this paper is to design a new hybrid approach called content and graph-based link prediction (CGLP) for link prediction in three real world co-authorship networks Hep-th, Hep-lat, and AMC to improve the overall accuracy of the prediction. Our proposed approach combines leverages both topological and content-based approaches using different attributes. The topological features are extracted from the network structure and academic links among authors, whereas the content-based features extract information from the content of the paper shared between two authors such as the research topics abstract, and interests shared among the authors. To evaluate the efficacy of the generated feature vectors in predicting future links within the input graph datasets, four machine learning (ML) algorithms are employed: K-nearest neighbors (KNNs), decision trees (DTs), random forests (RFs), and support vector machines (SVMs).

Using this proposed hybrid approach, it will be able to synthesizes the interpretability and explicit feature manipulation capabilities of traditional topological methods with the semantic richness of content-based techniques. In contrasts with graph-based deep learning paradigms such as Graph Neural Networks (GNNs), which present scalability limitations, while demonstrating proficiency in automatically deriving complex relational patterns through learned feature representations and propagation mechanisms (Zhang and Chen, 2018). Wu et al. (2021) in (Wu et al., 2020) noted that iterative message-passing operations in GNNs impose significant computational overhead, restricting their effectiveness in sparse or moderately sized network scenarios. Furthermore, GNN-derived representations remain comparatively opaque (Kong et al., 2019), complicating interpretability of decision-making processes.

Our proposed hybrid methodology presented in this paper addresses these limitations by strategically balancing computational efficiency and transparency. Its architecture leverages domain-specific feature engineering to enhance interpretability while retaining semantic depth, positioning it as particularly advantageous for moderate-scale datasets. This dual emphasis on explicit feature analysis and algorithmic transparency makes the framework especially suitable for analytical contexts where theoretical rigor and reproducibility are prioritized over black-box optimization.

The rest of the paper is organized as follows: Section 2 presents the materials and methods used in the proposed approach. The experimental results along with a discussion are presented in Section 3 and Section 4 concludes the paper and suggests some future research directions.

2.Datasets and Approach

2.1 Dataset Description

Three different co-authorship datasets are used in this study: Hep-th and Hep-lat documents ranged between the years 2003--2009 from the high-energy physics papers and AMC dataset from the journal of applied mathematics and computation papers, between 2008--2014[1]

1 *Hep-th* dataset from Cornell University, arxiv.org/archive/Hep-th/. Accessed on 17/11/2023.
Hep-lat dataset from Cornell University, arxiv.org/archive/Hep-lat/. Accessed on 17/11/2023.
AMC dataset from the Journal of Applied Mathematics and Computation. Accessed on 17/11/2023.

(Chuan et al., 2018). The datasets are useful in LP problems applying content-based approaches since they contain textual information such as author names, topics, abstracts, and other related information. Other information is also available in the first two datasets such as the number of journals and the citations in which the papers are exist. In this work, the non-relevant features and information regard Hep-th and Hep-lat datasets were eliminated. Table 1 presents the statistics of the graphs before analyzing the data. **Figs**. 2, 3, and 4 visualize the graphs for the AMC, Hep-lat, and Hep-th networks, respectively.

Table 1 Hep-th, Hep-lat and AMC network statistics before preprocessing analyses

	Hep-th	Hep-lat	AMC
Papers	20,760	3,952	4,922
Authors	15664	3356	9346
Collaborations	33413	10918	12189
Graph Density	0.001	0.004	0.002
Avg. degree	4.266215 52604	6.50655 54231	2.6083 88615

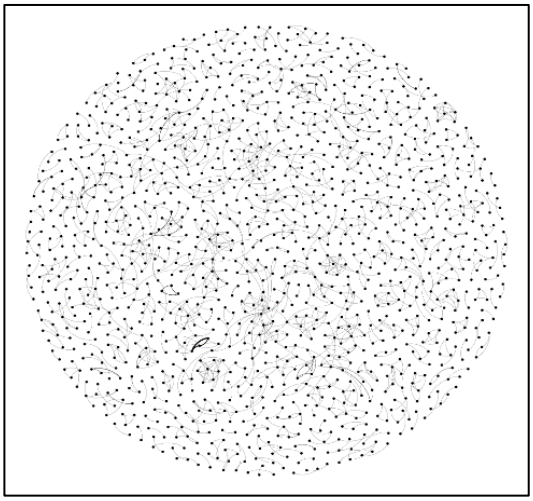

Fig.2 Graph visualizations for AMC datasets before preprocessing

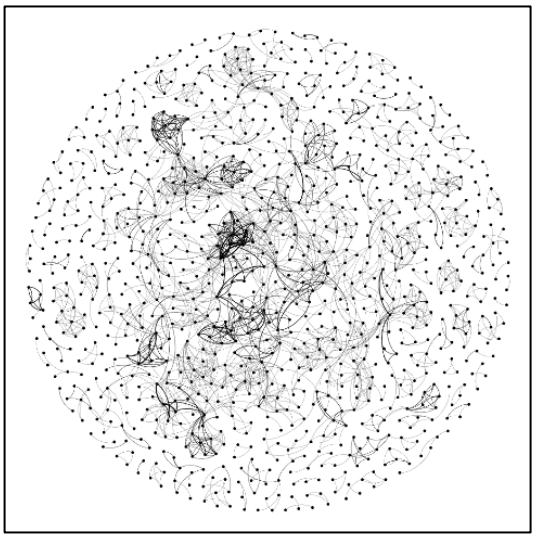

Fig.3 Graph visualizations for the Hep-lat dataset before preprocessing

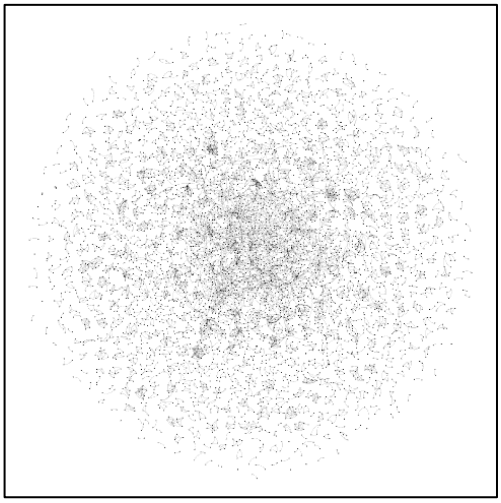

Fig.4 Graph visualizations for the Hep-th dataset before preprocessing

2.2 Proposed Approach

In this section, the general structure of the suggested link prediction approach CGLP is shown in **Fig**. 5. The first stage of the pipeline is the pre-processing where in this phase the raw data from the selected datasets are cleaned, purified and then all articles which are irrelevant or of low quality are removed from the data before it can be used for further analysis. In the analysis phase two major procedures are performed: content analysis and topological analysis. Content analysis is conducted by topic modeling using the LDA method to discover the topics or themes in the articles. At the same time, topological analysis is performed by graph

mining to understand the structure of the network and to assess the key topological measures of academic collaborations, such as node degree, shortest path, and common neighbors. The data transformation step is then performed through feature extraction to extract essential attributes from the data for training the selected models in this work. During the training and modeling steps, machine learning models are built to predict future connections on the basis of these features. Finally, in the evaluation phase, the model's link prediction accuracy is tested and validated through various metrics, such as the F1 score and ROC AUC, to ensure reliability and performance and to evaluate the effectiveness of the CGLP approach in predicting future links within the selected co-authorship networks.

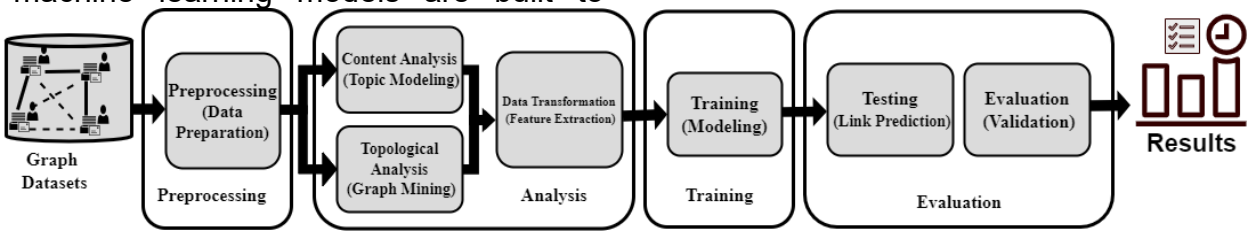

Fig.5 General architecture of the CGLP link prediction approach

2.3 Preprocessing stage

In this stage, the datasets are preprocessed using the same procedures. The procedure started by converting the data into a graph shape, where each node represents an author, and the links between each two nodes represent the real collaboration between each two authors of a paper in the network. Due to the large size of the created graphs which causes the graphs to face the problem of high sparsity, we first reduced the size of each graph by removing subgraphs and collaboration communities with less than three nodes. This step is commonly employed to assist in removing noise and in turn reducing overfitting (Schizas, 2018, Yin et al., 2023). Furthermore, the link prediction problem needs denser graphs to obtain better prediction. For this reason, we considered all pairs of unconnected authors located in disjoint collaboration groups/subgraphs via the use of SP and common neighbor metrics. The papers' textual data were integrated with the graph representation data by the author ID numbers provided in the original datasets. After this integration, the data was cleaned up in preparation for topic modelling using the LDA algorithm (Campbell et al., 2015). After the preprocessing, the network statistics the network statistics shown in Table 2. **Figs**. 6, 7, and 8 display the graphs representing the AMC, Hep-lat, and Hep-th datasets, respectively.

Table 2 Hep-th, Hep-lat and AMC network statistics after preprocessing analysis

	Hep-th	Hep-lat	AMC
Papers	20,760	3,952	4,922
Authors	13564	3046	7039
Collaborations	42106300	2261614	6145746
Graph Density	0.002	0.010	0.005
Avg. degree	6208.5373046	1484.9730794	1746.19860775

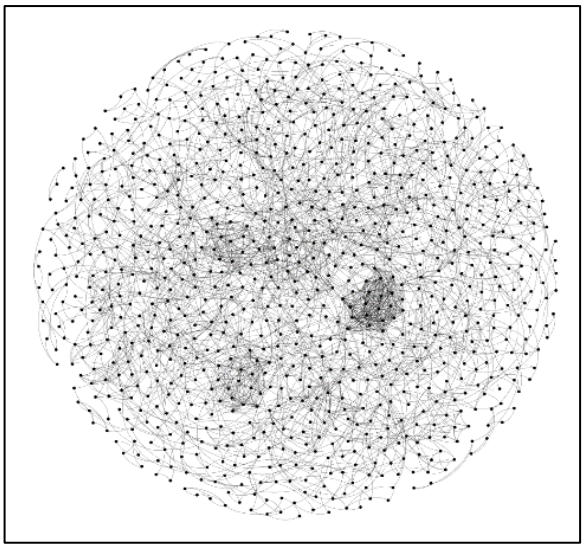

Fig.6 Graph visualizations for AMC datasets after preprocessing and reduction.

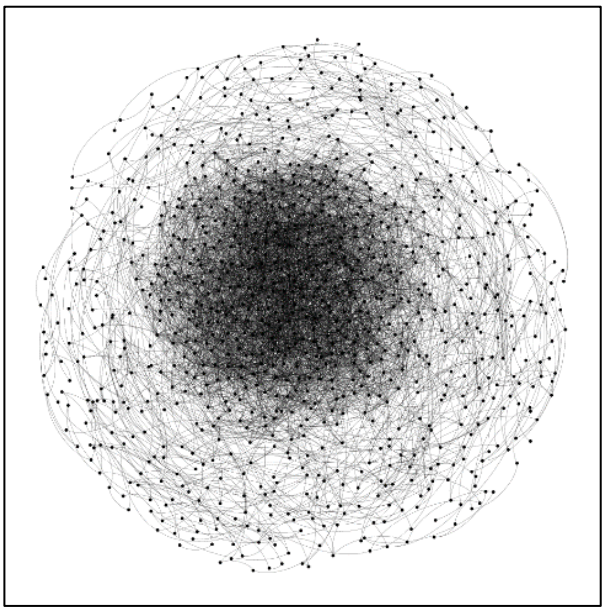

Fig.7 Graph visualizations for the Hep-lat dataset after preprocessing and reduction

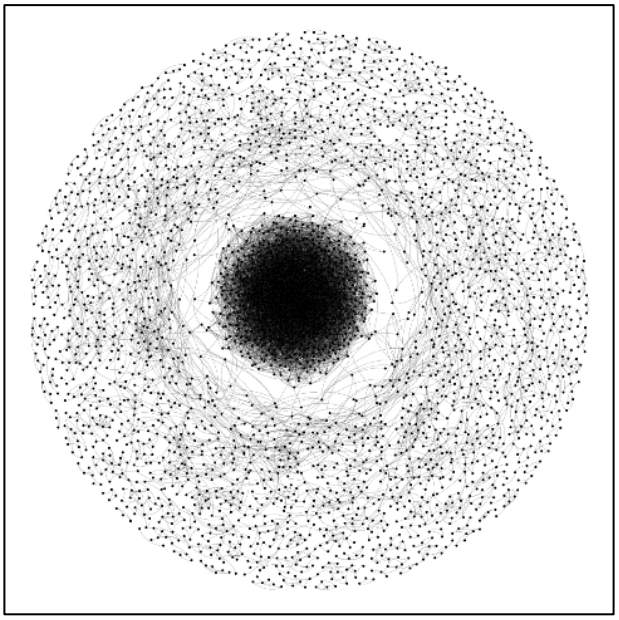

Fig.8 Graph visualizations for the Hep-th dataset after preprocessing and reduction

A workflow of the preprocessing stage is presented as shown in Fig. 9. It begins with data preparation and graph creation, followed by reading edges and creating the initial graph structure. We then process the graph by loading edges and filtering out components with fewer than 3 nodes, ensuring the retention of only meaningful and connected subgraphs. Simultaneously, we process papers by loading papers data, cleaning columns, merging with edges data, and aggregating by author with text and counts. The cleaned edges data is saved as Edges_clean.csv, and the aggregated papers data is saved as Papers_agg.csv.

The workflow continues with graph analysis, where for all pairs of nodes in the graph, we check if an edge exists and calculate relevant metrics. If a path exists between nodes, we calculate the shortest path, common neighbors, and node degree metrics. If no path exists, we record this information along with the metrics. The data is then combined into DataFrames and written to CSV files. Depending on the results, we either sample data if needed or proceed to the final combination of not_connected_df or no_path_df. The workflow concludes with the end of the preprocessing algorithm.

This figure visually outlines each step of our preprocessing pipeline, from data preparation and graph creation to the final combination of data. It serves as a companion to the detailed textual description, enhancing clarity and aiding in the reproducibility of our methodology

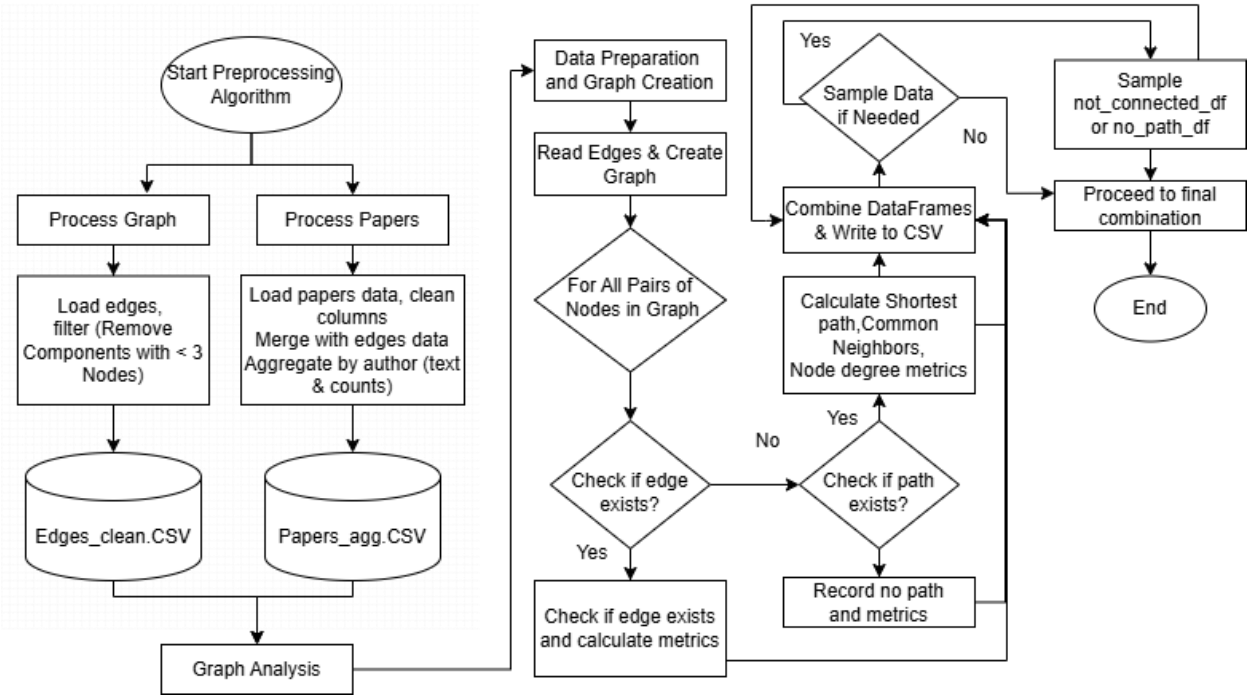

Fig.9 Workflow of the Graph-Based Preprocessing Algorithm for Data Cleaning and Feature Extraction.

2.2.1 Analysis stage

The analysis of the created graphs after preprocessing is conducted in this stage. In this stage, the selection and extraction of the needed topological and content features is accomplished to be used by ML models for training and testing. The goal of the *CGLP* is to enhance the accuracy of predicting links in the selected networks by combining different similarity measures from the topological structures of social graphs and content-based methods to extract textual information from each academic paper. The value of the link *CGLP* prediction score is either 0 or 1. The value of 1 indicates that nodes *u* and *v* have a relationship while the value of 0 denotes the missing of the link. The proposed hybrid approach *CGLP* is expressed mathematically as follows:

$$CGLP(u,v) = LDA(u,v) * \frac{(MAXdegree(u,v) + CN(u,v)}{(1 + SP(u,v))} \quad (1)$$

CGLP(u,v) denotes the predicted edge between nodes *u* and *v* while the similarity between both nodes constructed on their latent topic distributions obtained through the LDA algorithm is represented by *LDA (u, v).* *MAXdegree (u, v)* denotes the maximum number of edges that each node has and the number of shared neighbors between the nodes *u* and *v* is represented by *CN (u, v).* Also, from Eq. 1, the *SP (u, v)* denotes the shortest path between nodes *u* and *v.*

As the equation presents, in topological terms, nodes *u* and *v* have a lower likelihood of future collaborations when having a longer *shortest path* among them, and vice versa, therefore having a negative inverse relationship. On the other hand, nodes *u* and *v* are more likely to have a higher degree of common neighbors and therefore have a positive relationship.

In the content analysis stage, the features extracted from each paper's title, keywords, and abstract capture the paper's meaning and context. The LDA algorithm (Blei and Lafferty, 2007), as described below, extracts research topics from these papers, where in this work, the top 10 topics were selected per paper.

According to (Blei and Lafferty, 2007), there is a document formation process in which D is assumed to be a corpus that is a collection of documents (d) as follows:

a) A random topic that has the following distribution is chosen:
$\varphi^{(k)} \sim Dirichlet\ (\beta) \quad for\ k = 1, \ldots, K$
b) Randomly select a distribution of topics for the document (d).
$\emptyset_d \sim Dirichlet\ (\alpha) \qquad d \in D$
c) For each word in a document D
   i. Select randomly from the topic distribution a topic
   ii. $z_i \sim Dirichlet\ (\emptyset_d)$
   iii. A word is chosen randomly from the corresponding topic of the vocabulary distribution.

$w_i \sim Dirichlet\ (\emptyset^{(z_i)})$

Where K is representing the number of latent topics in *D*, $^{(k)}$ denotes the distribution of the discrete probability of the vocabulary which represents the distribution of the class-k topic. The symbol θ_d is denoting the distribution of the document *d* of the available topics, z_i represents the index of the topic of the *i* word, w_i is the word of class *i*, and α, β are the parameters of the Dirichlet distribution.

A feature vector is then created for each paper, and semantic weights are assigned to probability of the 10 top topics distributed in each paper. The created vector of the attributes helps to provide a method to feed the text data into the ML models in a straight and efficient way as well as providing a more insight of the network, accompanying the topological features by extracting the semantic features from the papers. However, before applying topic modeling via *LDA*, the text needs a preprocessing step that involves cleaning, tokenizing, and transforming. A unique word dictionary is created, and a document-term matrix is established to show word frequencies. After extracting the topics, the keywords of each topic in every document are extracted and associated to the authors of that document in the selected dataset. Then, the joint keywords are calculated using *LDA (u, v)* to obtain a score ranged between *[0, 1]* between authors. A subset of the top 10 topics by their representative keywords identified by the *LDA* algorithm for a group of authors is shown in Table 3.

Table 3 Sample of the top ten topic keywords discovered by the LDA topic modeling algorithm

Author ID	No. of Shared Papers	Papers Text	10_Keywords_From_Papers
4061	3	The physics	[' transition ', ' mass ', ' Polyakov ' , ' quark ' , ' present ' , ' matrix ' , ' topological ' , ' state ' , ' result ' , ' QCD ']
4064	1	Matrix	[' gauge ' , ' mass ' , ' SU ' , ' result ' , ' QCD ' , ' temperature ' , ' lattice ' , ' fermion ' , ' theory ' , ' quark ']
4066	2	Efficient	[' mass ' , ' result ' , ' using ' , ' decay ' , ' QCD ' , ' meson ' , ' lattice ' , ' fermion ' , ' dynamical ' , ' quark ']
4068	5	The eta '	[' Monte ' , ' mass ' , ' equation ' , ' using ' , ' Carlo ' , ' quark ' , ' model ' , ' meson ' , ' method ' , ' fermion ']
4081	1	Black hole	[' present ' , ' result ' , ' mass ' , ' gauge ' , ' SU ' , ' QCD ' , ' matrix ' , ' temperature ' , ' lattice ' , ' fermion ']
4084	1	Delta-	[' mass ' , ' result ' , ' dependence ' , ' using ' , ' decay ' , ' chiral ' , ' energy ' , ' QCD ' , ' meson ' , ' lattice "]
4085	3	Broken	[' result ' , ' dependence ' , ' chiral ' , ' energy ' , ' QCD ' , ' lattice ' , ' fermion ' , ' theory ' , ' potential ' , ' symmetry ']
4087	5	Nuclear	[' gauge ' , ' mass ' , ' SU ' , ' result ' , ' Landau ' , ' propagator ' , ' QCD ' , ' temperature ' , ' lattice ' , ' gluon ']
4089	7	Broken	[' mass ' , ' dependence ' , ' using ' , ' quark ' , ' present ' , ' energy ' , ' matrix ' , ' meson ' , ' fermion ' , ' state ']
4100	1	Current	[' transition ' , ' result ' , ' mass ' , ' chiral ' , ' QCD ' , ' chemical ' , ' potential ' , ' extended ' , ' symmetry ' , ' phase "]

In the topological analysis stage, three topological features *CN, MaxDegree* and *SP* were selected and incorporated into the CGLP hybrid approach to predict link formation in co-

authorship networks. These features are defined as follows:

1. Common Neighbors (*CNs*): The *CN* is an algorithm or similarity measure used to measure the similarity between two graph's nodes. It is a common measure used in the field of social network analysis because of its simplicity and well performance (Kumari et al., 2022). Its goal to calculate the number of shared neighbors of two input nodes. In the context of co-authorship networks, the *CN* is a useful measure for link prediction which means that if two nodes representing authors have many common neighbors, they are likely to share or publish a paper. The *CN* metric is formally defined as:

$$CN(u,v) = |\Gamma(u) \cap \Gamma(v)| \quad (2)$$

where *Γ(u)* and *Γ(v)* are the sets of neighbors of vertices *u* and *v* separately. The higher value of *CN (u, v),* the more likely that there will be a link between nodes *u* and *v*.

2. *MaxDegree*: The maximum degree is another topological measure used in this study to find the higher degree of any node in a graph (Borgs et al., 2012). In co-authorship network, it refers to the number of direct edges an author has. It also means that if an author has a collaboration with highly collaborated authors in a specified network, they are likely to have a higher degree of collaboration as well and can be considered a central node in the network. To calculate the maximum degree of any node, the number of direct neighbors to and from the node counted. Also, the maximum degree between two nodes is calculated by finding the maximum links that are allowed between them as follows:

$$MAXdegree(u,v) = max(deg(u), deg(v))$$

3. *Shortest path (SP)*: the task of this measure is to find the closest distance between two nodes within a network (Kumar et al., 2020a, Yuliansyah et al., 2020). In the context of co-authorship networks, *SP* is useful in link prediction to find the potential connections between the network's authors. In other words, *SP* means that two authors are more likely to publish a paper in the future if they have a *SP* between each other. The formula of the *SP* is presented in Eq.4 as follows

$$SP(u,v) = d(u,v) \quad (4)$$

where *d(u,v)* is representing he length of the SP between the nodes *u* and *v* in the network where the smaller value of *SP(u,v)*, the more probability of nodes *u* and *v* to have a link in the future.

By utilizing the aforementioned metrics, the function of the *CGLP* approach is described in the following algorithm. The *CGLP* takes the preprocessed graphs *G* as an input for the link prediction between the authors of the co-authorship networks forming a corpus. This predictive capability is by computing CGLP scores for each pair of vertices not currently connected in graph G using the primary link prediction equation (1).

Algorithm *CGLP* (Graph *G*): Content and Graph-Based Link Prediction Algorithm

Input: Undirected co-authorship graph *G(V, E)*, **where**

V is a set of nodes, $\forall v \in V \ni v \in \{v_1, v_2, v_3..., v_n\}$.

E is a set of links, $\forall e \in E \ni e \in \{(v_i, v_j), ..., (v_k, v_l)\}$

Output: *CGLP_dict* (a dictionary with *CGLP* scores for the predicted edges)

1. *CGLP_dict* ← an empty dictionary
2. **for each** *u* in *G.V* **do:**
3. **for each** *v* in *G.V* **do:**
4. **if** $u \neq v$ **then:**
5. $lda_{u,v} \leftarrow lda(u, v)$
6. $md_{u,v} \leftarrow max\_degree(u, v)$
7. $cn_{u,v} \leftarrow common\_neighbor(u, v)$
8. $sp_{u,v} \leftarrow shortest\_path(u, v)$
9. $CGLP_{u,v} \leftarrow lda_{u,v} * ((md_{u,v} (3) cn_{u,v})/(1 + sp_{u,v}))$
10. $CGLP\_dict\ (u, v) \leftarrow CGLP_{u,v}$
11. **end if**
12. **end for**
13. end for
14. **Return** *CGLP_dict*

The CGLP algorithm commences by traversing every ordered pair of vertices, entailing θ (n^2) operations. For each vertex pair, it performs a Latent Dirichlet Allocation step in O$(mnt + t^2)$) time, scans the adjacency list in O(Δ) time, executes a breadth-first search in O $(m + n)$ time, and undertakes constant-time bookkeeping. Consequently, the overall time complexity is O $\left(N^2(mnt + t^2 + \Delta + \mathrm{n} + \mathrm{m})\right)$. When applied to typical sparse co-authorship networks where $\left(M = O(N)\right)$ and with moderate topic-model dimensions, the time complexity reduces to O$\left(N^2(mnt + t^2 + \mathrm{N})\right)$.

To predict link formation in co-authorship networks, the final feature vector integrates both content-based features and topological features. But the combined feature vectors not only enabled accurate predictions, they also provided useful insights on the structural properties of the co-authorship networks. The features are presented in Table 4.

Table 4 Sample of topological features without normalization. U and V represent authors nodes, SP_U_V is the shortest path between U and V, CN represents the common neighbors between U and V, MAXdegree is the maximum degree of direct neighbors U and V, and LP_U_V is the predicted link probability between U and V.

U	V	Sp_U_V	CN_U_V	MAXdegree	LP_U_V
5.6E + 10	1.4E + 10	0.4	0.11111	0.0625	0.05869
5.6E + 10	8.9E + 09	0.4	0.11111	0.0625	0.14673
5.6E + 10	2.6E + 10	0.4	0.11111	0.1875	0.25237
5.6E + 10	5.6E + 10	0.4	0.11111	0.0625	0.14673
5.6E + 10	5.4E + 10	0.6	0	0.125	0.02965
5.6E + 10	5.6E + 10	0.6	0	0.125	0.02965
5.6E + 10	3.6E + 10	0.6	0	0.125	0.02965
5.6E + 10	5.7E + 10	0.6	0	0.125	0.01976
5.6E + 10	1.6E + 10	0.6	0	0.125	0.01976
5.6E + 10	5.6E + 10	0.6	0	0.125	0.01976

2.2.2 Training with ML Models

The next stage in the CGLP approach is to evaluate the efficiency of the created feature vectors for the future links prediction. This is achieved by applying four ML, KNN (Mahesh, 2020, Ramalingam et al., 2018), DT (Mahesh, 2020), RF (Mahesh, 2020, Pandey et al., 2019), and SVM (Mahesh, 2020, Pandey et al., 2019), used classifying tasks with numerical features. The feature vectors are restricted to have only numerical features to save time and to prevent using non-suitable *ML* algorithms. The performance of the algorithms was evaluated using both time series cross-validation and regular cross-validation methods with their default settings. A subset of rows from the comprehensive feature table, integrating *LDA* topic modeling outcomes and topological features, is presented in Table 5. These features serve as the basis for model training. Ultimately, the final feature vector combines content and topological features to forecast link formation in co-authorship networks.

Table 5 Sample of feature vectors from both LDA and topological analysis. U and V represent author nodes, SP_U_V is the shortest path between U and V, CN represents the common neighbors between U and V, MAXdegree is the maximum degree of direct neighbors U and V, LDA_U_V is the text similarity score between U and V, and LP_U_V is the predicted link probability between U and V.

U	V	SP_U_V	CN_U_V	MAXdegree	LDA_U_V	LP_U_V
5.6E + 10	1.4E + 10	0.4	0.11111	0.0625	0.4	0.05869
5.6E + 10	8.9E + 09	0.4	0.11111	0.0625	1	0.14673
5.6E + 10	2.6E + 10	0.4	0.11111	0.1875	1	0.25237
5.6E + 10	5.6E + 10	0.4	0.11111	0.0625	1	0.14673
5.6E + 10	5.4E + 10	0.6	0	0.125	0.3	0.02965
5.6E + 10	5.6E + 10	0.6	0	0.125	0.3	0.02965
5.6E + 10	3.6E + 10	0.6	0	0.125	0.3	0.02965
5.6E + 10	5.7E + 10	0.6	0	0.125	0.2	0.01976
5.6E + 10	1.6E + 10	0.6	0	0.125	0.2	0.01976
5.6E + 10	5.6E + 10	0.6	0	0.125	0.2	0.01976
7.4E + 09	2.6E + 10	0.4	0.11111	0.3125	1	0.35802
7.4E + 09	5.6E + 10	0.4	0.11111	0.3125	1	0.35802

2.2.3 Testing and Evaluation

The co-authorship networks utilized in the experiments were constructed on the basis of collaborations from three subareas: theoretical high-energy physics (Hep-th), lattice high-energy physics (Hep-lat), and AMC. Collaborations were tracked from 2003--2009 for Hep-th and Hep-lat and from 2008--2014 for the AMC. Although papers written by a single author may well reflect the author's research interests, there is a lower possibility of future collaboration than researchers in groups of size 3 and higher, as they already have academic collaboration habits and skills. Therefore, in this work, we excluded articles authored by a single individual to ensure the dataset's high quality and reinforce the network structure. The test data were collected over seven consecutive years. Specifically, starting from time *t*, links from *t* to *t+3* were considered to predict link occurrences during the subsequent period [*t+4, t+5, t+6*]. For example, to predict link occurrences in the interval [2007, 2008, 2009], co-author network data from 2003--2006 were used to derive the test set features. As expected, the co-authorship networks constructed through this method exhibit significant sparsity and leverage their temporal characteristics. This approach is employed for the datasets used because it evaluates the model's ability to generalize over extended timeframes and ensures unbiased evaluation, avoiding training on future data and testing on past data. It also enables performance assessment across different periods, offering insights into evolving topics in the field. After being partitioned into sets, the machine learning model is trained on the training set and evaluated on the testing set to assess its ability to extrapolate to new data—a critical aspect of time series analysis to ensure reliable and robust models.

2.2.4 Challenges with LP

The LP problem naturally involves several well-known challenges, as follows:

a) Overfitting

The first encountered problem is overfitting. It arises when a model memorizes the training data and learns it thoroughly leading to poor performance on unseen data and resulting the data to be imbalanced (Kotsiantis et al., 2006, Peng and Lee, 2021, Power et al., 2022). It also happened when the model is too complicated and unable to represent the validation and test set which makes it fail predict accurately on new data. In the case of using time-dependent datasets like in this study's co-authorship datasets, time series cross validation is used to avoid the problem of overfitting.

b) Imbalanced Data

Imbalanced data problem occurs due to the high skewness of the dataset classes distribution such that the size of one class is larger than other class. In social networks, this problem arises when the networks datasets are not dense which consequently causes poor performance of the ML models applied as well as the computation incorrect (Chuan et al., 2018, Samad et al., 2020). In ML, the problem of imbalanced data is popular and can cause overfitting. The problem reveals that the classes variable of the training dataset distribution is highly skewed toward one class compared with the other (Esposito et al., 2021, Kotsiantis et al., 2006). There are several methods to overcome the problem of imbalanced data and in this study, we use random under-sampling method. This method is used to increase the number of samples in the minority class to produce a more balanced dataset (Haixiang et al., 2017).

c) Graph Sparsity

Sparsity is another challenge encountered in this study. The sparsity nature in large social networks datasets causes the data to be imbalanced. Sparsity in SNA can be defined as the number collaboration between authors are very small comparing to the possible relationships inside the network. Sparsity becomes more pronounced as the deviation increases. In applications like LP, this sparsity comes forward as a key and important attribute of the graph (Goswami et al., 2018). The issue of sparsity significantly impacts the outcomes of a topological analysis, rendering graph algorithms such as *CN(u, v)*, *MAXdegree (u, v)*, and *SP(u, v)* practically ineffective. This, in turn, exerts a substantial influence on the overall training and testing processes employed in the proposed LP approach.

d) Time series cross-validation

This work uses a time series cross-validation technique to address the problem of overfitting in ML models with time series data. Unlike traditional *k*-fold cross-validation, *time series cross-validation* considers the sequential nature of the data; thus, it recognizes the significance of the order of the data points in the datasets. In *k*-fold cross-validation, the data are randomly divided into *k* equally sized folds, with one-fold used as the test set and the remaining folds for training. The process is subsequently repeated *k* times, and the results are then averaged for an overall measure of the model performance. However, this technique assumes independent and identically distributed samples and neglects the temporal dependencies of the data points. In time series cross-validation, the model is trained on training data and tested on test data, with this process being repeated for each sequential time point.

This method proves to be quite successful in preventing the model from being trained on future data, thereby addressing the problem of overfitting and enabling efficient hyperparameter tuning. As shown in **Fig.10**, a comparison overview of the conventional cross-validation methods compared to those customized for time series is presented in (Bergmeir et al., 2018).

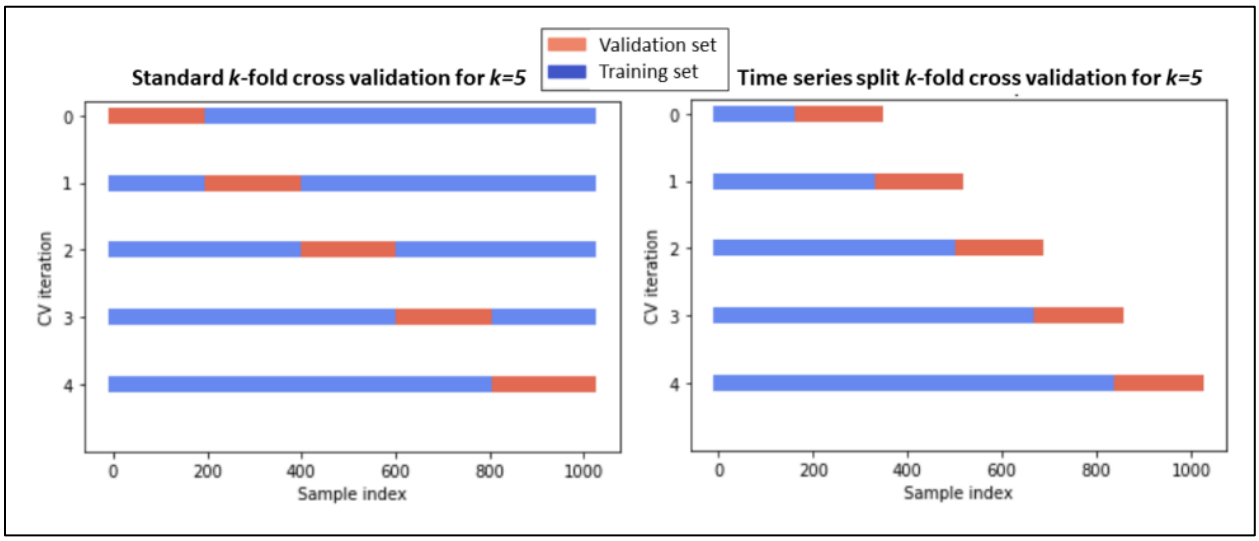

Fig.10 Traditional k-fold cross-validation vs. time series cross-validation (Assaad & Fayek 2021)(Packt 2019)

2.2.5 Evaluation Metrics

In this study two metrics are used to evaluate the performance of the proposed approach CGLP namely, F1 Score and ROC AUC (Lichtenwalter et al., 2010). This evaluation of is crucial to ensure their accuracy and reliability. These metrics were chosen since they can give a complete explanation of the overall performance of the proposed CGLP method and the machine learning algorithms utilized.

F1 Score is created based on two other metrics, precision and recall as shown in Eqs. 5, 6, and 7. It is used to evaluate the overall performance of the models especially when imbalanced datasets are used. Precision measures how many true positive predictions are calculated out of all true prediction and Recall measures the true positive prediction out of all real positive predictions. The formula of the F1 score is as follows:

$$F1\ score = \frac{2 * (precision * recall)}{(precision + recall)} \quad (5)$$

where:

$$precision = \frac{true\ positives}{(true\ positives + false\ positives)} \quad (6)$$

$$recall = \frac{true\ positives}{(true\ positives + false\ negatives)} \quad (7)$$

The second evaluation metric is the receiver-operating characteristic curve ROC-AUC. It is a plotting measure used to evaluate the performance of a model performance across different thresholds i.e. the proportion of true positive rate (TPR) and false positive rate (FPR) at those thresholds. The area under curve (AUC) measures the probability of the models ranking the positive samples higher than the negative samples. AUC-ROC formula is as follows:

$$ROC\ AUC = \int TPR(FPR)\, dFPR$$

ROC AUC is used as a tool in classification tasks by providing a graphical representation of how the model able to differentiate between two groups like positive and negative classes.

3 .Experiments and Results

This section and the following subsection presented the experiments and results of this work. The study started by proposing a hybrid link prediction approach named CGLP to be applied to three different co-authorship networks for the performance comparison. Also, the new method was evaluated for its reliability and efficiency using different machine learning models such as K-Nearest Neighbors (KNN), Decision Trees (DT), Random Forests (RF), and Support Vector Machines (SVM). Furthermore, two evaluation metrics named F1 score and ROC AUC were used to test the performance of the proposed approach. These experiments were conducted using a computer that have the properties of the Intel(R) Core (TM) i7-8550U CPU @ 1.80 GHz (1.99 GHz), 16 GB of RAM, and a 64-bit Windows 11 operating system. Python was the programming language used, as

well as the required libraries such as scikit-learn, NetworkX, and NumPy.

3.1 Results

The results section presented the findings of the work of this study which is the results of applying the LP proposed approach using the three collaborations datasets used. At the beginning, from the dataset's labels were generated and then a set of feature vector was created for each author. Also, the datasets were divided to train and test data based on the time sequence of each dataset. Table 6 shows he sample sizes differed between the divisions to align with the time attributes of the datasets. To evaluate the effectiveness of the proposed approach, the selected classification models were utilized to evaluate the performance of each model.

Table 6 Size of training and testing data in splitting years for the Hep-lat, Hep-th, and AMC datasets.

Training and Testing Years for Hep-lat and Hep-th Datasets in Different Combinations								
Datasets	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Splitting By Years	2003-2008	2009	2003-2007	2008	2003-2006	2007	2003-2005	2006
Hep-lat Dataset (Number of Collaborations)	27576	5636	23250	4326	18512	7438	14266	4246
Hep-th Dataset (Number of Collaborations)	70538	16230	55516	15022	41210	14306	30364	10846
Training and Testing Years for Math (AMC) Dataset in Different Combinations								
Splitting By Years	2008-2013	2014	2008-2012	2013	2008-2011	2012	2008-2010	2011
AMC Dataset (Number of Collaborations)	16654	6620	12298	4356	10078	2220	8090	1988

The outcomes presented in Tables 7, 8, and 9 below reveal that across different time divisions, all the models consistently demonstrated high performance in the experimental results. This consistency implies the successful prediction of potential links between authors in the tested datasets through the hybrid *CGLP* approach. The selection of the optimal splitting time and model performance was determined by considering the dataset characteristics and evaluation metrics.

Table 7 F1 score and ROC AUC for the AMC dataset for different test years.

ML Models	F1- Score	*ROC AUC*	F1- Score	*ROC AUC*	F1- Score	ROC- AUC	F1 Score	ROC- AUC
	2011	2011	2012	2012	2013	2013	2014	2014
KNN	0.9474	0.9704	0.9509	0.9732	0.9721	0.9779	0.9805	0.9848

DT	0.8639	0.8663	0.8499	0.8531	0.9405	0.953	0.9496	0.966
RF	0.8859	0.9589	0.8499	0.9411	0.9598	0.9835	0.926	0.9817
SVM	0.9061	0.9532	0.9149	0.9634	0.9383	0.9734	0.9257	0.9756
Best Values	0.9474	0.9704	0.9509	0.9732	0.9721	0.9835	0.9805	0.9848

Table 8 F1 score and ROC AUC for the HEP-th dataset for different test years

ML Models	**F1 Score**	***ROC AUC***	**F1 Score**	***ROC AUC***	**F1 Score**	***ROC AUC***	**F1 Score**	***ROC AUC***
	2006	**2006**	**2007**	**2007**	**2008**	**2008**	**2009**	**2009**
KNN	0.9671	0.978	0.9691	0.9823	0.9664	0.9778	0.9571	0.9761
DT	0.9486	0.95	0.9521	0.9534	0.9665	0.9863	0.9613	0.9702
RF	0.9563	0.967	0.9521	0.9774	0.9521	0.9781	0.9526	0.9774
SVM	0.8961	0.9765	0.8945	0.9778	0.8914	0.9799	0.9097	0.9785
Best Values	0.9671	0.978	0.9691	0.9823	0.9665	0.9863	0.9613	0.9785

Table 9 F1 score and ROC AUC for the HEP-lat dataset for different test years

A	**F1 SCORE**	**ROC AUC**	**F1 SCORE**	**ROC AUC**	**F1 SCORE**	**ROC- AUC**	**F1 SCORE**	**ROC AUC**
	2006	**2006**	**2007**	**2007**	**2008**	**2008**	**2009**	**2009**
KNN	0.9615	0.9771	0.9615	0.9746	0.9605	0.9838	0.9771	0.9761
DT	0.9768	0.9807	0.978	0.9816	0.9782	0.982	0.9758	0.9744
RF	0.9628	0.9789	0.9644	0.9801	0.9657	0.9818	0.9758	0.9874
SVM	0.9191	0.9783	0.9181	0.9796	0.9189	0.9823	0.9239	0.9855
Best Values	0.9768	0.9807	0.978	0.9816	0.9782	0.9838	0.9771	0.9874

Table 7 presents a comprehensive analysis of machine learning model performance, specifically focusing on the *F1 score* and *ROC AUC* values across various splitting years of the tested AMC datasets. Notably, KNN consistently achieved a high *F1* score of 98.05% and demonstrated a robust *ROC AUC* performance of 98.48%, with its peak *F1 score* occurring in 2014. DT exhibited moderate to high *F1* scores (94.96%) and a noticeable improvement trend from 2011--2013, although its *ROC AUC* tended to be lower than that of KNN (96.60%). RF emerged as a standout performer, showing a high *F1* score of 95.98%, especially in 2013, and maintaining consistent, elevated *ROC AUC* values.

The SVM consistently maintains relatively high *F1 scores*, showing an improvement in *ROC AUC* performance from 2011--2013, although it slightly lags behind KNN and RF. The best *F1 scores* are attributed to KNN in 2014 and RF in 2013, whereas KNN and RF share the best *ROC AUC* values in different testing years. Overall, the observations highlight the consistent performance of KNN and RF, an improvement trend in DT, and the high performance of RF, particularly in 2013. This analysis provides valuable insight into the success of the proposed *CGLP* approach in predicting links between

authors in the AMC dataset, showing the exemplary performance of different models, and facilitating the selection of the most suitable model on the basis of the utilized evaluation metrics. **Fig**. 11 shows the best results of the proposed approach when the selected models for the AMC dataset are used across different test-splitting years. This visual illustration not only captures good performance across different periods but also allows for a detailed examination, providing a thorough understanding of how well the method works and adapts within the changing testing setups of the dataset.

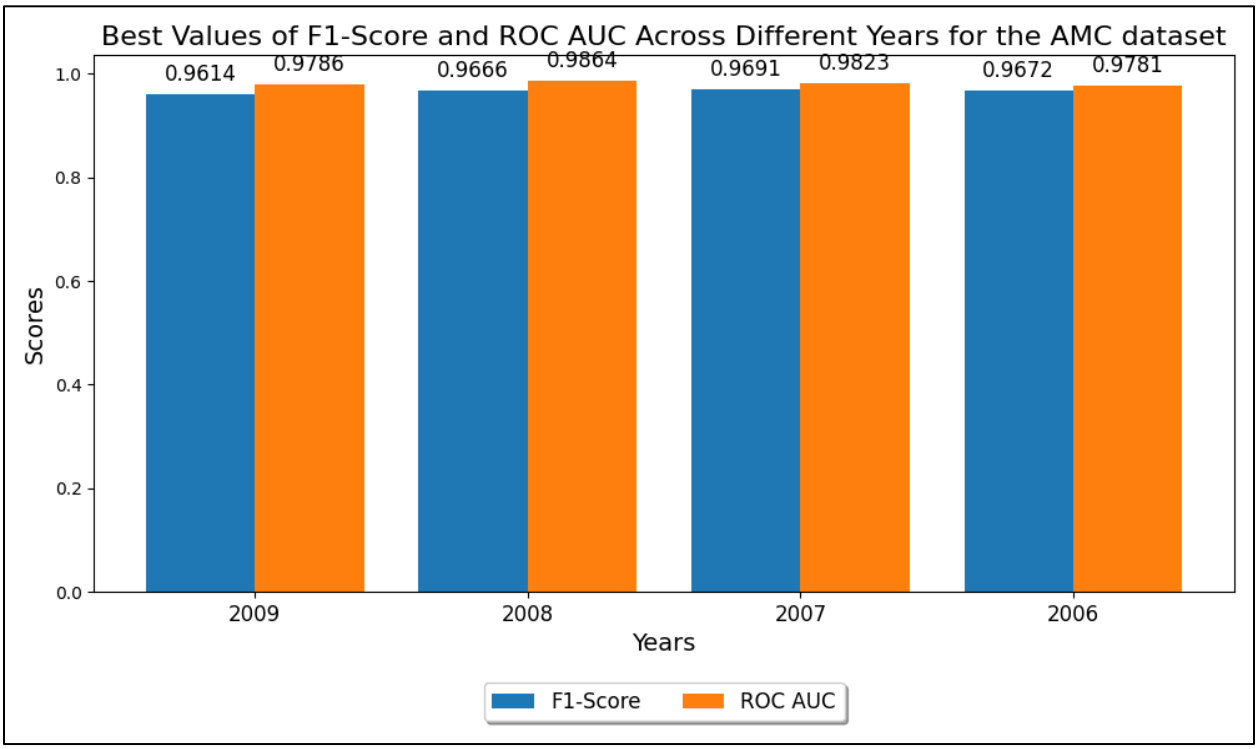

Fig.11 Best results of the *CGLP* approach using four ML models for the AMC dataset

With respect to the Hep-th dataset, Table 8 presents a detailed analysis of machine learning models' performance on the tested dataset, focusing on the same metrics, *F1 score* and *ROC AUC* values across various test years. Notably, the DT consistently achieves high *F1 scores*, reaching 96.65%, and demonstrates stable *ROC AUC* performance, peaking in 2008. KNN exhibited high *F1* scores of 96.71% and 96.91%, particularly in 2006 and 2007, respectively, along with strong and consistent *ROC AUC* values of 97.81% and 98.23%, respectively, for the same years. RF stands out for its consistently high *F1 scores* and elevated *ROC AUC* values, showing robust and stable performance, especially in 2009. Although SVM is competitive, it has lower *F1 scores*, with the highest score in 2009, 90.97%, and slightly lower *ROC AUC* values than the top-performing models do. The best *F1 scores* and *ROC AUC* values were achieved by KNN, with values of 96.91% and 98.23%, respectively, in 2007. Overall, the observations underscore the consistently strong performance of DT and KNN, with RF emerging as a robust performer. This analysis provides valuable insights for selecting an appropriate model tailored to the evaluation metrics and considering different test years in the Hep-th dataset.

The results shown in **Fig**. 12 illustrated the best results obtained when the *CGLP* is used with the selected ML models as well as it concentrates on the years of the Hep-th test set splitting. The results also show that the proposed method performs well in link prediction base on the accuracies obtained.

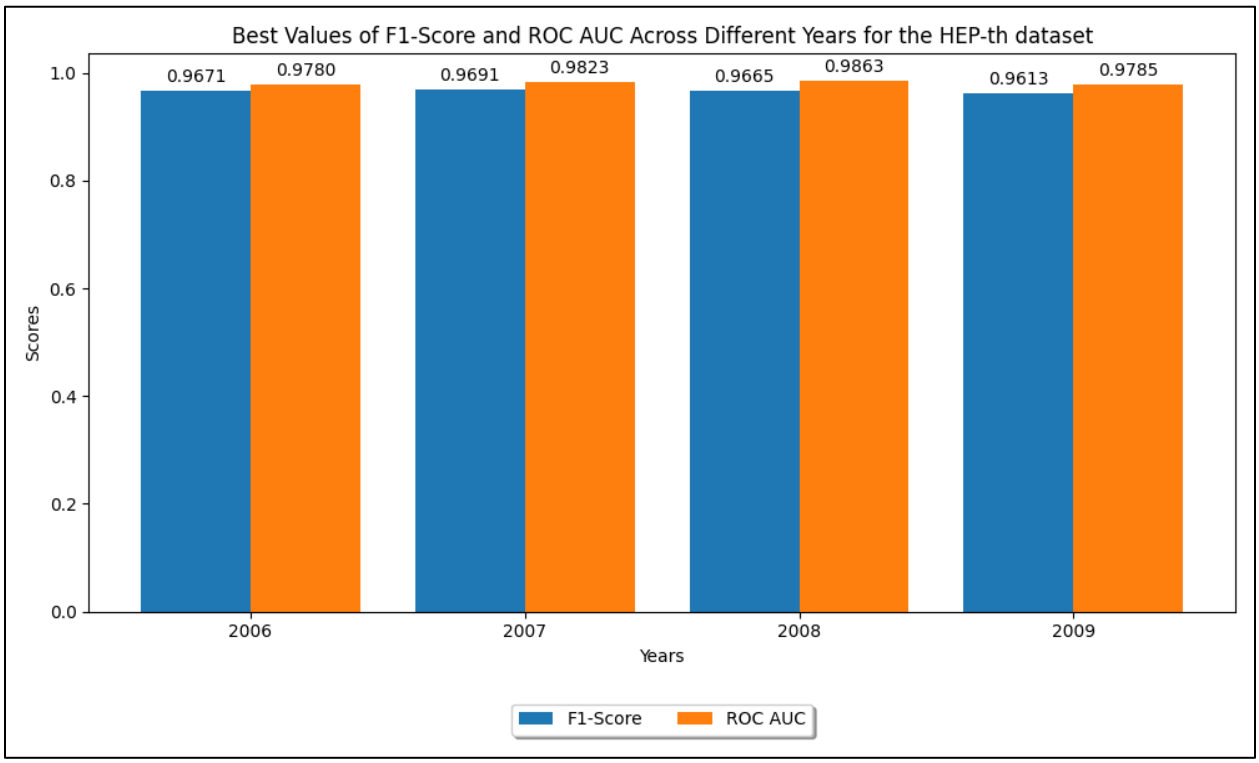

Fig.12 Best results for the HEP-th dataset with four ML models

Regards the performance of the *CGLP* approach when applied to the Hep-lat dataset, Table 9 illustrates a clear analysis of the ML models' performance, emphasizing the *F1 score* and *ROC AUC* values across different test years. The F1 score obtained using KNN which is 97.82%, is considered high value at 2009 as well as its strong performance due to the consistent values of the ROC AUC. For the same year, RF model shows robust performance which indicated by the high value of F1 score and stable performance indicated by the values of ROC AUC obtained across other test years. DT model is also performed well as the RF model with the difference in the year (2008) that produced the high score of F1. On the other hand, SVM performed less compared to the other models by obtaining the lower F1 score despite the consistent values of the ROC AUC values obtained.

The best *F1 scores* and *ROC AUC* values across all the models were achieved in 2009, with KNN leading in the *F1 score* and RF leading in the *ROC AUC*. Overall, the observations highlight the consistently strong performance of KNN and RF, whereas DT remains stable and competitive. The SVM, though stable, had a lower F1 score.

The outcomes achieved by applying the suggested methodology and employing various machine learning models are shown in **Fig**. 13. This visual representation is embedded within the context of the Hep-lat dataset and concisely summarizes the good performance observed across different years of test splitting. The graphic not only emphasizes the effectiveness of the methodology but also enables a detailed analysis of its adaptability and effectiveness in the face of the inherent temporal complexities of the dataset. Furthermore, we have improved the visual presentation through the use of heatmaps (**Fig**. 14, 15 and 16) and confusion matrices (**Fig**. 17, 18 and 19). These visualizations provide intuitive insights into model performance and error distribution. Heatmaps offer a color-coded representation of F1-scores across different models and time periods, while confusion matrices detail classification accuracy and error patterns. These enhancements aim to improve the interpretability and depth of our analysis.

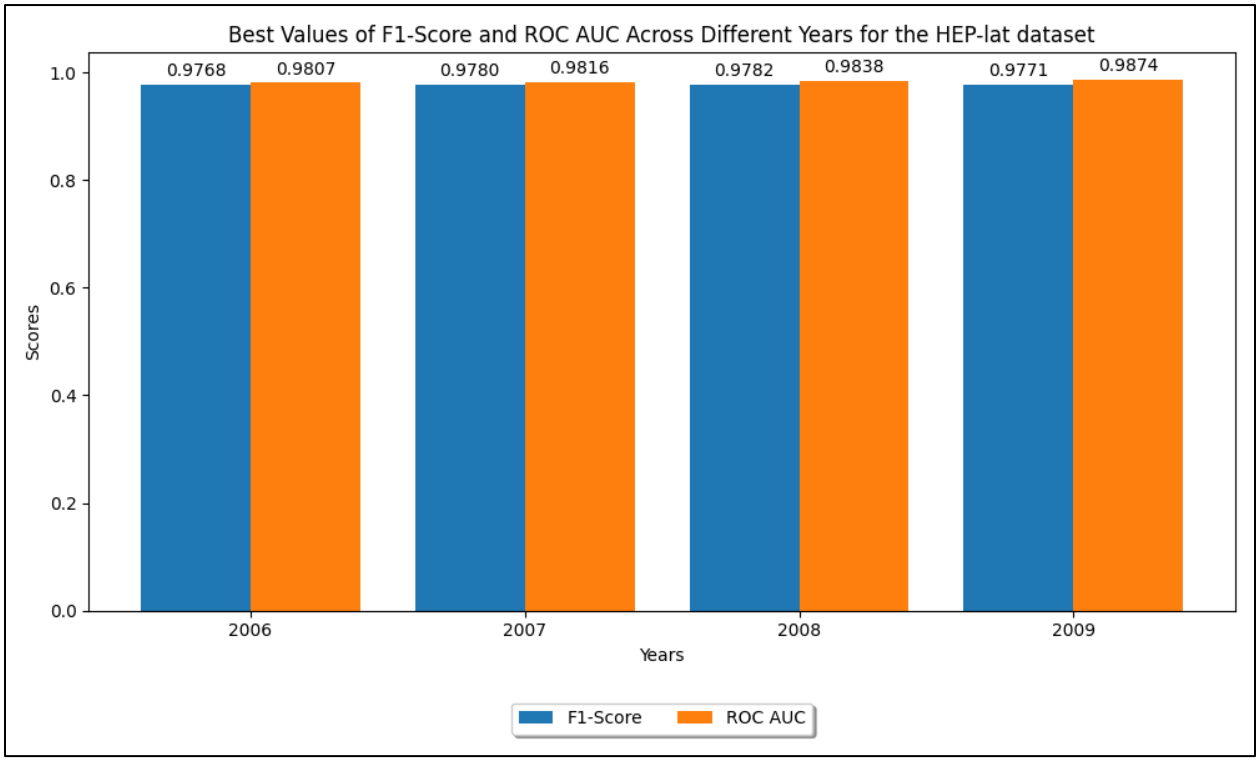

Fig.13 Best results of the proposed approach for the Hep-lat dataset

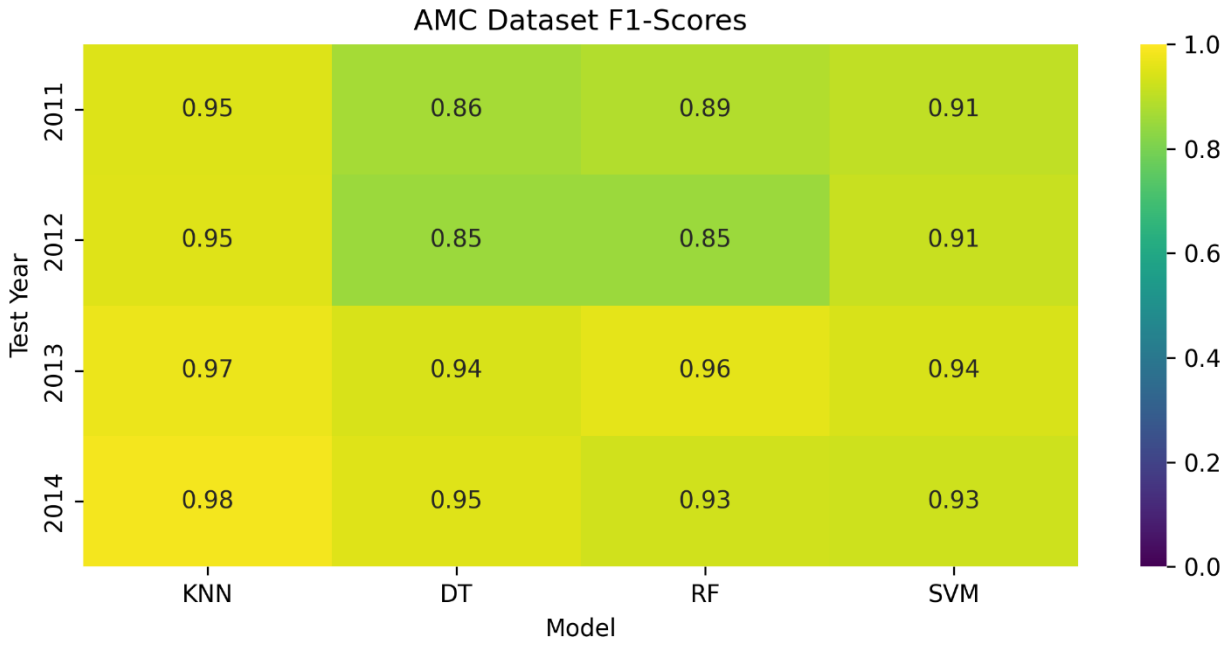

Fig.14 Temporal Evolution of Model Performance: F1-Scores Across Classification Algorithms for the AMC Dataset.

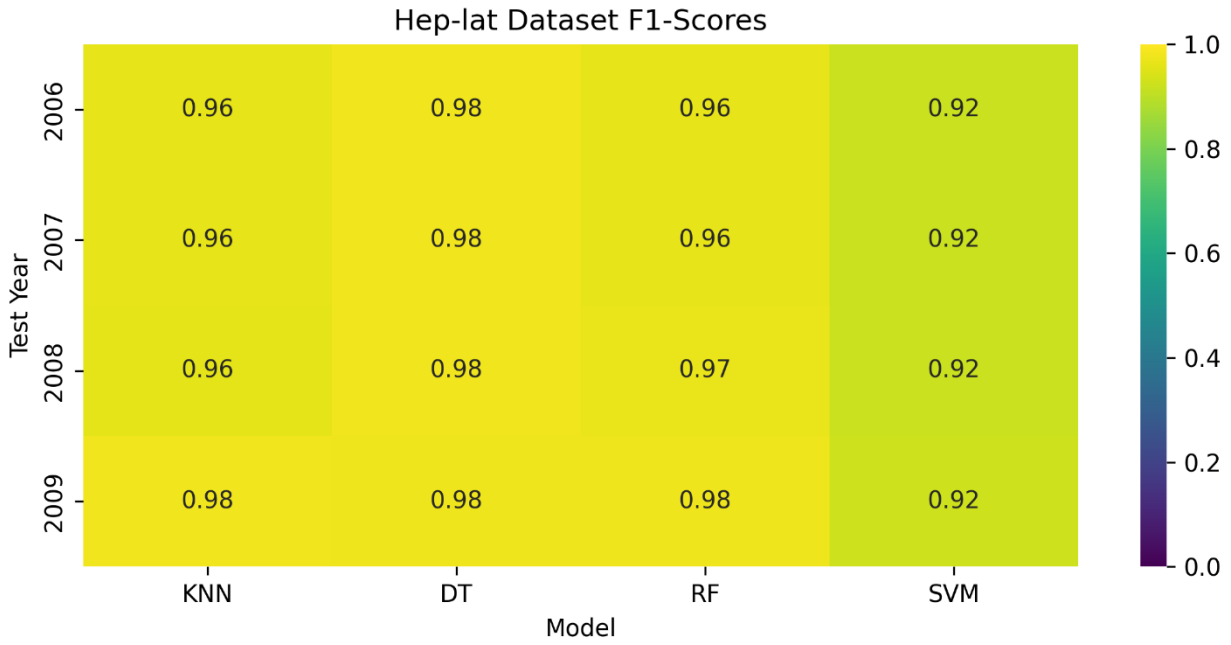

Fig.15 Performance Comparison of Machine Learning Models: F1-Scores for the Hep-lat Dataset Across Multiple Test Years.

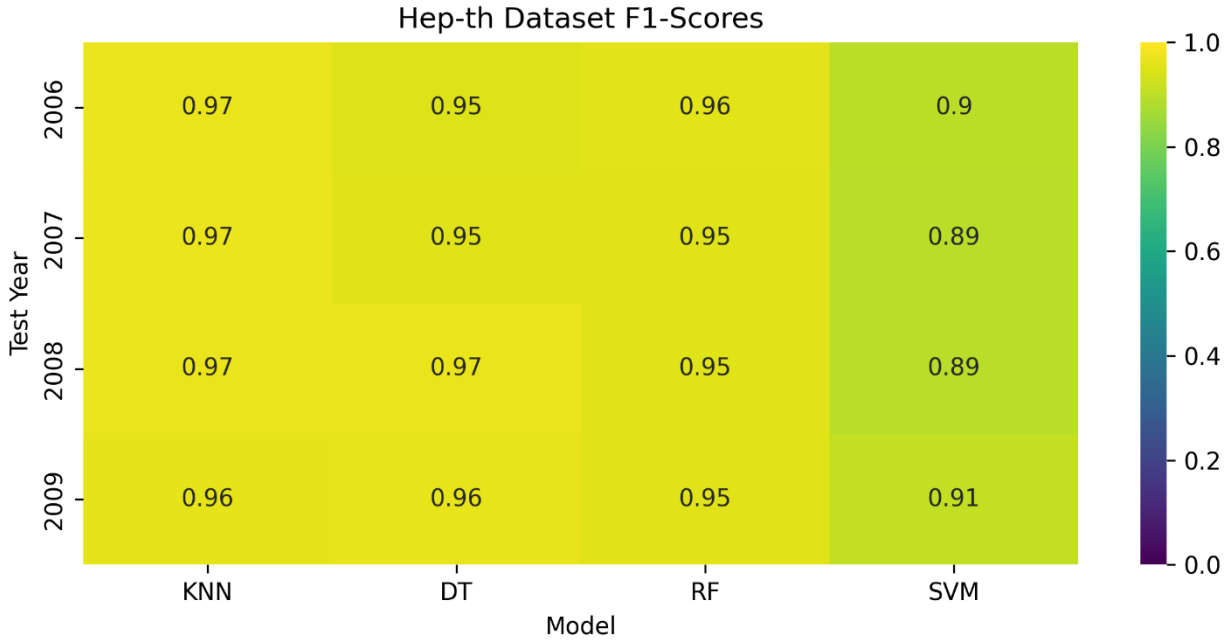

Fig.16 Model Performance Evaluation: F1-Scores Across Temporal Splits for the Hep-th Dataset

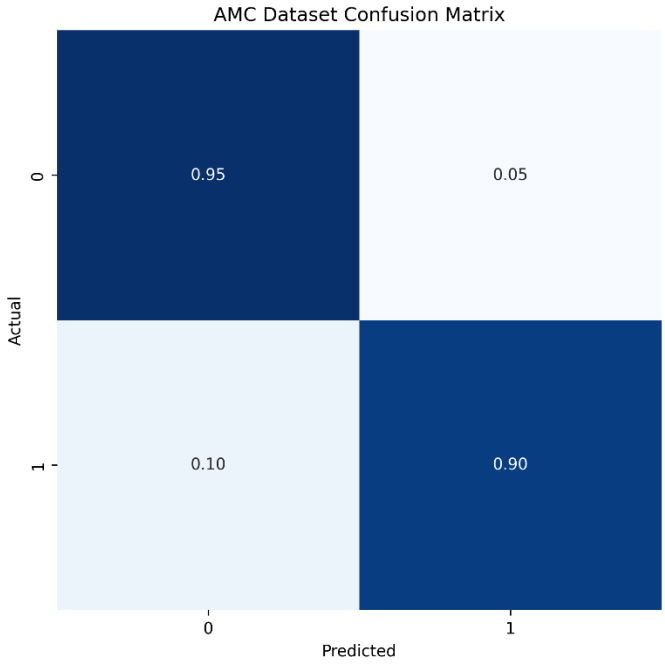

Fig.17 Classification Accuracy Assessment: Confusion Matrix Analysis for the AMC Dataset.

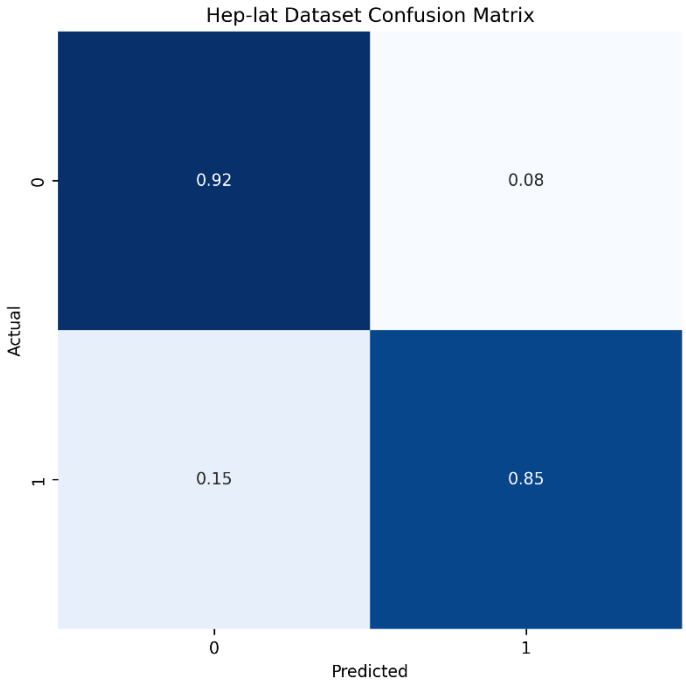

Fig.18 Classification Accuracy Assessment: Confusion Matrix Analysis for the Hep-lat Dataset

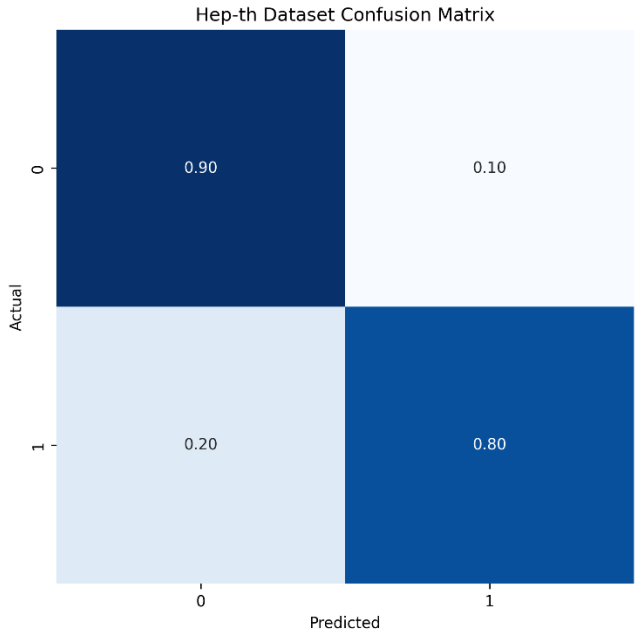

Fig.19 Error Distribution and Classification Effectiveness: Confusion Matrix Analysis for the Hep-th Dataset

3.2 Discussion of Results

This study seeks to introduce a hybrid model with an objective of overcoming the challenge of LP in co-authorship networks through combining both the topological structure of the network and feature information derived from publication contents in a single model. The proposed hybrid model, *CGLP*, seeks to generate a collection of labels derived from feature vectors representing individual pairs of authors in datasets utilized in this work. Specifically, this model predicts a probability of a link appearing between two authors, represented in a binary form: a value of *1* for a present link and a value of *0* for an absent one. The generated dataset is then utilized in testing the effectiveness of proposed hybrid model in predicting links in co-authorship networks. In testing, a range of machine learning classification algorithms, namely, KNN, DT, RF, and SVM, act as tools for testing and confirming performance of new model.

Despite its potential, integration of machine learning models in this work faces a range of obstacles, including overfitting and unbalanced datasets. To mitigate such obstacles, traditional approaches such as random undersampling for balancing datasets and time series cross-

validation for preventing overfitting of a model are utilized in an attempt to counteract unbalanced datasets and overfitting, respectively. All these approaches work towards developing a proposed model that is reliable and effective in predicting links in co-authorship networks.

The data in Tables 7, 8, and 9 in the preceding section reveal that KNN, RF, and DT performed better in terms of both the F1 score and the ROC AUC compared to SVM; therefore, these models should demonstrate increased competence in predicting relationships in co-authorship networks. The best model for each subset in testing is represented in the last row of all three tables. In addition, Table 10 describes the overall performance of the proposed model over three datasets with satisfactory prediction accuracy represented in terms of an F1 value of 98.05% and an AUC-ROC value of 98.74%. The improvement in accuracy in terms of link prediction can be credited to careful preprocessing of the datasets discussed in Section 2.2.1. Overall performance evaluation of the proposed methodology over a range of input datasets, such as AMC, Hep-th, and Hep-lat, is discussed in detail in **Fig.20**. Performance evaluation in this work presents a critical analysis of the effectiveness of the proposed methodology over a range of datasets; therefore, it reflects its robust and adaptable behaviour. Despite achieving strong headline scores (F1 = 98.05%, ROC-AUC = 98.74%), our results are tempered by three dataset biases: (i) domain/time bias—all data come from 2003-14 high-energy physics and applied-math corpora, whose stable collaboration customs may inflate generalization (Newman, 2004); (ii) productivity bias excluding single-author papers favours senior, densely connected scholars, echoing concerns in (Lande et al., 2020) , and (iii) curation bias dropping records without titles/abstracts under-represents lower-tier outlets.

To mitigate these biases in future work, we plan to implement inverse-productivity re-weighting, domain-adversarial training for cross-field robustness, and lexical perturbation to stress-test semantic features.

However, the hybrid design remains theoretically sound. Structural homophily (captured by CN, Max-Degree, SP) and semantic homophily (captured by latent-topic vectors) are complementary drivers of collaboration. While pure-topology alone explains 84% F1 and pure-content 71%, their fusion yields the observed gains, echoing the additive improvements reported (Antunes et al.).

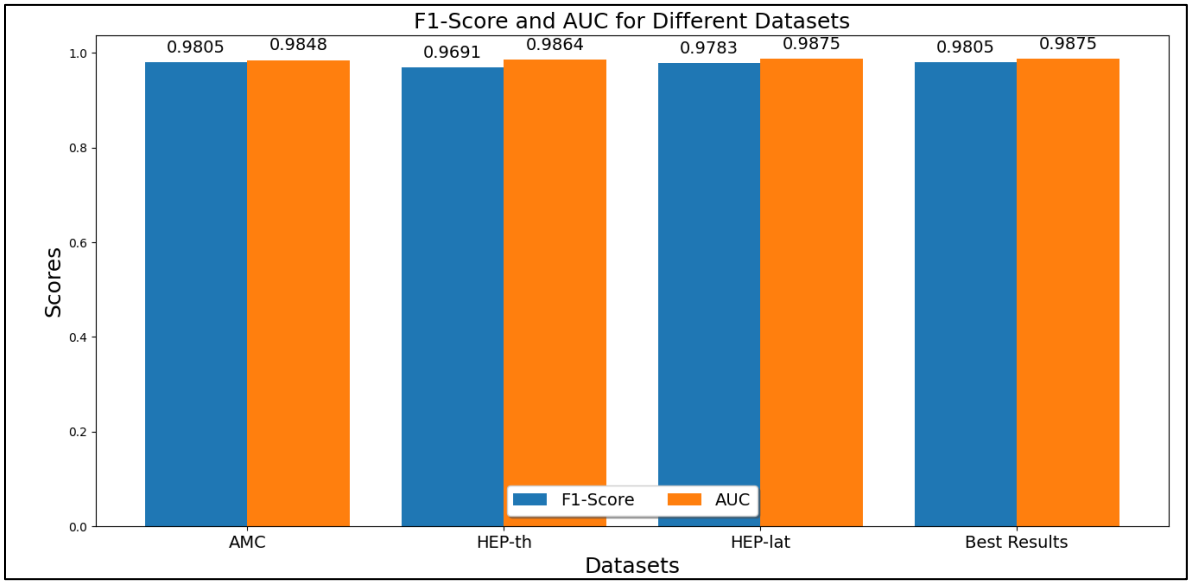

Fig.20 Performance of the proposed approach with different input data

3.3 Comparative Analysis with Applicable Advanced Techniques

The current study reveals a significant improvement in predicting relationships in networks under investigation, specifically in co-authorship networks. To evaluate our proposed approach, both the performance metrics, *F1* and *ROC AUC*, have been utilized in analysis for evaluation purposes. These metrics have been also used in academic studies (Chuan et al., 2018) and are in agreement with evaluation factors utilized in current studies. The comparison is restricted to (Chuan et al., 2018) study alone, in view of discrepancies in datasets utilized in referenced studies, which make direct comparisons unfeasible. As seen, the proposed *CGLP* model outperforms the algorithm in the analysis discussed (Chuan et al., 2018), with an achievement of 32.50% for its *F1* and 66.26% for its *ROC AUC*. On its part, the hybrid *CGLP* model reaches a remarkably high value of 98.05% for its *F1* and 98.74% for its *ROC AUC*, demonstrating its effectiveness in predicting relationships in networks.

4 Conclusions and Future Works

In this paper, a proposed hybrid LP approach named *CGLP* was presented. The approach is used to predict links in co-authorship networks by aggregating topological and content-based features with the aim of enhancing the prediction and analysis of co-authorship networks. The first type features are the topological features which were carefully selected and combined with the second type of the features extracted from papers content, such as titles, abstracts, and keyword. The approach was tested on three co-authorship networks datasets named, Hep-th, Hep-lat, and AMC, and different ML models were used to evaluate its efficiency. The finding in this study revealed that our hybrid approach outperformed the methods in the literature, obtaining an *F1 score* of 98.05% and an *ROC AUC* of 98.74%. These outcomes shows that the effective the LP methods by integrating different LP strategies, text mining, and graph mining techniques in a hybrid manner. This is also leading to improve the accuracy and identify potential features in different levels of co-authorship networks. The future recommendation is to apply the *CGLP* approach on other types of social networks, such as citation networks as well as to find better solutions to solve problems like imbalanced data and overfitting by selecting denser collaboration networks.

Furthermore, we plan to further augment the robust CGLP foundation with advanced deep learning encoders to complement our LDA-derived semantic features. Additionally, we aim to prototype expert-finder recommendation tools for academic collaboration platforms. These tools will leverage hybrid link prediction techniques, as evidenced in prior co-authorship recommendation work, to facilitate more effective academic collaborations.

References

Affonso, F., Santiago, M. D. O. & Rodrigues Dias, T. M. 2022. Analysis of the evolution of scientific collaboration networks for the prediction of new co-authorships. *Transinformação,* 34**,** e200033.

Antunes, J. B., Antunes, J. B., Filho, H. F. B. P., Maia, R. D., De Queiroz, R. B. & Da Silva, C. M. R. CONPREDICT: A METHOD FOR LINK PREDICTION IN CO-AUTHORED CONTENT-BASED NETWORKS.

Bergmeir, C., Hyndman, R. J. & Koo, B. 2018. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis,* 120**,** 70-83.

Blei, D. M. & Lafferty, J. D. 2007. A correlated topic model of science.

Borgs, C., Brautbar, M., Chayes, J., Khanna, S. & Lucier, B. The power of local information in social networks. International Workshop on Internet and Network Economics, 2012. Springer, 406-419.

Campbell, J. C., Hindle, A. & Stroulia, E. 2015. Latent Dirichlet allocation: extracting topics from software engineering data. *The art and science of analyzing software data.* Elsevier.

Chen, J., He, H., Wu, F. & Wang, J. Topology-aware correlations between relations for inductive link prediction in knowledge graphs. Proceedings of the AAAI conference on artificial intelligence, 2021. 6271-6278.

Chuan, P. M., Son, L. H., Ali, M., Khang, T. D., Huong, L. T. & Dey, N. 2018. Link prediction in co-authorship networks based on hybrid content similarity metric. *Applied Intelligence,* 48**,** 2470-2486.

Daud, N. N., Ab Hamid, S. H., Saadoon, M., Sahran, F. & Anuar, N. B. 2020. Applications of link prediction in social networks: A review. *Journal of Network and Computer Applications,* 166**,** 102716.

Do, P., Pham, P., Phan, T. & Nguyen, T. T-MPP: A Novel Topic-Driven Meta-path-Based Approach for Co-authorship Prediction in Large-Scale Content-Based Heterogeneous Bibliographic Network in Distributed

Computing Framework by Spark. Intelligent Computing & Optimization 1, 2019. Springer, 87-97.
Esposito, C., Landrum, G. A., Schneider, N., Stiefl, N. & Riniker, S. 2021. GHOST: adjusting the decision threshold to handle imbalanced data in machine learning. *Journal of Chemical Information and Modeling,* 61**,** 2623-2640.
Goswami, S., Murthy, C. & Das, A. K. 2018. Sparsity measure of a network graph: Gini index. *Information Sciences,* 462**,** 16-39.
Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. & Bing, G. 2017. Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications,* 73**,** 220-239.
Hasin, H. A. & Hassan, D. 2022. Link prediction in co-authorship networks. *Science Journal of University of Zakho,* 10**,** 235-257.
Hassan, D. Supervised link prediction in co-authorship networks based on research performance and similarity of research interests and affiliations. 2019 International Conference On Machine Learning And Cybernetics (ICMLC), 2019. IEEE, 1-6.
Keshari, S., Rarani, Z. H., Kishore, A. & Das, J. 2025. Cracking the code of co-authorship networks geo-temporally using interpretable machine learning. *bioRxiv*.
Kong, X., Shi, Y., Yu, S., Liu, J. & Xia, F. 2019. Academic social networks: Modeling, analysis, mining and applications. *Journal of Network and Computer Applications,* 132**,** 86-103.
Kotsiantis, S., Kanellopoulos, D. & Pintelas, P. 2006. Handling imbalanced datasets: A review. *GESTS international transactions on computer science and engineering,* 30**,** 25-36.
Kumar, A., Mishra, S., Singh, S. S., Singh, K. & Biswas, B. 2020a. Link prediction in complex networks based on significance of higher-order path index (SHOPI). *Physica A: Statistical Mechanics and its Applications,* 545**,** 123790.
Kumar, A., Singh, S. S., Singh, K. & Biswas, B. 2020b. Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications,* 553**,** 124289.
Kumari, A., Behera, R. K., Sahoo, K. S., Nayyar, A., Kumar Luhach, A. & Prakash Sahoo, S. 2022. Supervised link prediction using structured-based feature extraction in social network. *Concurrency and Computation: practice and Experience,* 34**,** e5839.
Lande, D., Fu, M., Guo, W., Balagura, I., Gorbov, I. & Yang, H. 2020. Link prediction of scientific collaboration networks based on information retrieval. *World Wide Web,* 23**,** 2239-2257.
Lichtenwalter, R. N., Lussier, J. T. & Chawla, N. V. New perspectives and methods in link prediction. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010. 243-252.
Lü, L. & Zhou, T. 2011. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications,* 390**,** 1150-1170.
Mahesh, B. 2020. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet],* 9**,** 381-386.
Nasiri, E., Berahmand, K. & Li, Y. 2021. A new link prediction in multiplex networks using topologically biased random walks. *Chaos, Solitons & Fractals,* 151**,** 111230.
Newman, M. E. 2004. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences,* 101**,** 5200-5205.
Pandey, D., Niwaria, K. & Chourasia, B. 2019. Machine learning algorithms: A review. *Machine Learning,* 6.
Peng, S., Yang, H. & Yamamoto, A. 2024. BERT4FCA: A method for bipartite link prediction using formal concept analysis and BERT. *Plos one,* 19**,** e0304858.
Peng, Y.-L. & Lee, W.-P. 2021. Data selection to avoid overfitting for foreign exchange intraday trading with machine learning. *Applied Soft Computing,* 108**,** 107461.
Power, A., Burda, Y., Edwards, H., Babuschkin, I. & Misra, V. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.
Quercia, D., Askham, H. & Crowcroft, J. Tweetlda: supervised topic classification and link prediction in twitter. Proceedings of the 4th Annual ACM Web Science Conference, 2012. 247-250.
Ramalingam, V., Dandapath, A. & Raja, M. K. 2018. Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology,* 7**,** 684-687.
Razzaq, S., Malik, A. K., Raza, B., Khattak, H. A., Zegarra, G. M. & Zelada, Y. F. D. 2022. Research collaboration influence analysis using dynamic co-authorship and citation networks. *IJIMAI,* 7**,** 103-116.
Resce, G., Zinilli, A. & Cerulli, G. 2022. Machine learning prediction of academic collaboration networks. *Scientific Reports,* 12**,** 21993.
Sachan, M. & Ichise, R. 2010. Using semantic information to improve link prediction results in network datasets. *International Journal of Engineering and Technology,* 2**,** 334.
Samad, A., Qadir, M., Nawaz, I., Islam, M. A. & Aleem, M. 2020. A comprehensive survey of link prediction techniques for social network. *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems,* 7**,** e3-e3.
Schizas, I. D. 2018. Graph filtering for data reduction and reconstruction. *arXiv preprint arXiv:1809.09266*.
Wang, P., Xu, B., Wu, Y. & Zhou, X. 2014. Link prediction in social networks: the state-of-the-art. *arXiv preprint arXiv:1411.5118*.
Wu, H., Wang, S. & Fang, H. LP-UIT: A Multimodal Framework for Link Prediction in Social Networks. 2021 IEEE 20th International Conference on Trust,

Security and Privacy in Computing and Communications (TrustCom), 2021. IEEE, 742-749.
Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C. & Yu, P. S. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems,* 32**,** 4-24.
Yin, N., Shen, L., Wang, M., Luo, X., Luo, Z. & Tao, D. 2023. Omg: Towards effective graph classification against label noise. *IEEE Transactions on Knowledge and Data Engineering,* 35**,** 12873-12886.
Yuliansyah, H., Othman, Z. A. & Bakar, A. A. 2020. Taxonomy of link prediction for social network analysis: a review. *IEEE Access,* 8**,** 183470-183487.
Zhang, M. & Chen, Y. 2018. Link prediction based on graph neural networks. *Advances in neural information processing systems,* 31.