

Course Project Report

**Link Prediction and influential Paper Detection in Citation Networks**

*Submitted By*

**Kesanam Ashinee (211AI023)**

**Jatti Deva Paul (211AI021)**

*as part of the requirements of the course*

**Social Computing (IT480) [Jan - Apr 2024]**

*in partial fulfillment of the requirements for the award of the degree of*

**Bachelor of Technology in Artificial Intelligence**

*under the guidance of*

**Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal**

*undergone at*



**DEPARTMENT OF INFORMATION TECHNOLOGY**

**NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL**

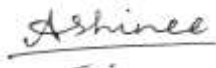
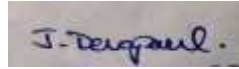
**JAN - APR 2024**

**DEPARTMENT OF INFORMATION TECHNOLOGY**  
**National Institute of Technology Karnataka, Surathkal**

**C E R T I F I C A T E**

This is to certify that the Course project Work Report entitled “**Link Prediction and influential Paper Detection in Citation Networks**” is submitted by the group mentioned below -

**Details of Project Group**

Name of the Student	Register No.	Signature with Date
Kesanam Ashinee	211AI023	
Jatti Deva Paul	211AI021	

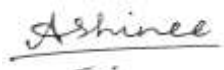
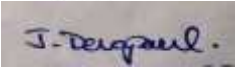
this report is a record of the work carried out by them as part of the course **Social Computing (IT480)** during the semester **Jan - Apr 2024**. It is accepted as the Course Project Report submission in the partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Artificial Intelligence**.

*(Name and Signature of Course Instructor)*  
**Dr. Sowmya Kamath S**

## DECLARATION

We hereby declare that the project report entitled “**Link Prediction and influential Paper Detection in Citation Networks**” submitted by us for the course **Social Computing (IT480)** during the semester **Jan - Apr 2024**, as part of the partial course requirements for the award of the degree of Bachelor of Technology in Artificial Intelligence at NITK Surathkal is our original work. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

### Details of Project Group

Name of the Student	Register No.	Signature with Date
Kesanam Ashinee	211AI023	
Jatti Deva Paul	211AI021	

Place: NITK, Surathkal

Date: 15/04/2024

# Link Prediction and influential Paper Detection in Citation Networks

Kesanam Ashinee <sup>1</sup>, Jatti Deva Paul <sup>2</sup>

**Abstract**— Link prediction in citation networks plays a vital role in understanding scholarly interactions and predicting future collaborations. In this study, we present a comprehensive approach for link prediction in citation networks. We start by extracting relevant node information, including abstracts, titles, and author names. Additionally, influential paper detection is integrated into our methodology, utilizing measures such as the number of citations, degree centrality, betweenness centrality, eigenvector centrality, and PageRank value. Feature engineering is then performed to extract various features such as temporal differences, common authors, TF-IDF similarities for titles and abstracts, along with network properties like centrality measures and preferential attachment. Subsequently, four different machine learning models, namely XGBoost, Logistic Regression, Decision Tree Classifier, and Random Forest Classifier, are trained on the extracted features. Evaluation is carried out using four metrics: F1-score, Accuracy, Recall, and Precision, to assess the models' predictive performance. Our approach demonstrates the effectiveness of combining machine learning algorithms with network analysis techniques for both link prediction and influential paper detection tasks in citation networks. The findings of this study contribute to enhancing our understanding of scholarly interactions and facilitating future collaboration predictions in academic domains.

**Keywords:** *Link prediction, Citation network, Network analysis, Centrality measures, XGBoost, Decision Tree, Logistic Regression, Random Forest, Influential Paper Detection*

## I. INTRODUCTION

Citation networks serve as intricate webs of scholarly communication, where academic papers are interconnected through citations, reflecting the flow of knowledge and ideas within various disciplines. Understanding the structure and dynamics of citation networks is crucial for unraveling the intricacies of scholarly interactions and predicting future collaborations. Link prediction, a key task in citation network analysis, aims to forecast potential connections between articles that have not yet been cited together.

In this era of burgeoning academic output, fueled by the exponential growth of scientific literature, the ability to predict future connections in citation networks has become increasingly important. Such predictions not only offer insights into the evolution of research fields but also facilitate the identification of emerging trends, influential papers, and potential research collaborations. Consequently, link prediction techniques have garnered significant attention across disciplines, ranging from bibliometrics and network science to machine learning and natural language processing.

Our study extends beyond traditional link prediction methods by incorporating influential paper detection into the analysis framework. In addition to forecasting future connections between articles, we aim to identify papers that exert significant influence within the citation network. This involves

leveraging measures such as the number of citations, degree centrality, betweenness centrality, eigenvector centrality, and PageRank value to assess the impact of individual papers on the network structure.

In our comprehensive approach for link prediction and influential paper detection in citation networks, we start by extracting relevant node information, including titles, abstracts, author names, and citation relationships. We then employ feature engineering techniques to derive informative features that capture the essence of scholarly interactions and relationships within the network. These features encompass textual similarities between titles and abstracts, author co-occurrence patterns, and network properties such as centrality measures and preferential attachment.

With our feature space constructed, we train four distinct machine learning models: XGBoost, Logistic Regression, Decision Tree, and Random Forest. These models learn complex patterns and relationships within the citation network, enabling accurate predictions of future citations and collaborations. Additionally, they assist in identifying influential papers that play a pivotal role in shaping the network dynamics.

Evaluation of our approach is conducted rigorously using the F1-score metric, providing a holistic assessment of both link prediction and influential paper detection capabilities. By comparing model predictions against ground truth data, we gain insights into the efficacy of our methodology and its ability to capture underlying patterns in the citation network.

Through our study, we seek to demonstrate the effectiveness of combining machine learning algorithms with network analysis techniques for both link prediction and influential paper detection tasks in citation networks. By accurately predicting missing links and uncovering influential papers, our approach aims to advance our understanding of scholarly interactions and facilitate the discovery of novel research collaborations. Ultimately, our research endeavors to empower stakeholders with valuable insights into the dynamics of scholarly communication networks, enabling them to navigate the ever-expanding landscape of scientific knowledge more effectively.

## II. LITERATURE SURVEY

The task of link prediction in citation networks is fundamental for understanding the dynamics of scholarly communication, identifying emerging trends, and predicting future collaborations. In this literature review, we explore various approaches and insights provided by seminal works in this domain.

Cui, Wang, and Zhai (2010) introduced the concept of citation networks as multi-layer graphs and explored link prediction within this framework. Their approach directly addresses the task of link prediction in citation networks by proposing novel algorithms and metrics for identifying missing links. The advantage lies in capturing the multidimensional relationships between papers, enhancing the accuracy of link prediction models. However, a potential disadvantage could be the complexity of analyzing multi-layer graphs, which may require specialized expertise and computational techniques. (Cui et al., 2010)

Batagelj (2003) contributed to the field by introducing efficient algorithms for citation network analysis. While not directly addressing link prediction, Batagelj's work laid the foundation for computational tools and methodologies that enable subsequent research. The advantage lies in the development of efficient algorithms, facilitating large-scale analyses of citation networks. However, one disadvantage could be the potential complexity of implementing these algorithms, especially for researchers without a strong background in computational methods. (Batagelj, 2003)

Hummon and Doreian (1989) explored the connectivity within citation networks, demonstrating its relevance for link prediction. By examining the development of DNA theory through citation patterns, they highlighted the flow of ideas and its impact on future research directions. The advantage lies in the insights gained into the structural properties of citation networks, which inform link prediction strategies. However, a potential disadvantage could be the difficulty in quantifying and generalizing connectivity patterns across different research domains. (Hummon and Doreian, 1989)

Kajikawa et al. (2007) focused on creating an academic landscape of sustainability science through citation network analysis. Although not explicitly addressing link prediction, their thematic analysis approach offers advantages in identifying influential papers and thematic clusters within citation networks. This can inform link prediction by uncovering latent relationships between scholarly works. However, a disadvantage could be the subjectivity involved in defining and identifying thematic clusters, which may vary depending on the researcher's perspective. (Kajikawa et al., 2007)

Leicht et al. (2007) investigated the large-scale structure and temporal dynamics of citation networks. Their approach provides insights into how relationships between papers evolve over time, which is valuable for link prediction. The advantage lies in understanding temporal patterns in citation networks, allowing for more accurate predictions of future links. However, a potential disadvantage could be the complexity of modeling temporal dynamics, which may require sophisticated methodologies and computational resources. (Leicht et al., 2007)

Newman (2003) offered a comprehensive review of complex networks, including citation networks. While not specifically focused on link prediction, Newman's work provides theoretical foundations for understanding the structural properties of citation networks. The advantage lies in the conceptual framework provided for analyzing citation net-

works, which informs link prediction models. However, a disadvantage could be the abstraction of theoretical concepts, which may not always translate directly into practical methodologies for link prediction. (Newman, 2003)

While each approach offers unique advantages for link prediction in citation networks, researchers must carefully consider the associated disadvantages and challenges. By integrating diverse methodologies and insights from seminal works, scholars can advance our understanding of scholarly communication and knowledge dissemination through more accurate link prediction models.

### III. PROBLEM STATEMENT AND OBJECTIVES

The aim of this work is to predict if one paper has cited the other paper and also to detect the influential paper in the citation network. The objectives include:

- To analyze the various properties of the citation network graphs
- To extract the relevant features from the available data and compile all the extracted information into unified data for training.
- Develop models that accurately predict future links between papers in the citation network based on historical citation patterns
- Detection the influential paper in the citation network.

### IV. METHODOLOGY

#### A. Dataset

The dataset taken is a physics paper citation network of 27,770 papers with 352,807 edges. The data covers papers in the period from January 1993 to April 2003 (124 months). The Node information for each paper out of 27,770, contains the paper (1) unique ID, (2) publication year (between 1993 and 2003), (3) title, (4) authors, (5) name of journal (not available for all papers), and (6) abstract.

The training data contains instances of two node ids and a binary value representing the edge(citation) between the nodes. These node ids in turn refer to the paper unique ID data in the node information.

The test data consists instances of two nodes for which we have to predict the existence of a edge(citation) between them.

#### B. Network Analysis

We can see from Fig 1, the distribution of paper publications that the number of papers published is increasing over the years and the distribution is almost exponential.

From the Node information of the network, we have observed that there are 284 unique journals in which the papers are published. The Fig 2 shows the top journals which contributed to the 95% of paper publications. From this we can say that any new paper coming in the network has a high probability of publishing in these journals.

We have observed that there are 16,350 unique sets(groups) of authors, with the highest collaboration among



Fig. 1: Number of Publications over the years

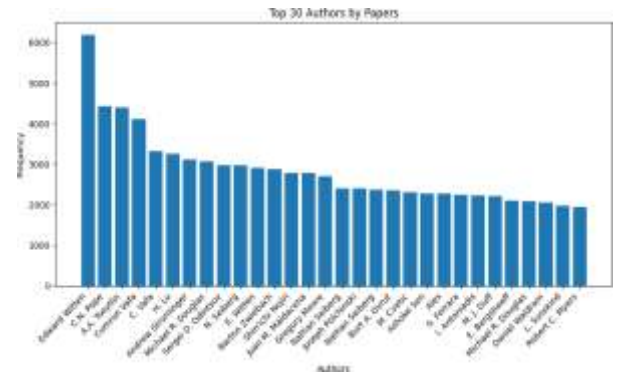


Fig. 4: Top 30 Authors according to number of papers

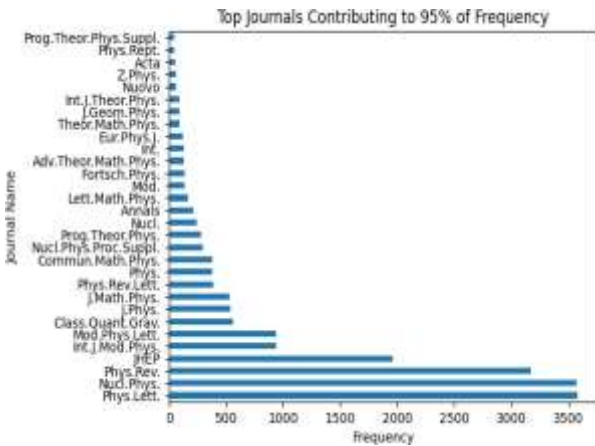


Fig. 2: Top Publishing journals

Shin'ichi Nojiri, Sergei D. Odintsov	38
C.M. Hull	27
Edward Witten	26
J.S.Dowker	19
Hiroshi Nishino	19
P. D. Jarvis (University of Tasmania), G. Rudolph (University of	1
Andrea Quadri (Max-Planck-Institut fuer Physik - Munich)	1
Marcin P. Flak, Krzysztof A. Meissner	1
A. I. Karanikas, E. N. Ktorides	1
F.Delduc, E. Ivanov, S. Krivonoz	1

Fig. 3: Authors and number of papers

a group of authors is 38 papers. Utilizing the authors' groups across all papers, we identified a total of 11,959 unique authors. For these 11,959 authors, we examined the number of papers published by each author. Figure 4 presents the top 30 authors ranked by the number of papers they have published. These authors can be regarded as the most influential authors in the field.

We can generally say that a new coming paper cites to an old(existing) paper. Year gaps within a citation network can provide insights into various aspects of the scholarly communication process, research trends, and the dynamics of knowledge dissemination. So the year gap between the papers is be observed for the existing papers. Through Fig 5 we can see that most of the papers are citing the recent papers  $\leq 3$  years.

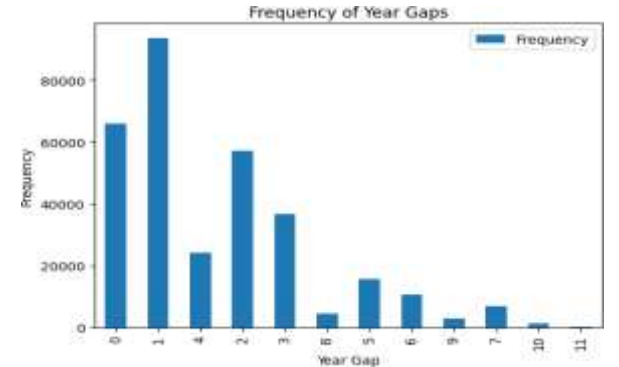


Fig. 5: Frequency the time gaps between cited papers

### C. Network properties

The average number of citations per each paper is seen as 25.4. Power Law Distribution: Also known as scale-free networks, most nodes have a low degree, but a few nodes, called hubs, have a very high degree. This phenomenon can be observed here.

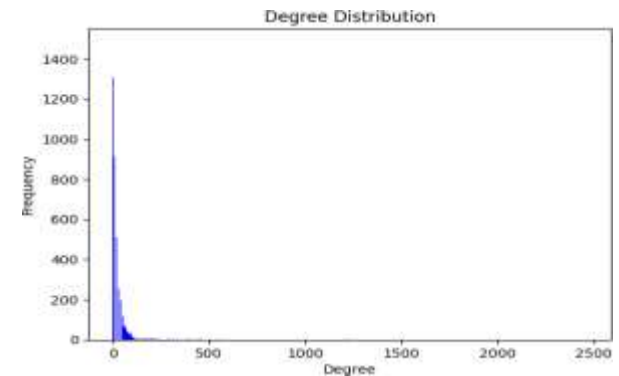


Fig. 6: Degree Distribution

The average shortest path length indicates that, on average, it takes approximately 9.8 steps to navigate from one node to another using the shortest path in the network. The average path length of approximately 3.27 is considerably shorter than the shortest path length. This discrepancy suggests the presence of a small-world phenomenon in the network.

TABLE I: Network Properties

S.No	Property	Value
1	Average degree	25.4
2	Degree distribution	Power Law
3	Path length - shortest, average, diameter	9.8, 3.2, 9
4	Geodesic path length (sample pair)	4
5	Clustering coefficient & average clustering coefficient (CC and Avg CC)	0.156
6	Number of Strongly Connected Components	20086
7	Number of Weakly Connected Components (WCC)	143
8	Giant component and coverage statistics	27400 nodes (98.67%)
9	Giant component properties shortest path, average path length, diameter, Avg CC	9.8, 3.27, 35, 0.18

The average clustering coefficient being relatively low suggests that the network's nodes have a moderate tendency to form local clusters or communities. However, compared to a fully connected or highly clustered network, the clustering level is relatively modest. The large number of strongly connected components indicates that the graph is highly fragmented, with many small clusters of nodes that are interconnected but not connected to the rest of the graph. The Giant network formed by the weakly connected components is the actual Giant network with 27400 nodes.

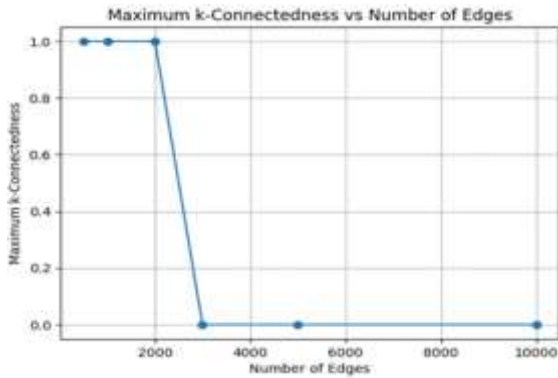


Fig. 7: K-Connectedness

The absence of k-connectedness beyond 2000 edges might indicate certain structural properties of the graph, such as increased complexity or a higher likelihood of disconnected components.

Degree, Betweenness, Closeness and Eigen vector centrality measure are calculated for the network for analysing the node importance. By leveraging these centrality measures collectively, we can gain insights into the structural importance of nodes within a network, identify key players, understand communication pathways.

Degree centrality measures how many connections a node has in a network. Nodes with a high degree centrality are highly connected to other nodes in the network. Nodes with a high degree centrality are often considered as influential or central in the network. The node with ID '9711200' is seen as the highest degree centrality node in this citation network graph. So this paper can be considered as the most influential paper among the citation network.

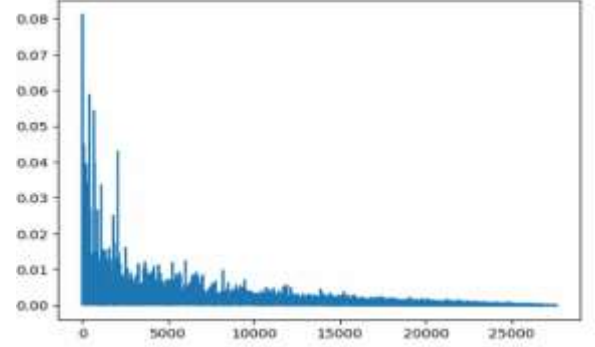


Fig. 8: Degree Centrality of nodes

Nodes with a high betweenness centrality are often considered as bridges or bottlenecks in the network, as they control the flow of information between different parts of the network. So the papers with high betweenness centrality value might be the bridge between two or more parts of network in the citation network.

Fig 10 shows the closeness centrality values of nodes. Nodes with high closeness centrality are considered central because they can reach other nodes in the network more quickly than nodes with lower closeness centrality.

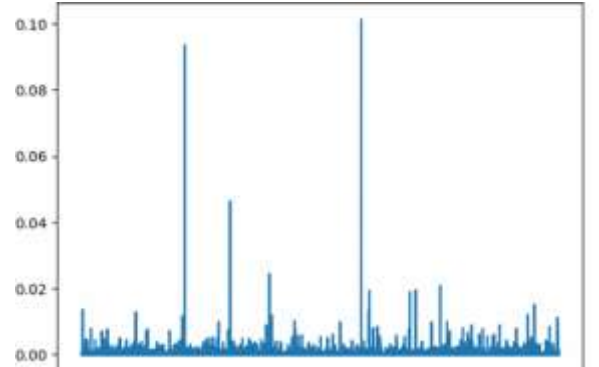


Fig. 9: Betweenness Centrality of nodes

Nodes with high eigenvector centrality are not only well-connected but are also connected to other nodes that are themselves important in the network. They represent nodes that have influence or prominence due to their connections to other influential nodes.

#### D. Data Preprocessing

Our methodology for link prediction in citation networks is a multi-step process designed to capture the intricate patterns

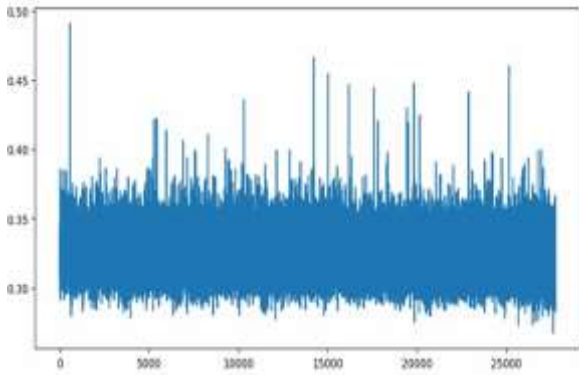


Fig. 10: Closeness Centrality of nodes

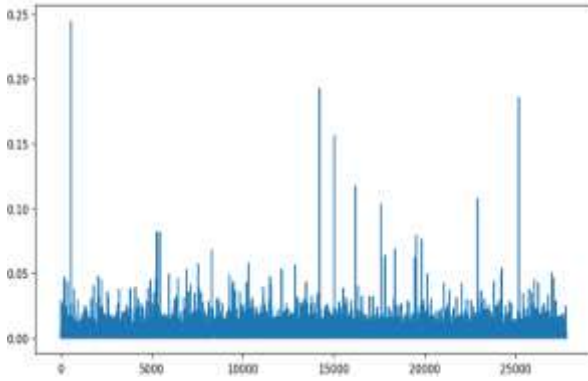


Fig. 11: Eigen Vector Centrality of nodes

of scholarly interactions and facilitate accurate predictions of future collaborations.

We first begin with preprocessing the data, we first convert the `node_information.csv` into `node_info` dataframe. Preprocessing involves text normalization techniques such as lemmatization, specifically applied to the 'title' and 'abstract' columns of the `node_info` dataframe. This step ensures consistency and enhances the quality of textual data, which is crucial for subsequent feature extraction.

#### E. Feature Engineering

As part of Feature Extraction we extract and engineer various features to enrich the training dataframe. Leveraging the `node_info` dataframe, we separate the node information such as publication years, author names, titles, journals, and abstracts for both source and target nodes. Furthermore, we delve into more intricate feature extraction, including year of publication differences, common authors between source and target nodes.

Based on TF-IDF for titles and abstracts, similarity measures are extracted. Additionally, we used the language model Roberta to extract embeddings for titles and abstracts, enabling us to gauge semantic similarity between nodes.

#### F. Graph Based Feature Extraction

Capturing Graph-based features are instrumental in the citation networks. Hence we computed several measures on

both directed and undirected graphs to encapsulate various network properties.

In directed graphs, we compute degree centrality measures such as in-degree, out-degree, and overall degree centrality for both source and target nodes. We also ascertain the core number for each node, indicative of its importance within the network. Preferential attachment, a measure of node attractiveness, is computed by evaluating the product of out-degree centrality for the source node and in-degree centrality for the target node.

For undirected graphs, we delve deeper into network analysis, calculating metrics such as the Jaccard index, common neighbors, Adamic Adar index, shortest path length, and target PageRank score. These metrics provide profound insights into the relational dynamics between nodes, facilitating a comprehensive understanding of the network's topology and connectivity patterns.

#### G. Data Modelling

The model is trained with the preprocessed and enriched dataset which consists of features such as year difference, common authors, common journal, title similarity, abstract similarity, title embeddings similarity, abstract embeddings similarity, source in centrality, source out centrality, target in central, source core number, target core number, preferential attachment for directed, jaccard index, preferential attachment for undirected, common neighbors, adamic adar, shortest distance between nodes, target pagerank.

As part of model training we employing an 80:20 split between training and testing data, we train four diverse classification models: XGBoost, Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. Each model is rigorously evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score to gauge its effectiveness in predicting linkages within citation networks.

#### H. Influential Paper Detection

The process of identifying influential papers within the citation network involves the summation of multiple metrics. **Number of citations** is used because papers with a high number of citations are often considered influential. Next **Degree centrality**, **Betweenness centrality**, **Eigenvector centrality** helps in identifying papers that are highly connected or influential within the network. and **PageRank** assigns a score to each paper based on the number and quality of citations it receives. Papers with higher PageRank scores are considered more influential. These metrics collectively contribute to assessing the influence of a paper within the network. All of these are used to compute a composite score which helps in identifying the Influence paper. (Farooq et al., 2018)

### V. EXPERIMENTAL RESULTS AND ANALYSIS

From the experimentation results, it is evident that the XGBoost classifier outperforms the other models in terms of accuracy, precision, recall, and F1-score. With an accuracy of 87.03%, precision of 87.76%, recall of 88.51%, and F1-score of 88.14%, XGBoost demonstrates robust predictive



TABLE II: Experimentation Results

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
XGBoost	87.03	87.76	88.51	88.14
Logistic Regression	58.62	58.37	83.55	68.73
Decision Tree Classifier	82.54	84.02	83.87	83.95
Random Forest Classifier	87.03	88.04	88.14	88.09

capabilities in identifying linkages within citation networks. This superior performance can be attributed to XGBoost's ability to handle complex datasets, effectively capturing intricate patterns and relationships among features.

In contrast, Logistic Regression exhibits relatively lower performance, achieving an accuracy of 58.62%, precision of 58.37%, recall of 83.55%, and F1-score of 68.73%. Despite its simplicity and interpretability, Logistic Regression struggles to capture the non-linear relationships inherent in citation networks, leading to suboptimal predictive performance.

The Decision Tree Classifier and Random Forest models deliver comparable results, with Decision Tree Classifier achieving an accuracy of 82.54%, precision of 84.02%, recall of 83.87%, and F1-score of 83.95%, and Random Forest achieving an accuracy of 87.03%, precision of 88.04%, recall of 88.14%, and F1-score of 88.09%. Both models leverage ensemble learning techniques to enhance predictive accuracy by aggregating the predictions of multiple decision trees. However, Random Forest exhibits slightly better performance due to its ability to mitigate overfitting and reduce variance, resulting in more stable and reliable predictions. Refer Table II

Overall, the experimentation results reveals that XGBoost performs well in link prediction tasks within citation networks. Its robustness, coupled with various important network features positions XGBoost as a formidable tool for advancing research in scholarly interactions and facilitating the identification of potential collaborations in academic domains.

In our analysis of the citation network, we observed that the metrics used to quantify the influence of papers exhibit a wide range of values. Specifically, the citation count ranged from 0 to 2294, while degree centrality varied from  $3.61e-05$  to  $0.0847$ . Similarly, betweenness centrality ranged from  $0.0$  to  $0.1043$ , eigenvector centrality ranged from  $2.22e-38$  to  $0.2527$ , and pagerank count ranged from  $0.0066$  to  $1.10e-05$ . Finally the Influential Composite score ranged from  $4.71e-05$  to  $2294.10$ . Utilizing these metrics, we identified the top three influential papers within the citation network. These papers, with their corresponding Influence Scores, are as follows: 1. Paper ID: 9711200, Influence Score: 2294.101; 2. Paper ID: 9802150, Influence Score: 1674.069; and 3. Paper ID: 9802109, Influence Score: 1560.065.

## VI. CONCLUSIONS AND FUTURE WORK

In this study, we proposed a comprehensive approach for link prediction in citation networks, combining machine learning algorithms with network analysis techniques.

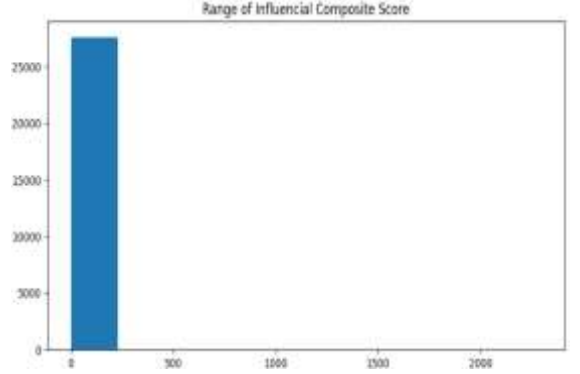


Fig. 12: Range of Influential Composite Score

Our experimentation results demonstrate promising performance across multiple models. XGBoost and Random Forest achieved the highest accuracy and F1-score, indicating their effectiveness in predicting future collaborations in scholarly interactions. However, Logistic Regression showed lower performance, suggesting the need for further investigation into its feature representation and model optimization. Decision Tree Classifier also exhibited competitive results, highlighting its potential in link prediction tasks. Through feature engineering and graph-based measures, we integrated various aspects of node information and network properties, contributing to a more nuanced understanding of scholarly interactions.

Apart from the work done, Next we can explore Dynamic Link Prediction and extend the approach to handle dynamic changes in citation networks over time. Developing models capable of adapting to evolving network structures and temporal dynamics could enhance the accuracy of future collaboration predictions. Techniques such as recurrent neural networks (RNNs) and graph neural networks (GNNs) may be suitable for capturing temporal dependencies and evolving relationships between scholarly entities.

## REFERENCES

- Batagelj, V. (2003). Efficient algorithms for citation network analysis. *arXiv preprint cs/0309023*.
- Cui, J., Wang, F., and Zhai, J. (2010). Citation networks as a multi-layer graph: Link prediction and importance ranking. *CS224W Project Report*.
- Farooq, A., Joyia, G. J., Uzair, M., and Akram, U. (2018). Detection of influential nodes using social networks analysis based on network metrics. In *2018 international conference on computing, mathematics and engineering technologies (icomet)*, pages 1–6. IEEE.

- Hummon, N. P. and Dereian, P. (1989). Connectivity in a citation network: The development of dna theory. *Social networks*, 11(1):39–63.
- Kajikawa, Y., Ohno, J., Takeda, Y., Matsushima, K., and Komiyama, H. (2007). Creating an academic landscape of sustainability science: an analysis of the citation network. *Sustainability Science*, 2:221–231.
- Leicht, E. A., Clarkson, G., Shedden, K., and Newman, M. E. (2007). Large-scale structure of time evolving citation networks. *The European Physical Journal B*, 59:75–83.
- Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.