A Machine Learning Model for Predicting Collaboration in Regional Co-Authorship Network

Jyoti Dua[1,*], Vivek Kumar Singh[2]

[1]Department of Computer Science and Engineering, Pranveer Singh Institute of Technology, Kanpur, Uttar Pradesh, INDIA.
[2]Department of Computer Science, Delhi University, Delhi, INDIA.

ABSTRACT

Scientific Research collaboration is one of the strengths in the research ecosystem due to its advantages in productivity and citation. Co-authorship network is one of the methods to analyze and evaluate the emerging research collaborations. Collaboration between pair of authors for the first time plays a vital role as the key to success for their collaboration in future. In this context, a focus on SAARC is highly justified, as fostering intra-regional scientific collaboration could help address shared challenges such as public health, climate change, and sustainable development, which demand collective scientific expertise. Therefore, the objective of this paper is to build a machine learning model for predicting new potential authors within South Asian Association for Regional Cooperation (SAARC) region who never collaborated for the last 20 years (2001-2020) using data from Web of Science (WoS). The co-authorship network was analyzed between two authors using structural and semantic similarities to predict whether the collaboration will happen in future or not. A proposed Meta-Learner Binary Classifier model is applied to the link prediction predictors after data pre-processing. The result shows structural and semantic features are good features to predict potential collaborators with 0.87 AUC before sampling and 0.99 AUC after sampling.

Keywords: Meta-Learner Binary Classifier Model, Co-authorship Network, Link Prediction, Machine Learning.

Correspondence:
Jyoti Dua
Department of Computer Science and Engineering, Pranveer Singh Institute of Technology, Kanpur-209305,
Uttar Pradesh, INDIA.
Email: jyotidua1984@gmail.com
ORCID: 0000-0002-4293-9062

Received: 12-08-2025;
Revised: 09-10-2025;
Accepted: 24-12-2025.

INTRODUCTION

Research collaboration is defined as two or more researchers working together on the same problem to achieve a common objective.[1] The research collaboration can be at an individual, country, institutional or disciplinary level. Research problems are now becoming multi-disciplinary and require multidisciplinary approaches to solve. Therefore, scientific collaboration networks are needed for solutions to social, political, economic, and technological problems. While collaborating, researchers share tacit knowledge, ideas, enhance productivity, and improve quality of research. However, establishing and maintaining the research collaboration is not an easy task. Therefore, researchers should find potential collaboration that succeeds in future.[2] During this process, researchers may face different uncertainties to find suitable and potential partners.

Research collaboration can be represented by graph or network. The collaboration graph refers to a finite non-empty set of vertices and finite set of edges. For instance, Newman[3] analyzed scientific collaboration networks, highlighting their structure and dynamics. He also studied co-authorship networks in various scientific areas and demonstrated how these networks display properties such as clustering and small-world phenomena. Barabási *et al.*,[4] introduced the concept of scale-free networks, exhibiting that many real-world networks, including collaboration graphs, have a small number of highly connected nodes, which play a vital role in the network's structure and connectivity.

The vertices are referred as actors, nodes or entities who are participants of that network such as authors, country, journal etc.,[5] whereas an edge is a connection between two actors showing collaborative relationship between them such as, authors in same paper, affiliated to the same organization etc. or based on their similar interest such as publishing papers on common research interest. Further, these collaboration networks are used to get insight into the collaboration in science. Thus, predicting associations between nodes or actors in a network is critical in network analysis.

Liben-Nowell and Kleinberg,[6] defined Link Prediction as the process of identifying the existence of a link between two nodes in a network by analysing the past collaboration or similarities. In a collaboration network there are possibilities that two authors

DOI: 10.5530/jscires.20251426

Publishing Partner : Manuscript Technomedia. [www.mstechnomedia.com]

might collaborate in future even if they never collaborated. Network topology could be used to find a considerable number of new potential/ prospective collaborations.

The present study focuses or centres around South Asian countries, namely Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka, which are all part of the SAARC intergovernmental organization. Under-development, population, poverty, and environmental degradation are some of the joint problems existing in SAARC nations.[7] Thus, the aim of the SAARC formation is to strengthen economic, cultural and social development and to work together as a team in scientific research and development. But, the South Asian Association for Regional Cooperation (SAARC) has encountered multiple challenges that have significantly reduced intra-collaboration among its member countries. One of the foremost challenges could be the ongoing geopolitical tensions. This is intensified by regional rivalries which further complicate collaboration. Furthermore, SAARC nations often prioritize their individual national interests, especially in economic and security issues, undermining regional initiatives, and favouring bilateral agreements over multilateral cooperation. Economic incorporation has also proven difficult due to persistent trade barriers, tariffs, and differing economic priorities preventing the formation of a common market. Political instability, marked by regime changes, internal conflicts, and security challenges in countries also hampers consistent regional cooperation. Poor infrastructure, including underdeveloped transport, energy, and communication networks, further restrict trade and investment flows, while significant disparities in economic development create difficulties in aligning the interests of more and less developed countries. Furthermore, varying commitment to SAARC, joined with weak enforcement mechanisms and a slow decision-making process, hinders the organization's effectiveness in achieving its goals. Internal security concerns, comprising terrorism, ethnic conflicts, and insurgencies, also cause nations to focus on domestic issues rather than regional cooperation. SAARC's administrative weaknesses, characterized by limited resources and insufficient institutional support, exacerbate these challenges, leading to stagnation in regional collaboration. Finally, poverty, climate change, and migration, highlight the need for joint action, yet diverging national policies often result in uneven responses rather than unified efforts.

The recent study by Dua *et al.*,[8] revealed that intra collaboration between SAARC countries is almost 1%. Therefore, it is necessary to increase and promote collaboration between the SAARC countries.

The paper has been specifically framed around a significant research question: predicting future collaborations in the SAARC co-authorship network using machine learning. This objective addresses the gaps left by earlier research that didn't pay enough attention to predicting collaboration in this understudied regional context. To conduct the study, a wide range of literature has been drawn that emphasizes key aspects such as data imbalance, structural and semantic feature extraction, and advanced classification techniques. These studies helped shape our methodological choices and provide a comparative framework for findings. Moreover, by explicitly targeting challenges like dataset imbalance and low intra-regional collaboration rates, the paper attempts to contribute actionable insights aligned with pressing regional and global research needs during 2001 to 2020.

Data

The eight countries that form the South Asian region are Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan and Sri Lanka. The publication records were downloaded for each of the member countries (SAARC) from the Web of Science (WoS). The data was collected over a period of 20 years (2001-2020). The search query employed was *CU='country_name' and LA='English' and PY=(2001-2020) and (DT='Article' or DT='Review')*, where CU field refers to country name, LA field refers to the language, PY field refers to publication year and DT to the Document Type. The country_name was replaced by the names of the eight South Asian countries one by one. The search was restricted only to Document types 'Review' and 'Article' as they incorporate the main research publications published in journals.

The downloaded research publication metadata for all the eight countries were combined, duplicate and erroneous records were removed based on DOI lookups, which resulted in a unique publications record. Further, intra-collaborated papers i.e., involving authors only within the SAARC region were retrieved. The intra-collaborated papers obtained in a total was 1,427.

METHODOLOGY

Co-Authorship Network Construction

The intra-collaborated records were analyzed to extract the author's full name using AF field and the co-authorship network was constructed with 4,827 nodes. The association between author pairs was recorded in the form of an edge table comprising of three columns. The first column presents the name of author1, the second column presents the name of author 2, and the third column presents a binary class label that is if there exists an edge between authors the class will be 1 and 0 otherwise. The number of connected edges obtained were 14,983 and unconnected edges were 11,63,2568. To test and validate/ evaluate the proposed link prediction algorithm some proportion (say, 30%) of already existing links were removed from the network provided that the graph does not get disconnected. After removing the 3,209 links, the new authorship network resulted in 11,774 connected and 11,63,5777 disconnected edges or links

Link prediction Approaches

Now, as we are intending to predict future potential collaboration, we focus on authors who never collaborated in a specified period of analysis. In our analysis, we focus on the information present or can be retrieved from the collaboration network. For that, we assume, firstly, the authors will collaborate if they have never collaborated and secondly, the authors are related or similar. Therefore, to determine the similarities between authors we apply a link prediction approach. The similarity scores between two authors can be determined in homogeneous or heterogeneous collaboration network. In homogeneous collaboration network the similarities can be determined by author's structural properties (say, distance) such as, Jaccard Similarity, Adamic-Adar, Common neighbors etc. On the other hand, in heterogeneous collaboration network the similarities between authors do not solely depend on structural properties but also on semantic similarities such as keywords, references, journal, conferences, etc.

This paper attempts to predict unforeseen linkages between SAARC authors by exploiting the collaboration network. To begin with, similarity scores were computed for potential author pairs. Various structural and semantic similarity measures over unconnected edges such as, Jaccard Similarity, Adamic-Adar, Common Neighbors, Preferential Attachment, Journal Similarity, and Keyword.

Machine Learning Classifiers

Moreover, these similarity scores were fed as features to machine learning algorithms. Before applying Machine Learning algorithms, the author's column was removed as machine learning algorithms are applied on numerical or categorical features. It is also important to check the multi-collinearity property for dependency, therefore, VIF strategy is used to check whether one feature is dependent on another feature or not. Finally, the features selected for predicting future links were based on Jaccard Similarity, Preferential Attachment, Journal Similarity and Keyword Similarity. Thus, removing dependent features and duplicate values resulted into 12,431 records. Further, MinMaxScaler () function is used to tune the feature values between 0 and 1 to avoid the biasness towards large values.

After the pre-processing the features, it was observed that the class labels were highly imbalance. The number of unconnected authors (0) were more than the number of connected authors (1). Initially, the number of negative labelled samples (0) was 11,285 and positive labelled samples (1) was 1,146. Therefore, the model may show biasness during classification process. Hence, the SMOTE technique is applied to oversample the minority class (1) with 11,285 samples to make it equal to the majority class (0). The data was split in two sets i.e. training and testing sets into 70:30 ratio which resulted into 8,701 train data and 3,730 test data before sampling whereas 15,799 and 6,771 into train data and test data respectively after sampling. Finally, the input features were fed into the supervised machine learning models.

Meta-Learner Binary Classifier

There are different weak/base learner models in machine learning such as SVM, KNN, DT, etc., which can produce different predictions for the same data. So, there should be a model that take predictions made by the base models as input and provide a smooth interpretation. Therefore, we apply a *Meta Learner Binary Classifier* presented in Figure 1 that is designed to improve model performance and predict the actual links efficiently. Meta Learner Binary classifier is used as stacking method in machine learning. The model improves predictions for the future by integrating different weak learners with Meta learners by ensemble them in parallel. In other words, *Meta Learner Binary Classifier* algorithm takes the output of base models as input and attempts to learn how to best combine the input prediction using meta-model to make a better output prediction. The paper uses KNN, SVM, Decision Tree, and XGBoost as base models and Logistic Regression as Meta model and applied 5-fold-cross-validation for the *Meta-Learner Binary Classifier.*

Related Work

There are two approaches widely used in link prediction which could be considered based on the type of information from the network and how they learn from existing relationships among the nodes to predict a non-existing link. Various authors have studied and implemented these approaches from different perspective. The first approach is featuring extraction-based methods which extracts and analyses the attributes from a network using similarity between nodes. For instance, in the collaboration network the node's attributes include research interest, conference venue, journals, keywords etc., for the collaboration network. A similarity function calculates similarity scores between two nodes. Greater the similarity score, greater the probability to form link between the two nodes. The study performed by Bhattacharyya, Garg and Wu[9] determined similarities among nodes for predicting link using keyword in a friendship network. The two major findings of their study were that correlations between direct friends are high regardless of number of hopes they are with each another, and individuals who are already friends has a higher similarity score than any other individual pair of users in the friendship network. Similarly, Akcora, Carminati and Ferrari[10] considered both network and profile similarity to predict link between the two nodes of social network (Facebook, YouTube, DBLP, and Epinion). The study intends to answer how graph size affect performance. Anderson *et al.*,[10] used user's interest overlap for measuring similarity. The authors estimated the similarity using two different features: distance metric to find similarity of interest overlap in the types of content and evaluating similarity of social ties in the set of people. The experiment was performed on three datasets: Wikipedia, Epinions, and Stack Overflow. The authors called former feature as tag similarity and later as social similarity.[11]

The similarities among the nodes for link prediction can be measured using graph topological properties including structural properties. There are two widely used graph topologies: local and global. In a research collaboration network node in a graph prefer to form new link with the node closer to it than the node far away in the network. Many different local topology-based metrics were designed in several studies. Some of the popular metrics are: Common neighbors, Preferential Attachment, Jaccard Coefficient, Resource Allocation, Adamic/Adar.[12-15] Global similarity index method also considers structural information for link prediction. The index includes katz index,[16] Shortest Path,[6] SimRank, Rooted PageRankHitting Time.[17-19]

The graph topology method in measuring similarity were used by different authors, Pavlov and Ichise,[2] analysed a Japanese co-authorship network by extracting structural properties such as Jaccard Similarity, Shortest Path, Common neighbours, etc, for predicting new collaboration among nodes. The study suggested/attributed structural properties as valuable information source finding collaboration. Topological and structural properties (clustering index, shortest path) and semantic features (keyword, Paper Titles and abstract) was jointly applied in study[20] on scholarly database Elsevier, BIOBASE, and DBLP in Computer Science and Biology fields. The authors performed experiments with the features and applied machine learning approaches for predicting links. In the analysis of the DBLP database, Sachan and Ichise[21] analyzed the structure of the network to increase the accuracy of the predictors by introducing a semantic approach and an event-based method. They employed an event-based methodology to more precisely identify potential cooperation by considering shared venues and journal information. They employed a variety of non-structural and event-based characteristics, such as common conference locations, common journals, and common terms in titles. To investigate the co-author relationship prediction using DBLP bibliography data, Sun. Y. *et al.*,[22] suggested the Path predict model. Here, the authors have considered topological features based on a meta-path to measure the similarity between author nodes and determine the significance of each in terms of predicting future author collaboration.

Probabilistic and Maximum-likelihood models optimize an objective feature comprised of many parameters. These models use mathematical techniques to design a model that matches the system and predict model parameters. Further, the resultant parameters are used to calculate likelihood of formation of non-existing links. Wang, Satuluri and Parthasarathy,[23] proposed a local probabilistic graphical method to estimate co-occurrence probability of nodes. These probability measure captures that information that are not captured by topological and semantic similarities. A Markov Random Field, local probabilistic graph model was proposed to compute co-occurrence probability. Clauset *et al.*,[24] designed a model that understands the topological relationships and predict the connections among the nodes in the hierarchical structure. The hierarchical structure was denoted by dendrogram in which similar connected pairs of nodes have the lowest common ancestors, which are lower in the tree than those of more distantly related pairs. The authors used three networks: Terrorist, metabolism and Grassland species and combined a maximum likelihood approach on all possible dendrograms, then averaged the corresponding probability to determine the mean probability over the sample dendrograms.

The second strategy makes use of feature learning-based approaches such as, random walks, matrix factorization, and neural network-based techniques. These learning strategies are based on the concept of learning a mapping that embeds nodes or complete (sub)graphs as points in a low-dimensional vector space.[25] The first learning method is matrix factorization which is represented in the form of row and column. In these approaches, vector representations of nodes corresponding to initial network are obtained by representing them in a low-dimensional space. The aim of such approaches is to reduce the dimensionality of this space while maintaining non-linearity and localization. Singular value decomposition and non-negative matrix factorization were the two methods used for matrix factorization. Laplacian Eigenmaps, Graph Factorization, GraRep are a few matrix-factorization techniques.[26-29] Node properties like node centrality and node similarity are investigated by

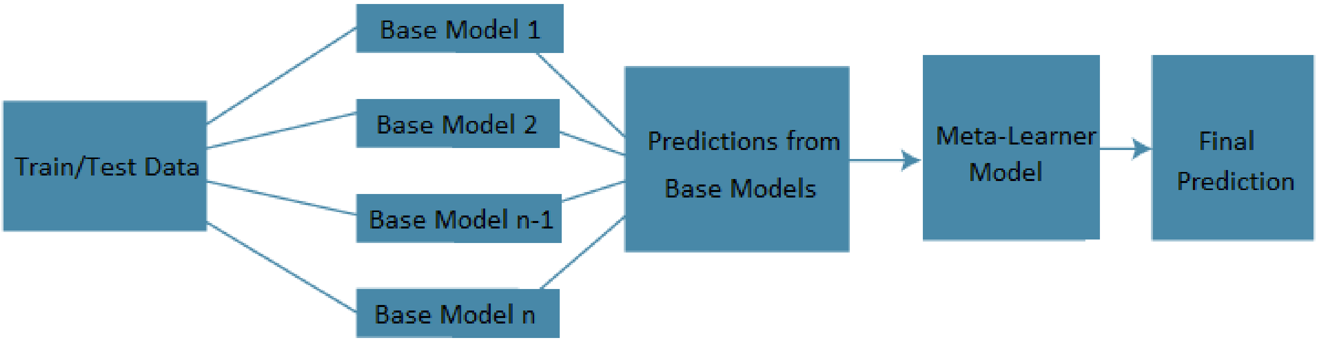

Figure 1: Architecture of Meta-Learner Binary Classifier.

Table 1: The performance of Meta-Learner Binary Classifier before and after sampling.

Base Models									
Model	Connected (Class Label 1)			Unconnected (Class Label 0)			Macro Average F1-score	Weighted Average F1-score	Accuracy
	P	R	F1-score	P	R	F1-score			
KNN (BS)	0.78	0.60	0.68	0.96	0.98	0.97	0.83	0.94	94.77
KNN (AS)	0.95	0.99	0.97	0.99	0.95	0.97	0.97	0.97	96.84
SVM (BS)	0.84	0.43	0.57	0.94	0.99	0.97	0.77	0.93	93.99
SVM (AS)	0.93	0.98	0.95	0.98	0.93	0.95	0.95	0.95	95.24
DT (BS)	0.93	0.64	0.76	0.96	1.00	0.98	0.87	0.96	96.25
DT (AS)	0.92	1.00	0.96	1.00	0.91	0.95	0.96	0.96	95.53
XGB (BS)	0.86	0.72	0.78	0.97	0.99	0.98	0.88	0.96	96.33
XGB (AS)	0.94	0.99	0.96	0.99	0.93	0.96	0.96	0.96	96.15
Meta-Learner Model									
Logistic Regression (BS)	0.82	0.74	0.78	0.97	0.98	0.98	0.88	0.96	96.76
Logistic Regression (AS)	0.96	0.98	0.97	0.98	0.96	0.97	0.97	0.97	96.87

graph exploration and sampling using random walks or search methods. The random walk-based methods include DeepWalk, Node2Vec, MetaPath2Vec, Graph SAGE, Watch Your Step (WYS), PathSim.[26-35] The Neural Network Based methods include strategies such as Graph Auto-encoders, Large Scale Information Network Embedding (LINE), Deep Neural Networks for Learning Graph Representations, Structural Deep Network Embedding (SDNE).[36-39] Makarov and Gerasimova[40] investigated the problem of predicting collaborations in co-authorship networks using a regression machine learning model. The task involves weighted edges connecting authors, formed by storing research papers. The model is evaluated on large AMiner co-authorship networks and the National Research University Higher School of Economics dataset. Results show better performance for the regression task on both networks. Resce, Zinilli, and Cerulli[41] examined the roles of network and non-network attributes contributing to the development of European university collaborations from 2011 to 2016 using four machine learning predictive algorithms. Results show that link formation accuracy is over 80%, public funding is crucial in Physical and Engineering Sciences (PE), Life Sciences (LS), network attributes count more than non-network attributes, and feature-importance scores differ across different scientific communities. Hasanzadeh, and Ghassemi[42] proposed a novel approach to temporal link prediction in dynamic networks, focusing on specific dynamics of each node rather than overall network dynamics. The approach improves accuracy and explainability in predicting future connections, with experimental results showing a 17.34% improvement in future collaboration efficacy in co-authorship networks. The approach also offers an interpretable layer over traditional methods.

RESULTS AND DISCUSSION

Initially, the structural and semantic features were extracted from the collaboration network. Table 1 shows the performance of different classification models evaluated on testing samples of author pairs to find relationship among them that might occur in future. Along with KNN, there are other models used such as, SVM, DT, and XGBoost. For test data, the proposed model's accuracy values are in the range 93% to 96% and 95% to 96% before (unbalanced) and after (balanced) sampling respectively. On accuracy, XGBoost model performs well with an accuracy of 96.33% before sampling. According to the Table 1, Except decision tree, all base models that were tested have precision of greater than 0.90 after sampling whereas less than 0.90 before sampling. The Meta classifier shows precision greater than 0.95 after sampling. This means that the features chosen to have a high discriminating capacity. Beside accuracy and precision, we could consider F1 score which is the measure of harmonic mean of precision and recall. On the other hand, the proposed model (Meta learner binary classifier) takes prediction of base models as input and yield an output from these predictions. Thus, improving the performance.

The results in Table 1 are consistent with existing research on link prediction methods. Feature extraction-based approaches, as discussed in related work, leverage graph topological and structural properties to enhance prediction accuracy. For instance, Bhattacharyya *et al.*,[9] and Akcora *et al.*,[10] highlight the

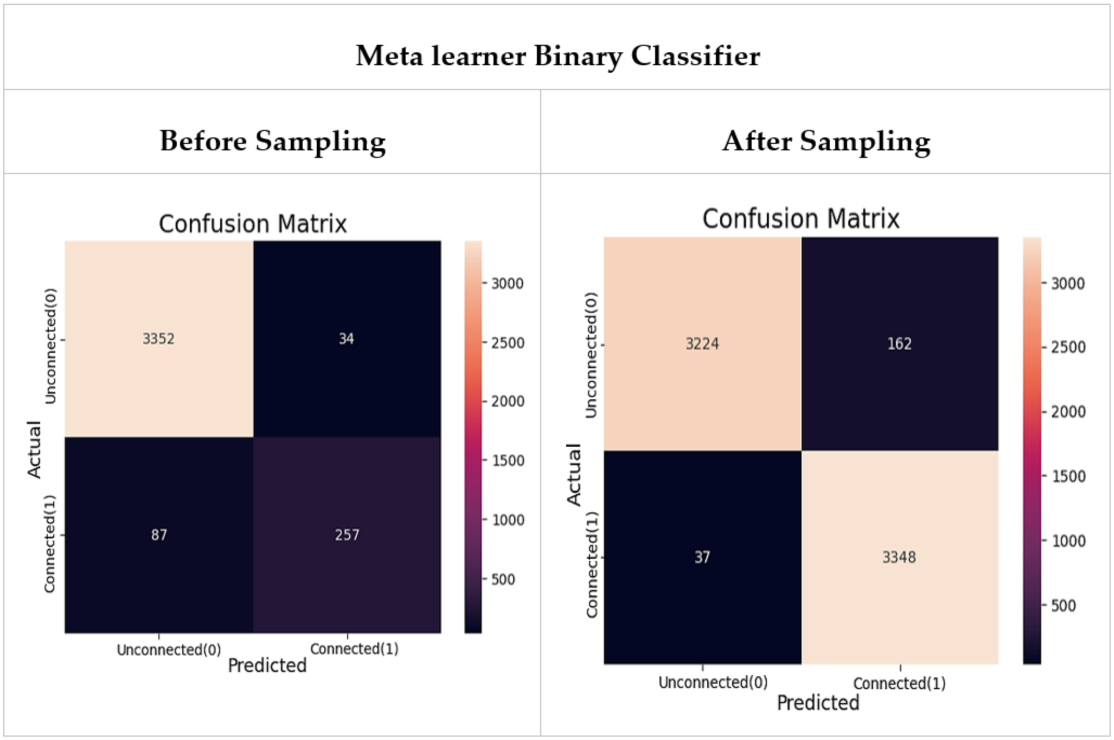

Figure 2: Confusion matrix of Meta-Learner Binary Classifier: Comparison of Performance Before and After Sampling.

use of structural features, such as node similarity and semantic attributes, which align with the features utilized in this research work. This research goes a step further by employing advanced classification models, such as XGBoost and the ensemble Meta-Learner, surpassing traditional probabilistic and machine learning methods like logistic regression and random walk-based techniques (e.g., DeepWalk, Node2Vec). Additionally, the integration of sampling techniques to address data imbalance and improve model performance complements recent advancements, such as the temporal link prediction model introduced by Hasanzadeh and Ghassemi.[42] The performance improvements highlighted in Table 1 also align with the growing adoption of ensemble learning methods to boost precision and recall, like the regression-based predictions in co-authorship networks by Makarov and Gerasimova.[40] Hence, Table 1 emphasizes the significance of balancing datasets and leveraging ensemble learning techniques for achieving high-accuracy link prediction, aligning closely with contemporary trends and advancements in related research.

The validation of a model is measured on recovery of actual links that were removed from the original graph. Therefore, Figure 2 represents confusion matrix before sampling and can be seen that out of 344 positive test data (class 1), 257 samples were predicted correctly. It was also found that only 34 author pairs were recommended to collaborate. On the other hand, after sampling, it can be observed that out of 3,385 positive test samples 3,348 predicted correctly and 162 author pairs were recommended.

The findings in Figure 2 align with patterns observed in related literature, emphasizing the key role of addressing data imbalance to enhance link prediction accuracy. While Bhattacharyya *et al.*,[9] and Akcora *et al.*,[10] focused on feature-based methods for calculating node similarity, they did not specifically tackle dataset imbalance, which can introduce biases in results. Makarov and Gerasimova[40] demonstrated that regression-based models performed more effectively when trained on comprehensive and well-structured datasets. Moreover, the performance improvements following sampling are consistent with the findings of Resce, Zinilli, and Cerulli,[41] who achieved over 80% link formation accuracy by applying machine learning models to datasets enriched with both network and non-network attributes. The introduction of balanced data in this paper resonates with the advancements of Hasanzadeh and Ghassemi,[42] who proposed a dynamic network approach that enhances prediction accuracy by targeting specific node dynamics. Hence, Figure 2 underscores the importance of balancing datasets to achieve high prediction accuracy, aligning with recent advancements in link prediction methodologies that emphasize robust data preprocessing and the application of advanced classification techniques.

ROC-AUC

AUC (Area under the Curve) value is typically used to assess the quality of link prediction algorithms using the Receiver Operating Characteristics (ROC) curve. Recall is also known as the True Positive Rate (TPR), which is defined as TP/ (TP+FN). The definition of False Positive Rate (FPR) is FP/TFP+TN). To better evaluate the performance at different classification thresholds

Figures 3a and 3b are plotted. Reducing the classification threshold increases the number of positive classifications, increasing the number of False positives and True positives. The optimal threshold that was identified before sampling (unbalanced data) and after sampling (balanced data) were 0.05 and 0.64 respectively. The AUC of the model before sampling (unbalanced dataset) was 0.87 but increases abruptly to 0.99 after sampling (balanced dataset). This indicates that model is performing best on the selected features for prediction and when data is balanced.

Compared to prior research, Bhattacharyya *et al.*,[9] and Akcora *et al.*,[10] emphasized that relying solely on similarity measures may not lead to optimal predictions unless data imbalance is addressed, as evidenced by their studies on social and collaboration networks. Similarly, Resce *et al.*,[41] pointed out that unbalanced datasets can compromise model sensitivity and recommended balancing strategies to enhance link prediction accuracy. The use of structural and topological features, as illustrated in Figure 3a, aligns with the work of Newman[3] and Barabási *et al.*,[4] where local and global graph metrics were employed for initial predictions. However, these methods also required improved processing to effectively handle class imbalances. While Figure 3a demonstrates satisfactory model performance, it underscores the importance of employing balancing techniques, as shown in Figure 3b, to enhance precision and recall across both classes. This

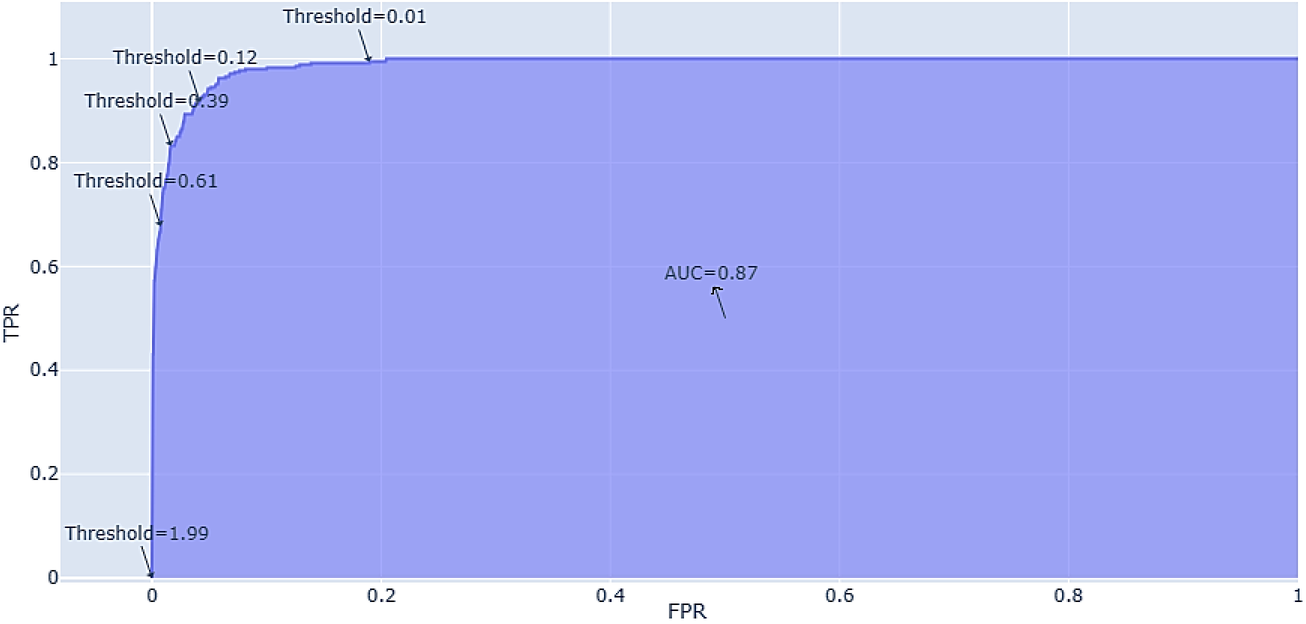

Figure 3a: ROC_AUC Performance of Meta-Learner Binary Classifier on unbalanced dataset (Before Sampling).

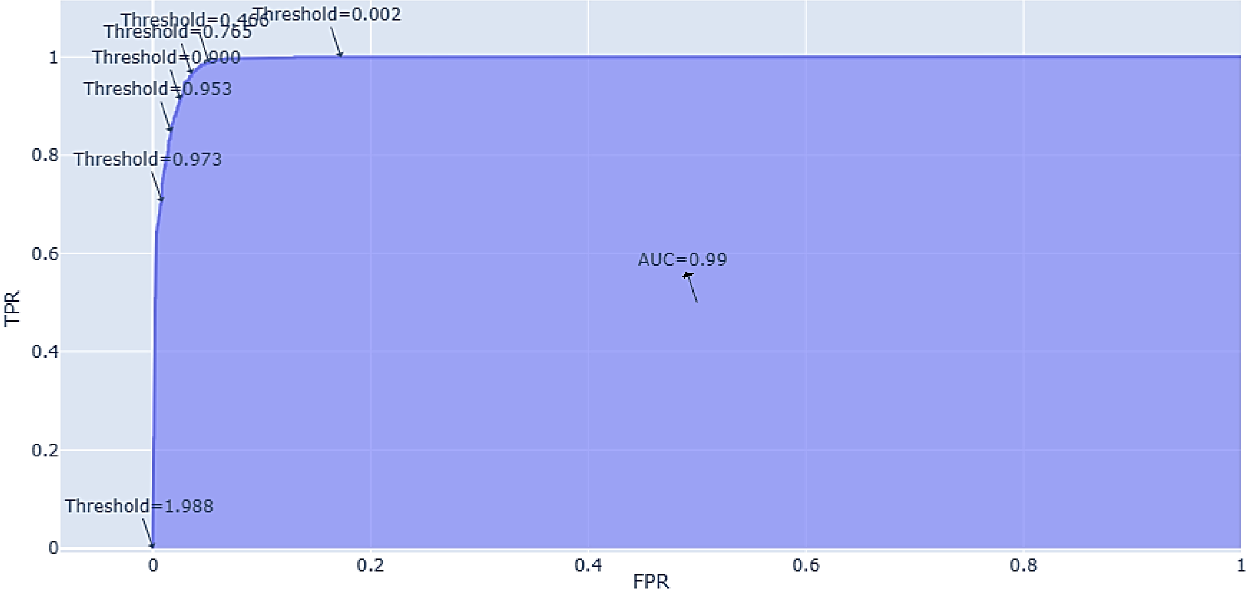

Figure 3b: ROC_AUC Performance of Meta-Learner Binary Classifier on balanced dataset (After Sampling).

approach aligns with best practices in the field for achieving more reliable link predictions. This approach aligns with the findings of Resce *et al.*,[41] who demonstrated that balancing datasets significantly improved link formation accuracy, achieving precision levels exceeding 80%. Similarly, Bhattacharyya *et al.*,[9] and Akcora *et al.*,[10] emphasized similarity-based methods but acknowledged that unbalanced datasets could compromise prediction reliability. In contrast, this paper integrates sampling techniques that parallels the dynamic link prediction strategies proposed by Hasanzadeh and Ghassemi,[42] who focused on node-specific dynamics within balanced datasets to enhance accuracy. By addressing data imbalance, Figure 3b highlights the benefits of combining ensemble methods with sampling strategies, representing a shift from traditional probabilistic and graph-based approaches, such as those by Newman[3] and Barabási *et al.*,[4] which relied primarily on local and global graph metrics without directly addressing class imbalances. Consequently, Figure 3b emphasizes the importance of balanced preprocessing in developing more reliable and robust link prediction methods.

CONCLUSION

The challenges faced by SAARC countries in terms of reduced intra-regional collaboration have significant practical implications for both researchers and policymakers. For researchers, geopolitical tensions, political instability, and limited infrastructure connectivity hinder opportunities for cross-border collaborations, knowledge sharing, and joint research projects. As a result, researchers are often forced to work in isolation, missing out on the potential for collective solutions to regional challenges like climate change, public health, and economic development. Policymakers, on the other hand, must navigate these complexities by creating frameworks that promote regional cooperation, address economic disparities, and foster trust-building initiatives. Practical measures, such as improving infrastructure, reducing trade barriers, and strengthening SAARC's institutional capacity, can create a conducive environment for collaboration. Furthermore, policymakers need to invest in joint research and development initiatives, ensuring that research outcomes are aligned with the region's collective needs, especially in areas like disaster management, poverty reduction, and regional security. Overcoming these challenges requires both research-driven insights and strong political will to implement reforms, paving the way for sustainable growth and cooperation in South Asia.

In this view, the paper presented a machine learning framework that utilizes structural and semantic features for link prediction within the SAARC co-authorship network, illustrating the effectiveness of the Meta-Learner Binary Classifier model. By mitigating data imbalance through SMOTE-based sampling, the model showed substantial performance enhancements, with the AUC value increasing from 0.87 to 0.99 after balancing. This highlights the key role of balanced data preprocessing in improving prediction accuracy. The proposed model emphasizes the value of integrating ensemble learning techniques to enhance precision and recall, offering a reliable and robust model for predicting potential collaborations in regional co-authorship network.

Despite these promising results, the research work has limitations. The relatively small dataset used may restrict the applicability of the results to larger or more complex networks. Additionally, the dependence on conventional machine learning models may limit the approach's scalability and adaptability to dynamic and heterogeneous datasets. The study's limitations primarily stem from the dataset's constraints and the scope of the methods employed. A key limitation is the relatively small dataset, which focuses on co-authorship data within the SAARC region from 2001 to 2020. While the dataset offers valuable insights into regional collaborations, its limited size and focus on a single region reduce the applicability of the findings to global contexts or other regional networks. Furthermore, the study relies heavily on structural and semantic features, which, though effective, may not fully capture the multifaceted factors influencing scientific collaborations, such as economic, social, or institutional dynamics. Although the methodological approach is robust, it also has its drawbacks. The machine learning classifiers and the Meta-Learner Binary Classifier yielded strong performance metrics, but their reliance on traditional machine learning methods may limit scalability for larger datasets or more complex networks. While the study addressed class imbalance using SMOTE, alternative techniques could further improve dataset representativeness, such as incorporating additional real-world factors like research funding, institutional affiliations, or geopolitical conditions.

To overcome these limitations, future research could expand the dataset to cover a broader timeframe and include data from other regions or global networks, enabling a more comprehensive understanding of co-authorship dynamics across diverse contexts. Advanced methodologies, such as deep learning models-particularly Graph Neural Networks (GNNs) or attention-based mechanisms-could also be explored to better capture intricate relationships in large-scale networks. Future studies could also integrate external socio-economic and geopolitical factors, such as disparities in research funding, institutional support systems, or cross-border researcher mobility, to develop a more holistic prediction model. Longitudinal studies examining the evolution of collaborations over time and testing model performance in real-world scenarios would also provide practical validation for these approaches. Moreover, the future work should address the ethical and practical considerations of predictive collaboration models to avoid reinforcing biases or inequities in the research ecosystem. A multidisciplinary approach involving collaboration with social scientists, policy experts, and domain specialists could ensure that these models are both equitable and impactful.

Future research will aim to overcome these limitations by applying the methodology to larger datasets and exploring advanced deep learning techniques, such as graph neural networks, to achieve more sophisticated and scalable link predictions. Further investigations could also incorporate temporal dynamics and additional non-structural factors, such as funding information and institutional affiliations, to improve the model's interpretability and broaden its applicability to diverse research collaboration scenarios.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

1. Katz JS, Martin BR. What is research collaboration?. Research policy. 1997;26(1):1-8.
2. Pavlov M, Ichise R. Finding experts by link prediction in co-authorship networks. FEWS. 2007;290:42-55.
3. Newman ME. The structure of scientific collaboration networks. Proceedings of the national academy of sciences. 2001;98(2):404-9.
4. Barabâsi AL, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T. Evolution of the social network of scientific collaborations. Physica A: Statistical mechanics and its applications. 2002;311(3-4):590-614.
5. Borgatti SP, Halgin DS. Analyzing affiliation networks. The Sage handbook of social network analysis. 2011;1:417-33.
6. Liben-Nowell D, Kleinberg J. The link prediction problem for social networks. In Proceedings of the twelfth international conference on Information and knowledge management 2003: 556-9.
7. Jha UC. Environmental issues and SAARC. Economic and Political Weekly. 2004: 1666-71.
8. Dua J, Lathabai HH, Singh VK. Measuring and characterizing research collaboration in SAARC countries. Scientometrics. 2023;128(2):1265-94.
9. Bhattacharyya P, Garg A, Wu SF. Analysis of user keyword similarity in online social networks. Social network analysis and mining. 2011;1(3):143-58.
10. Akcora CG, Carminati B, Ferrari E. User similarities on social networks. Social Network Analysis and Mining. 2013;3(3):475-95.
11. Anderson A, Huttenlocher D, Kleinberg J, Leskovec J. Effects of user similarity in social media. In Proceedings of the fifth ACM international conference on Web search and data mining 2012: 703-12.
12. Newman ME. Clustering and preferential attachment in growing networks. Physical review E. 2001;64(2):025102.
13. Jaccard P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. Bull Soc Vaudoise Sci Nat. 1901;37:241-72.
14. Zhou T, Lü L, Zhang YC. Predicting missing links via local information. The European Physical Journal B. 2009;71(4):623-30.
15. Adamic LA, Adar E. Friends and neighbors on the web. Social networks. 2003;25(3):211-30.
16. Katz S, Downs TD, Cash HR, Grotz RC. Progress in development of the index of ADL. The gerontologist. 1970; 10(1_Part_1):20-30.
17. Jeh G, Widom J. Simrank: a measure of structural-context similarity. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002; 538-43.
18. Chung F, Zhao W. PageRank and random walks on graphs. In Fete of combinatorics and computer science, 2010; 43-62. Berlin, Heidelberg: Springer Berlin Heidelberg.
19. Fouss F, Pirotte A, Renders JM, Saerens M. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. IEEE Transactions on knowledge and data engineering. 2007;19(3):355-69.
20. Al Hasan M, Chaoji V, Salem S, Zaki M. Link prediction using supervised learning. InSDM06: workshop on link analysis, counter-terrorism and security, 2006;30:798-805.
21. Sachan M, Ichise R. Using abstract information and community alignment information for link prediction. In2010 Second International Conference on Machine Learning and Computing 2010; 61-5. IEEE.
22. Sun Y, Barber R, Gupta M, Aggarwal CC, Han J. Co-author relationship prediction in heterogeneous bibliographic networks. In2011 international conference on advances in social networks analysis and mining 2011; 121-8. IEEE.
23. Wang C, Satuluri V, Parthasarathy S. Local probabilistic models for link prediction. In Seventh IEEE international conference on data mining (ICDM 2007), 2007; 322-31. IEEE.
24. Clauset A, Moore C, Newman ME. Hierarchical structure and the prediction of missing links in networks. Nature. 2008;453(7191):98-101.
25. Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. Advances in neural information processing systems. 2017; 30.
26. Belkin, M. and Niyogi, P., Laplacian eigenmaps and spectral techniques for embedding and clustering.*Advances in neural information processing systems*, 2001; 14.
27. Ahmed A, Shervashidze N, Narayanamurthy S, Josifovski V, Smola AJ. Distributed large-scale natural graph factorization. In Proceedings of the 22nd international conference on World Wide Web, 2013; 37-48.
28. Cao S, Lu W, Xu Q. Grarep: Learning graph representations with global structural information. In Proceedings of the 24th ACM international on conference on information and knowledge management 2015; 891-900.
29. Ou M, Cui P, Pei J, Zhang Z, Zhu W. Asymmetric transitivity preserving graph embedding. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining 2016; 1105-14.
30. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining 2014; 701-10.
31. Grover A, Leskovec J. node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining 2016; 855-64.
32. Dong Y, Chawla NV, Swami A. metapath2vec: Scalable representation learning for heterogeneous networks. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining 2017; 135-44.
33. Hamilton WL, Ying R, Leskovec J. Representation learning on graphs: Methods and applications. arXiv preprint arXiv:1709.05584. 2017.
34. Abu-El-Haija S, Perozzi B, Al-Rfou R, Alemi AA. Watch your step: Learning node embeddings via graph attention. Advances in neural information processing systems. 2018; 31.
35. Sun Y, Han J, Yan X, Yu PS, Wu T. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. Proceedings of the VLDB Endowment. 2011;4(11):992-1003.
36. Tran PV. Learning to make predictions on graphs with autoencoders. In2018 IEEE 5th international conference on data science and advanced analytics (DSAA), 2018; 237-45. IEEE.
37. Tang J, Qu M, Wang M, Zhang M, Yan J, Mei Q. Line: Large-scale information network embedding. In Proceedings of the 24th international conference on world wide web, 2015; 1067-77.
38. Cao S, Lu W, Xu Q. Deep neural networks for learning graph representations. In Proceedings of the AAAI conference on artificial intelligence, 2016;30(1).
39. Wang D, Cui P, Zhu W. Structural deep network embedding. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016; 1225-34.
40. Makarov I, Gerasimova O. Predicting collaborations in co-authorship network. In 2019 14th international workshop on semantic and social media adaptation and personalization (SMAP), 2019; 1-6. IEEE.
41. Resce G, Zinilli A, Cerulli G. Machine learning prediction of academic collaboration networks. Scientific Reports. 2022;12(1):21993.
42. Hasanzadeh Fard S, Ghassemi M. Temporal Link Prediction Using Graph Embedding Dynamics. arXiv e-prints. 2024 Jan:arXiv-2401.

Cite this article: Dua J, Singh VK. A Machine Learning Model for Predicting Collaboration in Regional Co-Authorship Network. J Scientometric Res. 2025;14(3):773-81.