IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.Doi Number

Hybrid GNN-based Link Prediction Model for Identifying Drug-related Organized Crime Groups on Twitter

Eun-Young Park[1], Hyeon-Woo Lee[2], and Jiyeon Kim[3]
[1, 2, 3]Department of Computer Engineering, Daegu University, Gyeongsan, South Korea

Corresponding author: Jiyeon Kim (e-mail: jyk@daegu.ac.kr).

This work was supported by 'Tech. Challenge for Future Program Policing([http://www.kipot.or.kr)'/]www.kipot.or.kr)' funded by Ministry of Science and ICT(MSIT, Korea) & Korean National Police Agency(KNPA, Korea). [Project Name : Development of Active Dark Web Information Collection, Analysis and Tracking Technology to Prevent Dark Web Crime / Project Number : RS-2023-00244362]

ABSTRACT In this paper, we propose a Graph Neural Network (GNN)-based link prediction model to analyze the interconnections of Drug-related Organized Crime Groups (DOCG) on X (formerly Twitter) and to identify their potential associations. To this end, we collected approximately one million tweets by employing slang terms commonly used for drug distribution and promotion. From these data, we constructed a knowledge graph in which tweets containing URLs to secure messengers such as Telegram and Snapchat, the accounts posting these tweets, and the hashtags included in them were modeled as nodes, thereby enabling the identification of accounts belonging to the same DOCG. In particular, to compare the impact of distinguishing node types (tweets, accounts, hashtags) on link prediction performance, we generated both homogeneous and heterogeneous knowledge graphs. For link prediction, we employed three GNN architectures: Graph Convolutional Network (GCN), Graph Attention Network (GAT), and Graph Isomorphism Network (GIN). We further developed single-model frameworks as well as hybrid models combining two architectures, with the objective of determining the most effective GNN for DOCG identification. Experimental results demonstrate that, across both homogeneous and heterogeneous settings, the GAT model achieved the highest F1-score among the single models, while the heterogeneous hybrid model combining GAT and GIN yielded the best overall performance. These findings indicate that GNN-based link prediction is effective in detecting latent associations among DOCG on social media platforms.

INDEX TERMS Graph Neural Network, Knowledge Graph, Link Prediction, Social Media, Cyber Crime

I. INTRODUCTION

Social media has become a central medium for information sharing, relationship building, and opinion formation in modern society. Platforms such as X (formerly Twitter), Facebook, Instagram, and Telegram exert broad influence due to their openness, immediacy, and anonymity, while at the same time giving rise to significant social risks. Social media has also been increasingly misused as a vehicle for cybercrime, with both the scale and modalities of such activities diversifying. Among these, the spread of drug trafficking through social media has emerged as a pressing social issue. Recent reports indicate that synthetic drugs such as LSD and MDMA are rapidly proliferating among young populations through dark web markets and social media spaces [1]. The National Drug Threat Assessment 2024 published by the U.S. Drug Enforcement Administration further reported that DOCG employ emojis, encrypted language, and private group functions to discreetly distribute drugs on social media, primarily targeting adolescents and young adults [2].

Social media–based drug trafficking demonstrates characteristics that differ markedly from traditional offline transactions: ease of access, guaranteed anonymity, and rapid dissemination of information. DOCG generate multiple accounts and exploit hashtags, URLs, and invitation codes to obscure their activities. They frequently employ slang, change accounts periodically, and redirect users to other platforms such as instant messengers, thereby complicating law enforcement monitoring. In particular, X is frequently exploited as a distribution channel because of its low posting barriers and relatively lenient regulatory environment. These

factors highlight the need for timely investigative responses and advanced analytical techniques.

Earlier studies on crime detection in social media have largely relied on static models such as analyzing individual posts, filtering keywords, or classifying content. However, DOCG operate not through single accounts or tweets but by constructing multilayered networks that connect multiple accounts, tweets, hashtags, and URLs. This evolution toward organized and sophisticated network-based activities calls for relational analysis models that can capture hidden associations among entities. In this paper, we propose a Graph Neural Network (GNN)-based link prediction model to analyze the internal structure of DOCG on X and to predict their potential inter-organizational connections. Specifically, we collected approximately one million drug-related tweets using slang terms associated with trafficking and promotion. From these data, we extracted tweets containing URLs that point to Telegram, Snapchat, and general websites. We then constructed a knowledge graph in which accounts, tweets, hashtags, and URLs were modeled as vertices and their interactions as edges.

For link prediction, we trained three GNN models: Graph Convolutional Network (GCN), Graph Attention Network (GAT), and Graph Isomorphism Network (GIN). In addition to single-model architectures, we developed hybrid models combining two GNNs, and compared their predictive performance to identify the most effective configuration for detecting latent DOCG associations. Unlike prior research that focused on analyzing the activity patterns of individual accounts or classifying tweet content using text mining and machine learning, the proposed model integrates drug-related tweets across multiple accounts, constructs a knowledge graph, and applies GNN-based link prediction to identify latent DOCG. We argue that our proposed model can assist law enforcement in capturing hidden organizational connections at an early stage and support proactive intervention.

The remainder of this paper is organized as follows. Section II reviews related work in social media analytics and cybercrime investigation. Section III details the dataset construction procedure, preprocessing steps, and the generation of the heterogeneous DOCG knowledge graph. Section IV presents the experimental results for both single and hybrid GNN models and analyzes their performance in the DOCG detection task. Section V provides a discussion of cross-domain validation results using an additional sex crime dataset and examines the generalizability of the proposed framework. Section VI concludes the paper and outlines directions for future research.

II. RELATED WORKS

Social media has become a critical platform for global information sharing and communication in modern society. Large volumes of user-generated content (UGC) and interaction data provide extensive information, and numerous studies have sought to analyze these data to understand and predict social phenomena. In particular, advances in graph-based analytical techniques and machine learning have contributed to the effective processing of social media data, enabling applications such as social network analysis, sentiment analysis, and the study of content diffusion patterns. In this section, we review prior research on the utilization and analysis of social media data, and examine studies focusing on the detection and analysis of criminal activities on social media, thereby situating the relevance of our work.

A. Studies on Non-crime-related Issues using Social Media

Social media continuously generates vast amounts of user-generated content, and the resulting big data have been widely utilized in fields such as social phenomenon analysis, policy development, and the study of online communication patterns. Prior research has focused on analyzing user behavior, identifying relationships within networks, and applying machine learning methods to extract meaningful insights from such data. Representative studies include the prediction of flood risks by integrating meteorological information with social media data [3], the use of Grove-model-based text embeddings with Graph Convolutional Networks for harmful content detection [4], and the optimization of social media marketing strategies for small and medium enterprises (SMEs) through the analysis of user response data combined with machine learning techniques [5]. Research has also examined science communication processes and user reactions on social media [6], as well as sentiment analysis with Heterogeneous Graph Neural Networks (H-GNN) for multimodal data fusion and mitigating missing information [7]. Further, studies have addressed digital content protection and copyright awareness [8], and investigated the influence of social media use on students' academic performance and emotional development [9]–[10].

Graph-based and deep learning approaches have also been applied in a variety of non-crime contexts. Rumor detection has been studied using GNNs with data augmentation [11], dual embeddings and TextGCN [12], and novel graph architectures [13]. Surveys of graph neural networks for time series [14] and dynamic link prediction [15] demonstrate the versatility of graph methods across domains. Sentiment classification has incorporated graph embedding techniques [16], while misinformation diffusion and fake news detection have been investigated through logic-based graph models [17] and graph-augmented transformer frameworks [18].

Overall, existing studies have primarily emphasized specific topics or user-level behaviors, often focusing on individual posts or content characteristics rather than organizational-level structures. In contrast, our study differentiates itself by employing knowledge graphs and GNN-based link prediction to identify latent DOCG networks, thereby addressing the structural aspects of organized activities that previous research has overlooked.

B. Studies on Crime-related Issues using Social Media

While social media has become central to everyday communication due to its accessibility, it has also been

increasingly exploited as a primary medium for online criminal activity. Criminal actors frequently utilize social media to plan offenses, conduct promotional activities, and expand organizational networks, with such cases increasing in prevalence. In response, diverse studies have sought to detect criminal behaviors by analyzing social media data. To situate our approach within this crime-focused literature, Table I contrasts representative studies along three generic dimensions—tasks (Classification/Organization Identification/Link Prediction), graph design (GNN; Heterogeneous; Multi-Relational), and methodology (Scope/Key Findings/Methodology). prior research has primarily addressed the legal, ethical, and educational issues of leveraging social media data for cyber investigations, as well as the development of forensic techniques for data collection and machine-learning–based models to identify criminal content.

Some studies have examined the legal and ethical considerations, as well as data authentication issues, related to the use of social media data in criminal investigations [19]. Others have analyzed police investigative strategies and case studies to assess the potential of social media as a tool for law enforcement [20]. Additional research has focused on the role of social media as a facilitator of crime and retaliatory violence [21], and highlighted the cybercrime risks associated with social media and the need for educational models for prevention [22]. From a technical perspective, digital forensic procedures and analytical methods have been proposed for collecting online activity evidence of offenders [23], and the contributions of machine learning and deep learning to social media forensics have been systematically discussed [24]. Further studies include data-mining–based methods for identifying criminal patterns [25], as well as models for early detection of cybercrime events using Twitter posts [26]. Other research has qualitatively analyzed seized online conversations among young offenders to illustrate how social media facilitates criminal organization in a stepwise manner [27], and graph-based approaches have also been attempted, for example applying intersection graphs to criminal networks for cybercrime mitigation [18]. In addition, more studies have emphasized graph-based and platform-specific methods. Investigations have examined how social media affordances shape drug dealer advertising practices [29], and official reports have detailed the growing threats of online drug trafficking via popular platforms [30]. Graph-based approaches such as CrimeGNN have been proposed for detecting community structures in criminal networks [31]. Comparative assessments of platform-level monitoring highlight differences in opioid-related surveillance across social media sites [32], and language- and region-specific illicit drug promotion has been examined in Thai-language posts on X [33]. From a methodological perspective, surveys have summarized privacy and adversarial risks in applying GNNs to social media [34], while new algorithms have improved fraud detection by integrating node features and structural information [35]. Forensic investigations leveraging artificial intelligence have also been explored [36], and broader surveys have mapped the techniques and attack trends in the emerging field of social cybersecurity [37]. In addition, deep learning–based community detection frameworks have been proposed to enhance the identification of covert organizations in online environments [38].

Overall, prior research has primarily addressed the legal, ethical, and educational issues of leveraging social media data for cyber investigations, as well as the development of forensic techniques for data collection and machine-learning–based models to identify criminal content. In contrast, our study identifies overlapping information across social media content, predicts account-level

TABLE I
COMPARISON OF REPRESENTATIVE WORKS ON SOCIAL-MEDIA CRIME ANALYTICS

Ref.	Classification	Community/ Organization Identification	Link Prediction	GNN	Heterogeneous Graph	Multi-Relational Edges	Scope	KeyFindings	Methodology
[4]	✓	✗	✗	✓	✗	✗	Troll detection (harmful-user classification) on social platforms	GCN-based features improve classification vs. text-only baselines	Deep feature extraction + GCN; supervised classification
[12]	✓	✗	✗	✓	✗	✗	Rumor classification with text graphs	Dual embeddings with TextGCN achieve competitive results	Text graph construction + GCN
[18]	✓	✗	✗	✓	✗	✗	Fake news classification (context-aware)	Hybrid BERT+GNN boosts performance by leveraging structure	Dual-stream Graph-Augmented Transformer (BERT + GNN)
[28]	✗	✓	✗	✗	✗	✗	Online criminal network analysis (community structure)	Intersection graphs reveal organization patterns	Graph analytics for community detection (non-GNN)
[31]	✗	✓	✗	✓	✗	✗	Community/organization discovery in criminal networks	GNN effective for community detection on crime graphs	GNN-based community detection
[35]	✓	✗	✗	✓	✗	✗	Fraud classification with graph signals	Combining topology + strong node features improves fraud detection	Supervised GNN with node/topology fusion
[36]	✗	✗	✗	✗	✗	✗	Forensic methods for social media investigations	Synthesizes investigative techniques; practice-oriented insights	Methodological/forensic review
[32]	✗	✗	✗	✗	✗	✗	Platform signals for opioid-crisis monitoring	Identifies platforms with stronger monitoring affordances	Empirical comparative assessment; public-health/forensic context
OURS	✓	✓	✓	✓	✓	✓	DOCG identification via link prediction over a heterogeneous social–content graph	Unifies organization detection and relation inference; operationally grounded	Hetero-GNN for link prediction with multi-relational schema; evaluation with standard metrics; tuning via Optuna; pipeline towards deployment

associations using GNN-based link prediction, and thereby contributes to the tracing of potential DOCG.

III. KNOWLEDGE GRAPH MODELING OF DRUG-RELATED TWEETS

In this section, we describe the construction of a dataset of drug-related tweets collected from X and present a modeling methodology for generating a DOCG knowledge graph. The graph is centered on tweets containing identical URLs, with accounts, hashtags, and URLs modeled as nodes and their connections represented as edges.

A. Construction of a Drug-related Tweet Dataset

To predict the relationships of Drug-related Organized Crime Groups (DOCG) within social media, it is essential to obtain account and tweet data where actual drug-related activities occur. In this paper, we collected tweets related to drug trafficking and promotion by selecting 40 drug-related terms and slang expressions, as listed in Appendix 1, and retrieved posts published within a 90-day window on X. This process yielded approximately one million tweets.

Due to the nature of keyword-based collection, however, identical tweets were often retrieved multiple times when they matched more than one keyword query. To address this, we performed a data refinement process to remove duplicate tweets collected through overlapping keywords. As a result, the final dataset consisted of approximately 470,000 unique tweets. Our dataset includes the elements summarized in Table II. These four elements are directly used as node types in the knowledge graph: N_{Tweet}, $N_{Account}$, $N_{Hashtag}$, and N_{URL}.

TABLE II
Elements of the Collected Tweet Dataset

Feature	Node	Description	Count
TweetID	N_{Tweet}	Unique identifier of a tweet	471,905
AccountID	$N_{Account}$	Unique identifier of the user who posted the tweet	169,334
Hashtag	$N_{Hashtag}$	Hashtags contained in the tweet	123,391
URL	N_{URL}	URLs contained in the tweet	40,151

Each TweetID serves as a unique identifier for a tweet and provides the basis for modeling its relationships with other elements such as hashtags and URLs. AccountID represents the unique identifier of the user who posted the tweet, and is used for analyzing account-level activities as well as network structures. Hashtag refers to the list of hashtags included in the tweet, contributing to topic classification and grouping of tweets. URL represents external links to websites or instant messenger channels, playing a critical role in identifying pathways directly associated with criminal activities, such as Telegram or dark web sites. The counts of hashtags and URLs represent unique values with duplicates removed. Furthermore, based on the number of tweets and accounts, the average number of posts per account was approximately three, indicating that a relatively small number of accounts were responsible for a large portion of the tweets.

For efficient analysis, all collected data were organized by account and stored in a database. Based on this structure, we computed account similarities and constructed networks by focusing on accounts sharing identical URLs. The resulting database was then utilized as the input for knowledge graph–based relational modeling and link prediction experiments, thereby ultimately yielding approximately 170,000 unique account records.

To filter out irrelevant information not associated with drug crimes, we preprocessed the data by analyzing the frequency distribution of URLs. As shown in Table III, the top seven URLs accounted for a disproportionately large share of occurrences; these were excluded to reduce noise and mitigate distortion in network construction.

TABLE III
HIGH-FREQUENCY URLS COLLECTED FROM SOCIAL MEDIA X

	URL	Description	Count
U_1	soci*****	Coca-Cola Christmas Eve official event page	9,601
U_2	soci*****	Coca-Cola Christmas official event page	7,702
U_3	pas*****	Pornographic content sharing site	4,664
U_4	zeal*****	Bitcoin-related community site	3,460
U_5	www.me****	French news website	2,758
U_6	blog.ip*****	Official blog of IPOR company	1,295
U_7	app.ip*****	Official application of IPOR company	1,295

When examining the overall frequency distribution of URLs, we observed that beyond the top seven entries, most remaining URLs corresponded to Telegram chat links (e.g., t.me/m***, t.me/E*****), which represent actual drug distribution channels. We therefore excluded only the top seven URLs.** In particular, the U_1–U_7 URLs were identified as advertising or promotional platform links with limited relevance to criminal activities. These links caused excessive clustering among multiple accounts sharing the same URL, which introduced two main issues. First, high-frequency URLs produced artificially strong connections among accounts with little or no actual criminal association. Second, the link prediction models tended to overfit these frequent URL-based connections, thereby reducing their ability to capture genuine relationships that indicate potential criminal associations. To mitigate such structural distortions and to ensure that meaningful relational patterns were preserved, we excluded the top seven URLs from the dataset.

B. Tweet-based Knowledge Graph Modeling

In this paper, we constructed a knowledge graph to analyze DOCG network structures on X and to identify potential associations. The knowledge graph is composed of nodes and edges, where each node corresponds to an entity and each edge represents a relationship between entities. As summarized in Table II, the dataset elements—TweetID, AccountID, Hashtag, and URL—are mapped to four node types: N_{Tweet}, $N_{Account}$, $N_{Hashtag}$, and N_{URL}. Specifically, $N_{Account}$ denotes the unique

identifier of a user, N_{Tweet} represents the unique identifier of a tweet, $N_{Hashtag}$ corresponds to hashtags embedded in tweets, and N_{URL} indicates external links contained in tweets. Since our objective is to predict associations between accounts, N_{URL} nodes were used only as intermediaries during preprocessing and were excluded from the final training graph. Based on these node definitions, the relationships among accounts, tweets, hashtags, and URLs were categorized into three edge types, as shown in Table IV.

TABLE IV
EDGE TYPES DEFINED IN THE KNOWLEDGE GRAPH

Feature	Direction	Description
E_{Posted}	$N_{Account} \rightarrow N_{Tweet}$	Connects an account to the tweet it posted
$E_{Contains}$	$N_{Tweet} \rightarrow N_{Hashtag}$	Connects a tweet to the hashtags it contains
	$N_{Tweet} \rightarrow N_{URL}$	Connects a tweet to the URLs it contains
$E_{SharedURL}$	$N_{Account} \rightarrow N_{Account}$	Connects accounts that shared the same URL

The E_{Posted} edge links an account to the tweets it created and is used to trace account activity sequences. Because a single account can generate multiple tweets, this relation establishes a structure that enables the behavioral patterns of accounts to be examined around their posted content.

The $E_{Contains}$ edge connects each tweet to the hashtags or URLs embedded within it. The hashtag connections are useful for detecting recurring keyword patterns frequently employed by DOCG, while the URL connections support aggregated analysis of tweets that reference the same external links.

The $E_{SharedURL}$ edge was constructed by grouping accounts that shared identical URLs and directly linking them within each group. In this process, as illustrated in Fig. 1, the URLs themselves were not retained as separate nodes; instead, direct account-to-account connections were established to reduce the number of nodes and improve the efficiency of the network model.

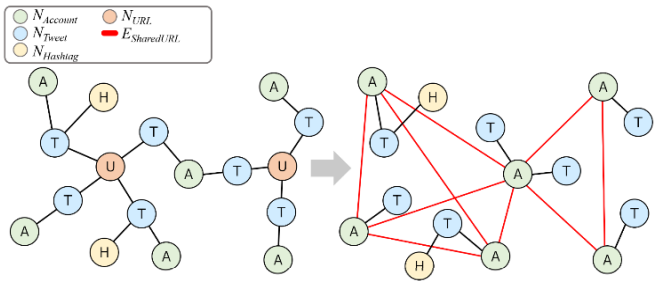


FIGURE 1. Example of link prediction based on a fully connected structure among accounts sharing the same URL.

For example, when accounts A, B, and C shared the same URL, a fully connected triad was formed among them, generating the links (A–B), (B–C), and (A–C). This model ensured direct connectivity between accounts and revealed hidden associations within drug distribution networks more clearly.

Furthermore, Fig. 2 illustrates the graph construction process with an example based on actual data. Account A posted a total of 13 tweets, one of which (TweetID 1744xxxx32) included hashtags such as Xan, adderall, and Mmemberville, along with Snapchat and Telegram URLs. Account B posted a total of seven tweets, one of which (TweetID 1746xxxx58) contained hashtags such as Weedsmokers, 420friendly, and Mmemberville, as well as the same external URL found in Account A's tweet. Although the two accounts operated independently without direct mentions or retweets, they repeatedly disseminated the same external URL, which resulted in their connection through the $E_{SharedURL}$ edge.

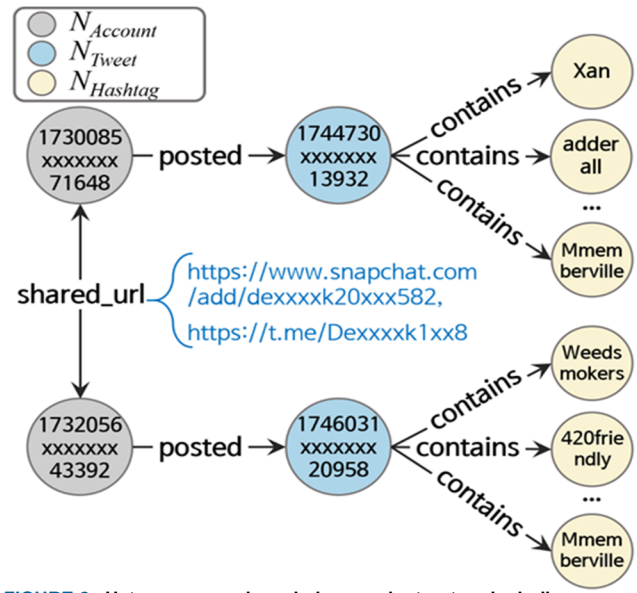


FIGURE 2. Heterogeneous knowledge graph structure including account-to-account connections based on shared URLs.

The constructed graph is a heterogeneous structure consisting of three node types ($N_{Account}$, N_{Tweet}, $N_{Hashtag}$) and three edge types (E_{Posted}, $E_{Contains}$, $E_{SharedURL}$). It was generated from approximately one million tweets. To enable the model to distinguish between actual and non-existent connections, we applied negative sampling by designating a subset of the $E_{SharedURL}$ edges as positive links and adding artificially generated negative links.

The tweet-based knowledge graph developed in this study goes beyond simple keyword analysis or user classification by structurally detecting hidden associations between accounts and providing a foundation for identifying key accounts. This model complements the limitations of traditional text-based detection methods and offers an effective basis for uncovering concealed relationships within DOCG networks, thereby supporting proactive investigative responses.

In Section IV, we present experiments conducted on the constructed knowledge graphs, distinguishing between homogeneous and heterogeneous settings, and compare link prediction performance across graph types. The heterogeneous knowledge graph developed in this study served as the input for GNN-based link prediction experiments. In the homogeneous graph experiments, only $E_{SharedURL}$ edges were included, allowing the models to learn direct account-to-account connections based on shared URLs. In the heterogeneous graph experiments, E_{Posted}, $E_{Contains}$, and $E_{SharedURL}$ edges were jointly incorporated, enabling integrated

learning of diverse behavioral patterns. Table V summarizes the types of knowledge graphs applied in the experiments.

TABLE V
TYPES OF CONSTRUCTED GNN KNOWLEDGE GRAPHS

Graph Type		Dataset Composition		Applied Models
		Node	Edge	
Single	Homogeneous	$N_{Account}$	$E_{SharedURL}$	GCN, GAT, GIN
	Heterogeneous	$N_{Account}$, N_{Tweet}, $N_{Hashtag}$	E_{Posted}, $E_{Contains}$, $E_{SharedURL}$	GCN, GAT, GIN
Hybrid		$N_{Account}$, N_{Tweet}, $N_{Hashtag}$	E_{Posted}, $E_{Contains}$, $E_{SharedURL}$	GCN-GAT, GCN-GIN, GAT-GIN, GAT-GCN, GIN-GCN

These types of knowledge graphs are utilized in the experiments presented in Section IV. We quantitatively analyze the impact of relationship types and node compositions on link prediction performance by comparing training results and evaluation metrics across different graph settings.

Before conducting the main experiments, we validate the academic contribution of our proposed GNN-based link prediction model by comparing its performance with prior benchmark studies. To this end, we utilize the BACRIM 2020 dataset, on which Contreras-Velasco et al. (2025) comparatively evaluated 14 link prediction algorithms, to verify whether our GNN models achieve competitive performance relative to prior work within the same criminal network domain, and furthermore, whether they demonstrate improved performance. This comparative experiment goes beyond simple model validation, aiming to quantitatively demonstrate the methodological superiority of our approach in the field of criminal organization network analysis.

The study by Contreras-Velasco et al. (2025) compared 14 link prediction algorithms, including node similarity indices, graph embedding methods, and GNN models, to predict covert alliance relationships among Mexican criminal organizations. The BACRIM 2020 dataset consists of 94 criminal organizations (nodes) and 92 alliance relationships (edges). We selected this dataset for the following reasons[39]: (1) as a publicly available dataset that has been sufficiently validated in prior research, it provides a reliable baseline for comparison; (2) it represents a real-world covert network containing incomplete observational data, presenting similar challenges (concealment, data sparsity) to those faced in DOCG detection; and (3) the prior study reported detailed performance metrics for the same GNN architectures used in our research, enabling direct performance comparison.

In this validation experiment, we focus solely on the GCN and GAT models among the 14 algorithms used in the prior study. This is because GCN and GAT are the core components of both the single models and hybrid models proposed in our research, and we aim to objectively evaluate the effectiveness of our proposed models through direct performance comparison between identical model architectures. While GIN is utilized in our hybrid models, we excluded standalone GIN performance from this validation, judging it to have limited direct relevance to our validation objectives given the simple structural characteristics of the BACRIM dataset.

It is important to note that the BACRIM 2020 dataset consists of a single node type (criminal organizations) and a single edge type (alliance relationships), making it structurally a homogeneous graph. Consequently, heterogeneous graph models designed to leverage the heterogeneity of multiple node types and diverse edge types cannot be applied to this dataset. Heterogeneous graph models exhibit their advantages when multiple node types and relationship types are present; therefore, in a dataset composed solely of a single type like BACRIM, there is no opportunity to leverage the structural benefits of heterogeneous graphs. Thus, in this validation experiment, we use only homogeneous GCN and GAT models to conduct direct performance comparison with the results of Contreras-Velasco et al.

Performance comparison centers on the F1 score. The F1 score, as the harmonic mean of precision and recall, is particularly suitable for evaluating link prediction in highly imbalanced networks where actual connections represent only a small fraction of all possible node pairs. Since both criminal alliance networks and DOCG networks exhibit sparse structures, the F1 score serves as an ideal metric for assessing how accurately models identify the minority class (actual links) without being misled by the overwhelming majority class (non-existent links). Moreover, since the prior study also used F1 score as a primary evaluation metric, objective performance comparison is possible through the same metric.

Fig. 3 shows the F1 scores comparing our GCN and GAT models with the reported results of prior work by Contreras-Velasco et al. on the BACRIM 2020 dataset. Our homogeneous GCN model achieved an F1 score of 0.968, which approaches the prior study's GCN result (0.980), demonstrating competitive performance. Our homogeneous GAT model recorded an F1 score of 0.928, which represents superior performance compared to the prior study's GAT result (0.904).

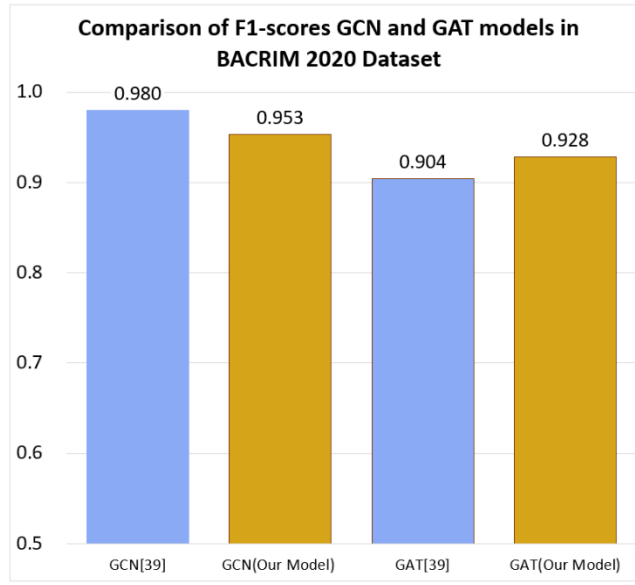


FIGURE 3. Comparative F1-score of GCN and GAT models in BACRIM 2020 Dataset.

Notably, our models achieved performance levels equal to or exceeding the near-perfect performance reported in the prior study. This demonstrates that the GNN models proposed in our research can reproduce or improve upon the best-performing baselines previously reported within the similar domain of criminal organization network analysis. Furthermore, by additionally proposing heterogeneous graph structures and hybrid GNN models not attempted in the prior study, our research presents academic contributions in terms of methodological expansion and practical applicability beyond simple performance reproduction.

Therefore, this comparative experiment supports that our research has achieved competitive performance comparable to the strongest prior results in the field of criminal network link prediction and, moreover, has presented an extended methodology applicable to complex social media-based DOCG networks.

As illustrated in Fig. 3, the relative behavior of GCN and GAT can be explained by the structural differences between the BACRIM dataset and the Twitter-based DOCG network. On the BACRIM dataset—which consists solely of criminal organizations with clearly defined alliance relationships—GCN achieved an F1-score of 0.980, whereas GAT reached 0.904. Because BACRIM is both highly curated and extremely small in scale, the uniform neighborhood aggregation of GCN is well suited to its clean graph structure and limited variability.

In contrast, on the Twitter DOCG network—where criminal accounts constitute only a small fraction of the overall graph and are interspersed among predominantly non-criminal users—GCN achieved 0.953 and GAT achieved 0.928, narrowing the performance gap considerably. This pattern suggests that attention mechanisms become more effective when crime-related signals are diluted within general user activity, as GAT explicitly computes neighbor-level importance scores rather than aggregating all neighbors uniformly. In environments where criminal nodes are sparsely connected and embedded among many benign neighbors, this selective weighting reduces the influence of structurally uninformative nodes and helps preserve discriminative patterns carried by minority criminal interactions.

IV. EXPERIMENTAL RESULTS

For graph-based link prediction aimed at identifying Drug-related Organized Crime Groups (DOCG), we used the constructed heterogeneous knowledge graph as input and experimentally applied various Graph Neural Network (GNN) models. The experiments are divided into two categories: single-model and hybrid-model settings.

In the single-model experiments, we applied representative GNN architectures—GCN, GAT, and GIN—individually. Each model was trained on both homogeneous and heterogeneous graphs, thereby enabling us to evaluate the effect of graph structural characteristics on predictive performance.

In the hybrid-model experiments, we applied configurations that combine two different GNNs. Each hybrid model was trained on heterogeneous graphs to examine how different model architectures interact with graph structures.

In this section, we classify the experiments according to graph type and model design, and we compare and analyze model performance and characteristics under each scenario.

A. Performance Analysis of Single Models

In this subsection, we present experimental results obtained from single Graph Neural Network (GNN) models. These experiments were designed to directly compare the structural properties and predictive performance of three representative architectures: GCN, GAT, and GIN.

The experiments were conducted under two graph configurations. In the homogeneous setting, only $E_{SharedURL}$ edges were included, enabling the models to learn direct account-to-account connections derived from shared URLs. In the heterogeneous setting, E_{Posted}, $E_{Contains}$, and $E_{SharedURL}$ edges were jointly incorporated, allowing the models to capture more complex and diverse relational patterns.

This design allowed us to evaluate differences in model behavior when trained on a single relation versus when trained on multiple relation types within the knowledge graph.

1) MODELS ON HOMOGENEOUS KNOWLEDGE GRAPH

In this subsection, we describe the homogeneous graph experiments, where only account nodes were considered while tweet and hashtag nodes were excluded. The graph was constructed by merging all entities ($N_{Account}$, N_{Tweet}, $N_{Hashtag}$) into a single node set, and account-to-account relationships were defined through the $E_{SharedURL}$ edge to represent positive links. Negative edges were generated using the negative sampling technique at a one-to-one ratio with the positive edges, ensuring balanced training for binary classification of link existence. The initial embeddings of all nodes were randomly initialized as trainable embeddings, and node representations were progressively optimized through parameter updates during training.

The three GNN models were configured as follows. GCN aggregates the features of neighboring nodes through averaging, enabling stable learning of global structural patterns. GAT incorporates an attention mechanism that learns the relative importance of neighboring nodes, assigning larger weights to more influential connections. GIN employs an MLP-based aggregation function, which provides strong representational power and enables the model to distinguish even subtle differences between structurally similar nodes. All models were designed with three layers, each followed by an ELU activation function, dropout, and batch normalization.

The performance of GNN models can vary substantially depending on hyperparameter settings, even under the same architecture. For instance, in the case of GAT, precision can be sensitive to the number of attention heads and learning rate adjustments, while GIN performance on recall may be affected by the combination of hidden channel size and weight decay. Therefore, drawing valid conclusions based solely on

architectural differences is insufficient, and a thorough search for optimal hyperparameter configurations is essential.

To this end, we employed Optuna, an automated hyperparameter optimization framework, to search for the best configurations. The search space is defined in Table VI, with the objective of maximizing F1-score on the validation set. Each trial was executed under an early stopping condition (patience = 15), and the models were retrained using the best parameter settings obtained from the search. Final performance was evaluated using precision, recall, F1-score, and AUC as the primary metrics.

TABLE VI
HYPERPARAMETER SEARCH SPACE FOR SINGLE-MODEL GNNS ON THE HOMOGENEOUS GRAPH USING OPTUNA

Parameter	Search Space	Search Method
Embedding Dimension	{64, 128, 256}	Categorical
Hidden Channels 1	{16, 32, 64}	Categorical
Hidden Channels 2	{8, 16, 32}	Categorical
Output Channels	{8, 16}	Categorical
Learning Rate (lr)	1e-4 -3e-3	Log-uniform
Dropout	0.2 - 0.6	Uniform
Weight Decay	1e-6 - 1e-3	Log-uniform
Attention Heads 1	{1, 2, 4}	Categorical
Attention Heads 2	{1, 2, 4}	Categorical

The search space for each parameter included embedding dimension, number of hidden channels, output channels, learning rate, dropout, weight decay, classification threshold, and the number of attention heads. Learning rate and weight decay were optimized using log-uniform sampling, allowing effective combinations to be explored over a wide range of values. Dropout and the threshold were tuned with uniform sampling, while embedding dimensions and hidden channels were restricted to categorical search spaces in order to systematically evaluate changes in model complexity.

Model performance was evaluated using precision, recall, F1-score, and AUC as the primary metrics. The results are summarized in Table VII.

TABLE VII
PERFORMANCE COMPARISON OF SINGLE GNN MODELS ON THE HOMOGENEOUS GRAPH

Model	Precision	Recall	F1-score	AUC
GCN	0.952859	0.996020	0.973961	0.998887
GAT	0.998943	0.978815	0.988776	0.999466
GIN	0.979839	0.554608	0.708302	0.982598

These results indicate that GAT achieved the best overall performance, with an F1-score close to 0.989, outperforming GCN. This highlights its strength in balancing precision and recall through attention mechanisms. GCN, while showing slightly lower precision, achieved the highest recall (≈0.996), demonstrating robust generalization of structural patterns. In contrast, GIN exhibited substantially lower recall and F1-score, reflecting its limitations in representation learning.

In addition, to evaluate model robustness under the imbalanced data environment of the homogeneous graph, we analyzed the Precision–Recall Curve and examined training stability through train/validation loss variations, as illustrated in Fig. 4.

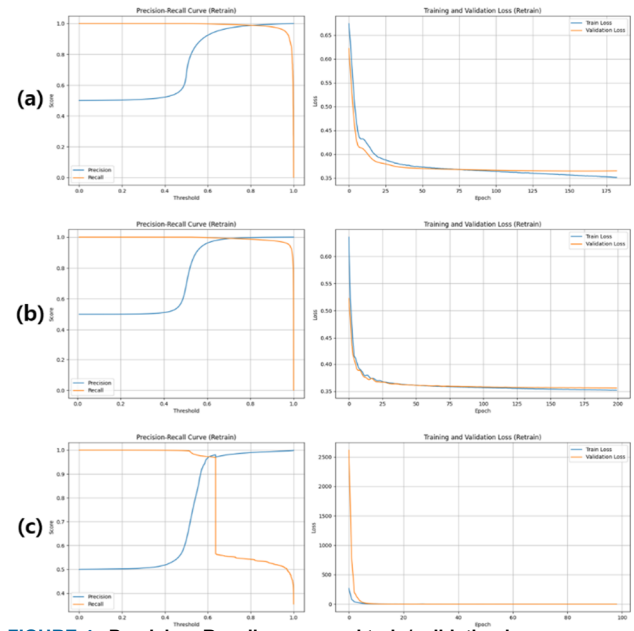


FIGURE 4. Precision–Recall curves and train/validation loss comparison for single GNN models on the homogeneous graph: (a) GCN, (b) GAT, (c) GIN.

For GCN, precision remained relatively stable across all threshold ranges, while recall showed a gradual decline beyond certain thresholds, indicating the potential omission of some reference nodes. The training and validation losses decreased sharply in the initial epochs and then converged stably, with no clear signs of overfitting.

For GAT, both precision and recall were maintained at consistently high levels across all thresholds, with balanced performance particularly evident in the mid-threshold range (0.6–0.8). The training and validation loss curves also converged in nearly identical patterns, indicating the most stable learning process among the models. This suggests that GAT performed robust representation learning even under the imbalanced data environment.

For GIN, recall initially modeled 1.0, demonstrating high sensitivity in the early range. However, the precision curve fluctuated sharply and showed instability in the later stages, reflecting a tendency toward overfitting to specific connection patterns. Although both training and validation losses decreased rapidly and converged, the larger oscillations and higher initial loss values suggested instability in capturing structural features during the early training process.

Overall, GAT achieved the best balance between precision and recall as well as the most stable learning process. GCN provided stable yet conservative detection performance, while GIN exhibited limitations due to unstable precision despite its high recall. Second, we compared the tendency of each model toward false positives using the confusion matrix. The results are presented in Fig. 5.

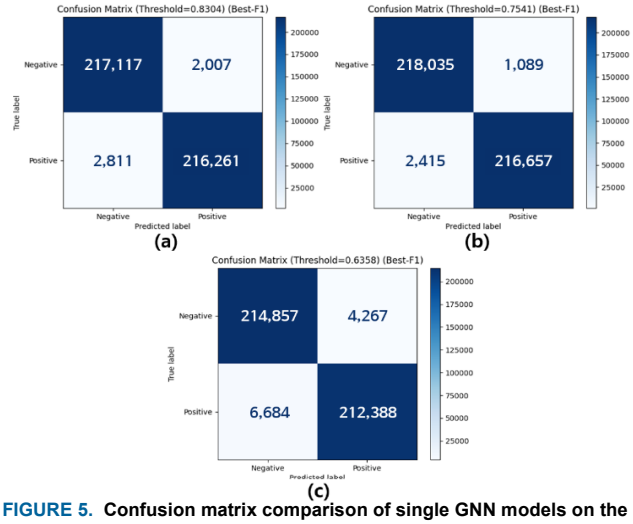


FIGURE 5. Confusion matrix comparison of single GNN models on the homogeneous graph: (a) GCN, (b) GAT, (c) GIN.

For GCN, the best performance was achieved at a threshold of 0.8304. The model produced 217,117 true negatives (TN) and 216,261 true positives (TP), indicating stable classification capability. False positives (FP) and false negatives (FN) were recorded at 2,007 and 2,811, respectively, showing a relatively balanced distribution of the two error types. This result suggests that GCN achieved balanced learning performance across both positive and negative classes.

For GAT, the optimal results were obtained at a threshold of 0.7341. The model achieved 218,035 TN and 216,657 TP, the highest correct classifications among the three models. Notably, FP was minimized to 1,089, demonstrating effective suppression of false detections in the negative class. Although FN was 2,415, this number was negligible relative to the overall dataset, indicating that the attention mechanism contributed to filtering out unnecessary connections.

For GIN, the optimal threshold was 0.6538. However, the error distribution differed from the other models. The model produced 214,857 TN and 212,388 TP, which were relatively lower, while FP and FN were 4,267 and 6,684, respectively. The considerable increase in FN indicates a larger number of missed actual connections, leading to degraded recall performance. These findings suggest that GIN exhibited overfitting tendencies during structural feature learning and failed to adequately capture certain relationships.

Overall, while all three models achieved high accuracy and F1-scores, their error patterns revealed clear differences. GCN exhibited a balanced distribution of FP and FN, ensuring stable classification performance. GAT most effectively reduced FP while also producing fewer FN, thereby achieving superior precision and recall. In contrast, GIN suffered from a substantial increase in FN, highlighting limitations in recall.

To illustrate these differences, we conducted a case study focusing on a representative node (AccountID: 1722170086922805248). Fig. 6 (a) presents a True Positive case from the GAT model, where actual $E_{SharedURL}$ connections surrounding the node were correctly predicted. In contrast, Fig. 6 (b) shows a False Negative case from the GIN model on the same node, where several existing $E_{SharedURL}$ connections were missed. In the visualizations, green nodes represent $N_{Hashtag}$, orange nodes represent N_{Tweet}, and blue nodes represent $N_{Account}$. Edge colors correspond to relation types: green edges $E_{Contains}$ denote connections between tweets and hashtags, blue edges E_{Posted} denote connections between accounts and tweets, and red edges $E_{SharedURL}$ denote shared-URL connections between accounts. Prediction outcomes are further distinguished by line style: red solid lines $E_{SharedURL}$(TP) indicate correctly predicted links (True Positives), whereas red dashed lines $E_{SharedURL}$(FN)

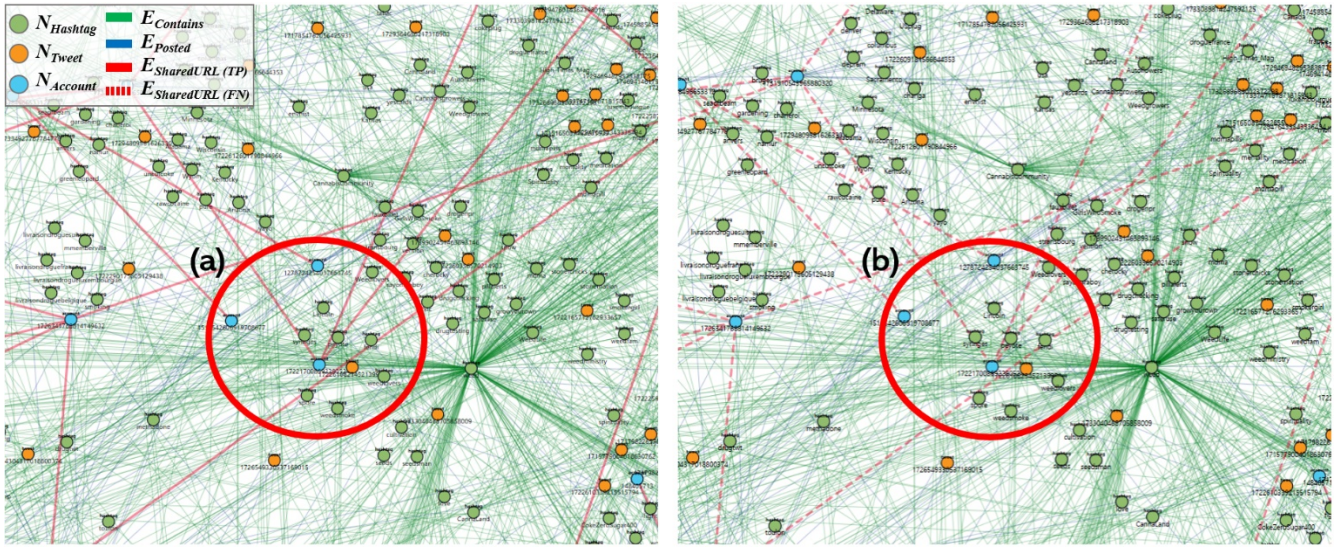


FIGURE 6. Comparison of GAT and GIN prediction results. (a) True Positives. (b) False Negatives.

indicate missed links (False Negatives). This node exhibits clear evidence of drug trafficking activity, repeatedly posting content with drug-related keywords (e.g., weed, cocaine, LSD, MDMA) and linking to a Telegram channel (t.me/liv*******beuh). Its tweets contain numerous drug-related hashtags (e.g., #weed, #cocaine, #lsd, #mdma, #pills, #drugtwt) and contact handles (e.g., @Pat******45, Snapchat IDs), creating strong structural and semantic signals that connect accounts via hashtags, tweets, and shared URLs.

As shown in Fig. 6 (a), GAT successfully leveraged these informative neighbors (e.g., drug-related hashtags and repeated shared-URL links) to correctly predict actual $E_{SharedURL}$ connections. In contrast, Fig. 6 (b) demonstrates that GIN, due to its uniform aggregation strategy, failed to capture these salient cues when mixed with weaker lifestyle or location-related hashtags (e.g., #fitness, #NYC, #illinois) and sparser connection patterns, resulting in missed predictions. These findings are consistent with the quantitative FN gap and demonstrate that attention-based aggregation is more effective at handling ambiguous structural patterns in homogeneous knowledge graphs derived from criminal networks.

2) MODELS ON HETEROGENEOUS KNOWLEDGE GRAPH

While the single-model experiments on the homogeneous graph provided a baseline evaluation of link prediction performance, they treated all nodes as a single type. This model failed to capture the differences among nodes with distinct semantic and structural attributes, such as accounts, tweets, and hashtags. In networks like drug trafficking organizations, where multiple actors and interactions coexist, it is particularly difficult to reflect the varying importance and meaning of different relationships.

To address this limitation, we experimented with single models on heterogeneous graphs, where node types were explicitly distinguished and relationship-specific characteristics were learned independently. As described in Section III, three edge types were defined: E_{Posted} (Account → Tweet), $E_{Contains}$ (Tweet → Hashtag), and $E_{SharedURL}$ (Account ↔ Account). Each relationship was modeled with a dedicated GNN layer and then integrated through the HeteroConv architecture. This design preserved the roles and meanings of different node types by separating relational learning paths, thereby enabling more accurate representation of complex network patterns.

Model training was conducted on heterogeneous graphs that retained the three node types, with the task formulated as a binary classification predicting whether an $E_{SharedURL}$ edge exists between account node pairs. Positive samples were defined as edges between accounts sharing the same URL, while negative samples were randomly sampled account pairs without connections. The training process incorporated heterogeneous nodes and relationships simultaneously, allowing the model to learn potential account associations. Hyperparameter optimization was performed using Optuna, with the search space defined in Table VIII. The optimization objective was to maximize the F1-score on the validation set, consistent with the previous experiments, and each trial was terminated under an early stopping condition (patience = 15).

TABLE VIII
HYPERPARAMETER SEARCH SPACE FOR SINGLE-MODEL GNNS ON THE HETEROGENEOUS GRAPH USING OPTUNA

Parameter	Search Space	Search Method
Embedding Dimension	{64, 128, 256}	Categorical
Hidden Channels 1	{32, 64, 128}	Categorical
Hidden Channels 2	{16, 32, 64}	Categorical
Output Channels	{8, 16, 32}	Categorical
Learning Rate (lr)	1e-4 -3e-3	Log-uniform
Dropout	0.2 - 0.6	Uniform
Weight Decay	1e-6 - 1e-3	Log-uniform
Attention Heads 1	{1, 2, 4}	Categorical
Attention Heads 2	{1, 2, 4}	Categorical
Attention Heads 3	{1, 2}	Categorical
Attribute Use	{True, False}	Categorical
Attribute Projection Dimension	{32, 64, 128}	Categorical
Classification Threshold	0.5-0.9	Uniform

The hyperparameter search space included embedding dimension, number of hidden channels, output channels, learning rate, dropout, weight decay, number of attention heads, attribute embedding projection dimension, and classification threshold. Learning rate and weight decay were optimized using log-uniform sampling to explore effective combinations over a wide range of values. Dropout and the classification threshold were tuned with uniform sampling. Embedding dimension, hidden channels, output channels, number of attention heads, attribute usage, and projection dimension were restricted to categorical searches in order to systematically assess the impact of model structural complexity.

Model performance was evaluated using precision, recall, F1-score, and AUC as the primary metrics. The results are summarized in Table IX.

TABLE IX.
PERFORMANCE COMPARISON OF SINGLE GNN MODELS ON THE HETEROGENEOUS GRAPH

Model	Precision	Recall	F1-score	AUC
GCN	0.973065	0.836729	0.899762	0.957234
GAT	0.987778	0.977966	0.982847	0.997379
GIN	0.957498	0.927973	0.942504	0.977957

These results indicate that GAT achieved the best overall performance, with an F1-score close to 0.983 and the highest AUC, highlighting its ability to balance precision and recall

effectively. GCN showed stable precision but relatively lower recall, suggesting limitations in capturing diverse node connections. GIN demonstrated more balanced performance than GCN, with recall above 0.928 and an F1-score above 0.943, although it tended to produce more false positives compared with GAT.

To further evaluate performance, we analyzed the Precision–Recall curves and the variations in train/validation loss, as illustrated in Fig. 7.

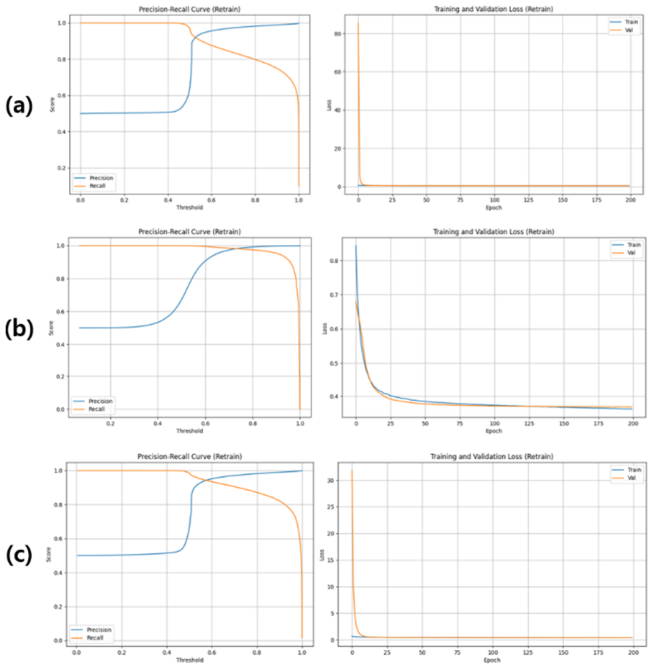


FIGURE 7. Precision–Recall curves and train/validation loss comparison of single GNN models on the heterogeneous graph: (a) GCN, (b) GAT, (c) GIN.

For GCN, the precision curve remained relatively stable across thresholds, while recall exhibited a gradual decline beyond the 0.6 threshold, indicating the possibility of missing certain reference nodes during detection. The training and validation losses decreased sharply during the initial epochs and then converged stably; however, slight fluctuations in validation loss suggested somewhat unstable generalization behavior.

For GAT, both precision and recall were consistently maintained at high levels across the entire threshold range, with precision showing a notable improvement in the 0.6–0.8 range. The training and validation loss curves followed highly similar trajectories with almost no signs of overfitting, producing the most stable learning curve among the models. This result indicates that the attention mechanism effectively emphasized critical neighbor information among diverse relations, enabling robust representation learning even in imbalanced data environments.

For GIN, recall initially modeled 1.0, indicating high sensitivity in the early range. However, the precision curve fluctuated considerably with threshold changes and dropped sharply in certain regions. The training curve showed a rapid decrease in loss during the early stages and eventually converged, but validation loss exhibited relatively large variations, reflecting a tendency toward overfitting to specific structural patterns.

Next, we compared the tendencies of each model regarding false positives and false negatives using the confusion matrices presented in Fig. 8.

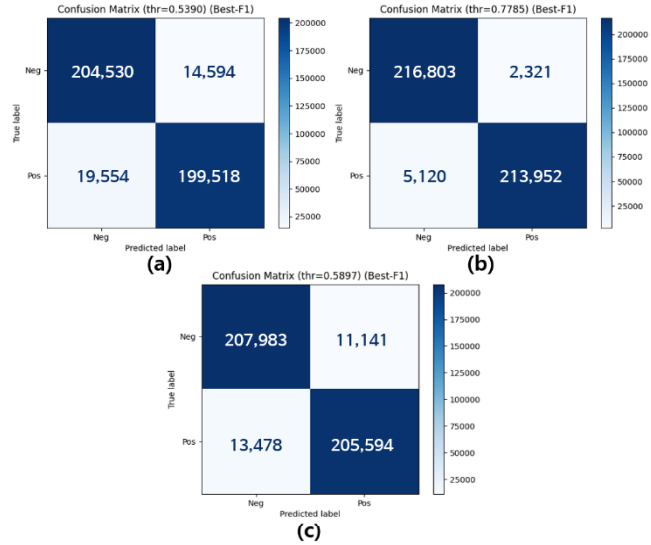


FIGURE 8. Confusion matrix comparison of single GNN models on the heterogeneous graph: (a) GCN, (b) GAT, (c) GIN.

The confusion matrix analysis revealed that GCN produced 14,594 false positives (FP), which was relatively low, while false negatives (FN) reached 19,554, indicating a loss in recall. Nevertheless, the model maintained stable performance in suppressing unnecessary connections. This outcome can be attributed to its structural property of uniformly reflecting relational information while avoiding excessive positive predictions.

For GAT, FP was minimized to 2,321, the lowest among the three models, while true positives (TP) reached 213,952, the highest overall. As a result, GAT achieved balanced performance in both precision and recall, leading to the best F1-score among the models. This can be explained by the attention mechanism, which effectively learned relationship-specific importance and assigned higher weights to critical neighbor information.

GIN maintained strong true positive detection with 205,594 TP; however, FP increased to 11,141, resulting in reduced precision. This is likely due to its structural property of aggregating all neighbor information equally, which led to positive classification of weakly meaningful or noisy connections.

In summary, within the heterogeneous graph single-model comparison, GCN demonstrated stable suppression of unnecessary links, GAT delivered the most balanced and superior overall performance, and GIN achieved strong true positive detection but at the cost of reduced precision. These findings indicate that the structural characteristics and relationship-handling strategies of each model directly

influence link prediction performance. The following section discusses experiments with hybrid models that combine the strengths of these single-model models.

B. Performance Analysis of Hybrid Models

In this section, we designed hybrid models to overcome the structural limitations of single models and to enhance predictive performance. As discussed earlier, homogeneous graph models ensured learning stability and error suppression for single relations but were limited in capturing fine-grained variations across diverse relationship types. In contrast, heterogeneous graph models effectively learned relationship-specific importance, but certain models exhibited over-detection and increased false positives.

To address these limitations, we implemented hybrid architectures that combine the complementary characteristics of GCN, GAT, and GIN on a common HeteroSAGE backbone. Specifically, we aimed to integrate the stable classification capability of GCN, the attention-based relational weighting of GAT, and the fine-grained pattern recognition of GIN. The goal of this design was to achieve a balanced trade-off between precision and recall under imbalanced data conditions, while simultaneously improving training stability and predictive reliability. The performance of the hybrid models was comprehensively evaluated and compared with that of the single models to assess the degree of improvement.

The hybrid model experiments were conducted on heterogeneous graph structures by combining GCN, GAT, and GIN. As in the single-model experiments, hyperparameter optimization was performed using Optuna, with the search space defined in Table X. The optimization objective was to maximize the F1-score on the validation set, and all trials were terminated under an early stopping condition (patience = 15).

TABLE X
HYPERPARAMETER SEARCH SPACE FOR HYBRID GNN MODELS ON THE HETEROGENEOUS GRAPH USING OPTUNA

Parameter	Search Space	Search Method
Embedding Dimension	{64, 128, 256}	Categorical
Hidden Channels 1	{32, 64, 128}	Categorical
Hidden Channels 2	{16, 32, 64}	Categorical
Output Channels	{8, 16, 32}	Categorical
Learning Rate (lr)	1e-4 -3e-3	Log-uniform
Dropout	0.2 - 0.6	Uniform
Weight Decay	1e-6 - 1e-3	Log-uniform
Attention Heads 1	{1, 2, 4}	Categorical
Attention Heads 2	{1, 2, 4}	Categorical
Attention Heads 3	{1, 2}	Categorical
Attribute Use	{True, False}	Categorical
Attribute Projection Dimension	{32, 64, 128}	Categorical
Classification Threshold	0.5-0.9	Uniform

The search space included factors such as embedding and hidden dimensions, as well as training stability parameters including learning rate, dropout, and weight decay. To evaluate the effectiveness of GAT-based modules, the number of attention heads was searched within the range of 1–4. Attribute usage and projection dimension were also incorporated as variables to examine the impact of heterogeneous node attribute information on the hybrid architectures. Finally, the classification threshold was searched within the interval [0.5, 0.9] to determine the optimal decision boundary for model output probabilities. Continuous variables were sampled using log-uniform or uniform distributions, while categorical parameters were selected through categorical search.

In this study, five hybrid model combinations were selected for experimentation. The GCN-GAT hybrid was designed to combine the global structural proximity learning of GCN with the neighbor-specific weighting mechanism of GAT, thereby reflecting both overall graph structure and localized neighbor interactions. The GIN-GAT hybrid integrated the strong structural discriminative capability of GIN with the attention mechanism of GAT, enabling the model to capture fine-grained structural differences while learning neighbor importance. The GCN-GIN hybrid leveraged the efficient global structural learning of GCN together with the discriminative power of GIN to account for both simple proximity patterns and complex relational distinctions. The GAT-GIN hybrid combined GAT's attention-based message passing with GIN's structural discrimination, strengthening both neighbor importance weighting and structural differentiation. Lastly, the GIN-GCN hybrid reversed the order, first learning detailed structural representations through GIN and then refining global proximity using GCN.

These five combinations were selected to maximize the complementary strengths of the models. In contrast, the GAT-GCN hybrid was excluded, as both models fundamentally rely on averaging mechanisms for structural proximity learning—GCN through simple averaging and GAT through attention-weighted averaging. This redundancy was expected to produce overlapping rather than complementary effects, leading to unnecessary parameter growth and reduced training efficiency. Therefore, the experiments focused exclusively on the five hybrid model combinations described above.

1) GCN-GAT HYBRID MODEL

The GCN-GAT hybrid model combines the global structural learning capability of GCN with the neighbor-specific weighting mechanism of GAT to perform link prediction on $E_{SharedURL}$ relationships between accounts. The integration of the two models is intended to jointly capture both global structural patterns and the relative importance of local neighbor interactions.

The performance of the GCN-GAT hybrid model achieved a precision of 0.940818, recall of 0.950432, F1-score of 0.945095, and AUC of 0.977580, showing balanced and satisfactory results across all four metrics. In particular, the F1-score remained above 0.94, confirming that the model was able to discriminate between positive and negative classes without excessive bias. Compared with the corresponding single heterogeneous models, the hybrid provided a meaningful improvement over GCN alone but did not surpass the performance of GAT. This indicates that while the stable structural aggregation of GCN was complemented by the attention-based representation of GAT, the hybrid did not completely overcome the inherent limitations of the individual models. Fig. 9 presents the training curves and classification results of the GCN-GAT hybrid model. The training curves visualize the variation of loss during training and validation, demonstrating the convergence process and stability of the model. The included confusion matrix illustrates the class-level prediction accuracy of the final classification results.

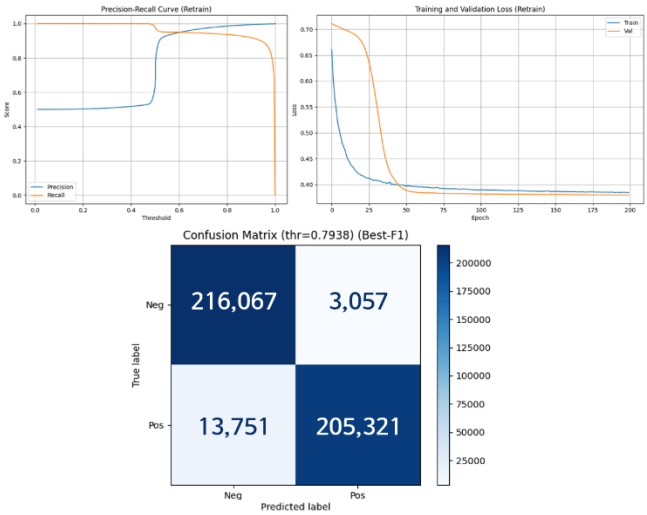


FIGURE 9. Precision–Recall curve, train/validation loss, and confusion matrix of the GCN-GAT hybrid model.

The Precision–Recall curve showed a relatively gradual decline across thresholds, with balanced performance observed in the mid-threshold region. The loss curves exhibited a steep decrease in the early training phase followed by stable convergence, with minimal divergence from validation loss, indicating limited signs of overfitting. According to the confusion matrix, 216,067 true negatives (TN) and 205,321 true positives (TP) were correctly classified, while 3,057 false positives (FP) and 13,751 false negatives (FN) were observed. The higher FN compared with FP suggests a conservative prediction tendency, which constrained recall performance. When compared with single heterogeneous models using the same dataset, the hybrid reduced both FP and FN significantly relative to GCN alone, demonstrating effectiveness in suppressing both false alarms and missed detections. However, compared with GAT alone, both error types increased, leading to a lower final F1-score. These findings indicate that while the GCN-GAT hybrid alleviated some limitations of GCN through attention-based weighting, it remained relatively conservative compared with the standalone GAT model.

2) GIN-GAT HYBRID MODEL

The GIN-GAT hybrid model combines the neighbor-weighting mechanism of GAT with the representations learned through the simple aggregation function of GIN. This hybrid is applied to predict $E_{SharedURL}$ relationships between accounts. The design preserves the basic structural learning capability of GIN while incorporating the relative importance of neighbors through GAT.

The performance of the GIN-GAT hybrid model achieved a precision of 0.940417, recall of 0.950432, F1-score of 0.945398, and AUC of 0.976490, showing overall balanced results. In particular, the F1-score remained above 0.94, confirming stable discrimination performance across both positive and negative classes. Compared with the single heterogeneous models, the hybrid achieved a notable improvement over GIN alone but did not reach the performance of GAT. This indicates that the hybrid effectively complemented the weaknesses of GIN, though it did not fully replicate the strengths of GAT. Fig. 10 presents the training curves and classification outcomes of the GIN-GAT hybrid model.

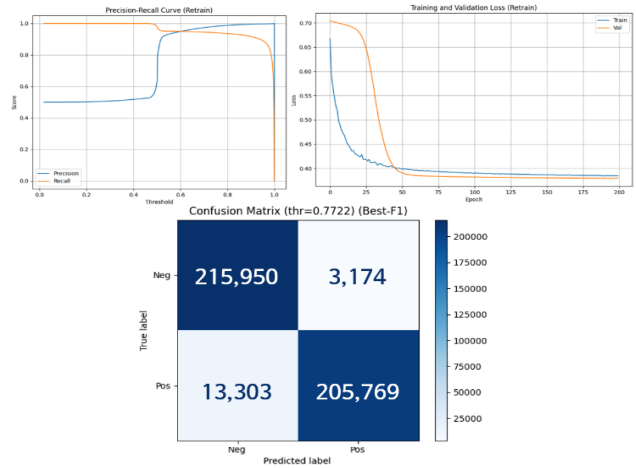


FIGURE 10. Precision–Recall curve, train/validation loss, and confusion matrix of the GIN-GAT hybrid model.

The Precision–Recall curve demonstrated stable precision and recall across varying thresholds, confirming that the model maintained consistent discriminative capability. The loss curves showed a rapid decrease in the early epochs followed by stable convergence. According to the confusion matrix, 215,950 true negatives (TN) and 205,769 true positives (TP) were correctly classified, while 3,174 false positives (FP) and 13,303 false negatives (FN) were observed. Although FN exceeded FP, slightly reducing recall, both FP and FN were significantly lower compared with the GIN single model. In contrast, relative to GAT alone, the hybrid produced higher error counts, resulting in a lower final F1-score. Overall, the GIN-GAT hybrid model substantially improved upon the single GIN model, but remained more conservative than the standalone GAT.

3) GCN-GIN HYBRID MODEL

The GCN-GIN hybrid model combines the global structural learning capability of GCN with the residual-based representation learning of GIN for predicting $E_{SharedURL}$ relationships between accounts. This combination is designed to capture both the overall structural context of the graph and the fine-grained relational information among individual nodes.

The performance results of the GCN-GIN hybrid model showed a precision of 0.933, recall of 0.901, F1-score of 0.917, and AUC of 0.951. These outcomes indicate that the strengths of GCN and GIN were leveraged in a complementary manner. GCN excels at learning global structural patterns, while GIN is effective in extracting detailed node-level features. By combining the two, the model achieved a balanced integration of structural information and fine-grained representations, resulting in stable precision and recall. Fig. 11 presents the Precision–Recall curve, training/validation loss curves, and confusion matrix of the GCN-GIN hybrid model.

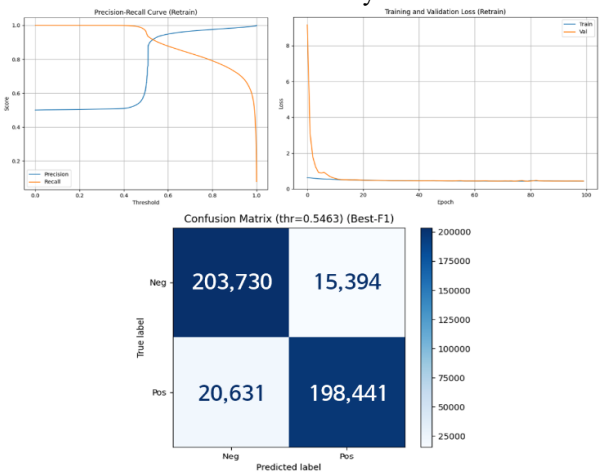


FIGURE 11. Precision–Recall curve, train/validation loss, and confusion matrix of the GCN-GIN hybrid model.

The Precision–Recall curve demonstrated stable precision and recall across threshold variations, indicating consistent discriminative capability. The loss curves decreased sharply during the early training phase and then converged stably, suggesting limited risk of overfitting. According to the confusion matrix, 203,730 true negatives (TN) and 198,441 true positives (TP) were correctly classified, while 15,394 false positives (FP) and 20,631 false negatives (FN) were recorded. Although FN was higher than FP, thereby constraining recall, both FP and FN were reduced compared with the single GCN and GIN models, resulting in improved performance overall. However, relative to GAT alone, FP and FN remained higher, leading to a lower final F1-score. In summary, the GCN-GIN hybrid model successfully complemented the weaknesses of the two single models and achieved overall performance improvements, but it remained more conservative than GAT-based models that explicitly learn neighbor importance.

4) GAT-GIN HYBRID MODEL

The GAT-GIN hybrid model combines the neighbor-weighting mechanism of GAT with the structural isomorphism discrimination capability of GIN for predicting $E_{SharedURL}$ relationships between accounts. This combination is designed to jointly capture neighbor importance and fine-grained structural differentiation.

The performance evaluation of the GAT-GIN hybrid model yielded a precision of 0.991824, recall of 0.934136, F1-score of 0.962116, and AUC of 0.976436, indicating consistently high and balanced results across all four metrics. In particular, precision modeled 0.99, demonstrating that false positives (FP) were extremely low, while recall remained above 0.93, confirming stable detection of the positive class. Compared with the single heterogeneous models, the hybrid showed a significant improvement over GIN alone and also outperformed GAT in terms of precision. This result suggests that the attention-based representations of GAT complemented the structural discriminative capability of GIN in a mutually reinforcing manner. Fig. 12 presents the training curves and classification results of the GAT-GIN hybrid model. The training curves visualize the changes in training and validation loss, providing insights into convergence and stability, while the included confusion matrix shows the class-level prediction accuracy of the final classification results.

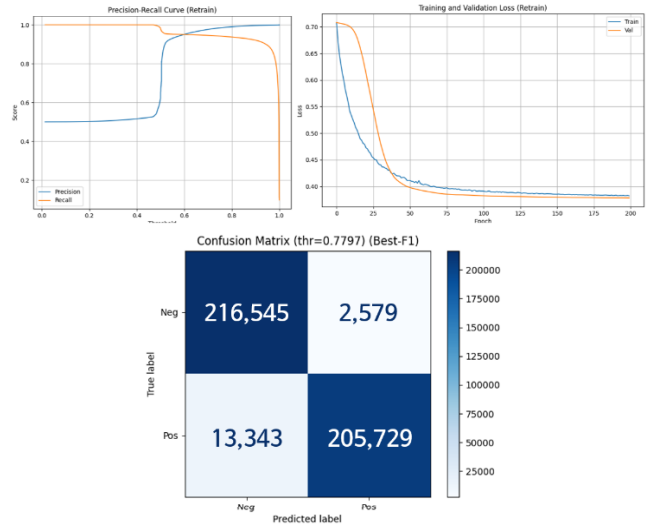


FIGURE 12. Precision–Recall curve, train/validation loss, and confusion matrix of the GAT-GIN hybrid model.

The Precision–Recall curve showed that both metrics remained high around the threshold of approximately 0.79, demonstrating stable predictive capability. The loss curves exhibited a steep decrease during the early training phase and then converged stably with training and validation losses at similar levels, indicating limited signs of overfitting. According to the confusion matrix, 216,545 true negatives (TN) and 205,729 true positives (TP) were correctly classified, while false positives (FP) were very low at 2,579, and false negatives (FN) were 13,343. Although a small number of missed positive cases were observed, the strong suppression of FP highlights the advantage of this hybrid model in minimizing false alarms, which is particularly beneficial in real-world operational environments.

5) GIN-GCN HYBRID MODEL

The GIN-GCN hybrid model combines the strong structural representation learning capability of GIN with the global aggregation property of GCN for predicting $E_{SharedURL}$ relationships between accounts. The purpose of this combination is to leverage GIN's ability to capture fine-grained structural differences while promoting generalization through the stable global aggregation provided by GCN.

The performance evaluation of the GIN-GCN hybrid model yielded a precision of 0.954260, recall of 0.936523, F1-score of 0.945309, and AUC of 0.982386, indicating overall stable and balanced results. In particular, the F1-score of 0.945 represented a substantial improvement compared with the single heterogeneous GCN and GIN models, suggesting that the weaknesses of the individual models were effectively complemented when combined. Fig. 13 illustrates the Precision–Recall curve, training loss curves, and confusion matrix results of the GIN-GCN hybrid model.

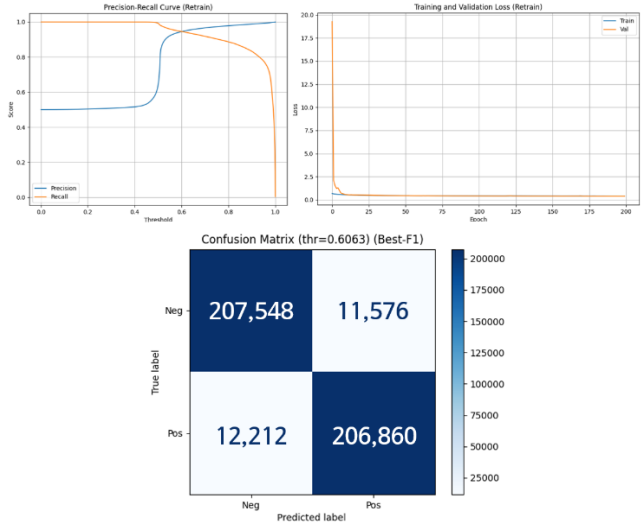


FIGURE 13. Precision–Recall curve, train/validation loss, and confusion matrix of the GIN-GCN hybrid model.

The Precision–Recall curve showed relatively moderate variation across thresholds, with both metrics maintaining stability and balance in the mid-threshold range. The loss curves for both training and validation decreased sharply in the early epochs and then converged rapidly, indicating limited risk of overfitting. According to the confusion matrix, 207,548 true negatives (TN) and 206,860 true positives (TP) were correctly classified, while 11,576 false positives (FP) and 12,212 false negatives (FN) were observed. FP and FN were distributed at similar levels, resulting in precision and recall being consistently maintained above acceptable thresholds.

Overall, the GIN-GCN hybrid model demonstrated significant performance improvements over the individual models, particularly in overcoming the limitations of GIN. However, when compared with GAT-based models, performance differences remained, suggesting that while GIN's structural learning capability complemented GCN's global aggregation, it did not achieve the same level of efficient weighting as provided by the attention mechanism in GAT.

C. Comparative Analysis of All GNN Models

In this section, we present a comprehensive comparison and analysis of the experimental results from both single and hybrid models. The structural characteristics and limitations of single models, as well as the complementary effects provided by hybrid models, are examined through quantitative metrics and visual outcomes. This integrated evaluation allows for a multidimensional assessment of the suitability of different GNN architectures for the link prediction task.

Table XI summarizes the performance results of all GNN models. Among the single models on homogeneous graphs, the performance differences were clear: GAT achieved the highest F1-score due to its neighbor-weighting mechanism, whereas GIN exhibited the lowest performance because of its structural limitations. Similar trends were observed on heterogeneous graphs, where the GAT model again achieved superior results, confirming that attention-based architectures remain effective even in complex environments involving multiple node types and relationships.

Furthermore, the performance of hybrid models generally showed improvements by compensating for the weaknesses of single models. For instance, the GCN-GIN hybrid reduced both FP and FN compared with the individual models, thereby

TABLE XI
OVERALL PERFORMANCE COMPARISON OF ALL GNN MODELS

Model			Precision	Recall	F1-score	AUC	TN	TP	FP	FN
Single	Homogeneous	GCN	0.952859	0.996020	0.973961	0.998887	217,117	216,261	2,007	2,811
		GAT	0.998943	0.978815	0.988776	0.999466	218,035	216,657	1,089	2,415
		GIN	0.979839	0.554608	0.708302	0.982598	214,857	212,388	4,267	6,684
	Heterogeneous	GCN	0.973065	0.836729	0.899762	0.957234	204,530	199,518	14,594	19,554
		GAT	0.987778	0.977966	0.982847	0.997379	216,803	213,952	2,321	5,120
		GIN	0.957498	0.927973	0.942504	0.977957	207,983	205,594	11,141	13,478
Hybrid		GCN-GAT	0.939818	0.950432	0.945095	0.977580	216,067	205,321	3.057	13,751
		GIN-GAT	0.940417	0.950432	0.945398	0.976490	215,950	205,769	3,174	13,303
		GCN-GIN	0.933212	0.900845	0.916743	0.951240	203,730	198,441	15.394	20,631
		GAT-GIN	0.991824	0.934136	0.962116	0.976436	216,545	205,729	2,579	13,343
		GIN-GCN	0.954260	0.936523	0.945309	0.982386	207,548	206,860	11,576	12,212

improving stability. Similarly, GCN-GAT and GAT-GIN demonstrated higher F1-scores relative to their single counterparts. However, none of the hybrid models surpassed the best performance achieved by the heterogeneous GAT model, suggesting that hybrid architectures can mitigate single-model limitations but may not necessarily exceed the optimal standalone model.

From a metric-wise perspective, most models-maintained precision above 0.93, while recall exhibited more variation due to differences in FN counts. The homogeneous GCN model showed weaker recall, but when combined with GAT in the hybrid setting, recall improved stably. AUC values remained consistently high (above 0.97) across all models, indicating reliable discriminative capability regardless of threshold variations. Finally, to highlight the best-performing models, the F1-scores of all GNN architectures are compared in Fig. 14.

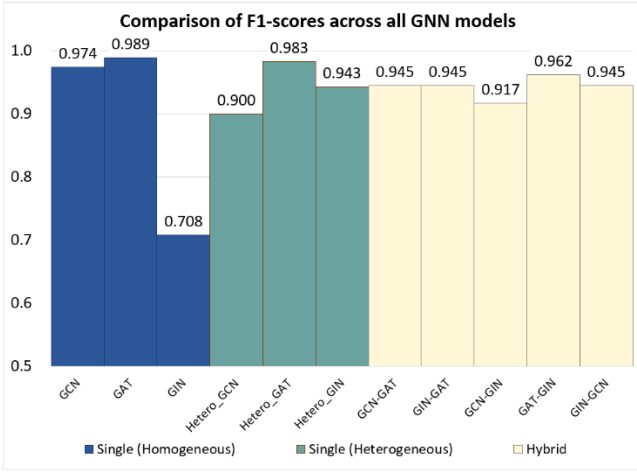


FIGURE 14. Comparative F1-score performance of all GNN models.

Overall, the results of this study demonstrate that single and hybrid models play complementary roles in GNN-based link prediction. The GAT family, leveraging the attention mechanism, consistently exhibited strong discriminative capability even in complex heterogeneous graph environments and achieved the best performance among the single models. In contrast, GIN was limited as a standalone model due to its structural simplicity; however, when combined with GCN in a hybrid form, it reduced both FP and FN, thereby improving performance and validating the effectiveness of the hybrid model.

Therefore, the selection of the final model should not be based solely on absolute metric values but rather on a comprehensive consideration of category-specific characteristics. Among the single models, GAT achieved the highest F1-scores in both homogeneous and heterogeneous graph experiments, while among the hybrid models, the GAT-GIN combination showed the best performance. Consequently, this study identifies GAT as the optimal single model and GAT-GIN as the optimal hybrid model for predicting connections among DOCG and identifying their potential inter-organizational associations.

V. DISCUSSION

To examine whether the behaviors observed in Sections IV.A–C are specific to drug-related DOCG networks or extend to other forms of illicit social media activity, we conducted an additional validation experiment using a dataset associated with the promotion of sexual exploitation content. The dataset was collected through the same keyword-based filtering procedure used for the DOCG dataset and contains posts with sexually exploitative promotional text and external URLs linking to platforms such as Discord and Telegram. Using the identical graph construction pipeline, we formed a heterogeneous network composed of 110,757 accounts, 110,757 tweets, 5,335 distinct hashtags, and 1,914 distinct external URLs, thereby maintaining structural alignment with the DOCG graph schema.

Although the semantic domain differs from drug trafficking, the resulting network preserves several structural characteristics that also appear in DOCG interactions: criminal and non-criminal accounts coexist within the same environment, illicit signals are embedded through domain-specific hashtags and repeated external contact channels, and associations among accounts arise primarily through shared URL usage rather than explicit social ties. These characteristics allow the dataset to serve as a suitable test for evaluating the generalizability of the proposed link prediction framework. The F1-scores of all models on the sex crime dataset are summarized in Fig. 15.

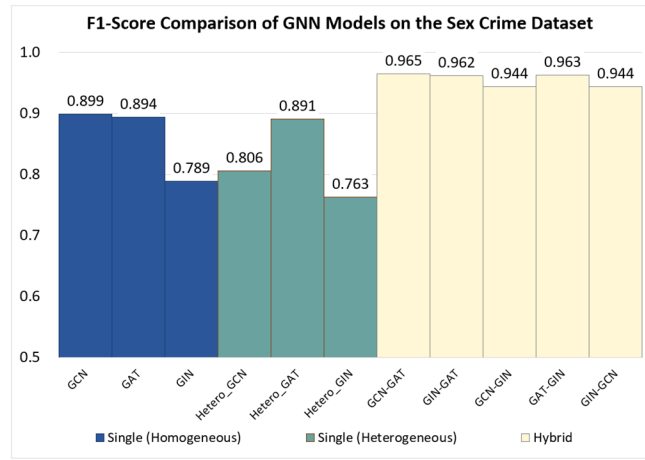


FIGURE 15. F1-score comparison of single and hybrid GNN models evaluated on the sex crime dataset.

As shown in Fig. 15, the performance patterns in this additional domain exhibit both consistencies and meaningful differences compared with the DOCG experiments. Among the homogeneous single models, GCN achieves the highest F1-score (0.8987), slightly surpassing homogeneous GAT (0.8936). This contrasts with the DOCG results, where attention mechanisms displayed more pronounced advantages. A key reason is the relatively higher structural coherence of the sex

crime dataset: examination of high-frequency URLs and linkage patterns indicates a more uniformly organized network, with less need for URL-level filtering than in the DOCG dataset. In such settings, GCN's uniform neighbor aggregation benefits from stable local neighborhoods and exhibits reduced oversmoothing, resulting in strong performance.

In the heterogeneous setting, GAT remains the strongest performer (0.8908), followed by heterogeneous GCN (0.8060). The structural diversity introduced by multiple node and edge types amplifies the utility of attention-based weighting, explaining why GAT maintains its relative advantage when applied to heterogeneous graphs. Hybrid architectures again produce the highest scores overall, consistent with the DOCG findings. GCN–GAT achieves the best performance (0.9650), followed closely by GAT–GIN (0.9628) and GIN–GAT (0.9615), with all hybrid models exceeding 0.94. These results confirm that combining complementary aggregation mechanisms—GCN's smoothing, GAT's selective neighbor weighting, and GIN's structural discrimination—yields robust embeddings across domains of differing density and noise characteristics.

The contrast with the BACRIM comparison further clarifies the architectural dynamics. The BACRIM dataset is small, fully criminal, and structurally clean, which favors GCN; accordingly, GCN outperformed GAT in that setting. The sex crime dataset, although much larger, also exhibits a relatively coherent structure compared to DOCG, which helps explain why homogeneous GCN performs strongly and why GAT's advantage is less pronounced than in DOCG networks. In environments where criminal signals are diluted within general user activity, as in DOCG, attention mechanisms help suppress irrelevant neighbors; in cleaner environments, such as BACRIM or the more structured sex crime dataset, uniform aggregation can be equally or more effective.

Taken together, these results demonstrate that the observed architectural tendencies are driven not by domain semantics but by underlying network structure. Hybrid models show the most stable and highest performance across all datasets, GAT retains robustness when signal dilution is severe, and GCN excels when neighborhood structures are clean or highly coherent. This indicates that the proposed framework can be transferred to other criminal domains with minimal modification, requiring only domain-specific data collection and hyperparameter adjustment.

VI. CONCLUSION

This paper proposes a GNN-based link prediction model to identify potential connections among DOCG using drug-related tweets collected from social media platform X. Approximately 470,000 unique tweets were collected, and preprocessing was applied to remove high-frequency URLs unrelated to criminal activity, thereby reducing structural distortions and improving analytical accuracy. The collected data were modeled as a heterogeneous knowledge graph, consisting of accounts, tweets, hashtags, and URLs as the primary nodes, and incorporating three edge types: E_{Posted}, $E_{Contains}$, and $E_{SharedURL}$. For comparison, a homogeneous graph structure was also constructed using only shared-URL relationships.

Experiments were conducted using single models based on GCN, GAT, and GIN, as well as five hybrid model combinations. The results showed that heterogeneous graphs generally outperformed homogeneous graphs on the link prediction task. In particular, GAT-based models that incorporate an attention mechanism achieved the highest F1-scores and AUC values. Among the hybrid models, the GAT–GIN combination achieved the highest precision, effectively suppressing false positives, while also providing stable performance improvements over the single-model baselines. To evaluate whether the proposed model maintains consistent performance on other crime-related network datasets and thereby assess its generalization capability, we conducted comparative experiments using the BACRIM dataset. The results indicated that the relative performance of GCN and GAT depends on the structural characteristics of the network. Because the BACRIM dataset is highly curated and very small in scale, it is favorable to uniform aggregation, and thus GCN exhibited stronger performance. The sex-crime dataset used as a comparative baseline in this study also consists of approximately 110,000 tweets and is relatively small; analysis of high-frequency URLs showed that no additional filtering was required, and this curated structure led to consistent and robust performance for both GCN and GAT.

However, both datasets are relatively small and heavily cleaned through the collection and preprocessing pipeline, which limits their ability to fully capture the characteristics of social media data encountered in real investigative environments. In contrast, the primary DOCG dataset used in this study is a large-scale corpus composed of approximately 470,000 unique tweets and contains substantial non-criminal activity and noise, thereby reflecting the characteristics of real-world social media data more faithfully. Despite being large and not fully curated, the proposed model achieved high accuracy on this dataset. Considering that data collected in actual criminal investigations typically contain a considerable amount of noise and irrelevant information, this result suggests that the proposed model has strong practical applicability in real investigative settings. Taken together, these observations indicate that differences in model behavior arise primarily from variations in data composition and relation sparsity rather than from domain-specific semantic properties, and they confirm that the proposed model maintains robust performance regardless of the level of data curation.

The proposed model goes beyond simple text-based analysis by leveraging complex structural relationships to detect hidden connections among accounts. Notably, it demonstrated the ability to identify potential inter-organizational links through shared URLs, even in the absence of direct mentions or retweets, highlighting its investigative utility. From a practical perspective, this study has the potential to serve as a core component of automated detection

systems targeting social media–based drug organizations, providing law enforcement agencies with essential tools for proactively understanding organizational structures and formulating response strategies. This framework further showed consistent behavior when tested on a separate sex-crime promotion dataset, suggesting its applicability across multiple illicit domains.

For future work, we plan to incorporate temporal information into a time-aware GNN framework to capture the dynamic evolution of relationships. Additional relationships such as mentions, retweets, and follows will be integrated to further enhance the structural diversity and representational power of the graph. Moreover, extending the analysis to multiple platforms, including Telegram and Instagram, will contribute to the development of a more comprehensive cybercrime detection system, which we aim to validate through empirical studies in real-world investigative contexts.

APPENDIX A
KEYWORD DEFINITIONS AND MEANINGS

Keyword	Description
2C-B	A synthetic psychedelic drug (phenethylamine class)
420Life	A slang term representing cannabis culture and lifestyle
acid	Slang for LSD (Lysergic acid diethylamide)
Adderall	Prescription stimulant containing amphetamine salts
Amphetamine	A stimulant drug affecting the central nervous system
benzos	Slang for benzodiazepines (e.g., Xanax, Valium)
bud	Slang for cannabis flower
carts	Slang for THC or cannabis oil cartridges for vaping
cocaine	A powerful stimulant drug derived from coca leaves
coke	Slang for cocaine
DMT	A hallucinogenic tryptamine drug (Dimethyltryptamine)
dmtvape	DMT consumed through a vape cartridge
drugs	General term for drugs or narcotics
drugstwt	Abbreviation for "Drug Twitter," an online community
drugtwt	Variant of "Drug Twitter" abbreviation
ecstacy	Common misspelling/slang for MDMA
edibles	Cannabis-infused food products
Fentanyl	A synthetic opioid analgesic
hash	Slang for hashish, concentrated cannabis resin
hashish	Cannabis product made from compressed resin
heroin	An opioid drug made from morphine
hydrocodone	A prescription opioid pain medication
ketamine	A dissociative anesthetic and recreational drug
kush	Slang for a high-quality strain of cannabis
lsd	A hallucinogenic drug (Lysergic acid diethylamide)
Marijuana	Cannabis, commonly used term for the plant
mdma	A psychoactive drug also known as ecstasy
Mescaline	A psychedelic compound from peyote cactus
meth	Slang for methamphetamine
methamphetamine	A powerful central nervous system stimulant
mushrooms	Slang for psilocybin-containing mushrooms
nufcfanscoke	Slang phrase combining "Newcastle fans" and cocaine
pills	General slang for tablets or capsule drugs
plug	Slang for a drug supplier or dealer
psilocybin	Hallucinogenic compound in certain mushrooms
psychedelics	General term for hallucinogenic substances
shrooms	Slang for psilocybin mushrooms
vegasplug	Slang for a drug dealer based in Las Vegas
weed	Slang for cannabis
Xanax	Brand name for alprazolam, a benzodiazepine

REFERENCES

[1] "LSD & MDMA: Laced by dark web spaces, grip of synthetic drugs tightens on youth," *Times of India*. [Online]. Available: https://timesofindia.indiatimes.com/city/bhopal/lsd-mdma-laced-by-dark-web-spaces-grip-of-synthetic-drugs-tightens-on-youth/articleshow/122410886.cms. Accessed: Aug. 25, 2025.

[2] U.S. Drug Enforcement Administration (DEA), *National Drug Threat Assessment 2024*. Washington, DC, USA, 2024. [Online]. Available: https://www.dea.gov/sites/default/files/2025-02/508_5.23.2024%20NDTA-updated.pdf. Accessed: Aug. 25, 2025.

[3] Q. Guo, S. Jiao, Y. Yang, Y. Yu, and Y. Pan, "Assessment of urban flood disaster responses and causal analysis at different temporal scales based on social media data and machine learning algorithms," *Int. J. Disaster Risk Reduct.*, vol. 117, Art. no. 105170, 2025.

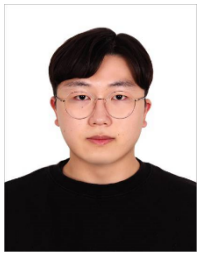
[4] M. Asif, M. Al-Razgan, Y. A. Ali, and L. Yunrong, "Graph convolution networks for social media trolls detection use deep feature extraction," *J. Cloud Comput.*, vol. 13, no. 1, p. 33, 2024.

[5] W. E. Kedi, C. Ejimuda, C. Idemudia, and T. I. Ijomah, "Machine learning software for optimizing SME social media marketing campaigns," *Comput. Sci. IT Res. J.*, vol. 5, no. 7, pp. 1634–1647, 2024.

[6] A. Coletti, R. McGloin, A. Oeldorf-Hirsch, and E. Hamlin, "Science communication on social media: Examining cross-platform behavioral engagement," *J. Social Media Soc.*, vol. 11, no. 2, pp. 236–263, 2022.
[7] J. Peng, Y. He, Y. Chang, Y. Lu, P. Zhang, Z. Ou, and Q. Yu, "A social media dataset and H-GNN-based contrastive learning scheme for multimodal sentiment analysis," *Appl. Sci.*, vol. 15, no. 2, p. 636, 2025, doi: 10.3390/app15020636.
[8] M. Rahman and M. Atikuzzaman, "Social media-based copyright awareness and knowledge-sharing practices among university students in Bangladesh," *Inf. Serv. Use*, vol. 44, no. 3, pp. 237–254, 2024.
[9] S. Samarasinghe and T. Chandrasiri, "The impact of social media on students' academic performance," *Am. Sci. Res. J. Eng. Technol. Sci.*, vol. 60, pp. 40–51, 2019.
[10] M. Chen and X. Xiao, "The effect of social media on the development of students' affective variables," *Front. Psychol.*, vol. 13, Art. no. 1010766, 2022.
[11] S. Patel, P. Bansal, and P. Kaur, "Rumour detection on benchmark Twitter datasets using graph neural networks with data augmentation," *Soc. Netw. Anal. Mining*, vol. 14, no. 1, p. 163, 2024.
[12] B. Pattanaik, S. Mandal, R. M. Tripathy, et al., "Rumor detection using dual embeddings and text-based graph convolutional network," *Discov. Artif. Intell.*, vol. 4, p. 86, 2024, doi: 10.1007/s44163-024-00193-6.
[13] T. Liu, Q. Cai, C. Xu, B. Hong, F. Ni, Y. Qiao, and T. Yang, "Rumor detection with a novel graph neural network approach," *arXiv preprint*, arXiv:2403.16206, 2024.
[14] M. Jin, et al., "A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.
[15] N. A. Holagh and Z. Kobti, "Survey of graph neural network methods for dynamic link prediction," *Procedia Comput. Sci.*, vol. 257, pp. 436–443, 2025.
[16] V. S. Anoop, C. S. Krishna, and U. H. Govindarajan, "Graph embedding approaches for social media sentiment analysis with model explanation," *Int. J. Inf. Manage. Data Insights*, vol. 4, no. 1, Art. no. 100221, 2024.
[17] V. Fionda, "Logic-based analysis of fake news diffusion on social media," *Soc. Netw. Anal. Mining*, vol. 15, no. 1, p. 59, 2025.
[18] H. R. Moorthy, et al., "Dual stream graph augmented transformer model integrating BERT and GNNs for context aware fake news detection," *Sci. Rep.*, vol. 15, no. 1, p. 25436, 2025.
[19] W. S. Spyt, "Social media and police investigations: Understanding the strategies that officers pursue when they encounter social media in their investigations," Ph.D. dissertation, Univ. Portsmouth, Portsmouth, U.K., 2017.
[20] S. A. Ajagbe, O. O. Adegboye, R. O. Abayomi, and A. A. Oyedele, "Artificial intelligence in social media forensics: A comprehensive survey and analysis," *IEEE Access*, vol. 12, pp. 40113–40134, 2024.
[21] B. G. Bokolo and Q. Liu, "Artificial intelligence in social media forensics: A comprehensive survey and analysis," *Electronics*, vol. 13, no. 9, p. 1671, 2024.
[22] C. M. Sinyangwe, Cybersecurity and Ethical Social Media Use: The Role of Guidance and Counseling in Zambia, Int. J. of Future Multidisciplinary Research (IJFMR), vol. 3, no. 5, pp. 1–8, 2023. [Online]. Available: https://www.ijfmr.com/research-paper.php?id=35982. [Accessed: Aug. 25, 2025].
[23] K. Maitra, "Digital forensic analysis of social media platforms for enhanced investigation and evidence collection," *Int. J. Innov. Appl. Stud.*, vol. 43, no. 2, pp. 261–271, 2024.
[24] A. Chauhan, "Social media analysis in criminal investigation," *Int. J. Sci. Res. Eng. Trends*, vol. 10, pp. 2060–2062, 2024, doi: 10.61137/ijsret.vol.10.issue5.256.
[25] A. Sayal, A. Gupta, C. Vasundhara, Y. BM, V. Gupta, and H. Maheshwari, "Crime detection using data mining techniques," in *Proc. 2024 6th Int. Conf. Comput. Intell. Commun. Technol. (CCICT)*, 2024, pp. 200–204.
[26] R. Rawat, A. S. A. Raj, R. K. Chakrawarti, K. S. Sankaran, S. K. Sarangi, H. Rawat, and A. Rawat, "Enhanced cybercrime detection on Twitter using Aho-Corasick algorithm and machine learning techniques," *Informatica*, vol. 48, no. 18, 2024.
[27] S. Van Berkel, E. Kleemans, and A. Mooij, "Social media and organizing violent crime against persons and properties: A qualitative analysis of online criminal communication between young offenders based on seized telephone information," *Trends Organized Crime*, pp. 1–22, 2025.
[28] N. Ohara, J. Ito, Z. Zhao, S. Osada, T. Sanda, N. Nakagawa, and M. Oguchi, "Graph-based analysis of criminal networks on social media: A novel approach using intersection graphs for cybercrime mitigation," in *Proc. 2025 19th Int. Conf. Ubiquitous Inf. Manage. Commun. (IMCOM)*, 2025, pp. 1–8.
[29] M. R. Haupt, et al., "The influence of social media affordances on drug dealer posting behavior across multiple social networking sites (SNS)," *Comput. Hum. Behav. Rep.*, vol. 8, p. 100235, 2022.
[30] U.S. Drug Enforcement Administration, "Social media drug trafficking threat overview," Tech. Rep., Washington, DC, USA, Feb. 2022. [Online]. Available: https://www.dea.gov/sites/default/files/2022-03/20220208-DEA_Social%20Media%20Drug%20Trafficking%20Threat%20Overview.pdf
[31] C. Yang, "Crimegnn: Harnessing the power of graph neural networks for community detection in criminal networks," *arXiv preprint*, arXiv:2311.17479, 2023.
[32] K. A. Carpenter, et al., "Which social media platforms facilitate monitoring the opioid crisis?," *PLOS Digit. Health*, vol. 4, no. 4, Art. no. e0000842, 2025.
[33] F. R. Lamy, S. C. Paek, and N. Meemon, "Online illicit drug distribution in the Thai language on X: Exploratory qualitative content analysis," *JMIR Infodemiology*, vol. 5, no. 1, Art. no. e71703, 2025.
[34] Y. Zhang, et al., "A survey on privacy in graph neural networks: Attacks, preservation, and applications," *IEEE Trans. Knowl. Data Eng.*, 2024.
[35] J. Chen, et al., "SCN_GNN: A GNN-based fraud detection algorithm combining strong node and graph topology information," *Expert Syst. Appl.*, vol. 237, Art. no. 121643, 2024.
[36] M. Arshad, et al., "Investigating methods for forensic analysis of social media data to support criminal investigations," *Front. Comput. Sci.*, vol. 7, Art. no. 1566513, 2025.
[37] M. Aos, B. Qolomany, K. Gyorick, J. Bou Abdo, M. Aledhari, J. Qadir, K. Carley, and A. Al-Fuqaha, "A survey of social cybersecurity: Techniques for attack detection, evaluations, challenges, and future prospects," Comput. Hum. Behav. Rep., vol. 18, 100668, 2025. doi: 10.1016/j.chbr.2025.100668.
[38] A. Shen and K.-P. Chow, "Community detection framework using deep learning in social media analysis," *Appl. Sci.*, vol. 14, no. 24, 2024.
[39] D. Contreras-Velasco, L. J. García-Vega, and R. Criado, "Uncovering hidden alliances in organized crime networks with machine learning: From node similarity to graph neural networks," Socio-Cognitive Systems, vol. 1, no. 1, pp. 1–18, 2025, doi: 10.1007/s42001-025-00429-0.

Eun-Young Park received the B.S. degree in computer engineering from Daegu University, Gyeongsan, South Korea, in 2025, where she is currently pursuing the M.S. degree in computer engineering. Her research interests include cybersecurity, artificial intelligence, and digital forensics.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3657961



Hyeon-Woo Lee received the B.S. degree in computer engineering from Daegu University, Gyeongsan, South Korea, in 2025, where he is currently pursuing the M.S. degree in computer engineering. His research interests include cybersecurity, artificial intelligence, and blockchain.

Jiyeon Kim received the B.S. and Ph.D. degrees in information security engineering from Seoul Women's University, Seoul, South Korea, in 2007 and 2013, respectively. Dr. Kim was a Postdoctoral Research Associate in the Department of Electrical and Computer Engineering, Carnegie Mellon University, United States, from 2014 to 2017. She is currently an Assistant professor in the Department of Computer Engineering, Daegu University, Gyeongsan, South Korea. Her research interests include cybersecurity, cybercrime investigation and cloud computing.