

Research on the correlation analysis of investigation clues and evidence chain reconstruction model based on multi-source data fusion

Jia Guo

International College, Krirk University, Bangkok, 10220, Thailand

Corresponding author: Jia Guo (e-mail: zhao20251018@163.com)

ABSTRACT To address the high discretization of multi-source heterogeneous data and the difficulty of reconstructing logically complete evidence chains in criminal investigations, this paper proposes a multi-source data fusion method for investigative clue association analysis. First, a spatiotemporal feature encoder maps text, image, and structured data into a unified semantic space. Second, a heterogeneous graph attention network is constructed with case entities as nodes to automatically discover deep associations among personnel, vehicles, communication devices, financial accounts, and locations. Third, a legal knowledge graph is introduced for logical pruning and consistency verification, transforming data-driven clue associations into legally interpretable evidence paths. Finally, a visualized closed-loop evidence-chain topology is generated to support judicial interpretability and investigator review. Experimental results show that the proposed model achieves 92.5% precision, 89.3% recall, a 93.5% implicit-connection detection rate, and an evidence-chain logical integrity score of 94.2, representing significant improvement over comparative methods. The model provides a technical framework for intelligent investigation, multimodal evidence fusion, and explainable evidence-chain reconstruction.

INDEX TERMS multi-source data fusion, investigative clue association, heterogeneous graph neural network, knowledge graph reasoning, evidence chain

I. INTRODUCTION

CRIMINAL investigation is a complicated job, which is highly dependent on information integration and logical reasoning. Its essence is to extract clues from massive heterogeneous information resources and establish a full chain of evidence [1-2]. In current criminal investigation practice, data sources are becoming more diverse, increasingly incorporating high-precision analysis of industrial fiber compositions found at crime scenes. The integration of these complex material datasets alongside traditional digital evidence necessitates more advanced fusion techniques to maintain the integrity of the investigative process. Communication records, surveillance videos, bank statements and onsite physical evidence are usually stored in separate systems with different formats and standards, presenting severe “data silos”. This discrete state makes it hard to discover implicit connections between clues, and establishing the full chain of evidence usually depends on experience of human beings, which is inefficient and even may lead to logic breaks or subjectivity missed by people, thus impacting the case detection rate and accuracy of judicial decision [3-4]. Therefore, how to realize deep integration

of multi-source heterogeneous data, automatically discover deep connections between clues, and reconstruct a legally valid closed-loop chain of evidence has become a key issue that needs to be urgently solved in intelligent investigation.

Domestic and foreign scholars have carried out extensive researches on the correlation analysis of investigation clues and the construction of the chain of evidence. Brown et al. studied the reasoning process of police investigators using forensic evidence such as fingerprints and DNA in common crimes, such as burglaries, in order to build a decision-making framework for the investigation of related cases [5]. Deslauriers-Varin and Fortin studied the direct and indirect relationships between artificial intelligence and supply chain turmoil and uncertainty, and the impact of artificial intelligence on supply chain resilience and performance [6]. Jurek et al. studied the relationship between evidence collection and case characteristics in sexual assault case investigations, and found that the victim’s testimony and cooperation significantly influenced the investigation process [7]. Rafizal and Sultan studied the impact of organizational culture, work environment and workload on the job satisfaction and performance of Pekanbaru police criminal investigators, and

found that none of the three had a great effect on job satisfaction [8]. However, most existing models focus on single data sources or static structures, making them insufficient for mining the spatiotemporal alignment and deep nonlinear correlations within multi-mode data—including the dynamic deformation patterns of high-performance industrial fibers.

To conquer these above issues, we attempt to construct investigation and judgment methods based on multisource data fusion and knowledge guidance. For instance, some methods utilize multimodal feature encoders to uniformly represent text and image data and integrate with graph convolutional networks for entity classification, but they are still poor in spatiotemporal alignment and dynamic relationship modeling. While another method attempts to introduce knowledge graph for reasoning. However, they ignore data-driven implicit association mining and cause the logical rules incompleteness problem. Therefore, in this paper, we propose a model for investigative clue association analysis and evidence chain reconstruction with the assistance of multimodal feature alignment, heterogeneous graph attention networks and legal knowledge graph reasoning. Firstly, spatiotemporal encoder realizes unified mapping of multi-source data. Secondly, heterogeneous graph attention network automatically mines deep relationship between entities. Finally, legal knowledge base logically prunes and reconstructs clue network and finally makes a leap from data relevance to legal evidence.

The main contributions of this paper can be summarized as follows: First, a multi-source heterogeneous data fusion framework for criminal investigation is constructed. The spatiotemporal encoder realizes unified mapping of multi-source data and solves data comparability problem. Second, a clue mining model based on heterogeneous graph attention network is designed. Heterogeneous graph attention network can automatically mine implicit connection between entities. Third, a legal knowledge graph is introduced. The legal knowledge graph logically verifies and reconstructs the evidence chain and makes the output chain logically rigorous and interpretable.

II. SPATIOTEMPORAL FEATURE ALIGNMENT AND SEMANTIC MAPPING OF MULTI-SOURCE HETEROGENEOUS DATA

The data collected in criminal investigations is usually highly multimodal. The call detail records and bank statements are tabular data, the transcripts of interrogations and case files are unstructured text, and the surveillance videos and photos from crime scenes are typical image data. Firstly, this paper conducts spatiotemporal normalization on all input data. For time information, the timestamps of different types of data are uniformly converted to Unix timestamp format. Recognizing that real-world investigative data sources often lack second-level precision—for example, bank records may only provide daily summaries, while witness statements may describe approximate time windows—the model adopts a flexible alignment strategy. Specifically,

high-precision data (e.g., surveillance timestamps) retain their original granularity, while low-precision or approximate timestamps are converted into time intervals with confidence weights. Temporal proximity between entities is then computed based on the overlap of these intervals rather than exact matching, allowing the model to logically handle missing or imprecise temporal information commonly encountered in heterogeneous investigative sources. For spatial information, all location information is mapped to the WGS-84 geographic coordinate system. For the precise coordinate information, such as surveillance points and base station sectors, they are converted into specific geographic coordinates according to the pre-defined mapping table; while for the non-precise location information, their precise coordinates can be estimated based on the coverage of base stations, and then the centroid of each base station sector is calculated. After the spatiotemporal base conversion, the data enters the feature extraction stage. For the textual data, we use pre-trained BERT model to encode the text and extract textual feature vectors with semantic information. For the image data, we use ResNet convolutional neural network to extract scene, target and behavior features. For the structured records, the discrete fields are transferred into dense vectors through embedding layer. On this basis, we design a Transformer-based cross-modal alignment module. This module takes time and spatial coordinates as query conditions to transfer the feature vectors from different modalities into the same high-dimensional semantic space. Then we select the candidate clues with high time and space overlap based on the similarity between features, which transfers a high-quality input set for subsequent association mining.

III. DEEP ASSOCIATION MINING OF CLUES BASED ON HETEROGENEOUS GRAPH ATTENTION NETWORK

After completing the spatiotemporal alignment and semantic mapping of multi-source data, this paper constructs a heterogeneous graph network with case entities as nodes and multiple relationships as edges. The specific definitions, representation symbols, and weight calculation basis of various nodes and edges are shown in Table 1.

TABLE 1. Node and Edge Type Definitions in Heterogeneous Graph Attention Networks

Node Type	Symbol	Number of Nodes (Example)	Edge Type	Edge Weight Calculation Basis
Person	P	156	Spatiotemporal Co-occurrence	Co-occurrence Frequency $\times$ Temporal Proximity
Vehicle	V	89	Ownership/Usage Relationship	Registration Information + Trajectory Matching Degree
Communication Device	D	203	Call/Message Contact	Frequency $\times$ Temporal Abnormality
Bank Account	A	112	Financial Flow	Amount $\times$ Frequency $\times$ Flow Direction
Location	L	45	Stay/Pass-by	Stay Duration $\times$ Frequency

Define the heterogeneous graph $G = (V, E, \phi, \psi)$, where V represents the set of nodes, E represents the set of edges, $\phi : V \rightarrow T_v$ is the node type mapping function, and $\psi : E \rightarrow T_e$ the edge type mapping function. Node types include personnel, vehicles, equipment, etc., and edge types include spatiotemporal co-occurrence, communication connections, fund transactions, etc.

For the central node v_i and its neighboring node set N_i^r of type r , the attention coefficient of node v_j to v_i is first calculated as follows (1):

$$e_{ij}^r = \text{LeakyReLU}(\mathbf{a}_r^T [\mathbf{W}_r \mathbf{h}_i \| \mathbf{W}_r \mathbf{h}_j]) \quad (1)$$

where $\mathbf{h}_i$ is the input feature of node v_i , $\mathbf{W}_r$ is the learnable transformation matrix corresponding to type r , $\mathbf{a}_r$ is the attention parameter vector, and $\|$ represents the vector concatenation operation. Then, the coefficients are normalized using the softmax function as follows (2):

$$\alpha_{ij}^r = \frac{\exp(e_{ij}^r)}{\sum_{k \in N_i^r} \exp(e_{ik}^r)} \quad (2)$$

The normalized attention weights reflect the importance of different types of neighboring nodes to the central node. The updated representation of node v_i is (3):

$$\mathbf{h}_i' = \sigma \left(\sum_{r \in T_e} \sum_{j \in N_i^r} \alpha_{ij}^r \cdot \mathbf{W}_r \mathbf{h}_j \right) \quad (3)$$

After propagation and aggregation through a multi-layer attention network, the final representation of each node not only includes its own features but also incorporates information from different types of neighbors within its multi-hop neighborhood. Compared to homogeneous graphs, heterogeneous graphs can more realistically reproduce the diversity of entities and relationships in an investigation scenario. However, different types of neighboring nodes have different meanings for the central entity. For example, for a suspect, a device node with which there is a high frequency of call records is obviously more valuable for analysis than a vehicle node that is only in the same location at a certain time. Therefore, a graph attention mechanism is introduced, enabling the model to dynamically learn the importance weights of different neighboring nodes. In detail, the HetGAT first computes attention coefficients for neighboring nodes across each relation type, applies a type-specific transformation to project features into a unified space, and

then assigns higher attention weights based on feature similarity. This allows the model to capture statistically salient patterns—such as frequent small-amount transactions or calls during atypical hours—as candidate signals for further investigation [9, 10]. It should be noted that these patterns serve as data-driven indicators of potential relevance, not as deterministic legal evidence. The final establishment of a legal relationship (e.g., conspiracy) relies on the contextual verification and logical pruning performed in Section 4.

IV. FUSING EVIDENCE LOGIC PRUNING AND REASONING FROM KNOWLEDGE GRAPHS

After propagation and aggregation of the heterogeneous graph attention network, the model generates a clue relationship network with multiple types of entities and diverse relationships. However, the objective of investigative analysis is not to reveal relationships but to establish an evidence chain that meets the requirements of law proof. In this paper, we build a legal knowledge graph for criminal investigation as a reasoning engine. This knowledge graph takes laws, regulations, judicial interpretations and investigative experience rules as knowledge sources, and formalizes abstract norms such as evidence rules and proof standards into computable knowledge representations. Concepts such as evidence types, facts to be proven and proof rules are represented as nodes in the knowledge graph, and logical rules define the relationships between concepts, such as “support,” “contradiction” and “reinforcement.” Such a structured knowledge graph provides a rule-based foundation for subsequent logical reasoning.

The first stage is logical pruning. The data-driven network of clues generated by the heterogeneous graph attention network is input into the knowledge graph, where each edge is assessed against a set of logical rules derived from legal norms and investigative experience. It is important to note that these rules are not used to predefine what constitutes a valid connection; rather, they serve as filters to remove associations that are logically impossible—such as a temporal contradiction where a person is recorded in two distant locations within an implausibly short time—while retaining statistically salient patterns that may represent novel implicit connections. This design ensures that the discovery process remains data-driven, with the knowledge graph acting as a post-hoc validator of logical consistency rather than a deterministic gatekeeper that restricts discovery to pre-programmed patterns. For example, associations generated

based on only one dimension (such as the same location) are regarded as weak and removed if unsupported by other dimensions. When conflicting evidence is encountered—such as a witness testimony contradicting a surveillance timestamp—the model does not simply prune either node; instead, it flags the contradiction as a logical inconsistency and retains both branches with a “conflict” relation, allowing the inconsistency to be highlighted for human review or resolved through additional evidence during the path reasoning stage in Section 4. This effectively reduces data noise, and the edges retained in the network are expected to have potential legal proof value. The second stage is path reasoning and evidence chain generation. In the pruned network, the model takes the core facts to be proven in the investigation as target nodes and uses a knowledge graph based path search algorithm to extract key paths that can form logical loops. In the search process, in addition to considering only the length and weight of paths, the model introduces logical consistency constraints, and when the supporting relationships between adjacent nodes in the paths conflict with logical rules, the model backtracks and reroutes the search. The final set of output paths are merged and deduplicated to generate several complete logical chains from the source evidence to the facts to be proven.

V. EVIDENCE CHAIN TOPOLOGY GENERATION FOR JUDICIAL INTERPRETABILITY

After logical pruning and reasoning based on the knowledge graph, the model selects legally supporting relational paths and forms several logical clues to the facts that need to be proven. However, these clues are still stored in the form of an internal graph structure or path set. It is neither intuitive for investigators to directly refer to their data-level output, nor can it be easily used as a reference for preparing legal documents or reporting cases.

Therefore, in this paper, we design an evidence chain topology generation module. The graph structure can intuitively display the supporting relationships and flow between various evidence nodes, and it conforms to the investigators’ cognitive characteristics of the “interlocking” relationships between evidence chains. In the process of generation, the module first merges and deduplicates multiple input logical paths. For example, the same physical evidence may support two facts to be proven: the time of the crime and the tool used to commit the crime. If they appear in two separate paths, only one node appears in the final topology graph in front of the investigator, preventing duplication and redundancy.

In terms of node design, we use different visual symbols to distinguish different types of evidence. Physical evidence, documentary evidence, witness testimony, electronic data, etc. all have different icons and colors. Allowing investigators to quickly identify them visually. Edges are represented by directed arrows, pointing from evidence nodes to the fact nodes to be proven. If there are multiple levels of support (e.g., witness testimony supports circumstantial evidence,

which in turn supports direct facts), they are connected together by multi-level arrows. To prevent circular reasoning, the generation module enforces the directed acyclic constraint graph in the generation process, ensuring that there is a clear logical flow. To improve interpretability, each node and edge in the topology graph support interactive clicking to view details. When clicking on an evidence node, it expands the original data summary: the time and location of the surveillance footage, the parties involved and time of communication records. When clicking on a supporting edge, it displays the legal rules or reasoning basis (e.g., “same time and space + communication contact” conforms to the rules of evidence for conspiracy). This allows the investigator to trace the generation logic of the model’s conclusions, understand why some clues were retained and why some connections were removed, thereby increasing trust in the model’s output.

VI. MODEL PERFORMANCE AND EVIDENCE CHAIN CONSTRUCTION EFFECT EVALUATION

A. EXPERIMENTAL SETUP

To fully verify the effectiveness of the multi-source data fusion investigation clue association analysis and evidence chain reconstruction model proposed in this paper, the experimental design revolves around three core dimensions: the accuracy and comprehensiveness of clue mining, the ability to discover implicit associations, and the quality of logical reconstruction of the evidence chain. The experimental data comes from a de-identified dataset of 100 closed cases provided by a municipal public security bureau, covering multisource heterogeneous data such as call records, checkpoint captures, bank statements, on-site records, and physical evidence information; at the same time, to test the model’s ability to perceive deep associations, an additional simulated case dataset containing artificially implanted implicit associations was constructed.

The baseline methods include: (1) traditional clue association methods based on keyword matching; (2) graph neural network models using only a single data source (such as communication records); (3) simple fusion models without spatiotemporal alignment; (4) GCN (Graph Convolutional Network) based on static graph structures; (5) pure data-driven graph models without knowledge graph reasoning; and (6) IBM i2 Analyst’s Notebook, a commercial association analysis tool. The evaluation metrics encompass objective quantitative indicators and subjective expert scores, specifically including precision, recall, F1 score, implicit association detection rate, evidence chain logical integrity score (calculated based on legal rule coverage and path closure), and interpretability and legal validity scores (1-5 points) given by senior criminal investigation experts. All experiments were conducted in the same hardware environment (NVIDIA Tesla V100 GPU, 32GB RAM), using five-fold cross-validation to ensure result stability.

B. MULTI-SOURCE DATA FUSION AND CLUE MINING PERFORMANCE EVALUATION EXPERIMENT

To verify the effectiveness of the model in mining clues from multi-source data, we design a comparative experiment on anonymized data of 100 closed cases, whose multisource data include communication, surveillance, and financial data. With the task of distinguishing the involved objects from the data, we make a comparison with the keyword matching, single-source GNN, unaligned fusion, and i2 tool in terms of precision, recall, and F1 score.

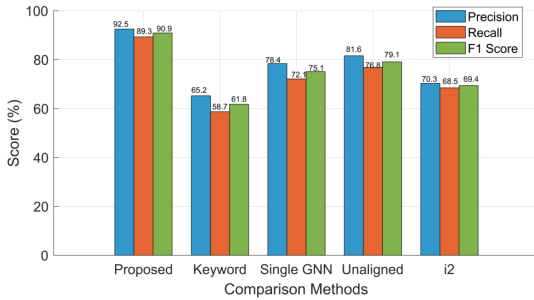


FIGURE 1. Performance Evaluation of Multi-Source Data Fusion and Clue Mining

As shown in Figure 1, the multi-source data fusion model proposed in this paper performs excellently in entity mining tasks, achieving a precision of 92.5%, a recall of 89.3%, and an F1 score of 90.9%. In contrast, the F1 score of traditional keyword matching methods is only 61.8%, and the F1 score of single-source graph neural network models, limited by information dimensionality, is 75.1%. Even when using multi-source data without spatiotemporal alignment, the F1 score is only 79.1%, lower than that of the model in this paper. This indicates that through spatiotemporal alignment and heterogeneous graph attention mechanisms, the model can effectively fuse multimodal data, significantly improving the accuracy and comprehensiveness of clue mining.

C. EXPERIMENTAL EVALUATION OF IMPLICIT ASSOCIATION DETECTION ABILITY

To explore the deep implicit associations mined by the model, we construct a simulated case dataset with second-order and higher-order associations implanted artificially. We compare the proposed model with static graph convolutional networks (GCN), heterogeneous graph models without attention mechanism and traditional rule tools (i2) from DeepMath paper by implicit association detection rate.

TABLE 2. Assessment of Latent Association Detection Capability

Methods	Hidden Relationship Discovery Rate (%)	Average Inference Time (s)
Proposed Model	93.5	2.1
Static GCN	78.2	1.2
Heterogeneous Graph without Attention	81.7	1.8
Rule-based Tool i2	62.4	0.8

In Table 2, our model achieves a latent association detection rate of 93.5%, which is 15.3 percentage points higher than the static GCN (78.2%), 11.8% higher than the heterogeneous graph model without attention mechanism (81.7%), and 31.1% higher than the rule tool i2 (62.4%). The results indicate that the heterogeneous graph attention mechanism can effectively capture nonlinear deep dependencies between nodes, significantly enhancing the ability to perceive hidden cues. Although the model's inference time is slightly higher than the comparison methods, it is within an acceptable range, demonstrating its superiority in complex association mining.

D. EVALUATION EXPERIMENT OF EVIDENCE CHAIN LOGICAL RECONSTRUCTION EFFECT

Using anonymized data involving 100 closed cases, a standard evidence chain was built by senior criminal investigation specialists as the standard. The model in this paper was compared with the evidence chains built by data association clustering only, GNN without knowledge graph, rule tool i2, and template generation method, by taking logical integrity score (based on legal rule coverage and path closure) and human scoring (1-5 points, experts score interpretability and legal validity) as the evaluation indicators. Each method was repeated 5 times. See Figure 2 for details:

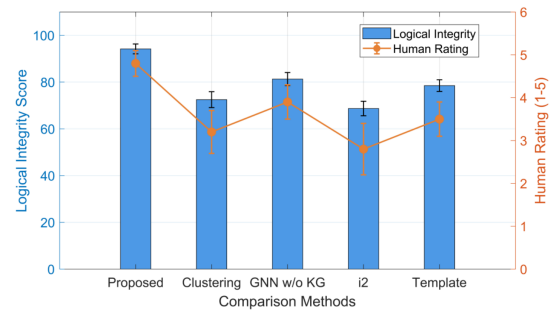


FIGURE 2. Evaluation of Evidence Chain Logical Reconstruction Effect

The logical integrity score of the model presented in this paper is 94.2 ± 2.1 , an improvement of 12.9 points compared to GNN without knowledge graph (81.3 ± 2.8), an improvement of 21.7 points compared to data association clustering only (72.5 ± 3.4), an improvement of 25.5 points compared to rule tool i2 (68.7 ± 3.1), and an improvement of 15.7 points compared to template generation method (78.5 ± 2.5). In terms of human scoring, our model achieved

a score of 4.8 ± 0.3 (out of 5), significantly higher than the scores of other methods (3.9, 3.2, 2.8, and 3.5). The results indicate that integrating knowledge graph reasoning can effectively verify the logical support relationships between evidence, significantly improving the rigor and legal validity of the evidence chain.

VII. CONCLUSION

This paper proposes a multi-source data fusion model for investigative clue association analysis and evidence chain reconstruction. By integrating spatiotemporal feature alignment, heterogeneous graph attention networks, and legal knowledge graph reasoning, the model effectively addresses the challenges of heterogeneous data integration and implicit relationship mining. Experimental results demonstrate that the proposed model achieves a precision of 92.5%, recall of 89.3%, implicit association detection rate of 93.5%, and evidence chain logical integrity score of 94.2, significantly outperforming baseline methods. The findings confirm that the model can enhance the accuracy and interpretability of evidence chain construction in criminal investigations. Limitations include the reliance on the completeness of the legal knowledge base and the relatively small sample size. Future work will focus on expanding the knowledge graph, incorporating few-shot learning for sparse data scenarios, and validating the model in real-world investigative settings.

AUTHOR CONTRIBUTIONS

Jia Guo designed, collected and analyzed the data, and drafted the manuscript. Jia Guo conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Jia Guo participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] S. Ageed Z, M. Zeebaree S R, M. Sadeeq M, et al., "Comprehensive survey of big data mining approaches in cloud systems," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 29-38, 2021, doi: 10.48161/qaj.v1n2a46.
- [2] R. S. A. Faqir, "Digital Criminal Investigations in the Era of Artificial Intelligence: A Comprehensive Overview," *International Journal of Cyber Criminology*, vol. 17, no. 2, pp. 77-94, 2023, doi: 10.5281/zenodo.4766706.
- [3] C. Horan and H. Saiedian, "Cyber crime investigation: Landscape, challenges, and future research directions," *Journal of Cybersecurity and Privacy*, vol. 1, no. 4, pp. 580-596, 2021, doi: 10.3390/jcp1040029.
- [4] R. Mahajan and V. Mansotra, "Correlating crime and social media: using semantic sentiment analysis," *International journal of advanced computer science and applications*, vol. 12, no. 3, pp. 309-316, 2021, doi: 10.14569/IJACSA.2021.0120338.
- [5] C. Brown, R. Julian, and M. Howes L, "'Strands in a cable': effective investigator decision-making using forensic identification evidence in volume crime investigations," *Policing and Society*, vol. 34, no. 5, pp. 417-433, 2024, doi: 10.1080/10439463.2023.2279063.
- [6] N. Deslauriers-Varin and F. Fortin, "Improving efficiency and understanding of criminal investigations: Toward an evidence-based approach," *Journal of Police and Criminal Psychology*, vol. 36, no. 4, pp. 635-638, 2021, doi: 10.1007/s11896-021-09491-6.
- [7] L. Jurek A, M. Kelley S, and W. Wells, "Criminal investigative effort in sexual assaults: Findings from a sample of cases with unsubmitted sexual assault kits," *Journal of police and criminal psychology*, vol. 36, no. 4, pp. 652-666, 2021, doi: 10.1007/s11896-021-09482-7.
- [8] J. Rafizal and M. Sultan F M, "Organizational Culture, Work Environment, and Workload on the Performance of POLRI Members (Case Study in Pekanbaru Police Criminal Reserve Unit)," *Journal of Applied Business and Technology*, vol. 3, no. 3, pp. 263-271, 2022, doi: 10.35145/jabt.v3i3.111.
- [9] Y. Liu, L. Fan, X. Wang, et al., "HGBER: Heterogeneous graph neural network with bidirectional encoding representation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 7, pp. 9340-9351, 2023, doi: 10.1109/TNNLS.2022.3232709.
- [10] J. Zhang and Q. Xu, "Attention-aware heterogeneous graph neural network," *Big Data Mining and Analytics*, vol. 4, no. 4, pp. 233-241, 2021, doi: 10.26599/BDMA.2021.9020008.