Semantic-Structure Based Graph Attention Networks for Dialogue Intention Classification in Fraud Detection

Yuji Nakashima
AAII, Faculty of Data Science
Musashino University
Tokyo, Japan
s2422018@stu.musashino-u.ac.jp

Virach Sornlertlamvanich
AAII, Faculty of Data Science
Musashino University
Tokyo, Japan
virach@musashino-u.ac.jp

Titipakorn Prakayaphun
AAII, Faculty of Data Science
Musashino University
Tokyo, Japan
titipakorn@musashino-u.ac.jp

Thatsanee Charoenporn
AAII, Faculty of Data Science
Musashino University
Tokyo, Japan
thatsanee@ds.musashino-u.ac.jp

2025 9th International Conference on Information Technology (InCIT) | 978-1-6654-7748-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/INCIT66780.2025.11276108

***Abstract*—Telephone fraud often involves criminals impersonating relatives or public officials in conversational contexts. Conventional sentence-level detection methods fail to capture the semantic flow and relational structure of such fraudulent dialogues. In this study, we construct a unique fraud dataset collected from official Japanese police websites, categorized into four classes: police, city hall, a relative, and a bank.**

We propose a Semantic-Structure Graph Attention Network (SS-GAT) that explicitly models conversational semantic relations between utterances. Unlike conventional transformers that treat conversations as linear text, our approach represents dialogue as a graph, where utterances are nodes and edges capture semantic and structural relations (e.g., question–answer, instruction–response). Graph Attention Network (GAT) mechanisms are applied to weight relevant contextual relations for intention classification.

In our experiments, using an architecture with a single GAT layer and frozen BERT, we achieved an average accuracy of 93.58% and F1-score of 94.59%. In comparison, baseline transformers failed to classify many classes. Notably, frauds involving impersonation of bank employees were the most difficult to detect due to their procedural, instruction-driven conversational style, which closely resembles frauds by other public institutions.

***Index Terms*—Fraud detection, conversational modeling, graph neural networks, GAT, semantic relations**

I. Introduction

A. Background

Telephone fraud has become increasingly sophisticated, with criminals impersonating trusted figures such as relatives, city hall employees, police officers, or bank officials. These conversations follow structured patterns: fraudsters establish credible identities, create urgent situations, and guide victims through step-by-step instructions. As of December 2024, reported special fraud cases reached 3,494 with losses totaling 15.31 billion yen, highlighting the critical need for early detection systems.

B. Problem Statement

Current detection methods analyze individual utterances in isolation, failing to capture the conversational flow and relational structures that characterize fraud interactions. While Bidirectional Encoder Representations from Transformers (BERT) and other transformer models excel at text classification, they model linear sequences rather than interactive dialogue structures. The challenge lies in detecting manipulative intent embedded in dialogue patterns rather than isolated fraudulent words.

C. Our Approach and Contributions

We propose a Semantic-Structure Graph Attention Network (SS-GAT) that represents conversations as graphs where utterances are nodes and edges capture semantic relations (e.g., Question–Answer (Q–A), instruction–response). Unlike prior binary fraud detection approaches, we focus on four-way typology classification using discourse-level patterns.

Our contributions are threefold:

1) We construct a novel dataset from official Japanese police sources with four impersonation categories (relative, city hall, police, bank) and balance it via carefully controlled Large Language Model (LLM) augmentation.
2) We propose SS-GAT that represents conversations as utterance graphs connected by semantic and structural relations, allowing attention to focus on relational cues indicative of manipulative intent.
3) We benchmark SS-GAT against transformer baselines and show significantly higher accuracy and F1-scores for four-way typology classification.

This research contributes to the growing field of conversational AI and its application to social good by offering a practical solution for early fraud detection.

II. Related Work

This work proposes SS-GAT, an original model that integrates a Japanese BERT with a Graph Attention Network (GAT). Unlike approaches that focus on extracting keywords or key sentences, we define explicit relations between utterances, construct a conversational graph, and model the semantic flow across the dialogue. We organize prior work into three groups—key-sentence approaches, BERT/Transformer-based approaches, and GAT-based approaches—and clarify how our work relates to and differs from each.

A. Key-Sentence-Based Approaches

Gao et al. (2024) [1] introduce a Multi-Level Dynamic TextRank (MDTR) algorithm to extract semantically coherent key-sentence summaries from long dialogues, then embed the summarized text with BERT and enhance features via hierarchical attention. They also design a keyword list and compute keyword frequencies for fraud and non-fraud data, derive a "degree of correlation" for each keyword, and use highly correlated keywords for fraud detection. Earlier telephone conversation linguistic fraud modeling using linguistic features is reported in [2]. In addition, recent studies have proposed cost-sensitive graph models, RoBERTa-based approaches, dynamic sparse attention, real-time AI detection, voice phishing detection, and call content understanding for fraud detection [3]–[8].

B. Transformer-Based Approaches

BERT is a bidirectional pre-trained encoder that achieves strong performance across tasks via masked language modeling (MLM) and next sentence prediction (NSP) [9], [10]. Wang et al. (2021) [11] fine-tune a multi-label BERT for outbound fraud-risk detection, using the final hidden state of the [CLS] token for classification. Recent work also includes RAG-based LLMs [12], real-time phone scam detection [13], [14], and large audio language models for telecom fraud [15]. Gao et al. (2024) [1] present HDRIN, combining DialogBERT, sliding windows, MDTR-based summarization, and Hierarchical Attention Networks (HAN) to capture dual-role interactions. Li et al. (2024) propose RoBERTa-MHARC, coupling RoBERTa with multi-head attention and residual connections to better capture multifaceted contextual cues. Qin et al. (2020) report that their Co-GAT framework benefits further when coupled with pre-trained encoders such as BERT, RoBERTa, or XLNet. Lin et al. (2024) [16] introduce FraudGT, a Graph Transformer with attention mechanisms and edge-aware components that addresses limitations of message-passing GNNs.

C. GAT-Based Approaches

Recent GAT-based approaches include joint dialog act recognition and sentiment classification using co-interactive graph layers with multi-head attention [17]–[19]. Gao et al. (2024) [1] present RCGN, combining GCN, GAT, and GraphSAGE to detect fraud leaders in telecom networks, using GAT to assign adaptive weights to neighbors and mitigate error propagation. Hu et al. (2023) [20] survey graph-based methods (GCN, GAT, GraphSAGE) for mobile network fraudster mining, highlighting efficiency in financial and social fraud scenarios. Wang et al. (2019) develop a semi-supervised GAT for financial fraud detection. Liang et al. (2021) evaluate RGAT (relation-aware GAT) for conversational emotion recognition, leveraging relation attention and relative positional encoding. Lin et al. (2024) [16] benchmark GAT as a strong baseline when introducing FraudGT. Hu et al. (2024) propose GAT-COBO, a cost-sensitive boosted GAT that addresses class imbalance in telecom fraud detection by focusing attention on the most informative parts. Other notable works include emotion recognition and speaker/position-aware graph models for conversation analysis [21]–[23].

III. Dataset

The dataset for this study is based on audio data published on official Japanese police websites. These audio recordings are authentic accounts of fraud cases and serve as a reliable resource for training and evaluation. The dataset includes the following four fraud categories:

Relation annotation was conducted by a native Japanese annotator (the first author) following a concise, example-based guideline tailored to telephone conversations. We specified relation labels for each sequential dialogue—Introduction/Greeting, Explanation/Guidance, Request/Demand, Question/Confirmation, Fraud Action, Denial/Refusal, Response/Reply, Switch/Transition, and Emotional Expression—using minimal decision cues (e.g., interrogative forms, request constructions, and confirmation patterns), and provided dialogue examples in Table I. Formal assessment of inter-annotator agreement was not undertaken due to resource limitations.

1) Relative impersonation fraud: the perpetrator pretends to be a family member in trouble.
2) City hall impersonation fraud: the criminal claims to represent municipal offices.
3) Police impersonation fraud: the criminal pretends to be a law enforcement officer.
4) Bank impersonation fraud: the criminal poses as a financial institution employee.

The dataset contains 22 data instances for each class.

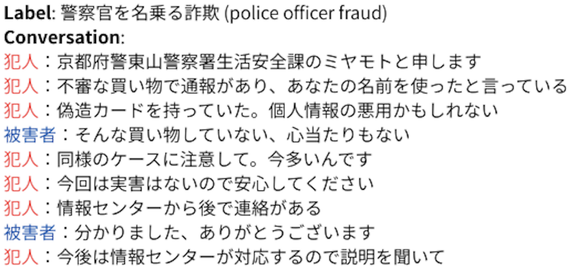
IV. Methodology

A. Input Processing and Text Encoding

The SS-GAT framework processes dialogue data beginning with individual **Utterances** as the fundamental input units. Each utterance represents a conversational turn containing both textual content and metadata. The **Text Encoder** utilizes a **pre-trained BERT** model (cl-tohoku/bert-base-japanese-v3) to transform each utterance into semantic representations. This pre-trained model is essential because it provides rich contextual understanding of Japanese language patterns, eliminating the need to learn basic language semantics from scratch on limited fraud detection data.

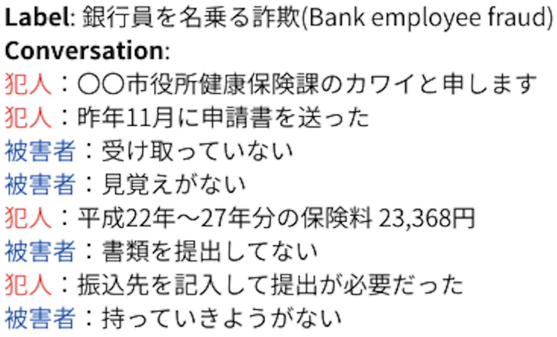
TABLE I
Utterance Pairs and Relations

Relation	Utterance (Speaker: Content)
Self-introduction / Greeting	Perpetrator: I am Kawai from the Health Insurance Division of XX City Hall.
Explanation / Guidance	Perpetrator: We sent the application form last November.
Denial / Refusal	Victim: I haven’t received it.
Denial / Refusal	Victim: I don’t recall it.
Fraudulent Action	Perpetrator: Insurance premiums for Heisei 22–27: 23,368 yen.
Fraudulent Action	Perpetrator: You didn’t submit the documents.
Request / Demand	Perpetrator: You were required to fill in the transfer destination and submit.



Label: 警察官を名乗る詐欺 (police officer fraud)
Conversation:
犯人：京都府警東山警察署生活安全課のミヤモトと申します
犯人：不審な買い物で通報があり、あなたの名前を使ったと言っている
犯人：偽造カードを持っていた。個人情報の悪用かもしれない
被害者：そんな買い物していない、心当たりもない
犯人：同様のケースに注意して。今多いんです
犯人：今回は実害はないので安心してください
犯人：情報センターから後で連絡がある
被害者：分かりました、ありがとうございます
犯人：今後は情報センターが対応するので説明を聞いて

Fig. 1. Dataset distribution analysis showing the four fraud categories (relative, city hall, police, bank) with balanced class representation. Each category contains 22 authentic fraud case recordings from official Japanese police websites, totaling 88 instances for the base dataset.



Label: 銀行員を名乗る詐欺(Bank employee fraud)
Conversation:
犯人：〇〇市役所健康保険課のカワイと申します
犯人：昨年11月に申請書を送った
被害者：受け取っていない
被害者：見覚えがない
犯人：平成22年～27年分の保険料 23,368円
被害者：書類を提出してない
犯人：振込先を記入して提出が必要だった
被害者：持っていきようがない

Fig. 2. Dialogue length distribution and utterance statistics across fraud categories. The average dialogue length is 18 utterances with standard deviation of 4.3, showing consistent conversational patterns across different fraud types that enable systematic graph-based analysis.

The BERT model processes each utterance through its transformer architecture, with the final [CLS] token representation serving as a 768-dimensional semantic vector that captures the utterance’s complete contextual meaning. To address the challenge of small-scale datasets prone to overfitting, these high-dimensional features undergo compression from 768 to 128 dimensions through a linear transformation followed by ReLU activation and layer normalization.

B. Multi-Head Attention Enhancement

The compressed text representations are enhanced through Multi-head attention mechanisms that capture intra-dialogue dependencies before graph processing. This attention layer is crucial for identifying relationships between utterances within the same conversation, allowing the model to understand how different parts of a dialogue relate to each other semantically. The multi-head design enables parallel attention computation across different representation subspaces, capturing diverse types of utterance relationships simultaneously.

Following the attention computation, Add & Norm operations are applied, implementing residual connections combined with layer normalization. These operations are essential for training stability, preventing gradient vanishing problems, and allowing information from the original compressed representations to flow through alongside the attention-enhanced features.

C. Speaker and Relation Encoding

The framework incorporates multimodal information through specialized encoders. The Speaker Encoder processes Speaker ID information, distinguishing between perpetrator and victim roles through learnable 128-dimensional embeddings. This speaker-aware modeling is critical for fraud detection because fraudulent conversations exhibit distinct patterns based on participant roles, with perpetrators typically following specific deception strategies while victims display characteristic response patterns.

The Relation Encoder processes Edge relations that capture semantic relationships between utterances. Ten distinct relation types are supported: Introduction/Greeting, Explanation/Guidance, Request/Demand, Question/Confirmation, Fraud Action, Denial/Refusal, Response/Reply, Switch/Transition, and Emotional Expression. Each relation type receives a learnable scalar weight that modulates attention computation during graph message passing, allowing the model to automatically emphasize fraud-indicative relations such as "Fraud Action" while de-emphasizing routine conversational elements.

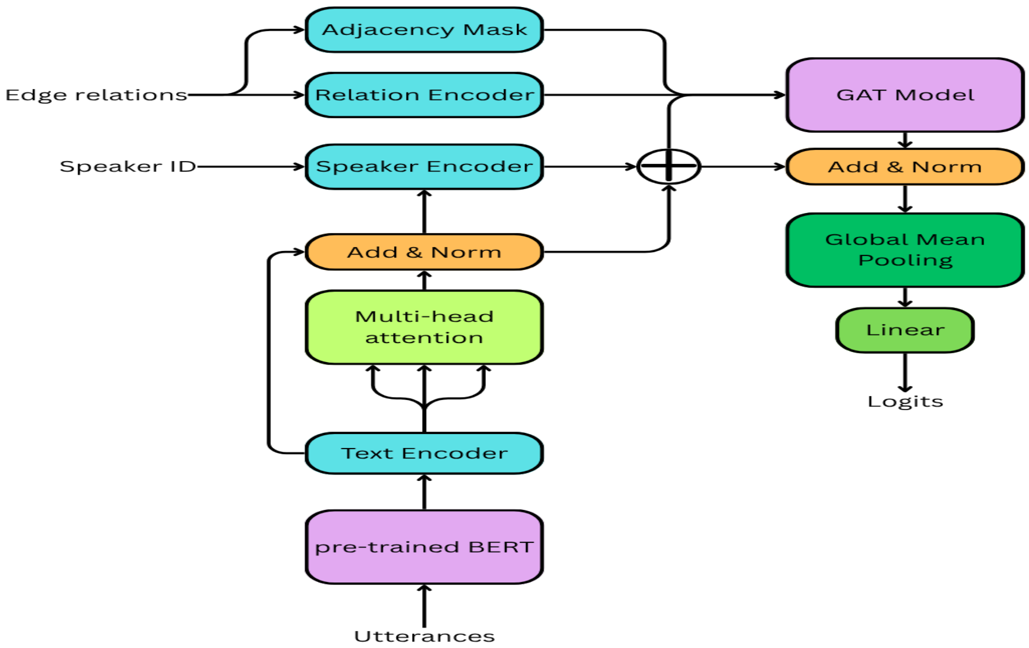


Fig. 3. SS-GAT framework overview showing the processing pipeline from input dialogue through graph construction to classification. The architecture consists of three main components: (1) Text-First Processing module that enhances BERT embeddings with multi-head attention, (2) Graph Construction module that integrates speaker and relation information with hard adjacency constraints, and (3) GAT Processing layers with explicit edge-based attention for fraud detection.

D. Hard Graph Construction

The core innovation lies in the hard graph structure where nodes represent individual utterances and edges represent explicit adjacency relationships. An Adjacency Mask enforces structural constraints derived from human-annotated graph structures, preventing attention computation over non-connected utterance pairs. This hard constraint approach is essential because it ensures the model respects actual conversational flow and turn-taking patterns, which are critical indicators in fraud detection scenarios.

Unlike soft attention mechanisms that can learn spurious relationships, the adjacency mask guarantees that the GAT Model focuses only on meaningful conversational connections. This constraint is particularly important for fraud detection where the sequence and structure of deceptive tactics follow predictable patterns that must be preserved during learning.

E. Graph Attention Network Processing

The GAT Model implements the core graph reasoning through hard attention constraints. The model computes attention weights between connected utterances while completely ignoring non-adjacent pairs, as defined by the adjacency mask. Each attention head in the multi-head architecture captures different aspects of utterance relationships, with relation-specific weights modulating these attention scores based on edge types.

The GAT layer processes the combined node features (text & speaker information) and propagates information through the graph structure, allowing each utterance to aggregate relevant information from its connected neighbors. This message passing mechanism enables the model to understand complex conversation dynamics where the meaning of individual utterances depends on their context within the broader dialogue structure.

F. Classification and Output Generation

Following graph processing, node representations undergo residual connections and layer normalization to maintain training stability. Global Mean Pooling then aggregates the variable-length dialogue sequences into fixed-size representations, computing the average of all node embeddings to create a dialogue-level feature vector. This pooling operation is necessary because dialogues contain varying numbers of utterances, but the classification layer requires fixed-dimensional input.

The final Linear classification layers transform the pooled dialogue representation into class-specific scores. A two-layer MLP architecture with ReLU activation processes the 128-dimensional pooled features, ultimately producing Logits for the four fraud categories: bank employee impersonation, police officer impersonation, family member impersonation, and city hall employee impersonation. These logits represent unnormalized class probabilities that are subsequently processed through softmax and cross-entropy loss for training.

G. Architectural Advantages

The SS-GAT architecture addresses key challenges in fraud detection through its design choices. Frozen BERT parameters reduce overfitting risk while maintaining semantic understanding, compressed dimensions (768→128) match the small dataset scale, and hard graph structure prevents spurious relationship learning. The combination of text-first processing with explicit graph constraints enables robust performance on limited fraud detection datasets while maintaining interpretability through explicit attention mechanisms and relation-aware edge modeling.

V. Experiments

We employ a frozen pre-trained Japanese BERT as the text encoder, projecting its 768-dimensional embeddings to 128 via a linear layer with ReLU and layer normalization to reduce overfitting on the small-scale fraud detection dataset. The graph component is a single-layer GraphAttentionNetwork with one attention head, residual connections, and layer normalization; attention is constrained by explicit adjacency masks from human-annotated dialogue structures, and ten relation types each have a learnable scalar weight that modulates attention during message passing.

Training uses AdamW with a learning rate of 1×10^{-3}, L2 weight decay of 1×10^{-5}, gradient clipping at a max norm of 1.0, dropout of 0.1 for GAT and 0.3 for baselines, a batch size of 16, and 20 epochs. Evaluation follows stratified five-fold cross-validation, reporting mean and standard deviation across folds; all experiments are implemented in PyTorch on an NVIDIA GPU.

A. Results

TABLE II
Comparison with Transformer baseline.

Model	Accuracy	F1 score
SS-GAT	**93.58%**	**94.59%**
Transformer	84.67%	83.54%
dialogue BERT	80.00%	80.73%
key sentence BERT	55.38%	55.37%
LSTM	89.67%	84.98%
BiLSTM	92.25%	91.12%

Table II summarizes the comparative results across baseline and proposed models. SS-GAT achieves the best performance, reaching 93.58% accuracy and 94.59% F1, substantially outperforming all baselines. Among transformer-based baselines, the vanilla Transformer attains 84.67% accuracy and 83.54% F1. The dialogue BERT model, which processes inputs at the granularity of individual sentences, yields 80.00% accuracy and 80.73% F1, while the key sentence BERT model, operating at the granularity of individual words, performs markedly worse with 55.38% accuracy and 55.37% F1. Recurrent baselines show competitive results: LSTM achieves 89.67% accuracy and 84.98% F1, and BiLSTM improves further to 92.25% accuracy and 91.12% F1. Overall, SS-GAT delivers a clear margin over both transformer-based and recurrent alternatives, highlighting the effectiveness of the proposed approach.

VI. Discussion

A. Effectiveness of Relation Information

Incorporating relation information into the graph attention mechanism enhances structural modeling of dialogue. The model assigns learnable scalar weights to ten relation types (e.g., greeting/identification, explanation/guidance, request, question/confirmation, fraudulent action, denial/rejection, response, topic shift, emotion expression, unlabeled). This approach enables differential emphasis on interaction types that are salient for fraud detection; for instance, fraudulent actions can be amplified during feature aggregation, improving sensitivity to key manipulative moves.

The simplicity of scalar weighting offers practical advantages over more complex relation embeddings by reducing parameterization and risk of overfitting in low-resource settings. However, two constraints limit effectiveness. First, the use of a hard graph structure prevents attention to nodes without annotated edges, which curbs spurious connections but may omit meaningful long-range dependencies typical of conversational manipulation. Second, performance is sensitive to annotation quality: mislabeled or missing relations propagate as structural errors that degrade inference.

Empirical comparison with a strong text-only baseline shows modest gains, indicating that relation types contribute useful but incremental information. This outcome suggests that current label granularity may be insufficient to capture the nuanced progression of fraud. Future work should evaluate hierarchical taxonomies and richer relation representations that encode intensity, temporality, and compositionality.

B. Effectiveness of Contextual Information

The architecture emphasizes contextual understanding through a pipeline that leverages pretrained language representations, intra-dialogue attention, and structure-aware propagation. Pretrained Japanese language representations provide robust semantics while mitigating overfitting in a small data regime. Dimensionality reduction serves as regularization, retaining salient fraud-related cues while improving generalization.

Multi-head attention supports the modeling of diverse contextual phenomena, such as turn-taking dynamics, topical coherence, and pragmatic signals. Nonetheless, limited head count and reduced hidden dimensionality constrain capacity to capture heterogeneous patterns prevalent in complex dialogues. A further limitation is the absence of explicit positional or temporal encoding beyond the implicit sequential structure of pretrained representations. Fraud conversations often unfold through phases (trust-building, information elicitation, monetary request); explicit modeling of temporal progression could strengthen sensitivity to staged manipulation.

C. Effectiveness of Speaker Information

Explicit representation of speaker roles (fraudster and victim) adds an interaction-centric signal that complements textual content. Speaker-aware modulation allows identical utterances to be interpreted differently depending on the source, reflecting role-specific pragmatics and intent. This is particularly relevant in fraud contexts, where power dynamics, turn control, and persuasion strategies differ systematically between participants.

Effectiveness is tempered by simplifying assumptions. Binary role assignment may not capture multi-party interactions or shifting roles over time. Moreover, a basic integration mechanism can underrepresent complex speaker-conditioned patterns such as strategic politeness, hedging, or escalation. A more expressive design—e.g., role hierarchies, dynamic role inference, or adaptive gating that learns when to prioritize speaker versus textual cues—may yield stronger gains. Notably, speaker information can meaningfully guide information flow in the graph, for example by emphasizing victim-centric context when tracing manipulative tactics.

D. Effectiveness of Graph Attention Networks

A hard-constraint graph attention formulation brings structural rigor to dialogue modeling by restricting attention to annotated edges. This design embodies the intuition that conversational structure—turn adjacency, topical links, and rhetorical relations—should delimit information propagation. The resulting improvements over a text-only baseline are consistent yet modest, suggesting that while structural signals are beneficial, much of the useful context is already captured by pretrained sequential models.

The hard-constraint paradigm enhances interpretability, as attention weights can be analyzed in relation to explicit discourse relations. However, it limits discovery of latent or implicit links, which are often crucial in detecting subtle manipulation that spans distant turns. Model depth and capacity are intentionally constrained to prevent overfitting in a small dataset, trading expressive power for stability. These choices are appropriate for the current data scale but may cap performance.

Generalization remains a challenge: if test dialogues exhibit relation patterns unseen during training, the model has limited means to adapt. Hybrid designs that combine hard structural constraints with learnable soft attention for unannotated or uncertain links could improve robustness. Finally, the approach is computationally efficient due to sparsity induced by the structural mask, making it suitable for practical applications.

Overall, relation signals, contextual modeling, and speaker information each contribute complementary strengths. Their combined effect is positive but bounded by data scale, annotation fidelity, and model capacity. Future research should prioritize richer and more consistent annotations, explicit modeling of temporal progression, more expressive speaker role representations, and hybrid structural mechanisms. Larger corpora would support deeper graph architectures and more nuanced relation models, enabling stronger gains in fraud detection performance.

VII. Conclusion

This work shows that combining semantic understanding with explicit dialogue structure improves fraud detection. Our proposed SS-GAT (semantic-structure-based Graph Attention Network) integrates pretrained language representations with relation- and speaker-aware graph reasoning. The improvements over strong text-only baselines are modest but consistent and interpretable, supporting the view that modeling discourse relations and participant roles adds value beyond sequential text processing.

SS-GAT has three main strengths: it unifies semantic and structural signals, offers transparent behavior through structured attention, and remains computationally efficient. Nonetheless, performance is limited by small datasets, coarse relation labels, and shallow graph capacity for capturing long-range, implicit manipulation.

Future work should focus on:

1) **Richer annotations:** expand relation types and improve labeling quality.
2) **Temporal modeling:** add explicit timing and dialogue phase indicators.
3) **Role modeling:** move beyond binary roles and allow roles to change over time.
4) **Hybrid attention:** combine hard structural constraints with learnable soft links to discover unseen relations.
5) **Scalable architectures:** explore deeper or hierarchical GAT variants with careful regularization.
6) **Data expansion and transfer:** build larger, more diverse corpora and study domain adaptation.
7) **Interpretability:** provide clearer explanations of how relations and roles influence predictions.

Overall, SS-GAT is a practical and extensible approach that aligns model decisions with dialogue structure. With better annotations, temporal cues, richer role representations, and hybrid attention, future versions can deliver stronger performance while preserving interpretability and efficiency.

References

[1] P. Gao, J. Zheng, C. Shuai, and L. Zhang, “A hierarchical dual-role interaction network for telephone conversation fraud detection,” vol. 12, pp. 174 122–174 132.

[2] N. Bajaj, T. Goodluck Constance, M. Rajwadi, J. Wall, M. Moniri, N. Cannings, C. Woodruff, and J. Laird, “Fraud detection in telephone conversations for financial services using linguistic features,” 12 2019.

[3] P. Gao, Z. Li, D. Zhou, and L. Zhang, “Reinforced cost-sensitive graph network for detecting fraud leaders in telecom fraud,” *IEEE Access*, vol. PP, pp. 1–1, 01 2024.

[4] J. Li, C. Zhang, and L. Jiang, “Innovative telecom fraud detection: A new dataset and an advanced model with roberta and dual loss functions,” *Applied Sciences*, vol. 14, p. 11628, 12 2024.

[5] B. Hong, T. Connie, T. S. Ong, and A. Teoh, “Classifying scam calls through content analysis with dynamic sparsity top- k attention regularization,” *IEEE Access*, vol. PP, pp. 1–1, 01 2025.

[6] R. Oļeiņiks, “Real-time fraud detection and prevention based on artificial intelligence tools,” *Baltic Journal of Modern Computing*, vol. 13, 01 2025.

[7] M. K. Moussavou Boussougou and D.-J. Park, “Attention-based 1d cnn-bilstm hybrid model enhanced with fasttext word embedding for korean voice phishing detection,” *Mathematics*, vol. 11, p. 3217, 07 2023.

[8] Q. Zhao, K. Chen, T. Li, Y. Yang, and X. Wang, “Detecting telecommunication fraud by understanding the contents of a call,” *Cybersecurity*, vol. 1, 12 2018.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding.” [Online]. Available: https://arxiv.org/abs/1810.04805

[10] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, “Very deep convolutional networks for text classification,” 2017. [Online]. Available: https://arxiv.org/abs/1606.01781

[11] Z. Wang, M. Yang, C. Jin, J. Liu, Z. Wen, S. Liu, and Z. Zhang, “Ifdds: An anti-fraud outbound robot,” vol. 35, no. 18, pp. 16 117–16 119. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/18030

[12] G. Singh, P. Singh, and M. Singh, “Advanced real-time fraud detection using rag-based llms,” 2025. [Online]. Available: https://arxiv.org/abs/2501.15290

[13] Z. Shen, S. Yan, Y. Zhang, X. Luo, G. Ngai, and E. Y. Fu, “"it warned me just at the right moment": Exploring llm-based real-time detection of phone scams,” 2025. [Online]. Available: https://arxiv.org/abs/2502.03964

[14] Z. Shen, K. Wang, Y. Zhang, G. Ngai, and E. Y. Fu, “Combating phone scams with llm-based detection: Where do we stand?” 2024. [Online]. Available: https://arxiv.org/abs/2409.11643

[15] Z. Ma, P. Wang, M. Huang, J. Wang, K. Wu, X. Lv, Y. Pang, Y. Yang, W. Tang, and Y. Kang, “Teleantifraud-28k: An audio-text slow-thinking dataset for telecom fraud detection,” 2025. [Online]. Available: https://arxiv.org/abs/2503.24115

[16] J. Lin, X. Guo, Y. Zhu, S. Mitchell, E. Altman, and J. Shun, “Fraudgt: A simple, effective, and efficient graph transformer for financial fraud detection,” pp. 292–300.

[17] L. Qin, Z. Li, W. Che, M. Ni, and T. Liu, “Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification.” [Online]. Available: https://arxiv.org/abs/2012.13260

[18] C. Cerisara, S. Jafaritazehjani, A. Oluokun, and H. Le, “Multi-task dialog act and sentiment recognition on mastodon,” 2018. [Online]. Available: https://arxiv.org/abs/1807.05013

[19] Z. Chen, R. Yang, Z. Zhao, D. Cai, and X. He, “Dialogue act recognition via crf-attentive structured network,” 2017. [Online]. Available: https://arxiv.org/abs/1711.05568

[20] X. Hu, H. Chen, H. Chen, X. Li, J. Zhang, and S. Liu, “Mining mobile network fraudsters with augmented graph neural networks,” *Entropy*, vol. 25, p. 150, 01 2023.

[21] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, “Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation,” 2019. [Online]. Available: https://arxiv.org/abs/1908.11540

[22] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” 2018. [Online]. Available: https://arxiv.org/abs/1710.10903

[23] C. Liang, C. Yang, J. Xu, J. Huang, Y. Wang, and Y. Dong, “S+page: A speaker and position-aware graph neural network model for emotion recognition in conversation,” 12 2021.

[24] R. Johnson and T. Zhang, “Deep pyramid convolutional neural networks for text categorization,” 01 2017, pp. 562–570.

[25] H. Kumar, A. Agarwal, R. Dasgupta, S. Joshi, and A. Kumar, “Dialogue act sequence labeling using hierarchical encoder with crf,” 2017. [Online]. Available: https://arxiv.org/abs/1709.04250