A hybrid structural-contextual framework for malicious URL detection based on graph attention networks and knowledge distillation

Nguyen Trung Hieu[1*],Pham Viet Cuong[2], Tran Nguyen Ngoc[3]

[1]The People Police Academy, Pham Van Nghi, Dong Ngac, Hanoi, Vietnam;
[2]National Defence Academy, 93 Hoang Quoc Viet, Nghia Do, Hanoi, Vietnam;
[3]Le Quy Don Technical University, 236 Hoang Quoc Viet, Nghia Do, Hanoi, Vietnam.
[*]Corresponding author: hieunt.dcn@gmail.com
Received 21 Mar. 2026; Revised 22 May 2026; Accepted 26 May 2026; Published 25 Aug. 2026.
DOI: https://doi.org/10.54939/1859-1043.j.mst.113.2026.150-157

ABSTRACT

Malicious URLs are evolving with increasingly sophisticated obfuscation techniques, posing a persistent threat to global cybersecurity. Traditional deep learning methods still struggle to identify the intricate topological characteristics and hidden structural dependencies inherent in URL sequences. This paper proposes a structural-semantic integration framework leveraging Graph Attention Networks and Knowledge Distillation for robust multi-layer malicious URL detection. We conceptualize raw URLs as character-level directed graphs to extract intricate structural patterns, fused with high-level lexical context distilled from a pretrained DistilBERT model. A hybrid loss integrating Kullback-Leibler divergence and Negative Log-Likelihood balances computational efficiency and predictive performance. Experimental results on large-scale, imbalanced datasets show our framework achieves 98.86% accuracy and 98.16% F1-score.

Keywords: Malicious URLs; Graph attention networks; Graph convolutional network; Knowledge distillation.

1. INTRODUCTION

The rapid evolution of cyber-attacks has transformed malicious URLs into sophisticated vectors that leverage complex obfuscation and dynamic redirection. Large Language Models (LLMs) such as Llama 3 or GPT 4 can decode the semantic nuances of URLs beyond traditional signature-based system [22, 23], the balance between computational cost and accuracy remains a gap in research [24]. Lighter sequential models (LSTM, BERT [1], DistilBERT [2]) capture lexical semantics well but often miss the topological significance of individual characters exploited in obfuscation techniques like homograph attacks or character-level padding. Conversely, graph-based models such as GCN and GAT [3] capture structural dependencies effectively but lack the deep semantic understanding needed to interpret sub-domain or path vocabularies [4].

To bridge this gap, we propose a novel Hybrid Structural-Contextual Framework that unites character-level structural integrity with high-level semantic intelligence. Each URL is represented as a character-level directed graph, from which GAT layers extract topological features, while a Knowledge Distillation mechanism transfers "dark knowledge" from a large-scale Teacher model (DistilBERT/ELECTRA) into our compact Student architecture. This yields a model that is both semantically aware and lightweight enough for real-time cyber defense deployment.

Our main contributions in this paper:

- A solution for detecting malicious URLs that inherits knowledge from a large language model (ELECTRA) via knowledge distillation, reducing the model's size to less than half of ELECTRA-base while maintaining detection capabilities.

- Rigorous experiments on an imbalanced dataset demonstrate the superiority of the proposed model, with a Recall of 97.69%, confirming its resistance to overfitting and robustness against sophisticated obfuscation compared to other traditional and modern methods.

The following sections of the paper are structured: Section 2 reviews related works, Section 3 presents the proposed method, Section 4 reports experiments and Section concludes.

2. RELATED WORK

Transformer-based models such as BERT [1], DistilBERT [2], and ELECTRA [7] have achieved strong results in malicious and phishing URL detection by capturing contextual and long-range dependencies in URL sequences [4, 6, 7], with several studies refining BERT embeddings or applying byte-pair encoding to improve accuracy and robustness against evasion techniques [10-13]. However, the approaches, exemplified by URLTran [11], typically require large training data and substantial computational resources, which limits their practicality in resource-constrained settings.

Graph neural networks (GNNs) have achieved considerable success in malicious URL detection [14, 15]. Veličković et al. [3] introduced an attention mechanism into graph neural networks to weigh the importance of neighbouring nodes, allowing the model to focus on the most relevant relationships rather than treating all neighbouring nodes equally. Hossain et al. [16] combined GAT with LSTM for multi-class URL classification (benign, defacement, phishing, malware), achieving 98.06% accuracy and a weighted F1 score of 98.04%.

Knowledge Distillation, introduced by Hilton et al. in 2015 [5], transfers knowledge from a large Teacher model to a compact Student, reducing the model size while improving the generalization on imbalanced data. Prior work has applied this idea to phishing and malicious URL detection using RoBERTa/DitilRoBERTa [19] and ELECTRA-based [20] teacher-student pairs, reporting accuracies above 94% and F1-scores above 99% on binary classification tasks. Zaimi et al. [20] distilled DistilBERT with statistical features into a CNN-LSTM student on the same dataset [17], achieving an F1-score of 97.26% - a result we compare against directly in Section 4.2.3.

3. METHODOLOGY

In this section, we present the proposed model for malicious URL detection. Our architecture is based on knowledge distillation [5] and a hybrid approach that combines GAT [3] with DistilBERT [2].

3.1. Graph representation of URLs

Within the proposed framework, each raw URL is transformed into a character-level directed graph, avoiding manual feature engineering and enabling the model to learn intrinsic structural representations automatically.

We define a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v_i \in \mathcal{V}$ represents a unique character at position i within the URL string, and edges form a bidirectional sequential structure ($i \leftrightarrow i + 1$). This topology enables GAT's multi-head attention to weight dependencies between consecutive character pairs. For instance, the attention weight for a dot "." connected to "c" in ".com" would be substantially higher than that of an arbitrary pair like "a" and "b" in a random string.

To maximize representation capability, each node is represented by a hybrid feature vector $x_i \in \mathbb{R}^d$, which integrates identity-based information with malicious contextual features:

- Identity Features: One-hot encoding for the 69 most frequent characters, identifying character identity at each position.

- Global Contextual Features: three z-score normalized statistics per node: Character Entropy, Sparse N-gram Ratio, and Percent-Encoding Ratio (an obfuscated indicator).

Embedding global features into each node lets the GAT's attention mechanism "perceive" the overall malicious context of the URL while computing local connection weights (Figure 1), which is the key to the Student model's sensitivity toward sophisticatedly obfuscated variants.

We apply Z-score normalization ($\mu = 0$, $\sigma = 1$) so that the three global features, which have disparate scales, do not disproportionately bias learning:

$$Z = \frac{x - \mu}{\sigma} \quad (1)$$

where x is the raw feature value, μ and σ are the mean and standard deviation of that feature computed over the training set. This prevents data leakage and ensures a consistent normalization scale on unseen data.

- The GCN+GAT layers process the URL at the character level. The final output is an aggregated graph embedding that represents the structural properties of the entire URL sequence.

- DistilBERT processes the URL at the standard token level. The final output is a global representation vector of the entire sequence, typically derived from the [CLS] token embedding.

- The Fully Connected (FC) classification head is fed by concatenating the aggregated GCN-GAT graph vector with the DistilBERT [CLS] embedding.

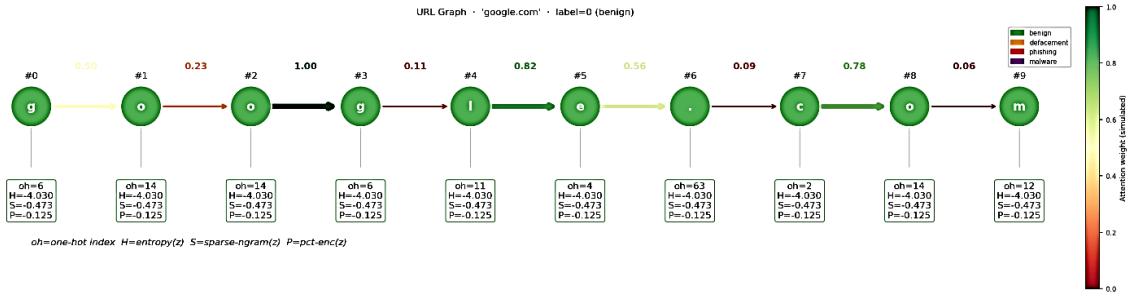

Figure 1. *Transforming the URL into a graph representation with learned attention weights.*

3.2. Knowledge distillation system framework

The Knowledge Distillation [5] conveys contextual understanding from the Teacher LLM to a hybrid Student capable of handling graph structures, unlocking class-relationship knowledge that hard ground-truth labels alone cannot capture (Figure 2).

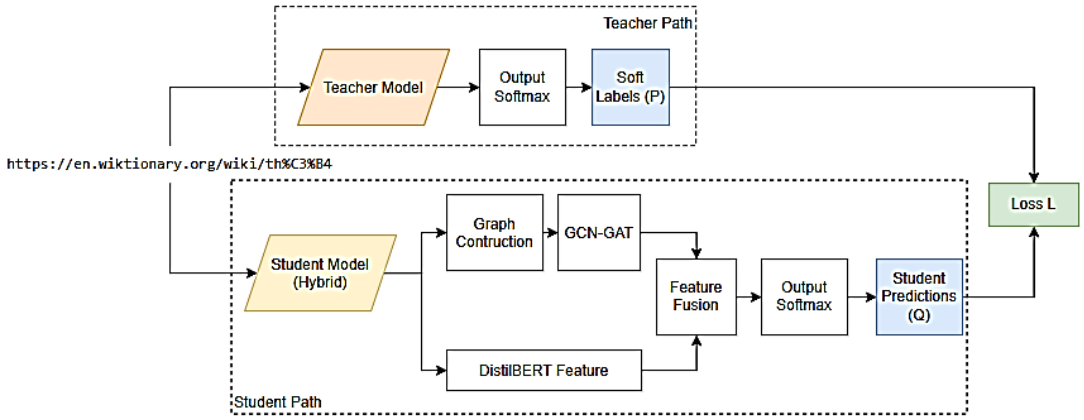

Figure 2. *Teacher-student architectural framework.*

Rather than training solely on hard ground-truth labels, the Student also learns from Teacher's soft-target probability distribution, adopting LLM-level generalization at a much lower computational cost.

Teacher Model: We use ELECTRA [7] as a context extraction expert, providing soft probability distributions for each URL, which optimizes resource use for real-time URL detection tasks.

Student Model: The Student is a hybrid GCN+GAT and lightweight DistilBERT architecture (Figure 3) that learns from both ground-truth labels and the Teacher's predictions, inheriting the Teacher's generalization while remaining sensitive to character-level structures.

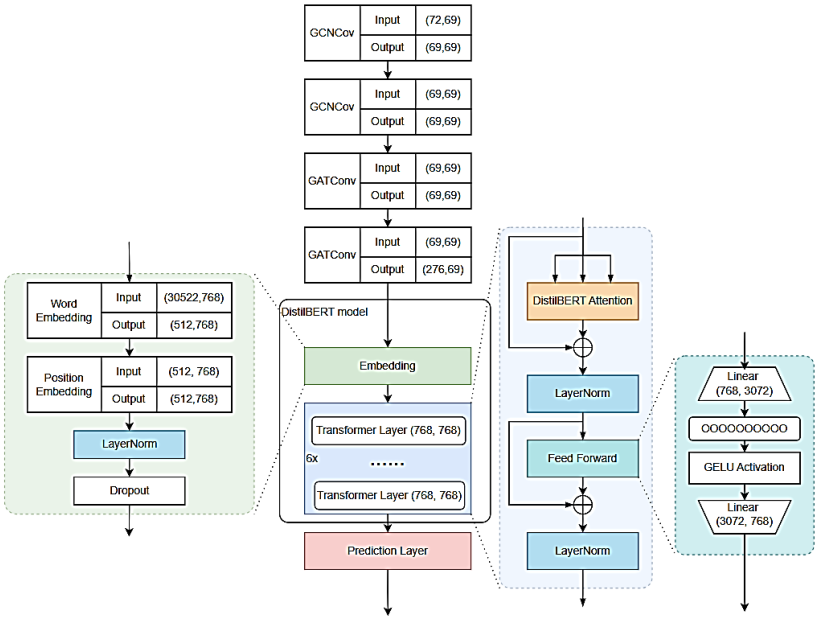

Figure 3. *Architecture of the GCN-GAT-DistilBERT model.*

The Temperature Scaling facilitates knowledge transfer: for each URL, the Teacher's logit vector is converted to a soft probability via a modified Softmax with temperature T:

$$p_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \tag{2}$$

Here, T is a hyperparameter controlling the smoothness of the distribution: when $T > 1$, the distribution softens and reveals correlations between classes (e.g., a URL exhibiting both Phishing and Malware traits), which the Student learns to emulate by minimizing KL-divergence.

To train the Student within the Knowledge Distillation framework, we use a hybrid loss balancing ground-truth labels against the Teacher's behavior.

The total loss function, L_{total} is defined as a weighted combination of two components: (L_{CE}) and (L_{KL}):

$$L_{total} = \alpha \cdot L_{KL} + (1 - \alpha) \cdot L_{CE} \tag{3}$$

Here, α weights the reliance on the Teacher's knowledge. The first term is the Kullback-Leibler (KL) Divergence, (L_{KL}), measuring the discrepancy between the Student's (q) and the Teacher's (p) soft distributions after temperature scaling:

$$L_{KL} = T_{KL}^2 \cdot KLDiv(p_s, p_t) \tag{4}$$

A key advantage of this term is the smoothness it introduces to the loss landscape: instead of sparse binary signals (0 or 1), the Student captures the nuanced inter-class correlations from the Teacher, which is especially valuable for data-scarce classes such as Malware.

The second component is Cross-Entropy loss (L_{CE}), which compels the Student's predictions to align with the ground-truth hard labels y:

$$L_{CE} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{5}$$

where C is the number of classes (Benign, Phishing, Malware, Defacement). This term keeps the model anchored to accurate classification, while the KL term regularizes against the overconfidence NLL alone can cause.

4. EXPERIMENTS AND DISCUSSION

4.1. Dataset

The dataset [17] comprises 651,191 URLs across four classes: 428,103 benign, 96,457 defacement, 94,111 phishing, and 32,520 malware (Table 1).

The dataset exhibits a severe class imbalance with Malware class accounting for only ~5% of the samples versus over 65% for Benign URLs, reflecting the real-world rarity of new malicious URLs relative to benign traffic and posing a major challenge for practical detection.

Table 1. *Statistical distribution of URL classes in the dataset.*

Class	Number of URLs	Percentage (%)
Benign	428103	65.741541
Defacement	96457	14.812398
Phishing	94111	14.452135
Malware	32520	4.993927
Total	**651191**	**100%**

A preprocessing pipeline normalizes the raw labels (benign, defacement, phishing, and malware) into numerical classes (0-3) via graphical representation and feature extraction.

The dataset is split 80/10/10 (train/val/test) via stratified sampling to preserve class imbalance. Raw URLs are one-hot encoded (69-dimensional); global features are computed only on the training set to prevent data leakage (Table 2).

Table 2. *Global statistical feature values computed on the train_dataset.*

Statistical features	Mean	Std
character_entropy	4.1727	0.3787
sparse_ngram_ratio	0.7126	0.0972
percent_encoding_ratio	0.0027	0.0216

4.2. Experimental results

This section presents the experimental results evaluating the loss functions and the synergy between components in the hybrid architecture, using an Intel(R) Xeon(R) Gold 5320 2.2GHz, NVIDIA A100 GPU (40 GiB), 512GiB RAM running Ubuntu 20.04.

4.2.1. Experimental fine-tuning of T and α parameters

This experiment aims to find the optimal parameters T and α for the knowledge distillation model. We conducted the experiment with temperature T= {3,4,5} and α={0.2, 0.5, 0.7, 0.9} on the train/val dataset. The experimental results are described in Table 3

Table 3. *Summary of experimental results.*

T	α	F1-score	Precision	Recall	Accuracy
T = 5	**0.2**	0.9845	0.9879	0.9811	0.9901
	0.5	0.9831	0.9845	0.9818	0.9898
	0.7	0.9842	0.9888	0.9798	0.9902
	0.9	0.9837	0.9880	0.9795	0.9898
T = 4	**0.2**	0.9841	0.9882	0.9801	0.9903
	0.5	0.9840	0.9874	0.9807	0.9898
	0.7	**0.9844**	**0.9885**	**0.9805**	**0.9902**
	0.9	0.9839	0.9864	0.9815	0.9898
T = 3	**0.2**	0.9836	0.9860	0.9813	0.9897
	0.5	0.9830	0.9870	0.9792	0.9896
	0.7	0.9833	0.9868	0.9798	0.9901
	0.9	0.9842	0.9874	0.9811	0.9903

We choose $T = 4$, $\alpha = 0.7$ as the optimal parameter for the proposed model. Here, the model achieves a competitive F1-score (0.9844). This choice is based on the high stability of the indices across multiple trials and the ability to maximize semantic knowledge from the Teacher model ($\alpha = 0.7$), while maintaining an ideal balance between Precision and Recall.

Table 4 compares Teacher and Student ($T = 4$, $\alpha = 0.7$) on the validation dataset: accuracy rises from 0.9886 to 0.9902 and F1-score from 0.9813 to 0.9844, confirming the effectiveness of the distillation approach. Despite requiring only 66M parameters versus ELECTRA's ~109M (a ~40% reduction), the Student remains highly effective on imbalanced classes such as malware.

Table 4. *Comparison of the teacher model and Student model.*

Model	F1-score	Precision	Recall	Accuracy
Teacher	0.9813	0.9811	**0.9816**	0.9886
Student	**0.9844**	**0.9885**	0.9805	**0.9902**

4.2.2. Evaluating component contributions and benchmark comparisons

Using $T = 4$, $\alpha = 0.7$, we further examined each component's contribution to accuracy (Table 5).

Table 5. *Experimental results for the classes.*

	Precision	Recall	F1-score	Support
benign	0.99	0.99	0.99	42808
defacement	1.00	1.00	1.00	9531
phishing	0.96	0.97	0.96	9408
malware	0.97	0.97	0.97	2365
accuracy			0.99	64112
macro avg	0.98	0.98	0.98	64112
weighted avg	0.99	0.99	0.99	64112

We further evaluate the proposed model against independent baselines on an out-of-sample test set (Table 6).

As shown in Table 6, GCN+GAT alone performs poorly (F1=0.5707) despite a deceptively high accuracy (0.9113) – with benign URLS at 65% of the data, structural information alone is insufficient without DistilBERT's semantic understanding, and the model likely overlooks minority classes.

Table 6. *Comparison of the proposed model and components.*

Model	F1-score	Precision	Recall	Accuracy
DistilBERT only	0.9732	0.9774	0.9714	**0.9936**
GCN+GAT only	0.5707	0.7218	0.5473	0.9113
GCN+GAT+DistilBERT	**0.9844**	**0.9885**	**0.9805**	0.9902

DistilBERT alone already performs strongly (F1 = 0.9732), and integrating GCN+GAT via knowledge distillation further improves this to F1 = 0.9844, confirming the benefit of combining structural and semantic features.

4.2.3. Compare with other models

We also compared our model against recent studies using the same dataset [17], including a LoRA-fine-tuned BERT classifier from Western Sydney University [18] covering the same four classes (benign, defacement, phishing, malware).

We also compared Zaimi's model [21] and Clark's model [7] as Teacher models on same dataset [17]; parameter count is also a key factor in deployment efficiency.

Table 7. *Comparison of novel URL detection performance between the two teacher models.*

	F1-score	Precision	Recall	Accuracy
Gleyzie Tongo [18]	0.9698	0.9780	0.9643	**0.9937**
Clark [7]	0.9713	0.9753	0.9694	0.9931
Zaimi [21]	0.9726	0.9822	0.9632	0.9819
Our proposed	**0.9844**	**0.9885**	**0.9805**	0.9902

5. CONCLUSIONS

This study validates a hybrid GAT and Knowledge Distillation framework for detecting malicious URLs. By fusing structural and semantic features, our approach adapts well to sophisticated obfuscation and rare malware variants under class imbalance, while remaining robust to domain shift. Future work will extend this to multi-stage distillation covering threats like DNS tunneling and botnet traffic, and optimize it for real-time edge deployment.

REFERENCES

[1]. J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "*BERT: Pre-training of deep bidirectional transformers for language understanding*", arXiv, (2019).

[2]. V. Sanh, L. Debut, J. Chaumond and T. Wolf, "*DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*", arXiv, (2020).

[3]. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò and Y. Bengio, "*Graph attention networks*", arXiv, (2018).

[4]. A. Yilmaz and R. Das, "*A novel hybrid approach combining GCN and GAT for effective anomaly detection from firewall logs in campus networks*", Computer Networks, Vol. 259, p. 111082, (2025).

[5]. G. Hinton, O. Vinyals and J. Dean, "*Distilling the knowledge in a neural network*", arXiv preprint arXiv:1503.02531, (2015).

[6]. Z. Y. Lim, Y. HanPang, E. C. KahJun and S. Y. Ooi, "*Contra-KD: A lightweight transformer model for malicious URL detection with contrastive representation and model distillation*", Future Internet, Vol. 18, No. 157, (2026).

[7]. K. Clark, M.-T. Luong, Q. V. Le and C. D. Manning, "*ELECTRA: Pre-training text encoders as discriminators rather than generators*", Eighth International Conference on Learning Representations, Addis Ababa, Ethiopia, (2020).

[8]. A. Aljofey, S. A. Bello, J. Lu and C. Xu, "*BERT-PhishFinder: A robust model for accurate phishing URL detection with optimized DistilBERT*", IEEE Transactions on Dependable and Secure Computing, Vol. 22, No. 4, pp. 4315–4329, (2025).

[9]. O. Niyaoui and O. Reda, "*Malicious URL detection using transformers' NLP models and machine learning*", International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD'2023), Switzerland, (2024).

[10]. M.-Y. Su and K.-L. Su, "*BERT-based approaches to identifying malicious URLs*", Sensors, Vol. 23, No. 20, (2023).

[11]. P. Maneriker, J. W. Stokes, E. G. Lazo, D. Carutasu, F. Tajaddodianfar and A. Gururajan, "*URLTran: Improving phishing URL detection using transformers*", (2021).

[12]. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "*RoBERTa: A robustly optimized BERT pretraining approach*", (2019).

[13]. C. Wang and Y. Chen, "*TCURL: Exploring hybrid transformer and convolutional neural network on phishing URL detection*", Knowledge-Based Systems, Vol. 258, (2022).

[14]. C. Wang, Z. Liu and Y. Zeng, "*Detecting phishing URLs with GNN-based network inference*", 2024 International Conference on Networking and Network Applications (NaNA), Yinchuan, China, (2024).

[15]. A. Chattopadhyay, D. Reti and H. D. Schotten, "*GNN-based anomaly detection for encoded network traffic*", arXiv, (2024).

[16]. M. I. Hossain, K. A. A. Arafat, B. Shepard, K. Craig and I. Parvez, "*A graph-attentive LSTM model for malicious URL detection*", arXiv, (2025).

[17]. M. Siddhartha, "*Malicious URLs Dataset*", Kaggle, (2025). Available online: Malicious URLs Dataset.

[18]. G. Tongo, F. Farid, A. Al-Areqi and F. Ahamed, "*Fine-tuned BERT for malicious URL detection*", Western Sydney University. Available online: Fine-Tuned BERT for Malicious URL Detection.
[19]. P. H. Hussan and S. M. Mangj, "*BERTPHIURL: A teacher-student learning approach using DistilRoBERTa and RoBERTa for detecting phishing cyber URLs*", Journal of Future Artificial Intelligence and Technologies, Vol. 1, No. 4, pp. 417–428, (2025). DOI: 10.62411/faith.3048-3719-71.
[20]. K. S. Jishnu and B. Arthi, "*Real-time phishing URL detection framework using knowledge distilled ELECTRA*", Journal for Control, Measurement, Electronics, Computing and Communications, Vol. 65, No. 4, (2024). DOI: 10.1080/00051144.2024.2415797.
[21]. R. Zaimi, K. S. Eljil, H. Mohamed, M. Lamia and F. Nait-Abdesselam, "*An enhanced mechanism for malicious URL detection using deep learning and DistilBERT-based feature extraction*", The Journal of Supercomputing, Vol. 81, No. 2, (2025). DOI: 10.1007/s11227-024-06908-x.
[22]. G. Goldenits, P. König, S. Raubitzek and A. Ekelhart, "*Small language models for phishing website detection: Cost, performance, and privacy trade-offs*", Journal of Cybersecurity and Privacy, Vol. 6, No. 2, p. 48, (2026). DOI: 10.3390/jcp6020048.
[23]. S. E. Blake, "*Phishsense-1B: A technical perspective on an AI-powered phishing detection model*", arXiv:2503.10944, (2025).
[24]. Phan Thi Hai Hong, Truong Thi Thu Hang, Dang Van Giap and Ta Huu Vinh, "*EB-UNet++: An enhanced crack segmentation network combining EfficientNet-B2 and UNet++ with boundary extraction module*", Journal of Military Science and Technology, No. IITE, pp. 148–159, (2025).

TÓM TẮT

Một khung cấu trúc-ngữ cảnh lai để phát hiện URL độc hại dựa trên mạng lưới chú ý đồ thị và chắt lọc tri thức

Các URL độc hại tiếp tục phát triển với các kỹ thuật xáo trộn, gây ra mối đe dọa đáng kể cho an ninh mạng toàn cầu. Bài báo này giới thiệu một khung kết hợp cấu trúc-ngữ nghĩa tận dụng Mạng chú ý đồ thị (GAT) và Chắt lọc tri thức (KD) để phát hiện URL độc hại đa lớp. Phương pháp chuyển đổi các URL thô thành đồ thị có hướng ở cấp độ ký tự, kết hợp với ngữ cảnh ngôn ngữ cấp cao chưng cất từ DistilBERT, dùng hàm mất mát lai (KL và NLL) để cân bằng chi phí tính toán và hiệu năng. Kết quả thực nghiệm trên tập dữ liệu lớn, không cân bằng cho thấy khung đề xuất đạt độ chính xác 98.6% và điểm F1 98.16%.

Từ khóa: Phát hiện URL độc hại; Mạng chú ý trên đồ thị (GAT); Mạng tích chập trên đồ thị; Chưng cất tri thức.