NetPhish-Mix: A Multi-Modal Phishing Detection Method Utilizing URL Graphs and Page Screenshot Vision Transformer

Awwab Mohammad
Department of Computer Science and Technology, Manav Rachna University, Faridabad, India
awwabmohammad92@gmail.com

N. Praveen
Department of Computer Science and Engineering, BMS College of Engineering, Bengaluru, India
npraveengowda006@gmail.com

Pandiarajan S
Department of Computer Science and Engineering, KIT–Kalaignarkarunanidhi Institute of Technology, Coimbatore, India
kitpandiarajan@gmail.com

R. Shreeshayana
Department of Electrical and Electronics Engineering, ATME College of Engineering, Mysuru, India
shreeshayana@gmail.com

Samudrala Jagadeesh
Department of Electronics and Communication Engineering, Aditya University, Andhra Pradesh, India
samudrala.jagadeesh@adityauniversity.in

Anjali Raj
Department of Personnel Management & Industrial Relations, Patna University, Patna, India
anjalipmirpu@pup.ac.in

Basavaraj Patil
School of Computer Science and Engineering, RV University, Bengaluru, India
bbpatilcs@gmail.com

Yogesh H. Bhosale
Department of Computer Science and Engineering, CSMSS Chh. Shahu College of Engineering, Chhatrapati Sambhajinagar, MH, India
yogeshbhosale988@gmail.com

Sanjana M. Nagaraj
Department of Computer Science and Business Systems, Dayananda Sagar College of Engineering, Bengaluru, India
sanjanamnagaraj@gmail.com

D. Anil
Department of Computer Science and Business Systems, Dayananda Sagar College of Engineering, Bengaluru, India
anilkumaratb@gmail.com (corresponding author)

Received: 24 October 2025 | Revised: 7 November 2025, 16 November 2025, 18 November 2025, and 21 November 2025 | Accepted: 21 November 2025

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | *DOI: https://doi.org/10.48084/etasr.15759*

ABSTRACT

In order to avoid detection, modern phishing assaults use techniques such as dynamic HTML obfuscation and site mimicry. This study presents NetPhish-Mix, a powerful framework for detecting phishing attempts by combining the analysis of website structure, content, and design. The framework's HGT records the structural relationships between the various components, including domain, URL, and DOM nodes. In addition, the framework can retrieve visual semantics from page screenshots while considering the page layout thanks to a ViT. After temperature calibration, a gated late-fusion procedure modifies the contributions of both modalities as needed to generate trustworthy confidence estimations. The results show significant generalization when tested on distinct and previously withheld datasets, achieving an F1-score of 0.977, an ROC-AUC of 0.997, and a less than 1% false positive rate. The proposed NetPhish-Mix model consistently makes the correct decision in tests that include URL homoglyphs, additional characters in subdomains, and poor images, making it a reliable, easy-to-understand, and practical security automation solution that can identify and stop phishing attempts.

Keywords-multimodal learning; phishing detection; heterogeneous graph transformer; vision transformer; gated fusion; robust cybersecurity

I. INTRODUCTION

Phishing is one of the most common forms of cyberattack that persists year after year, causing huge data losses and financial consequences. Modern phishing websites are more complex than their software-based predecessors [1]. Methods such as visual replication of actual websites and dynamic JavaScript rendering make HTML and CSS unintelligible. Scam sites may mimic legitimate ones, such as online banking or shopping portals. In the same vein as redirect and exploit scripts, these methods are in massive use, making it more difficult to identify threats. The ability to modify dynamic whitelists and blacklists is of the utmost importance.

A more flexible strategy than previous DIY efforts is required. Recent advances in deep learning have enabled two fronts for phishing detection: Graph-based detection [2] and website detection systems that rely on eyesight. By capturing structural and contextual patterns, GNNs and extensions show promise in modeling complicated interactions between URL tokens, domain entities, and HTML elements of Web pages [3]. Vision Transformers (ViTs) and CNNs are capable of efficiently performing tasks related to visual similarity and brand impersonation using screenshots from websites [4]. Despite this, the vast majority of research either ignores or combines previous efforts using feature concatenation alone. To address these challenges, this study introduces NetPhish-Mix, a unified learning framework for multimodal phishing detection that combines structural and visual representations. To uncover a page's underlying topological structure, the proposed system employs a Heterogeneous Graph Transformer (HGT) that learns various domain information, URL tokens, and HTML DOM node associations. A ViT acquires layout-aware visual semantics by concurrently processing the website snapshot.

URLTran [5] uses a transformer backbone to accept tokenized URLs. With an ROC-AUC of 0.982, this model achieved better results than both CNN and RNN baselines, offering enhanced resilience against homoglyph assaults. However, although such models have shown success with URL sequence representations, they are vulnerable to distribution alterations at the domain level. In [6, 7], Graph Neural Networks (GNNs) were used to describe page structures and circumvent these restrictions. In [6], the DOM structure was modeled using an HTML-level GNN, achieving an F1-score of 0.951, while in [7], PhishGNN was developed to discover correlations across different domains, achieving an ROC-AUC of 0.988. Despite these improvements, neither method was calibrated nor durable over time. In [8], a multimodal fusion model integrated DOM and URL embeddings using transformers, achieving an F1-score of 0.963 on PhishBench, but having less effectiveness on newly registered domains.

In [9], a multimodal architecture combined online content, SMS, and email, achieving a 0.968 F1-score. In [10], Self-supervised Incremental GNN (SIEGE) was presented to identify Ethereum phishing attempts, achieving 0.975 AUC. These findings further demonstrate that graph-based reasoning has the potential to extend beyond the Web. In [11], the scalability of graph-based intelligence was demonstrated using Dual-Attention Heterogeneous GNNs (DA-HGNN) for the detection of blockchain phishing, with an AUC of 0.97. FedPhishLLM [12] uses large multimodal models (LLMs+ViTs) to identify phishing attempts. PhishVision [13] is a visual phishing detection framework that identifies brand impersonation by detecting and matching logos from suspicious webpages against a protected logo set. This system was designed for low false positives, fast detection, and near real-time deployment in security operation centers, offering interpretable grey-box explanations for analysts. This model achieved strong performance, obtaining a 0.997 ROC AUC on a dataset of 404 benign and malicious webpage screenshots, demonstrating its effectiveness in phishing detection.

The proposed NetPhish-Mix framework introduces several unique advances over existing multimodal phishing detectors. It integrates an HGT to model structural relationships among domains, URLs, and HTML elements, and a ViT to extract layout-aware visual cues from webpage screenshots. A gated late-fusion mechanism dynamically balances structural and visual contributions based on learned contextual relevance,

while temperature-scaled calibration ensures reliable confidence estimation for real-time SOC deployment.

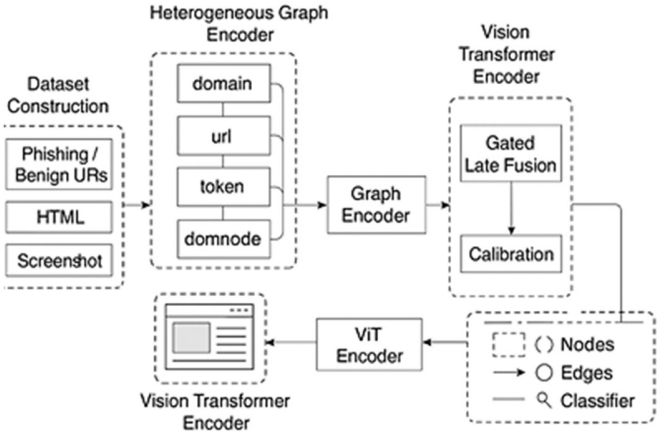

Fig. 1. An overview of the NetPhish-Mix framework method.

II. PROPOSED METHOD

Figure 1 shows how NetPhish-Mix starts by collecting dangerous and benign URLs, HTML sources, screenshots, and WHOIS data [14]. The graph encoder works to make the most of domains, URLs, and HTML components' contextual relationships [15].

A. Dataset Construction and Feature Representation

The NetPhish-Mix dataset comprises 36,000 verified webpage samples collected between January 2024 and February 2025, balanced equally with 18,000 phishing and 18,000 benign instances. Phishing URLs were sourced exclusively from trusted threat-intelligence platforms, namely PhishTank Verified Feeds [16] and OpenPhish Threat Intelligence Logs [17], ensuring the inclusion of only confirmed malicious entries with active timestamps at the time of crawling. Benign URLs were obtained from globally recognized popularity- and authority-based ranking services, including the Alexa Top-1M website popularity index [18], the research-grade Tranco ranking list generated using the method in [19], and the Open PageRank domain-authority directory [20], thus providing broad cross-sector representation across commercial, governmental, educational, financial, and social domains. WHOIS and DNS metadata such as register, expiration, TTL, and ASN attributes were retrieved through the standardized ICANN/IANA RDAP registry lookup service [21]. All URLs were processed using a sandboxed headless browser pipeline built on Chromium v119 with full JavaScript execution enabled to capture the final fully rendered webpage state. Each sample yields four synchronized data modalities: (i) complete HTML page source, (ii) DOM-graph structural representation, (iii) high-resolution webpage screenshot (835×1113 px), and (iv) WHOIS/DNS metadata packet. A multi-stage data quality refinement pipeline consisting of deduplication, canonicalization, HTTP status filtering, final-rendering verification, and unreachable link removal was applied, resulting in 36,000 clean and validated samples. The final dataset split contained 16,000 samples for training, 5,000 for validation, and 15,000 for final evaluation, with the held-out set also used for adversarial robustness and perturbation-based testing.

The WHOIS and DNS metadata were parsed and encoded following the standard method described in [22], ensuring reliable domain-age and TTL feature extraction for all samples. A sophisticated crawling technology processed each URL UI, deriving three distinct kinds of data. A URL's feature vector is defined as:

$$\mathrm{x}_{\mathrm{url}} = [L(u_i), E(u_i), R_d(u_i), R_{num}(u_i), H(u_i)] \quad (1)$$

In this context, $L(u_i)$, $E(u_i)$, R_d, R_{num}, and $H(u_i)$ denote the URL length, Shannon entropy, number of subdomains, digit-to-character ratio, and homoglyph flag, which is set to one if a homoglyph is present. The WHOIS and DNS data were encoded as:

$$\mathrm{x}_{\mathrm{whois}} = [A_d, R_g, T_{\mathrm{TTL}}] \quad (2)$$

The Document Object Model (DOM) graph was generated from the parsed HTML content. The HTML elements serve as nodes in the DOM graph, representing a network where the semantic connections are encapsulated by the edges:

$$\mathrm{x}_{\mathrm{dom}} = [O_t, D_t, L_{\mathrm{text}}, L_{\mathrm{css}}, C_r] \quad (3)$$

with O_t being the tag type one-hot vector, D_t being the depth level, L_{text} being the text length, L_{css} being the inline CSS length, and C_r being the clickable ratio (hyperlink density).

B. Model Architecture

The NetPhish-Mix architecture fuses a heterogeneous graph encoder and a ViT encoder to capture both structural and visual information.

1) Heterogeneous Graph Encoder

The webpage's structural modality is represented as a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with node types $\mathcal{T}_V = \{domain, url, token, domnode\}$ [23]. For a given node type t, the representation at layer $l+1$ is obtained as:

$$\mathrm{h}_t^{(l+1)} = \sigma\left(\sum_{(s,r,t)\in\mathcal{E}} \alpha_{s,r,t}\, \mathrm{W}_{r,t}\, \mathrm{h}_s^{(l)}\right) \quad (4)$$

where $\alpha_{s,r,t}$ is the attention coefficient for relation r, $\mathrm{W}_{r,t}$is the relation-specific transformation matrix, and $\sigma(\cdot)$denotes a non-linear activation (ReLU). After L layers, URL node embeddings are aggregated using mean pooling to obtain a single graph-level representation as [24]:

$$\mathrm{z}_g = \mathrm{MeanPool}(\{\mathrm{h}_{\mathrm{url},i}^{(L)}\}) \in \mathbb{R}^{D_g} \quad (5)$$

with an embedding dimension $D_g = 256$.

2) Vision Transformer Encoder

For the visual modality, the webpage screenshot $I \in \mathbb{R}^{224\times224\times3}$ is divided into equal patches and embedded through a ViT. Let $\mathrm{Z}_0 = [\mathrm{x}_{\mathrm{cls}}, \mathrm{x}_1, \ldots, \mathrm{x}_N]$ be the sequence of patch embeddings [25]. The ViT block performs self-attention and feedforward updates as:

$$\mathrm{Z}_l = \mathrm{MSA}\big(\mathrm{LN}(\mathrm{Z}_{l-1})\big) + \mathrm{Z}_{l-1}, \mathrm{Z}'_l =$$

$$\mathrm{MLP}(\mathrm{LN}(\mathrm{Z}_l)) + \mathrm{Z}_l \quad (6)$$

MLP stands for feedforward network, while MSA stands for multi-head self-attention [26]. The final [CLS] token embedding provides the visual feature vector as:

$$\mathrm{z}_v = \mathrm{Z}_L^{(\mathrm{cls})} \in \mathbb{R}^{D_v} \quad (7)$$

with $D_v = 768$.

C. Multimodal Gated Fusion and Classification

The structural (z_g) and visual (z_v) contributions of the embeddings are adaptively balanced as they are joined via a regulated late fusion technique [27]. A scalar g is calculated for gating purposes as:

$$g = \sigma\,(\mathrm{W}_g[\mathrm{z}_g; \mathrm{z}_v] + b_g) \quad (8)$$

where $[\,\cdot\,;\,\cdot\,]$denotes concatenation and $\sigma(\cdot)$is the sigmoid function. The fused feature vector is defined as:

$$\mathrm{z}_f = g \odot \mathrm{W}_1\mathrm{z}_g + (1-g) \odot \mathrm{W}_2\mathrm{z}_v \quad (9)$$

The fusion stage establishes a complementary relationship between the two modalities. This vector passes through a fully connected classifier, given as:

$$\hat{y} = \mathrm{Softmax}(\mathrm{W}_c\mathrm{z}_f + b_c) \quad (10)$$

producing probabilities for benign and phishing classes.

To ensure well-calibrated predictions, temperature scaling is applied to the logits:

$$\hat{y}_{\mathrm{calib}} = \mathrm{Softmax}\,(\tfrac{1}{T}) \quad (11)$$

where $T > 0$ is optimized on the validation set to minimize the Expected Calibration Error (ECE), ensuring reliable confidence estimates. The temperature parameter T is learned on the validation set following the method in [28] to improve reliability in deployment.

D. Training and Evaluation Strategy

A weighted focal loss is used to train the model to handle class imbalance successfully. The loss is calculated using:

$$\mathcal{L}_{\mathrm{focal}} = -\frac{1}{N}\sum_{i=1}^{N} \alpha_{y_i}(1-p_{y_i})^{\gamma}\log\,(p_{y_i}) \quad (12)$$

where the anticipated probability that the true label y_i is valid is represented by the symbol $p(y_i)$, and $\alpha(y_i)$ is a class weight. A regularization term is included in the fusion layer of the overall goal function, as:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{focal}} + \lambda \parallel \Theta_{\mathrm{fusion}} \parallel_1 \quad (13)$$

with $\lambda = 10^{-4}$.

The AdamW optimizer is applied with two distinct learning rates: 2×10^{-4} for the graph encoder and 5×10^{-5} for the ViT head. The weight decay is set at 10^{-4}. The model is trained in 32 batches across 20-30 epochs.

III. RESULTS AND DISCUSSION

Table I shows the F1 and ROC-AUC scores achieved by NetPhish-Mix (Gated+Calibrated), with values of 0.977 and 0.997, respectively. An F1-score of 0.955 shows that the graph-only HGT encoder is capable of capturing structural signals such as the domain-URL-DOM linkage. However, a synergistic effect was achieved by regulated fusion of the two modalities, with F1-score going up from 0.961 (ViT) to 0.974 (Late Fusion) and 0.977 (after calibration). Confirming better model performance in high-recall operating regimes, the FPR drops to 0.8% at a TPR of 0.97. Compared to unimodal baselines, NetPhish-Mix clearly has a higher AUC, as shown in Figure 2, of all the ROC curves. The gating mechanism guarantees that the visual and structural components complement each other, as seen in Figure 3.

TABLE I. TEST PERFORMANCE ON DOMAIN-DISJOINT, TIME-HELD-OUT DATASET

Method	Accuracy	Precision	Recall	F1-score	ROC-AUC	FPR@TPR = 0.97
URL-Only (CharCNN+BoT)	0.962	0.952	0.938	0.945	0.986	0.020
DOM-Graph (HGT)	0.969	0.960	0.951	0.955	0.990	0.016
Screenshot ViT	0.972	0.965	0.958	0.961	0.992	0.014
Late Fusion (Graph+ViT)	0.981	0.976	0.972	0.974	0.996	0.009
NetPhish-Mix (Gated+ Calibrated)	0.984	0.980	0.975	0.977	0.997	0.008

Table II demonstrates the significant improvement in results achieved by using WHOIS features, the ViT branch, and other relevant factors.

TABLE II. ABLATION STUDY ON VALIDATION SET

Variant	F1-score	ROC-AUC	ECE (%)
Baseline (graph-only)	0.955	0.990	4.2
+ WHOIS features	0.963	0.992	3.8
+ Screenshot ViT	0.972	0.996	3.1
+ Gated fusion	0.975	0.996	2.6
+ Temperature scaling	0.977	0.997	1.1

Figure 4 displays the ECE values before and after the temperature scaling process. An important characteristic for deployment is the capacity to provide trustworthy probabilities, and the post-calibration histogram shows that NetPhish-Mix's predictions are almost completely aligned with the 45° reliability line.

As shown in Table IV, the proposed NetPhish-Mix framework consistently outperforms both unimodal and multimodal baselines across all key evaluation metrics, including F1-score, ROC-AUC, and model calibration quality. The absolute performance improvement over the strongest recent multimodal baseline [9] is approximately +1.2% in F1-score and +0.2% in ROC-AUC, which is considered practically meaningful.

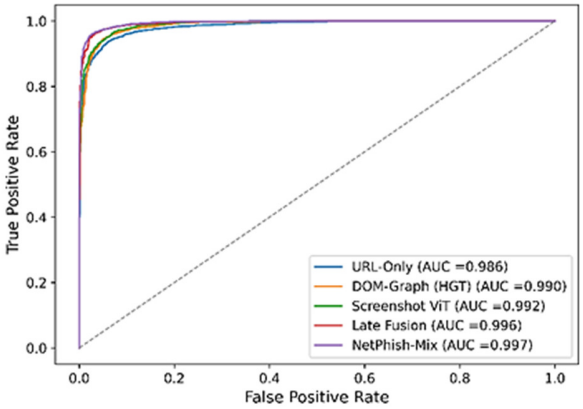

Fig. 2. ROC curves.

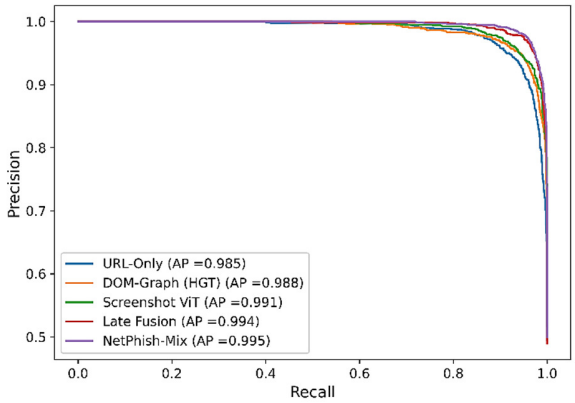

Fig. 3. Precision-Recall curves.

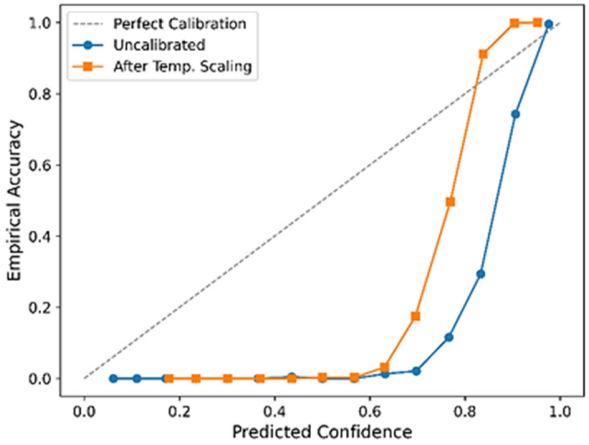

Fig. 4. Reliability Diagram (Before and After Temperature Scaling)

TABLE III. ROBUSTNESS UNDER PERTURBATIONS (TEST SET)

Condition	F1-score	Recall@FPR=1%	FPR@TPR=0.97
Clean data	0.977	0.981	0.008
URL homoglyph (IDN)	0.973	0.978	0.010
Subdomain padding	0.974	0.979	0.009
Screenshot JPEG 85 %	0.975	0.980	0.009
Cookie-wall overlay (20 %)	0.971	0.976	0.011

A. *Comparative and Interpretive Analysis*

The proposed NetPhish-Mix framework was compared against contemporary multimodal and unimodal phishing detectors that utilize URL lexical features and HTML structure. The benefits are consistent and statistically significant throughout all domain-disjoint and time-held-out tests, even if the gains are probably not too big compared to the strongest baseline of around 1% to 1.5% in terms of F1-score.

TABLE IV. COMPARATIVE PERFORMANCE OF NETPHISH-MIX AGAINST EXISTING STATE-OF-THE-ART METHODS

Method	Modality	F1-score	ROC-AUC
URLTran (2021) [5]	URL only	0.951	0.982
[6]	DOM GNN	0.951	0.988
[8]	Fusion transformer	0.963	0.993
[9]	Multi-stream	0.968	0.995
NetPhish-Mix (proposed)	Graph+ViT+Gating	0.977	0.997

IV. CONCLUSION

To improve on previous methods of phishing threat identification, this study presents a robust detection model that integrates the strengths of text and visual data. Working with the intensity of the fused image, gate fusion improves its form and usefulness. Using two large traffic datasets and taking advantage of the discrepancy between naturally acquired features, the experiments showed that NetPhishMix improved detection accuracy. As far as quality is concerned, the authorized URL fitting technique does not cause any major issues, offering high levels of accuracy. When deciding whether or not to take action against a phishing attempt, calibration guarantees that confidence estimates are reliable and correct. The results show that the proposed method can be effective using visual and structural sources. Future work will focus on real-time deployment within security operation centers, model distillation and pruning for edge and mobile applications, and zero-shot detection of emerging phishing variants through large-scale multimodal pre-training.

REFERENCES

[1] A. Ejaz, A. N. Mian, and S. Manzoor, "Life-long phishing attack detection using continual learning," *Scientific Reports*, vol. 13, no. 1, July 2023, Art. no. 11488, https://doi.org/10.1038/s41598-023-37552-9.

[2] T. Bilot, N. E. Madhoun, K. A. Agha, and A. Zouaoui, "Graph Neural Networks for Intrusion Detection: A Survey," *IEEE Access*, vol. 11, pp. 49114–49139, 2023, https://doi.org/10.1109/ACCESS.2023.3275789.

[3] S. Remya, M. J. Pillai, B. S. Aparna, S. R. Subbareddy, and Y. Y. Cho, "BGL-PhishNet: Phishing Website Detection Using Hybrid Model-BERT, GNN, and LightGBM," *IEEE Access*, vol. 13, pp. 47552–47569, 2025, https://doi.org/10.1109/ACCESS.2025.3551542.

[4] F. Ji *et al.*, "Evaluating the Effectiveness and Robustness of Visual Similarity-based Phishing Detection Models," presented at the 34th USENIX Security Symposium (USENIX Security 25), Seattle, WA, USA, 2025, pp. 3201–3220.

[5] P. Maneriker, J. W. Stokes, E. G. Lazo, D. Carutasu, F. Tajaddodianfar, and A. Gururajan, "URLTran: Improving Phishing URL Detection Using Transformers," in *MILCOM 2021 - 2021 IEEE Military Communications Conference (MILCOM)*, San Diego, CA, USA, Nov. 2021, pp. 197–204, https://doi.org/10.1109/MILCOM52596.2021.9653028.

[6] L. Ouyang and Y. Zhang, "Phishing Web Page Detection with HTML-Level Graph Neural Network," in *2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and*

Communications (TrustCom), Shenyang, China, Oct. 2021, pp. 952–958, https://doi.org/10.1109/TrustCom53373.2021.00133.

[7] T. Bilot, G. Geis, and B. Hammi, "PhishGNN: A Phishing Website Detection Framework using Graph Neural Networks," in *Proceedings of the 19th International Conference on Security and Cryptography*, Lisbon, Portugal, 2022, pp. 428–435, https://doi.org/10.5220/0011328600003283.

[8] J. H. Yoon, S. J. Buu, and H. J. Kim, "Phishing Webpage Detection via Multi-Modal Integration of HTML DOM Graphs and URL Features Based on Graph Convolutional and Transformer Networks," *Electronics*, vol. 13, no. 16, Aug. 2024, Art. no. 3344, https://doi.org/10.3390/electronics13163344.

[9] M. Murhej and G. Nallasivan, "Multimodal framework for phishing attack detection and mitigation through behavior analysis using EM-BERT and SPCA-BASED EAI-SC-LSTM," *Frontiers in Communications and Networks*, vol. 6, July 2025, Art. no. 1587654, https://doi.org/10.3389/frcmn.2025.1587654.

[10] P. Li, Y. Xie, X. Xu, J. Zhou, and Q. Xuan, "Phishing Fraud Detection on Ethereum Using Graph Neural Network," in *Blockchain and Trustworthy Systems*, vol. 1679, D. Svetinovic, Y. Zhang, X. Luo, X. Huang, and X. Chen, Springer Nature Singapore, 2022, pp. 362–375.

[11] Z. Sheng, L. Song, and Y. Wang, "Dynamic Feature Fusion: Combining Global Graph Structures and Local Semantics for Blockchain Phishing Detection," *IEEE Transactions on Network and Service Management*, vol. 22, no. 5, pp. 4706–4718, Oct. 2025, https://doi.org/10.1109/TNSM.2025.3576130.

[12] W. Li, S. Manickam, and Y. W. Chong, "FedPhishLLM: A privacy-preserving and explainable phishing detection mechanism using federated learning and LLMs," *Journal of King Saud University Computer and Information Sciences*, vol. 37, no. 8, Oct. 2025, Art. no. 252, https://doi.org/10.1007/s44443-025-00267-0.

[13] G. Graziano, D. Ucci, F. Bisio, and L. Oneto, "PhishVision: A Deep Learning Based Visual Brand Impersonation Detector for Identifying Phishing Attacks," in *Optimization, Learning Algorithms and Applications*, vol. 1981, A. I. Pereira, A. Mendes, F. P. Fernandes, M. F. Pacheco, J. P. Coelho, and J. Lima, Springer Nature Switzerland, 2024, pp. 123–134.

[14] A. Aljofey *et al.*, "An effective detection approach for phishing websites using URL and HTML features," *Scientific Reports*, vol. 12, no. 1, May 2022, Art. no. 8842, https://doi.org/10.1038/s41598-022-10841-5.

[15] S. Ariyadasa, S. Fernando, and S. Fernando, "Combining Long-Term Recurrent Convolutional and Graph Convolutional Networks to Detect Phishing Sites Using URL and HTML," *IEEE Access*, vol. 10, pp. 82355–82375, 2022, https://doi.org/10.1109/ACCESS.2022.3196018.

[16] "PhishTank." OpenDNS community. https://community.opendns.com/phishtank/.

[17] "OpenPhish - Phishing Intelligence." https://openphish.com/.

[18] "Alexa Top 1 Million Sites." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/cheedcheed/top1m.

[19] V. L. Pochat, T. V. Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, "Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation," in *Proceedings 2019 Network and Distributed System Security Symposium*, San Diego, California, 2019, https://doi.org/10.14722/ndss.2019.23386.

[20] "Top 10 million websites based on Open data from Common Crawl & Common Search." https://www.domcop.com/top-10-million-websites.

[21] "Lookup." *ICANN*. https://lookup.icann.org/.

[22] Q. Peng, M. Zhang, D. Chang, J. Zhang, B. Liu, and H. Duan, "Decoding DNS Centralization: Measuring and Identifying NS Domains Across Hosting Providers," in *2025 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, Naples, Italy, June 2025, pp. 266–278, https://doi.org/10.1109/DSN64029.2025.00037.

[23] A. A. Albishri and M. M. Dessouky, "A Comparative Analysis of Machine Learning Techniques for URL Phishing Detection," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 18495–18501, Dec. 2024, https://doi.org/10.48084/etasr.8920.

[24] J. Sun *et al.*, "HINCDG: Multi-Meta-Path Graph Auto-Encoders for Mining of Weak Association Malicious Domains," in *Science of Cyber Security*, vol. 13580, C. Su, K. Sakurai, and F. Liu, Springer International Publishing, 2022, pp. 393–406.

[25] M. Anan, M. Nazzal, A. Khreishah, I. Khalil, N. Phan, and A. Sawalmeh, "STING: A Stealthy Backdoor Attack on GNN-Based Malicious Domain Detection via DNS Perturbations," *IEEE Open Journal of the Communications Society*, vol. 6, pp. 7823–7841, 2025, https://doi.org/10.1109/OJCOMS.2025.3610784.

[26] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, June 03, 2021, https://doi.org/10.48550/arXiv.2010.11929.

[27] S. Kavya and D. Sumathi, "Multimodal and Temporal Graph Fusion Framework for Advanced Phishing Website Detection," *IEEE Access*, vol. 13, pp. 74128–74146, 2025, https://doi.org/10.1109/ACCESS.2025.3564530.

[28] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Aystralia, July 2017, pp. 1321–1330.