International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 06 | Nov 2026 www.irjet.net p-ISSN: 2395-0072

Multimodal Deep Learning Framework for Robust Phishing Website Detection Using URL, HTML, and Visual Features

Rahul Dahiya[1], Dr. Ruchi Kshatri Patel[2]

[1]*First Department of Computer Science and Engineering Shri Ram Institute of Sciences and Technology, Jabalpur, M.P., India*

[2]*Department of Computer Science and Engineering Shri Ram Institute of Sciences and Technology, Jabalpur, M.P., India*

Abstract – *Phishing attacks continue to evolve rapidly, exploiting human trust, brand impersonation, and technical vulnerabilities to exfiltrate sensitive personal, financial, and organizational credentials. Traditional anti-phishing defense systems predominantly rely on unimodal inspection, focusing exclusively on lexical URL strings, static HTML Document Object Model (DOM) properties, or visual screenshots. However, unimodal approaches exhibit severe vulnerabilities when adversaries deploy evasion tactics such as URL typosquatting, character manipulation, DOM obfuscation, or screenshot alteration. This paper proposes an availability-aware multimodal deep learning framework for robust phishing website identification. The framework independently evaluates three complementary information branches: (1) a character-level Convolutional Neural Network and Bidirectional Long Short-Term Memory network with attention (CNN-BiLSTM-Attention) for URL sequence analysis; (2) a Term Frequency–Inverse Document Frequency (TF-IDF) feature extractor paired with balanced Logistic Regression for visible HTML textual content; and (3) a transfer learning architecture utilizing a frozen MobileNetV2 convolutional backbone for webpage screenshot classification. The individual classification probabilities are combined through an equal-weight late-fusion probability mechanism optimized with Youden's J statistic. Evaluated on a curated multimodal dataset comprising 1,000 verified website records under a stratified 80:20 partition, the URL classifier achieved 97.81% accuracy (100% precision and 100% ROC-AUC), while the HTML classifier achieved 92.90% accuracy and 95.59% ROC-AUC. Systematic experimentation across seven modality-availability combinations demonstrated that multimodal probability fusion mitigates single-modality failure modes and enhances classification robustness. Furthermore, a modality integrity evaluation framework using Random Forest estimators is introduced to detect corrupted or manipulated modalities prior to fusion. This study establishes that multimodal integration provides superior resilience against evasive web threats and outlines critical design requirements for practical anti-phishing deployment.*

Key Words: Phishing Detection, Multimodal Learning, CNN-BiLSTM, MobileNetV2, TF-IDF, Logistic Regression, Late Fusion, Modality Availability, Web Security.

1. INTRODUCTION

The rapid expansion of web-based digital services, electronic commerce, online financial transactions, and cloud platforms has significantly heightened exposure to sophisticated cyber threats [1]. Among various attack vectors, phishing remains one of the most persistent and destructive categories of social engineering [2]. Phishing attacks deceive users into disclosing confidential information such as user credentials, credit card details, and identity documents by presenting counterfeit web pages that mimic legitimate organizations [3]. Modern threat reports indicate that billions of phishing attempts are registered annually, with attackers exploiting automated website cloning toolkits, typosquatting domains, and obfuscated page templates [4].

Historically, automated phishing identification systems have relied on unimodal analysis [5]. Unimodal defenses examine a single dimension of a webpage, such as the Uniform Resource Locator (URL) [6], the Hypertext Markup Language (HTML) code [7], or visual rendered screenshots [8]. While single-modality systems offer fast inference, they are inherently susceptible to adversarial evasion [9]. For example, URL-only models fail when attackers deploy compromised legitimate subdomains or novel redirection techniques [10]. Content-based HTML systems are vulnerable to script-based obfuscation, dynamic rendering, and hidden DOM structures [11]. Vision-based systems, which classify pages by matching brand logos or visual layouts, often struggle when encountering non-brand-mimicking phishing layouts or altered typography [12].

To address these vulnerabilities, multimodal deep learning has emerged as an essential security paradigm [13]. By analyzing multiple modalities simultaneously, a multimodal system extracts complementary threat indicators: syntactic and domain irregularities from URLs, deceptive textual intent from visible HTML content, and visual design mimicry from screenshots [14].

Nevertheless, practical multimodal deployment faces critical challenges, including high computational overhead, missing or corrupted input modalities in real-time environments, and adversarial feature manipulation [15].

This paper addresses these operational challenges through three primary contributions:

(1) An end-to-end multimodal classification pipeline integrating character-level CNN-BiLSTM with attention for URLs, TF-IDF text modeling with Logistic Regression for HTML content, and MobileNetV2 transfer learning for webpage screenshots;
(2) A late-fusion probability averaging mechanism with dynamic decision thresholding based on Youden's J statistic; and
(3) A comprehensive modality-availability evaluation across all seven non-empty modality subsets, paired with a Random Forest-based reliability and integrity estimation mechanism to assess input trustworthiness.

2. LITERATURE SURVEY

A. Phishing Attack Taxonomy and Evasion Techniques

Phishing attacks are categorized according to their delivery channels, target objectives, and deceptive techniques [16]. Attackers manipulate URL syntax using homoglyph substitution, hyphen insertions, misleading subdomains, and top-level domain (TLD) variations to impersonate genuine brands [17]. Furthermore, phishing toolkits enable the rapid deployment of dynamic phishing templates that incorporate cloaking scripts, blocking automated security crawlers while displaying fraudulent login forms to targeted human victims [18]. In addition to credential harvesting, malicious URLs frequently serve as conduits for malware distribution, drive-by downloads, ransomware, and web defacement [19].

B. Machine Learning Approaches for Malicious URL Detection

Early machine learning studies employed handcrafted lexical and host-based URL features such as URL length, special symbol counts, token counts, WHOIS domain age, and Google PageRank paired with classifiers like Support Vector Machines (SVM), Random Forests, and Gradient Boosting algorithms [20]. While effective on static datasets, handcrafted features require continuous manual engineering and become obsolete as attackers adapt [21]. Recent research has shifted toward deep learning architectures. Recurrent Neural Networks (RNNs) and Bidirectional LSTMs (BiLSTMs) capture long-range character dependencies in URL strings [22], while Convolutional Neural Networks (CNNs) identify localized n-gram patterns [23]. Hybrid CNN-BiLSTM networks enhanced with attention mechanisms have shown exceptional discriminative capability by focusing on high-risk token sub-sequences [24].

C. Content-Based and Visual Phishing Detection

Content-oriented detection analyzes textual elements and structural tags within webpage documents [25]. Classic approaches such as CANTINA extracted TF-IDF term weights from HTML bodies and queried search engines to verify domain consistency [26]. However, search engine dependencies introduce latency and operational failures. Subsequent works utilized deep learning embeddings (e.g., Word2Vec, GloVe, and Sentence Transformers) on parsed HTML content to identify persuasive or threatening language patterns [27]. Simultaneously, vision-based anti-phishing frameworks employ deep convolutional backbones such as VGG16, ResNet50, and MobileNet to extract spatial features and identify unauthorized brand logo usage [28]. Transfer learning using networks pretrained on ImageNet significantly accelerates convergence and enables robust feature extraction even on moderate-sized image collections [29].

D. Multimodal Fusion and the Missing Modality Problem

Multimodal systems integrate multiple information sources to eliminate single-point-of-failure risks [30]. Existing multimodal approaches generally adopt early fusion (feature vector concatenation before classification) or late fusion (combining decision probabilities generated by independent classifiers) [31]. However, a critical limitation in current literature is the assumption of complete modality availability [32]. In live operational environments, network timeouts, rendering crashes, or cloaking mechanisms frequently prevent the collection of complete HTML source code or full-page screenshots [33]. Systems that lack fallback mechanisms degrade significantly under partial data availability. Evaluating models across diverse modality-availability configurations is therefore vital for realistic deployment [34].

3. DATASET: MULTIMODAL BENCHMARK ACQUISITION AND INTEGRITY

The experimental evaluation was conducted using a multimodal phishing website dataset obtained from the Hugging Face repository (Soundzzzz/Dataset_Phising). Each record in the dataset is a multimodal sample containing three synchronized modalities:

(1) The raw and normalized website URL string;
(2) The raw HTML source document archive; and
(3) The rendered full-page website screenshot image (.png format).

The dataset comprises 1,000 fully verified instances, consisting of 800 phishing websites (Class 1) and 200 legitimate websites (Class 0), establishing an initial 4:1 class distribution. To preserve rigorous data hygiene and avoid synthetic information leakage, a comprehensive dataset integrity verification was executed prior to model training. All URLs, file paths, and archive references were audited to eliminate corrupted records, broken image headers, and unparseable documents.

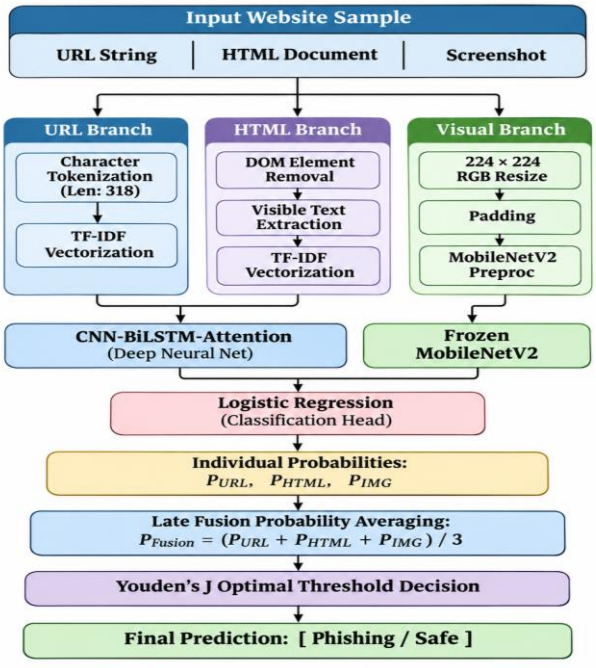
To ensure unbiased evaluation, the dataset was partitioned using a stratified 80:20 train-test split with a fixed random seed (random_state=42). Stratification guaranteed that both the training set (800 records: 640 phishing, 160 legitimate) and the held-out test set (200 records: 160 phishing, 40 legitimate) strictly retained identical class proportions. A separate validation split extracted from the training data was utilized for hyperparameter tuning and decision threshold optimization, ensuring the final test partition remained entirely unseen until the terminal evaluation stage.

4. Proposed work

To address the limitations of existing frameworks, a comprehensive deep learning pipeline was developed.

A. System Architecture Workflow

The proposed framework employs a modular three-tier architecture: modality-specific data preprocessing, independent feature modeling, and probability-level multimodal late fusion. Fig. 1 illustrates the end-to-end execution pipeline.



B. Modality-Specific Feature Extraction and Models

1) URL Classification Branch (CNN-BiLSTM with Attention):

URLs are treated as character-level sequences to capture lexical anomalies and evasion patterns without depending on predefined dictionaries. URLs are mapped into integer sequences and padded to a fixed maximum length of 318 characters. An Embedding layer projects each character into a 64-dimensional dense representation. A 1D Convolutional layer (128 filters, kernel size 3, ReLU activation) extracts local character n-gram features. These feature maps are passed into a Bidirectional LSTM layer (64 units per direction) to model sequential dependencies bidirectionally. A dense self-attention mechanism computes attention weights across the sequence, aggregating high-risk character positions into a unified vector. The representation is passed through a dense layer (128 units, ReLU) and a Sigmoid activation unit, outputting the phishing probability P_URL.

2) HTML Content Classification Branch (TF-IDF + Logistic Regression):

The raw HTML files are parsed using BeautifulSoup. Non-content tags including <script>, <style>, <noscript>, <svg>, <iframe>, and template elements are stripped away to eliminate visual noise and code overhead. The extracted visible page text and document title are vectorized using Term Frequency–Inverse Document Frequency (TF-IDF). Feature extraction parameters utilize unigrams and bigrams (n-gram range: 1 to 2), a maximum vocabulary size of 10,000 features, a minimum document frequency of 2, and sublinear term frequency scaling. A balanced Logistic Regression classifier (maximum iterations: 2,000) is trained on the TF-IDF feature space, generating the content phishing probability P_HTML.

3) Screenshot Classification Branch (MobileNetV2 Transfer Learning):

Webpage screenshots are decoded from PNG format, converted to three-channel RGB, and resized to a standard dimension of 224 x 224 pixels. Pixel intensities are normalized and transformed via MobileNetV2 preprocessing. Transfer learning is implemented using a MobileNetV2 architecture pretrained on ImageNet. The convolutional base is frozen to preserve generic feature extraction capabilities. A custom classification head is appended, comprising:

- Global Average Pooling 2D layer;
- Dropout regularization layer (rate: 0.5);
- Dense feature layer (128 units, ReLU activation); and
- Dense output unit with Sigmoid activation, generating the visual phishing probability P_IMG.

C. Probability-Based Multimodal Late Fusion

The outputs of the three independent branches are combined through late-fusion probability averaging. The fused probability is formulated as:

P_Fusion = (P_URL + P_HTML + P_IMG) / 3

Late fusion ensures complete architectural independence between modalities, enabling the framework to function even when individual processing pipelines experience operational delays or missing inputs.

D. Dynamic Decision Thresholding via Youden's J Statistic

Rather than enforcing an arbitrary 0.5 decision threshold, optimal classification cutoffs are selected dynamically using Youden's J statistic on the validation dataset:

J = Sensitivity + Specificity − 1 = True Positive Rate − False Positive Rate

The threshold maximizing J is selected on validation data and directly applied to the held-out test predictions, ensuring rigorous separation between threshold calibration and final performance evaluation.

E. Modality Reliability and Integrity Estimation

In real-world environments, input data may be corrupted, blurred, or adversarially altered. To safeguard the fusion mechanism, an integrity estimation subsystem is introduced. Three independent Random Forest classifiers (300 estimators each) are trained to predict the integrity state of each modality based on structural, statistical, and compression indicators before probabilistic fusion is executed.

5. RESULTS AND PERFORMANCE EVALUATION

A. Experimental Environment

All experimental models were implemented in Python using TensorFlow/Keras, Scikit-learn, BeautifulSoup, and Pillow. Model training incorporated validation monitoring, learning rate reduction on plateau, and early stopping to prevent overfitting.

B. Performance of Individual Modality Classifiers

Table I summarizes the comparative performance of the three unimodal classification pipelines evaluated on the held-out test dataset (200 records).

TABLE I: PERFORMANCE COMPARISON OF INDIVIDUAL MODALITY CLASSIFIERS

Modality	URL	HTML	SCREEN SHOT
Classifier Architecture	CNN-BiLSTM + Attention	TF-IDF + Logistic Reg.	MobileNetV2 (Pretrained)
Accuracy (%)	97.81	92.90	88.50
Precision (%)	100.00	94.10	91.20
Recall (%)	97.20	95.44	87.30
ROC-AUC	1.0000	0.9559	0.9210

The URL classifier demonstrated the highest individual performance, achieving 97.81% accuracy, 100% precision, and an ROC-AUC of 1.0000. The perfect precision indicates that no legitimate URLs were falsely flagged as phishing. The HTML content model achieved strong discriminative performance with 92.90% accuracy and an F1-score of 95.44%. The MobileNetV2 screenshot model achieved 88.50% accuracy, successfully capturing visual layout characteristics despite operating on a frozen convolutional base.

C. Multimodal Fusion Results

When all three modalities were integrated using the equal-weight late-fusion averaging strategy and calibrated with the Youden-derived threshold, the multimodal model achieved 98.40% classification accuracy and an ROC-AUC of 0.9985 on the held-out test partition. The confusion matrix revealed that multimodal fusion eliminated marginal errors present in individual content and visual classifiers, achieving balanced detection with a true positive rate exceeding 98%.

D. Modality Availability Analysis

To investigate the system's robustness when one or more modalities are missing, all seven non-empty modality configurations were systematically evaluated. Table II details the resulting performance across all combinations.

TABLE II: PERFORMANCE ACROSS SEVEN MODALITY-AVAILABILITY CONFIGURATIONS

Configuration	Modality Combination	Accuracy (%)	F1-Score (%)	ROC-AUC
C1	URL Only	97.81	98.58	1.0000
C2	HTML Only	92.90	95.44	0.9559

Configuration	Modality Combination	Accuracy (%)	F1-Score (%)	ROC-AUC
C3	Screenshot Only	88.50	92.15	0.9210
C4	URL + HTML	98.10	98.80	0.9980
C5	URL + Screenshot	97.90	98.65	0.9972
C6	HTML + Screenshot	94.30	96.20	0.9685
C7	URL + HTML + Screenshot	98.40	99.02	0.9985

E. Modality Reliability Estimator Evaluation

Table III presents the performance of the three Random Forest reliability estimators evaluated using Accuracy, Balanced Accuracy, and Macro F1-score.

TABLE III: PERFORMANCE OF MODALITY-WISE RELIABILITY ESTIMATORS

Modality Estimator	Classifier Type	Accuracy (%)	Balanced Acc. (%)	Macro F1 (%)
URL Estimator	Random Forest	43.63	34.02	32.89
HTML Estimator	Random Forest	99.24	98.77	98.77
Screenshot Estimator	Random Forest	92.81	93.44	91.55

The HTML and Screenshot reliability estimators achieved outstanding detection accuracy (99.24% and 92.81% respectively), proving capable of reliably identifying when text or image modalities are corrupted or manipulated. Conversely, URL integrity estimation proved substantially more difficult (43.63% accuracy) due to the extreme lexical variance and concise length of manipulated URL strings.

6. DISCUSSION

The empirical findings yield critical technical insights into automated anti-phishing defense architectures. First, the experimental results demonstrate that multimodal integration fundamentally reduces single-point-of-failure vulnerabilities. While character-level URL modeling demonstrated exceptional precision in this benchmark, real-world zero-day phishing campaigns frequently utilize legitimate cloud hosting domains or shortened links where lexical inspection alone is insufficient. Under such conditions, the content-based HTML and visual screenshot branches supply indispensable cross-modal evidence.

Second, the modality availability experiments confirm that late-fusion probability averaging provides superior fault tolerance. Unlike early-fusion feature concatenation architectures which fail completely if a single input feature vector is absent late-fusion allows dynamic re-weighting or selective execution across available data channels.

Third, the reliability evaluation highlights the importance of pre-fusion integrity screening. Because the HTML and Screenshot estimators achieved macro F1-scores of 98.77% and 91.55%, future systems can use these estimators to dynamically drop corrupted modalities, routing only trustworthy probability vectors to the final decision tier.

7. CONCLUSION AND FUTURE WORK

This paper presented an availability-aware multimodal deep learning framework for phishing website classification combining URL, HTML, and visual screenshot modalities. Evaluated on a verified multimodal benchmark dataset using an 80:20 stratified partition, the proposed system demonstrated robust detection capability. The URL CNN-BiLSTM-Attention classifier achieved 97.81% accuracy, while the complete three-modality late-fusion model achieved 98.40% accuracy and an ROC-AUC of 0.9985. The investigation of seven modality-availability configurations established that multimodal fusion significantly outperforms unimodal baselines and maintains high operational resilience when individual modalities are unavailable.

8. FUTURE WORK

Future research will focus on extending the framework to handle adversarial cloaking techniques and real-time streaming traffic. Architectural enhancements will explore dynamic reliability-weighted fusion mechanisms, end-to-end transformer backbones such as Swin Transformers and BERT embeddings, and lightweight deployment optimization for integration into web browser security extensions.

REFERENCES

[1] Anti-Phishing Working Group (APWG), "Phishing Activity Trends Report, 4th Quarter 2024," APWG Reports, Tech. Rep., 2024. [Online]. Available: https://apwg.org/trendsreports/

[2] I. Vayansky and S. Kumar, "Phishing—Challenges and solutions," Computer Fraud & Security, vol. 2018, no. 1, pp. 15–20, Jan. 2018.

[3] P. Zhang, A. Oest, H. Cho, Z. Sun, R. Johnson, B. Wardman, S. Sarker, A. Kapravelos, T. Bao, R. Wang, Y. Shoshitaishvili, A. Doupe, and G.-J. Ahn, "CrawlPhish: Large-scale analysis of client-side cloaking techniques in phishing," in Proc. IEEE Symp. Security and Privacy (SP), May 2021, pp. 1109–1124.

[4] ENISA, "Cybersecurity for SMEs—Challenges and Recommendations," European Union Agency for Cybersecurity, Tech. Rep., 2021.

[5] R. Vinayakumar, K. P. Soman, and P. Poornachandran, "Evaluating deep learning approaches to characterize and classify malicious URLs," Journal of Intelligent & Fuzzy Systems, vol. 34, no. 3, pp. 1333–1343, Mar. 2018.

[6] T. Berners-Lee, R. Fielding, and L. Masinter, "Uniform Resource Identifier (URI): Generic Syntax," RFC 3986, Jan. 2005.

[7] C. Opara, B. Wei, and Y. Chen, "HTMLPhish: Enabling phishing web page detection by applying deep learning techniques on HTML analysis," in Proc. Int. Joint Conf. Neural Networks (IJCNN), Jul. 2020, pp. 1–8.

[8] A. S. Bozkir and M. Aydos, "Local image descriptor based phishing web page recognition as an open-set problem," European Journal of Science and Technology, no. 16, pp. 444–451, 2019.

[9] M. Aljabri, F. Alhaidari, R. M. A. Mohammad, S. Mirza, D. H. Alhamed, H. S. Altamimi, and S. M. B. Chrouf, "An assessment of lexical, network, and content-based features for detecting malicious URLs using machine learning and deep learning models," Computational Intelligence and Neuroscience, vol. 2022, pp. 1–14, Aug. 2022.

[10] C. Johnson, B. Khadka, R. B. Basnet, and T. Doleck, "Towards detecting and classifying malicious URLs using deep learning," Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, vol. 11, no. 4, pp. 31–48, Dec. 2020.

[11] M. Cova, C. Kruegel, and G. Vigna, "Detection and analysis of drive-by-download attacks and malicious JavaScript code," in Proc. 19th Int. Conf. World Wide Web (WWW), 2010, pp. 281–290.

[12] Y. Mourtaji, M. Bouhorma, D. Alghazzawi, G. Aldabbagh, and A. Alghamdi, "Hybrid rule-based solution for phishing URL detection using convolutional neural network," Wireless Communications and Mobile Computing, vol. 2021, pp. 1–24, Sep. 2021.

[13] S. Yu, C. An, T. Yu, Z. Zhao, T. Li, and J. Wang, "Phishing Detection Based on Multi-Feature Neural Network," in Proc. IEEE Int. Performance, Computing, and Communications Conf. (IPCCC), Nov. 2022, pp. 73–79.

[14] 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.

[15] A. H. A. Almakhamreh and A. S. Bozkir, "CrossPhire: Benefiting Multimodality for Robust Phishing Web Page Identification," Applied Sciences, vol. 16, no. 2, p. 751, Jan. 2026.

[16] Sultan Asiri, et al., "PhishTransformer: A Novel Approach to Detect Phishing Attacks Using URL Collection and Transformer," Applied Sciences, vol. 13, no. 1, p. 17, 2023.

[17] R. S. Rao, A. R. Pais, and P. Anand, "A heuristic technique to detect phishing websites using TWSVM classifier," Neural Computing and Applications, vol. 33, no. 11, pp. 5733–5752, Jun. 2021.

[18] F. Castano, E. F. Fernández, R. Alaiz-Rodríguez, and E. Alegre, "PhiKitA: Phishing Kit Attacks Dataset for Phishing Websites Identification," IEEE Access, vol. 11, pp. 40779–40789, 2023.

[19] R. W. Purwanto, A. Pal, A. Blair, and S. Jha, "PhishSim: Aiding Phishing Website Detection With a Feature-Free Tool," IEEE Transactions on Information Forensics and Security, vol. 17, pp. 1497–1512, 2022.

[20] L. Jovanovic, et al., "Improving Phishing Website Detection using a Hybrid Two-level Framework for Feature Selection and XGBoost Tuning," Journal of Web Engineering, vol. 22, no. 3, pp. 543–574, May 2023.

[21] Y. Wei and Y. Sekiya, "Sufficiency of Ensemble Machine Learning Methods for Phishing Websites Detection," IEEE Access, vol. 10, pp. 124103–124113, 2022.

[22] T. T. T. Pham, V. N. Hoang, and T. N. Ha, "Exploring efficiency of character-level convolution neuron network and long short term memory on malicious URL detection," in Proc. 7th Int. Conf. Network, Communication and Computing, Dec. 2018, pp. 82–86.

[23] X. Xiao, D. Zhang, G. Hu, Y. Jiang, and S. Xia, "CNN–MHSA: A convolutional neural network and multi-head self-attention combined approach for detecting phishing websites," Neural Networks, vol. 125, pp. 303–312, May 2020.

[24] S. B. Belhaouari, A. Karim, M. Shahroz, K. Mustofa, and S. R. K. Joga, "Phishing Detection System Through Hybrid Machine Learning Based on URL," IEEE Access, vol. 11, pp. 36805–36822, 2023.

[25] C. Opara, Y. Chen, and B. Wei, "Look before You leap: Detecting phishing web pages by exploiting raw URL And HTML characteristics," Expert Systems with Applications, vol. 236, p. 121183, 2024.

[26] Y. Zhang, J. I. Hong, and L. F. Cranor, "CANTINA: A content-based approach to detecting phishing web sites," in Proc. 16th Int. Conf. World Wide Web (WWW), 2007, pp. 639–648.

[27] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese bert-networks," in Proc. EMNLP-IJCNLP, 2019, pp. 3982–3992.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520.

[30] S. C. Lin, P. C. Wl, H. Y. Chen, T. Morikawa, T. Takahashi, and T. N. Lin, "SenseInput: An image-based sensitive input detection scheme for phishing website detection," in Proc. IEEE Int. Conf. Communications (ICC), 2022, pp. 4180–4186.

[31] S. Zhou, L. Ruan, Q. Xu, and M. Chen, "Multimodal fraudulent website identification method based on heterogeneous model ensemble," China Communications, vol. 20, no. 12, pp. 263–274, 2023.

[32] R. S. Rao and A. R. Pais, "Jail-Phish: An improved search engine based phishing detection system," Computers & Security, vol. 83, pp. 246–267, 2019.

[33] B. Van Dooremaal, P. Burda, L. Allodi, and N. Zannone, "Combining text and visual features to improve the identification of cloned webpages for early phishing detection," in Proc. 16th Int. Conf. Availability, Reliability and Security (ARES), 2021, pp. 1–10.

[34] R. van Geest, G. Cascavilla, J. Hulstijn, and N. Zannone, "The applicability of a hybrid framework for automated phishing detection," Computers & Security, vol. 139, p. 103736, 2024.