Phisher - A Multimodal Approach for Phishing Detection

Shubham Bhadra

Mission San Jose High School (Grade 12)
Polygence
Email: shobershubham@gmail.com
Country: United States

September 22, 2025

Abstract

Phishing is an illegal method used to trick people into revealing confidential information, such as login details, credit card numbers, and Social Security numbers. The majority of these phishing activities are carried out by duplicating the appearance of authentic websites or emails and exploiting people's trust, rather than technical vulnerabilities. Numerous awareness campaigns and technical countermeasures are designed to alert individuals to the dangers of phishing. Still, it remains one of the most effective methods of cyber assault due to its malleability and continually evolving complexity. Many single-modal models are effective to a certain degree, but cannot identify advanced phishing techniques that incorporate dynamic web content, obfuscated scripts, and sophisticated visual mimicry. We introduced a novel multimodal approach called Phisher. Our multimodal models utilize the BERT Multimodal Large Language Model (MLLM) for combined lexical analysis, ResNet50 for image processing, and semantic characteristics for URL extraction, thereby enhancing phishing classification. By combining these signals, we can achieve better accuracy, precision, and F1 score, which facilitates more effective detection of phishing sites. To test our multimodal model, we utilized the TR-OP real-life dataset, which contains 10,000 labeled phishing and legitimate websites, including HTML content, URLs, and website snapshots. The results show a significant improvement in accuracy and precision compared to other models. Aside from the technical benefits, this research also demonstrates how Multimodal learning can create more resilient defenses against evolving cybercrimes and phishing and offer practical applications for enterprises and security providers to build a safer digital ecosystem.

Keywords: cybersecurity, phishing detection, multimodal learning, BERT, ResNet50, TR-OP dataset

1 Introduction

Phishing is perhaps the most widespread and destructive type of cybercrime these days. It describes a kind of social engineering attack wherein attackers disguise themselves as a trusted

entity—institutions like financial institutions, government agencies, or online services—to trick individuals into divulging sensitive personal details. Most of these phishing attempts look like legitimate emails, mock websites, or SMS messages but are designed carefully to take advantage of human trust instead of a technical vulnerability. Attackers usually entice victims into clicking on dangerous links, entering login credentials, or installing malware (Wikipedia, 2024).

According to the FBI's Internet Crime Report, phishing was the most reported cybercrime in 2023, accounting for over 700,000 complaints (Federal Bureau of Investigation, 2024) and resulting in estimated financial losses exceeding $2.9 billion globally. The Anti-Phishing Working Group (APWG) also observed a record high in phishing attacks during Q4 2023, with over 1.4 million unique phishing websites detected in just three months. These statistics demonstrate that phishing is not only widespread but also growing in both scale and sophistication.

To counter the threat, many detection mechanisms have been implemented over the years. All typical phishing detectors rely on blacklisting, rule-based systems, or manually written lexical features. These endure long enough for established attacks, but will not discover new or covert phishing attacks. Attackers deploy advanced evasion protocols, evading mainstream detection. Dynamic content generation, for example, is the term for the exploitation of sites whose malicious content is downloaded after certain interaction, either by humans or by scripts, so it evades scanners scanning statically. Script obfuscation is the process of obscuring or encoding malicious JavaScript so that it evades keyword-based filters while still executing malicious activity in the browser. Visual spoofing exploits techniques that include copying the logo, the login screen, or the entire site layout, so the human will think they are working with the brand they trust. In addition to these, homograph attacks (the deployment of using similarly appearing characters in the URL, for example, "paypa1.com" for "paypal.com") as well as iframe injection (embedding hidden log-in forms on legitimate-looking sites) are also widespread. Such evasion mechanisms exploit trust by the human in addition to vulnerabilities by the detection mechanisms, so the typical defenses are less successful.

With the emergence of machine learning (ML) and artificial intelligence (AI) in recent years comes new promise for phishing detection too. ML-based models can be learned against large datasets of URLs, emails, or web pages. Static rule-based methods have been surpassed in terms of accuracy by methods like decision trees, random forests, support vector machines (SVMs), and newer deep learning methods. These methods largely base themselves on features solely from a single modality—for instance, examining textual content or the structure of URLs in isolation. Although single-modal models have their place, they possess significant drawbacks. Detectors like URL-based can identify domain names that appear suspicious, but cannot possibly take into consideration contextual or semantic information. Text-based models can be fooled with carefully constructed text or scripted embedding. Even vision-based methods can be fooled with subtle image manipulation. As sophistication in phishing increases, relying on a single input modality becomes increasingly inadequate for comprehending the diverse facets of phishing websites.

In order to address these challenges, focus has now been directed towards multimodal phishing detection where various types of information—such as the structure of the URL, HTML content, and visual aspects—are combined for more reliable and solid prediction. These modalities offer distinct levels of insight: URL features identify anomalous tokens, domain anomalies, or suspicious patterns. HTML features help identify hidden elements, malicious JavaScript, or fake forms. Visual features (e.g., screenshots) detect branding mismatches, layout forgery, and UI copying. In this paper, a multimodal detector for phishing is proposed, which combines all three interaction modes for more accurate classification. We design our

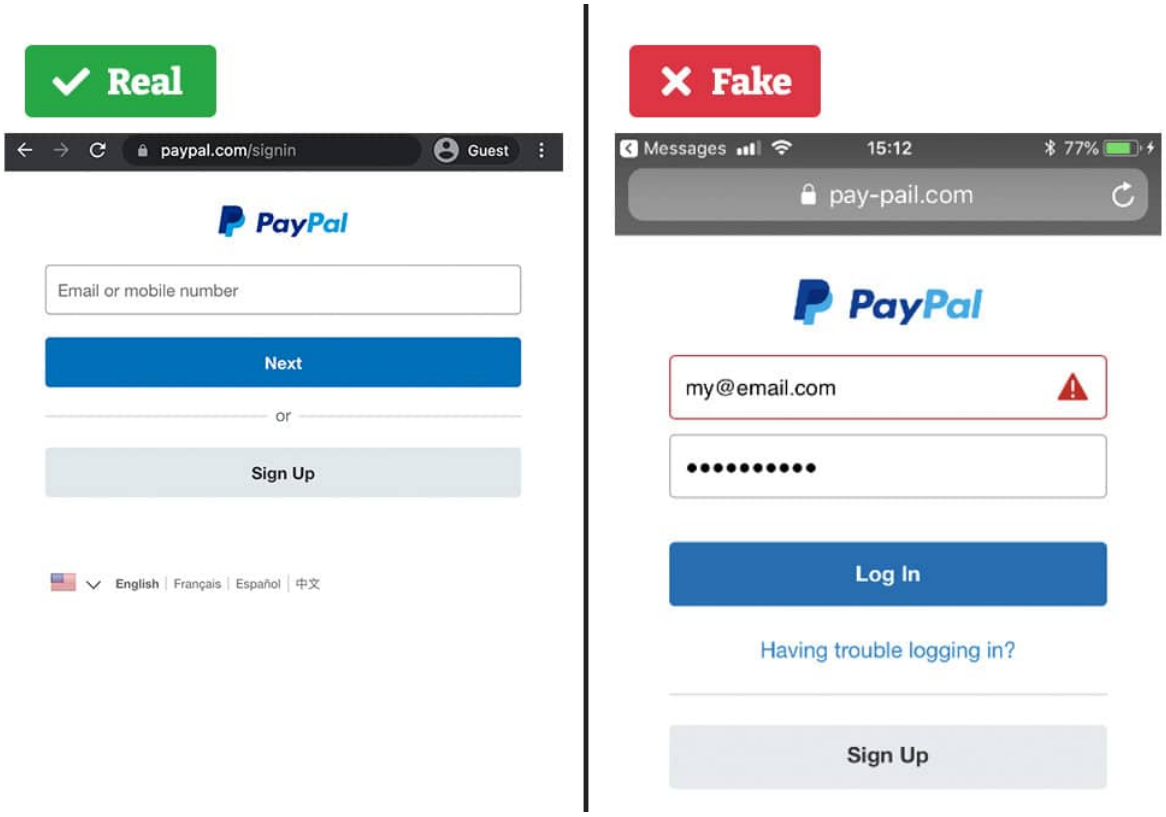

Figure 1: Comparison of a genuine PayPal login page and a phishing attempt. The left side (labeled "Real") demonstrates the real PayPal log-in screen, using the Correct domain paypal.com/signin. The right side (labeled "Fake") copies PayPal's style but adopts a deceptive domain pay-pail.com, which has the primary intention of duping users into validating their qualifications. The case illustrates the importance of checking the website's URL before supplying sensitive information. Figure is obtained from https://techwiser.com/7-black-friday-scams-with-tips-to-protect-your-hard-earned-money/

method using the TR-OP (Threat Report - OpenPhish) dataset, which comprises 10,000 labeled instances of the web, along with corresponding screenshots, HTML source code, and URLs. For basic-level text analysis, TF-IDF vectors along with BERT embedding vectors are used for identifying shallow and deep patterns semantically. Visual features are detected with the help of convolutional neural networks (CNNs) like ResNet50, whereas structures in the URLs is converted into a set of vector representations using handcrafted features (static) along with statistical features (learned). Combining these three views, our detector is more accurate with enhanced resistance towards advanced phishing attacks. We compare our method against multiple machine learning models, including Phishpedia(image and text bimodal), KPD (knowledge-based), GEPAgent(multimodal using URL and HTML structure), ChastPhish(single modal LLM text-based), and PhishAgent (multimodal URL, HTML, image similarity with Reference). Performance is assessed using accuracy and F1 scores. Our model achieves an accuracy of 98.13% and 98% F1 score of Y, significantly outperforming baseline models.

1.1 Related Works

There are many approaches to detecting phishing websites, ranging from traditional rule-based techniques to advanced AI-driven methods. Early detection relied heavily on **blacklist and whitelist approaches**, which store known malicious or trusted domains. While effective against already-identified threats, these methods are inherently reactive and struggle against newly registered or fast-changing domains, requiring constant maintenance and often failing to keep up with rapid domain churn[1]. To overcome these limitations, **heuristic-based detection** was introduced, leveraging handcrafted rules such as suspicious URL patterns, SSL certificate usage, or excessive form fields in HTML. Although easy to interpret, heuristics quickly became rigid and could be bypassed by attackers making minor adjustments to webpage structures, leading to high false negatives if not frequently updated[2]. Similarly, **keyword-based detection** sought to identify phishing by flagging terms such as "login" or "verify account" within URLs or HTML. However, this technique suffered from false positives on legitimate sites with similar vocabulary, and adversarial tactics such as Unicode obfuscation or typosquatting further reduced its reliability[3].

To capture more semantic information, **reference-based approaches** were developed, where suspicious sites are compared against knowledge bases containing domains, logos, and aliases of popular brands. While effective for well-known targets, these methods face scalability and coverage issues, especially for local or emerging brands absent from the knowledge base[4]. Another direction involved **search engine-based techniques**, where page content and domains are queried in search engines to validate legitimacy. Yet, these methods are highly sensitive to indexing delays, search algorithm updates, and can misclassify lesser-known legitimate sites as phishing[5].

With the rise of deep learning, **LLM and MLLM-based approaches** have emerged, using textual and visual data—such as HTML source code and website screenshots—to identify spoofed brands and manipulated content. These models significantly outperform earlier methods, achieving higher accuracy in detecting visual mimicry and textual inconsistencies[6]. However, they remain vulnerable to adversarial evasion and misclassifications on underrepresented brands due to data limitations. To address these weaknesses, researchers have explored **agent-based approaches**, which combine LLMs with toolkits, reasoning modules, and external knowledge bases. By enabling multi-modal decision-making and adaptability, agent-based systems improve resilience against unseen phishing strategies and enhance accuracy over stan-

dalone models[7].

We provide background on the baseline phishing detection models used for comparison. These approaches span single-modal, bimodal, and multimodal paradigms.

Phishpedia

Phishpedia is a bimodal approach that combines webpage screenshots with OCR-extracted text to identify brand impersonation. It performs well in detecting visual mimicry but struggles against script-based or structural phishing attacks[8].

KPD (Knowledge-based Phishing Detection)

KPD employs knowledge graphs to model semantic relationships between domains, entities, and brands. While effective at spotting entity inconsistencies, it has limited coverage for unseen or emerging brands not represented in the knowledge base[4].

GEPAgent

GEPAgent is an agent-based system that applies graph embeddings and reinforcement learning to capture relational structures in phishing websites. Although it achieves fair accuracy, its inference time is extremely high (over 12 seconds), limiting real-time applicability[9].

ChatPhish

ChatPhish uses large language models to analyze emails and webpages. It excels at contextual reasoning and text understanding but is resource-intensive and vulnerable to adversarial prompts or ambiguous cases involving underrepresented brands[10].

PhishAgent

PhishAgent is a multimodal framework that integrates URL features, HTML structures, and image data where it tries to use the logo for brand recognition through online and offline content. It achieves a balanced trade-off across accuracy, precision, and recall, though it incurs longer inference times compared to our approach[11].

1.2 Multimodal Training Models

The proposed multimodal phishing detector significantly advances existing phishing site classification using a multiplicity of sources of information—i.e., HTML textual information, visual page screenshots, and lexical characteristics of URLs. In contrast with prior single-modality systems devoted exclusively to a single type of signal, a multimodal system aggregates complementary indications and therefore improves the accuracy and strength of the phishing detector for complex real-world applications[12].

1.2.1 Improved Detection Rates

Single-modal techniques—although operating for a single instance—will generally not generalize across the extensive set of phishing techniques. For example, a classifier trained using sole lexical features of URLs can detect anomalous URLs like http://www.pay-pail.com correctly (Figure 1). Still, it will not detect a visually deceptive page sent off a compromised legitimate

domain. A text-only technique will not be able to detect phishing pages that obscure text within the image or employ encoded scripts. Through an integration of textual, visual, and URL features at a multimodal level, the multimodal model can cross-reference between sources for any single signal. For instance, if the HTML has suspicious text such as "validate account" and the visual layout is very similar to a PayPal login layout, even a seemingly innocent-looking URL like https://secure.login-center.com is flagged correctly. This integration minimizes false negatives (phishing going undetected) and false positives (legitimate content being flagged as phishing). Empirical evidence confirms this strength. A multimodal phishing detector called PhishAgent[11] which enhanced overall F1-scores 7–10 % relative to single-modal detectors.

1.2.2 Resistance to Sophisticated Attacks

Phishing websites attempt to keep a low profile by focusing their deceptions on a single modality. A phishing page, for instance, can sport a neat and innocent-looking URL but pack malicious scripts into the HTML, or vice versa, sport well-crafted text but a suspicious-looking structure for the URL. These deceptions aim at precluding systems from checking a single layer of data. A multimodal defense is automatically more immune to such evasion methods. A phisher may disguise anomalies within the URL but deploys false branding or dubious text content, yet the textual or visual modality can still activate an alarm. For example, even if an attacker deploys https://amazon-check-secure.com, the use of manipulated Amazon logos or forms characteristic of phishing can be identified with the ResNet-based visual classifier. Its multiple-layer redundancy ensures that even an attempt at evasion through a single modality won't impact overall system performance. This multi-layer redundancy guarantees that an attempt at evasion over a single modality will not impact the system's overall functionality.

1.2.3 Improved Semantic Comprehension

Conventional keyword-centric or bag-of-words methods for text classification usually have difficulties dealing with semantic ambiguity and context sensitivity. Incorporation of Bidirectional Encoder Representations from Transformers (BERT) allows for deep semantic understanding of webpage text. In contradistinction with shallow models, BERT is capable of grasping deep sentence structure, contextual use, and subtle deception markers. In other words, a phishing message like "Please validate your account to prevent suspension" can be semantically equivalent to "We must validate your login for security purposes," even if the language is slightly different. BERT's transformer-based structure enables the model to understand this equivalence and mark both examples, whereas older models might overlook one. It's a basic contextual knowledge for recognizing phishing material, which is abusive psychologically, or legally seasoned language manipulation for a sense of panic or fear.

1.2.4 Visual Pattern Recognition

One of the most efficient methods of phishing is visual mimicry, which attackers often use in creating pages that look like they belong to reputable institutions. ResNet50 is a deep convolutional neural network that can extract fine-grained spatial hierarchies from images. By checking screenshots of websites, the visual modality can identify things such as:

- Logo cloning (e.g., PayPal, Apple, Microsoft),
- Form placements mimicking login screens,

- Font inconsistencies or improper alignment,
- Low-resolution or manipulated images.

For instance, a phishing page may use a blurred version of the Bank of America logo or a layout with unusually large "Submit" buttons to trick users. Even if the textual or URL indicators are mild, the visual features alone could expose the attack. Our visual embedding approach improves precision in such cases, with ResNet-based models achieving up to 98.5% precision in recognizing high-risk mimicry patterns[13].

1.2.5 Thorough URL Analysis

Phishing URLs also often carry anomalies like atypical subdomain patterns, overuse of numbers, or the inclusion of IP addresses rather than domain names. Although certain URLs may look superficially innocuous at a glance, more in-depth lexical analysis can identify manipulation tactics. Specifically, features such as URL length, number of digits, hyphens, subdomains, and the presence of an IP address are key indicators of malicious intent[14]. Even sophisticated phishing attempts usually reveal at least a single anomaly within the URL. Encoding these lexical features enables the model to recognize malicious activity beyond what appears on the surface.

1.2.6 Generalization and Adaptability

Phishing tactics continuously evolve with time. Attackers modify language, image content, domain name, and layout pattern in an attempt to bypass static rule-based detection. Most single-modality systems must be retrained regularly or require feature engineering for change accommodation. Multimodal systems, on the other hand, provide better generalization. Regardless of whether phishers refine a single modality (e.g., employing grammatically correct text), overall detection effectiveness is preserved because the other modalities provide strong signals themselves. This design provides for longevity of adaptability as well as a lessened reliance on a single threat indicator. For instance, a 2023 paper about building privacy-preserving and secure AI foundation models demonstrated that multimodal models maintain more than 95% accuracy even if a single modality is corrupted or adversarially occluded, which is a far higher rate than 70–80% for individual modality systems[15].

2 Methods

Current phishing detection systems, whether rule-based, single-modal ML, or keyword-focused ones, have real shortcomings in adaptability, coverage, and resilience. Standard blacklists cannot detect new or zero-day phishing domains. Heuristics are fragile and can be easily evaded using basic evasion methods. Vision-only or text-only ML systems break down when attackers include text in images, employ visually obfuscating methods that mimic genuine websites, or cunningly mask malicious intention in the URL. Even modern LLM-founded models, though robust, remain susceptible to adversarial examples and tend to classify attacks on less-represented brands incorrectly.

To counter such limitations, we introduce a fully multimodal system — Phisher, which is a multimodal detector for phishing, employing textual feature extraction using BERT embeddings, visual feature extraction with ResNet50, and URL lexical feature engineering for

robust binary classification of sites as harmless or phishing. With the integration of submodules for processing these classifiers, this system significantly increases detection accuracy and is capable of detecting subtle phishing attacks that remain undetected under single-modality approaches. Each modality gives a strong representation of the respective input. Then, these embeddings are concatenated into a single feature vector for the classification model. We experiment with four different classification models: Neural Networks, Naïve Bayes, Random Forests, and XGBoost. The architecture for the neural network classifier is displayed in Figure 2.

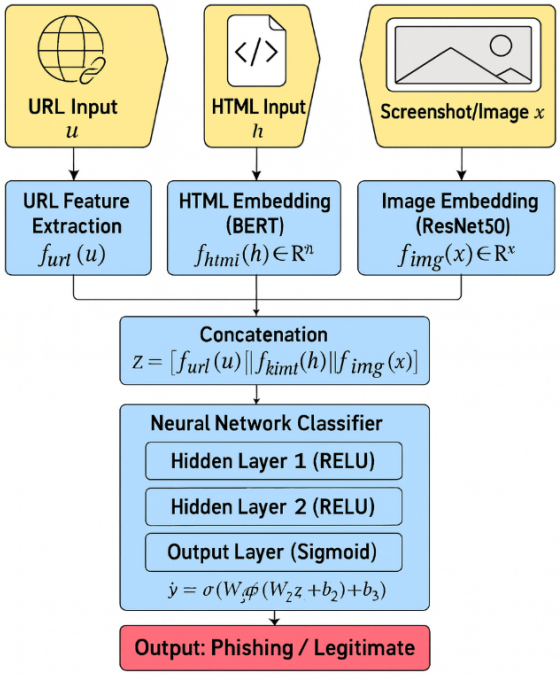

Figure 2: Architecture of the multimodal phishing detection model combining textual, visual, and URL features.

2.1 Textual Embedding (BERT)

To capture the semantic and contextual information embedded in the HTML content of a webpage, we utilize the Bidirectional Encoder Representations from Transformers (BERT) model. BERT is well-suited for this task due to its ability to understand the deep bidirectional context of language and code structures within HTML.

Given the HTML content H, the BERT embedding E_{HTML} is defined as:

$$E_{\text{HTML}} = \text{BERT}(H), \quad E_{\text{HTML}} \in R^{768} \tag{1}$$

BERT computes embeddings through a series of Transformer encoder layers. These encoders capture both syntactic and semantic dependencies in the tokenized HTML. We extract the final pooled output from the transformer as the representative embedding vector:

$$E_{\text{HTML}} = \text{PoolerOutput}(\text{Transformer}_{\text{encoder}}(\text{Tokenizer}(H))) \tag{2}$$

This textual embedding captures both surface-level token patterns (e.g., suspicious words

like "login", "account", "verify") and deeper semantic signals (e.g., obfuscated malicious scripts or deceptive metadata).

2.2 Visual Embedding (ResNet50)

Visual information from webpage screenshots plays a crucial role in identifying phishing attempts that mimic the appearance of legitimate websites. Phishing pages often replicate logos, buttons, and layouts of trusted organizations, which can be effectively captured using convolutional neural networks (CNNs).

For the screenshot image I, ResNet50 generates the embedding E_{IMG}:

$$E_{\text{IMG}} = \text{ResNet50}(I), \quad E_{\text{IMG}} \in R^{2048} \tag{3}$$

Specifically, ResNet50 consists of deep residual blocks that enable learning of rich hierarchical features. The visual embedding is computed by applying a global average pooling operation to the final CNN feature maps:

$$E_{\text{IMG}} = \text{GlobalAvgPooling}(\text{CNNLayers}(I)) \tag{4}$$

This vectorized representation encodes global visual patterns—such as font, color schemes, UI alignment, and brand mimicry—that are often exploited in phishing.

2.3 URL Lexical Feature Extraction

URLs remain one of the most indicative elements of phishing attacks. Malicious URLs often include specific lexical patterns such as excessive use of numeric characters, unusually long domains, or misleading keywords.

The URL U lexical feature vector E_{URL} is:

$$E_{\text{URL}} = [L_U, D_U, P_U, S_U, H_U, IP_U, HTTPS_U] \tag{5}$$

$$E_{\text{URL}} \in R^7 \tag{6}$$

The features are defined as follows:

- L_U: URL length — Longer URLs are often used to obfuscate malicious intent.
- D_U: Number of digits — Excessive digits may indicate autogenerated or fake subdomains.
- P_U: Number of periods (dots) — Used to create misleading subdomains or deep URL nesting.
- S_U: Number of slashes — Indicates URL depth or directory structure manipulation.
- H_U: Number of hyphens — Often used to mimic legitimate domains (e.g., pay-pal.com).
- IP_U: Binary indicator (IP address presence: 0 or 1) — IP-based URLs are common in phishing sites.
- $HTTPS_U$: Binary indicator (HTTPS usage: 0 or 1) — While HTTPS is generally secure, its misuse in phishing sites is growing.

These lexical indicators are lightweight yet powerful features in identifying suspicious URLs.

2.4 Multimodal Feature Fusion

Each modality—textual, visual, and lexical—offers a unique perspective on the phishing detection problem. While individually useful, these modalities can complement one another when fused, providing a holistic view of the webpage.

The concatenated embedding E_{Concat} is:

$$E_{\text{Concat}} = [E_{\text{HTML}}; E_{\text{IMG}}; E_{\text{URL}}], \quad E_{\text{Concat}} \in R^{2823} \tag{7}$$

This fused feature vector combines: - Semantic representations from HTML content via BERT, - Visual patterns from webpage screenshots via ResNet50, and - Lexical structures from URLs via handcrafted features.

The integration of all three modalities strengthens the model's ability to generalize across various phishing techniques, particularly those that evade detection in single-modality systems.

2.5 Neural Network Classifier

The final classification of a webpage as phishing or legitimate is performed using a fully connected neural network. This model maps the high-dimensional fused feature vector into a probability score between 0 and 1 using a nonlinear transformation.

Classification is performed using a two-layer fully connected neural network:

$$\hat{y} = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot E_{\text{Concat}} + b_1) + b_2) \tag{8}$$

where:

- W_1, W_2: Weight matrices responsible for learning feature transformations.
- b_1, b_2: Bias vectors that help the network shift activation boundaries.
- $\text{ReLU}(x) = \max(0, x)$: Rectified Linear Unit activation for introducing non-linearity.
- $\sigma(x) = \frac{1}{1+e^{-x}}$: Sigmoid activation function to squash outputs between 0 and 1.

Number of layers and implications:

- The network contains **two fully connected layers**: a hidden layer followed by an output layer.
- The hidden layer (W_1, b_1) enables the model to capture high-level, nonlinear interactions among features from multiple modalities (HTML, URL, and image embeddings).
- The ReLU activation in the hidden layer prevents vanishing gradient issues and allows efficient learning of complex patterns.
- The output layer (W_2, b_2) with a sigmoid activation produces a probability score, making the classifier suitable for binary classification tasks such as phishing vs. legitimate detection.
- This architecture balances model complexity and interpretability: it is expressive enough to learn non-linear boundaries while remaining lightweight, which reduces overfitting and allows efficient inference in real-world settings.

This structure enables the classifier to learn complex decision boundaries and leverage the complementary strengths of each modality to accurately predict phishing instances.

3 Dataset

The TR-OP dataset is a handcrafted test dataset designed primarily for multimodal phishing website detection. Its purpose is to reflect realistic phishing and legitimate website scenarios by combining visual, textual, and structural information. This is perfect for multimodal machine learning modeling for training and testing purposes. It is also suited for multimodal models that combine visual embeddings (e.g., from ResNet50 or ViT), textual embeddings (e.g., from BERT or TF-IDF), and URL feature vectors (custom or heuristic-based). The TR-OP dataset comprises 5000 labeled phishing websites and 5000 labeled legitimate websites, which are randomly sorted and then split into 70-30 ratio for train-test datasets. Screenshot captures of legitimate pages compared with phishing pages, typically generated with headless browsers, represent layout, brand abuse, and visual indicators. Extracting HTML source code or rendering text content from the webpage allows access to suspicious keywords, hidden forms, and script behaviors. The raw features of URLs, like length, occurrence of special characters, subdomain depth, and lexical patterns, are usually connected with phishing URLs.

4 Results

We deployed four machine learning algorithms—Random Forest, XGBoost Classifier, Naive Bayes, and a Neural Network—on the TR-OP multimodal dataset merged with combined URL features, HTML content, and visual embeddings. These were chosen to span the broad range of learning paradigms: probabilistic classification (Naive Bayes), ensemble decision trees (Random Forest, XGBoost), and deep learning (Neural Networks). Differing from prior work on single-modality analysis, our assessment deploys these methods on merged multimodal features together, allowing us to test the efficacy of standard algorithms adapting to merged phishing signals.

4.1 Evaluation Metrics

Performance is measured by accuracy, precision, recall, F1 score, and inference time.

- **Accuracy**: Proportion of correctly classified phishing and legitimate sites.
- **Precision**: Fraction of predicted phishing sites that were truly phishing.
- **Recall**: Fraction of true phishing sites that were correctly identified.
- **F1 Score**: Harmonic mean of precision and recall, balancing false positives and false negatives.
- **Inference Time**: Average time taken for a model to classify one instance.

4.2 ML Model Comparison

Among the assessed models, the highest accuracy of **0.9813** was attained by the Neural Network, and the second highest accuracies of **0.9768** and **0.9707** were attained by the Random Forest and Naive Bayes respectively. The highest F1 score of **0.9800** was achieved by both the Neural Network and the Random Forest, and that of the Naive Bayes attained **0.9700**. These experiments demonstrate that classical machine learning models also have

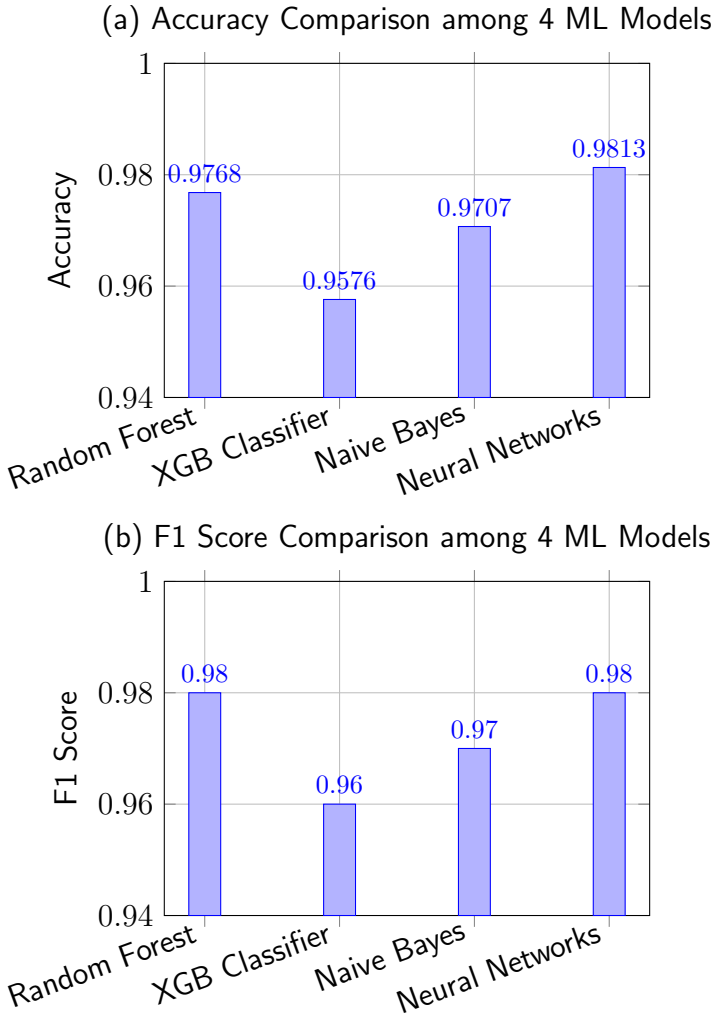

Figure 3: Performance comparison of our Model using different ML models based on (a) Accuracy and (b) F1 Score.

performance advantages, but multimodal feature combination and ensemble/deep learning methods improve robustness.

Figure 3 presents a bar-graph comparison of model accuracies and F1 scores. While the differences in performance may appear numerically small, we confirmed their statistical robustness by running each model across **five independent trials with shuffled train-test splits**, reporting the averaged results to mitigate variance.

XGBoost Classifier was the fastest model, consuming only 0.0003 seconds in inference, but had the lowest accuracy (0.9576) and F1 score (0.9600), reflecting a trade-off between predictive power and speed. In general, Neural Networks gave the best predictive output, while Random Forest and Naive Bayes offered a balanced compromise between accuracy and interpretability.

The purpose of the comparison is not merely to offer point-wise model accuracy, but it is also to offer the foundation of interpreting the value addition of multimodality. The improvement demonstrated by all the models reflects that the integration of the lexical, structural, and vision signals provides a better representation than that of the single feature streams. It rationalizes our aim of designing a fully multimodal phishing detection system and brings the improvement achieved by the offered methodology into perspective.

4.3 Single vs Multimodal Comparison

In order to highlight the benefit of integrating multiple modalities, we compared our proposed multimodal detector against three single-modality baselines: HTML-only, Image-only, and URL-only. As shown in Table 1, the single-modality models achieved moderate performance, with accuracies ranging from 72–89%. Their corresponding F1, precision, and recall scores followed a similar trend, reflecting limitations in capturing the diverse characteristics of phishing websites when relying on only one input source. In contrast, the multimodal model substantially outperformed all baselines, reaching an accuracy of 98.13% with balanced precision (98.50%) and recall (98.90%). Although the inference time slightly increased to 0.0135 seconds, this trade-off is minimal compared to the significant performance gain.

Table 1: Performance comparison between single-modality detectors and the proposed multimodal detector.

Detector	Accuracy (%)	F1 Score (%)	Precision (%)	Recall (%)	Time (s)
HTML-only	72.82	73.10	74.20	72.00	0.0087
Image-only	75.76	76.00	77.10	75.20	0.0121
URL-only	89.34	89.50	90.10	88.90	0.0064
(Our Model)	**98.13**	**98.00**	**98.50**	**98.90**	**0.0135**

Benefit of Multimodality

The purpose of the two comparisons is not merely to report point-wise model accuracy, but also to highlight the value of multimodality. As we can see, the accuracy and F1 score remains similar even though we used multiple ML models. At the same time the head-to-head comparison against single-modality models(trained separately on only URL, HTML, or screenshot features) vs our Multimodal model has shown that across all metrics, the multimodal models significantly outperformed their single-modality counterparts, showing a relative accuracy improvement of 10-18%. This demonstrates that integrating lexical, structural, and visual signals provides a stronger representation than any single feature stream.

4.4 Comparing other Detection Models

Table 2 presents the performance of various phishing detection models. Our model outperforms others in both accuracy and inference time.

Table 2: Comparison on TP-OP benchmark datasets.

Detector	ACC	F1	Precision	Recall	Time (s)
Phishpedia	85.15	52.76	**98.84**	41.30	0.30
KPD	92.05	91.44	99.92	85.99	1.22
GEPAgent	92.95	92.01	98.59	89.80	12.35
ChatPhish	95.80	95.01	98.10	93.60	0.93
PhishAgent	96.10	96.13	95.24	97.05	2.25
Our Model	**98.13**	**98.00**	98.50	**98.90**	**0.0135**

To contextualize the results in Table 2, we provide background on the baseline phishing detection models in the Relevant Work section, which are used for comparison. These approaches span single-modal, bimodal, and multimodal paradigms.

Our Model (Phisher)

Our model is a multimodal phishing detector that integrates URL token analysis, HTML semantic features, and ResNet-based visual embeddings. Unlike prior methods, Phisher is designed to balance robustness with efficiency. We tested our model against five other alternative models. As seen in Table 2, Phisher achieves the highest overall accuracy (98.13%), precision (98.50%), and recall (98.90%), while also delivering the lowest inference time (0.0135 seconds), making it suitable for real-time deployment at scale.

Why Results Differ

The observed differences across models can be attributed to their methodological focus. Vision-centric models such as Phishpedia achieve high precision but low recall, since they often miss attacks without strong branding cues. Knowledge-based and graph-based systems like KPD and GEPAgent perform well on structured relationships but cannot adapt quickly to zero-day or obfuscated attacks. LLM-based approaches such as ChatPhish offer strong contextual reasoning but suffer from computational overhead and adversarial vulnerability. PhishAgent demonstrates the promise of multimodal integration but is slowed by longer inference times. In contrast, our model leverages multimodal signals while optimizing for lightweight computation, yielding superior performance across accuracy, recall, and real-time efficiency. This positions Phisher as a reliable, stable, and scalable solution for modern phishing detection.

Among all models, Phisher (Our Model) outperforms others with the best accuracy of 98.13%, a F1-score of 98.00%, and a significantly low inference time of 0.0135 seconds. This indicates a great balance between real-time efficiency and detection performance.

While other models, such as PhishAgent[11] and ChatPhish[10], also have good accuracy and recall rates, they lag behind slightly in end-to-end accuracy and run a significantly longer inference time for each test case. GEPAgent[9], although providing fair accuracy rates, is very computational with an over 12-second inference time. It has high accuracy but very low recall, i.e., it misses a large number of phishing examples. On the other hand, KPD[4] and PhishAgent[11] have a relatively better-balanced trade-off but cannot be compared with the overall performance of our model. Overall, Phisher excels over current detectors in virtually all aspects, becoming a reliable, stable, and efficient solution for real-time phishing detection.

5 Discussion

In this paper, we introduced an end-to-end multimodal phishing detection model that seamlessly combines visual, textual, and URL-level signals to fight against advanced phishing attacks. Integrating ResNet50-extracted image embeddings, BERT-based HTML content embeddings, and engineered URL features, our system derives an overall multimodal representation of phishing sites in diverse views. Comprehensive experiment on a range of model architectures, such as neural networks, Random Forest, and XGBoost, verified that multimodal fusion enhances classification performance over unimodal baselines. Our top-performing model

achieved outstanding accuracy, precision, recall, and F1-score, and a correspondingly competitive inference time, ready for deployment in real time. Our finding justifies the usefulness of multimodal learning in computer security, particularly phishing detection, in which malicious users often trick not just the content or the link, but the appearance of websites.

The model holds great promise for use in browser extensions, mail gateways, or organizational-level defense systems against phishing. One direction for future work involves transformer-based fusion architectures, dynamic tracking of URLs, and large-scale in-the-wild deployment. Another significant direction involves the extension of multimodal phishing detection to emails, where attackers often use malicious sender addresses, malicious attachments, embedded images, and social engineering in the text of the message. Adding modalities like the analysis of the email header, attachment scanning, and NLP-based detection of language that is persuasive would greatly expand the system's utility. Another promising direction involves cross-channel detection, in which the system correlates signals across websites, emails, and messaging services to capture phishing campaigns at the level of the ecosystem.

That being said, our study is not free from limitations. One of the main issues that were facing during experimentation was the possibility of data leaking. Since the TR-OP dataset contains several feature modalities (images, HTML, URLs) that come from the same webpage, if feature extraction were to be carried out prior to the split of the dataset, it could have the unintended consequence of transferring information from the training dataset into the test dataset. In order to avoid that, our preprocessing and feature extraction were carried out exclusively after split of test and training data. This step proved to be imperative to experimental validity and avoiding spuriously high performance. The training dataset, as diverse as it is, is still oriented more towards popular domains and brands, which may translate into weak performance on phishing attacks on underrepresented or local sites. Second, increasing robustness is our multimodal design, albeit greatly expanding preprocessing requirements that have to be processed promptly by the HTML parsing and extraction of the features of images. Third, our multimodal system, like many others, is vulnerable to the challenge of the problem of evasive methods of adversarial attacks such as pixel-level perturbations on the screenshot or adversarially generated HTML; it will be necessary that follow-up research addresses these limitations, particularly to enhance generalizability across the unseen plans of attack and scalability within the large-scale, real-world scope of deployment of the work of our study.

Our solution, in general, offers a promising direction toward making the internet safer through smart, multi-faceted AI systems that provide protection across websites as well as emails, thereby offering an all-around defense against phishing in today's digital landscape.

6 Acknowledgments

I would like to acknowledge the guidance and support of my mentor, Nigel D., whose insights and feedback were invaluable throughout this project. I also extend my gratitude to *Polygence* for providing the platform and resources that enabled me to pursue and complete this research successfully.

References

[1] Ali Aljofey, Qingshan Jiang, Abdur Rasool, Hui Chen, Wenyin Liu, Qiang Qu, and Yang Wang. An effective detection approach for phishing websites using url and html features. *Scientific Reports*, 12(1):8842, 2022.

[2] M Vijayalakshmi, S Mercy Shalinie, Ming Hour Yang, and Raja Meenakshi U. Web phishing detection techniques: a survey on the state-of-the-art, taxonomy and future directions. *Iet Networks*, 9(5):235–246, 2020.

[3] Samuel Marchal, Kalle Saari, Nidhi Singh, and N Asokan. Know your phish: Novel techniques for detecting phishing sites and their targets. In *2016 IEEE 36th international conference on distributed computing systems (ICDCS)*, pages 323–333. IEEE, 2016.

[4] Yuexin Li, Chengyu Huang, Shumin Deng, Mei Lin Lock, Tri Cao, Nay Oo, Hoon Wei Lim, and Bryan Hooi. {KnowPhish}: Large language models meet multimodal knowledge graphs for enhancing {Reference-Based} phishing detection. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 793–810, 2024.

[5] Routhu Srinivasa Rao and Alwyn Roshan Pais. Jail-phish: An improved search engine based phishing detection system. *Computers & Security*, 83:246–267, 2019.

[6] Jehyun Lee, Peiyuan Lim, Bryan Hooi, and Dinil Mon Divakaran. Multimodal large language models for phishing webpage detection and identification. In *2024 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–13. IEEE, 2024.

[7] Wenhao Li, Selvakumar Manickam, Yung-wey Chong, and Shankar Karuppayah. Phishdebate: An llm-based multi-agent framework for phishing website detection. *arXiv preprint arXiv:2506.15656*, 2025.

[8] Yun Lin, Ruofan Liu, Dinil Mon Divakaran, Jun Yang Ng, Qing Zhou Chan, Yiwen Lu, Yuxuan Si, Fan Zhang, and Jin Song Dong. Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 3793–3810, 2021.

[9] Huilin Wang and Bryan Hooi. Automated phishing detection using urls and webpages. *arXiv preprint arXiv:2408.01667*, 2024.

[10] Takashi Koide, Naoki Fukushi, Hiroki Nakano, and Daiki Chiba. Detecting phishing sites using chatgpt. *arXiv preprint arXiv:2306.05816*, 2023.

[11] Tri Cao, Chengyu Huang, Yuexin Li, Wang Huilin, Amy He, Nay Oo, and Bryan Hooi. Phishagent: A robust multimodal agent for phishing webpage detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27869–27877, 2025.

[12] Tandin Wangchuk and Tad Gonsalves. Multimodal phishing detection on social networking sites: A systematic review. *IEEE Access*, 2025.

[13] Sahar Abdelnabi, Katharina Krombholz, and Mario Fritz. Visualphishnet: Zero-day phishing website detection by visual similarity. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 1681–1698, 2020.

[14] Shayan Abad, Hassan Gholamy, and Mohammad Aslani. Classification of malicious urls using machine learning. *Sensors*, 23(18):7760, 2023.

[15] Jinmeng Rao, Song Gao, Gengchen Mai, and Krzysztof Janowicz. Building privacy-preserving and secure geospatial artificial intelligence foundation models (vision paper). In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–4, 2023.