DeepPhish-X: Multi-Modal Feature Engineering for Phishing Detection Using Hybrid Models of Computer Vision, Natural Language Processing, and Graph Neural Networks

Program Name:	**Masters of Science (Data Science)**
Subject Name:	**Project 1 (MCST1043)**
Student Name:	Cui ZhiWen
Metric Number:	MCS241040
Student Email & Phone:	cuizhiwen@graduate.utm.my
Project Title:	DeepPhish-X: Multi-Modal Feature Engineering for Phishing Detection Using Hybrid Models of Computer Vision, Natural Language Processing, and Graph Neural Networks
Supervisor 1:	
Supervisor 2 / Industry Advisor(if any):	

2. Related Works

Phishing detection has long been a critical area of cybersecurity research. As summarized in Table 2, various approaches have been developed over the years to tackle this issue.[1]

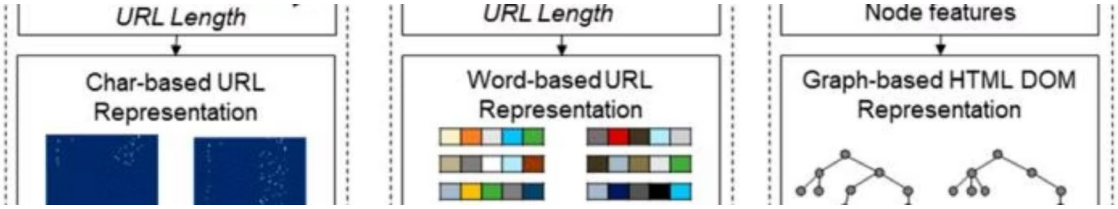
URL-Based Methods

Early methods primarily focused on analyzing URL features, considering factors like URL length, suspicious substrings, and domain reputation.[1] For example, the Texception model uses convolutional layers to analyze both character-level and word-level information of URLs, achieving notable performance on large datasets.[1] Advancements in phishing detection have seen the integration of multiple machine learning techniques, such as MOE/RF, which combines multi-objective evolution optimization with Random Forest, yielding high accuracy and recall.[1] Similarly, GramBeddings employs a four-channel architecture with CNN, LSTM, and attention layers, demonstrating significant accuracy on various datasets.[1] The use of adversarial examples, as seen in URLBUG, highlights the challenges posed by adversarial attacks, which degrade the performance of machine learning models. Notably, URLBUG's performance was lower compared to other models because it tested on adversarial URLs generated to deliberately evade detection, showcasing the difficulties in maintaining robustness when dealing with generated data.[1] For example, one method combines multiple machine learning techniques to analyze the lexical features of URLs and web-scraped content, integrating URL structure and web content for a more comprehensive detection approach.[1] Another method explores embedding URL components and testing against adversarial attacks, enhancing model robustness and making it more effective in responding to sophisticated evasion techniques.[1]

HTML-Based Methods

While URL-based methods offer valuable insights, they often fail to capture the full context of phishing attacks. HTML-based approaches, such as PhishSim, address this limitation by analyzing the content and structure of the webpage, achieving high detection rates.[1] Recent

research has increasingly focused on integrating URL and HTML features for a more comprehensive detection strategy. For example, a method integrates MLP for structured data with NLP models for HTML content, fusing embeddings to improve detection accuracy.[1]

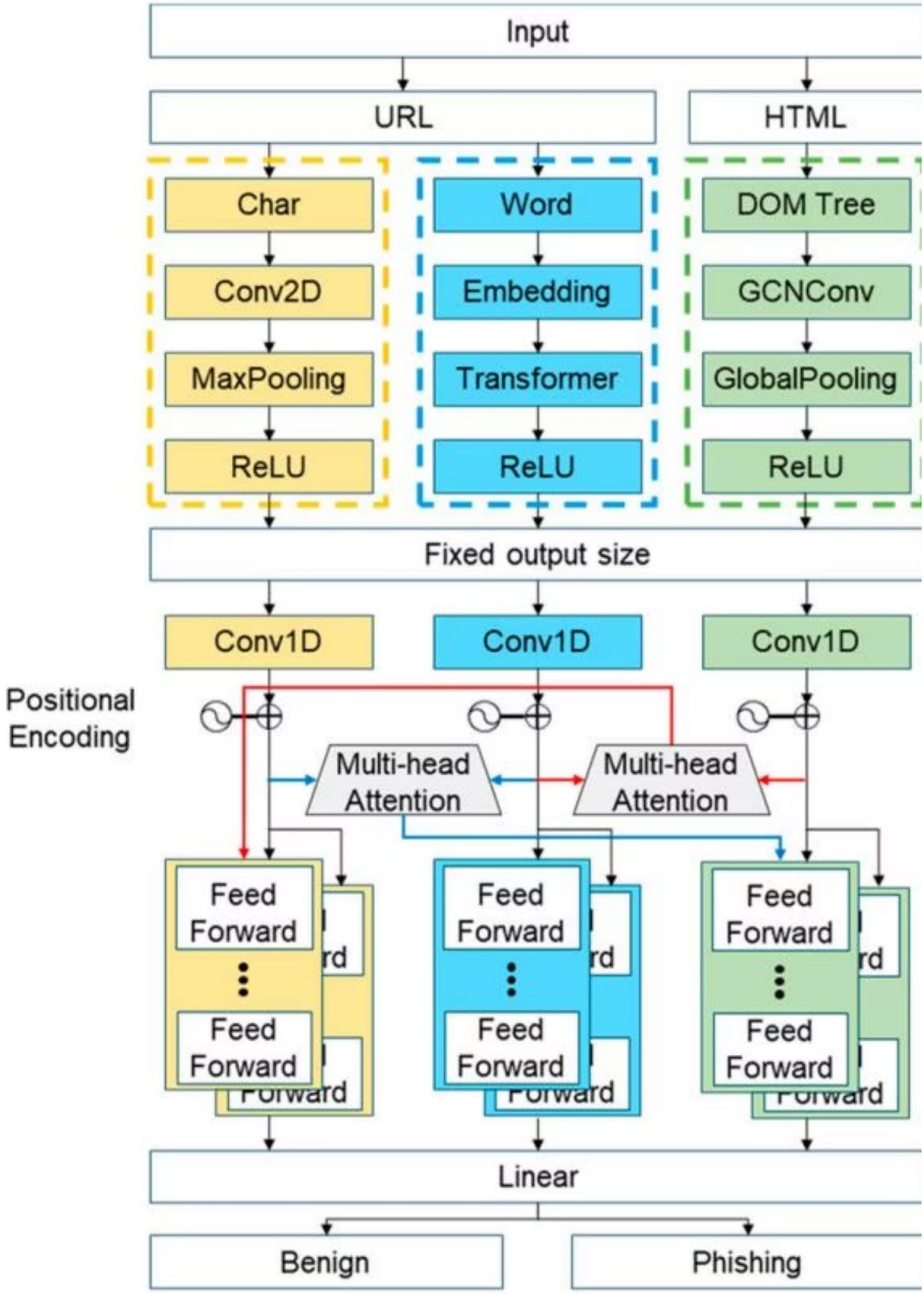


Multi-Modal Methods

The WebPhish framework combines raw URL and HTML content analysis, achieving high accuracy, and demonstrating the effectiveness of multi-modal approaches.[1] The PhiUSIIL framework leverages URL similarity indexing and incremental learning to adapt to real-time threats, achieving near-perfect accuracy, precision, and recall, which underscores the potential of real-time adaptive models in phishing detection.[1] Another notable approach integrates raw URL, HTML tags, and image analysis using word embeddings and convolutional layers, achieving high accuracy and demonstrating the benefits of incorporating multiple data types for phishing detection.[1]

The evolution of related research, from simple URL feature analysis to complex HTML and multi-modal integration, directly reflects the escalating sophistication of phishing attacks. Initially, simple URL checks might have sufficed, but as attackers learned to mimic URLs and adopt more complex webpage structures, detection methods also had to adapt accordingly. The emergence of adversarial examples and the decline in URLBUG model performance, further underscore the continuous "arms race" between attackers and defenders. This suggests that cybersecurity research is inherently reactive and adaptive. The development of increasingly sophisticated deep learning models (such as those utilizing GCNs and Transformers) is a necessary response to the adaptive and innovative nature of cybercriminals. This implies a need for continuous acceleration in research and development to stay ahead of new attack vectors, obfuscation techniques, and adversarial operations, emphasizing that no single, static defense mechanism can remain effective indefinitely.

DeepPhish-X distinguishes itself from existing research by leveraging Graph Convolutional Networks (a **Graph Neural Network** technique) to effectively model the complex dependencies among HTML tags within the DOM structure, thereby optimizing feature representation for phishing detection.[1] Furthermore, DeepPhish-X employs a Transformer network (a


Input
URL
HTML
Char
Word
DOM Tree
Conv2D
Embedding
GCNConv
MaxPooling
Transformer
GlobalPooling
ReLU
ReLU
ReLU
Fixed output size
Conv1D
Conv1D
Conv1D
Positional Encoding
Multi-head Attention
Multi-head Attention
Feed Forward
Feed Forward
Linear
Benign
Phishing

Natural Language Processing model) to integrate URL features with HTML DOM Graph features, enabling the model to selectively attend to and extract complementary relationships among these multi-modal features.[1] This "complementary" nature means that these disparate pieces of information, when fused, provide a richer and more robust understanding of the true nature of a webpage than any single modality could offer. This precise feature integration enhances the overall detection accuracy and robustness. This principle of synergy—where the combined intelligence of diverse data types and specialized processing models yields superior results—is key to designing effective AI systems in complex adversarial domains like cybersecurity.

Table 2: Overview of Recent Studies on Phishing Detection, Categorized by URL-Based, HTML-Based, and Multi-Modal Approaches, Including Methodologies, Data Representations, Datasets, and Performance Metrics [1]

Method	Representation	Description	Dataset	Performance
Multimodal Classification	Raw URL, HTML content	Combines raw URL and HTML content analysis, using embeddings and convolutional layers	Alexa (benign): 22,687 Phishtank (phishing): 22,687	Accuracy: 98.1% Precision: 98.2% Recall: 98.1% F1 Score: 98.1%
	Image, Raw URL, HTML Tags	Combines raw URL, HTML tags and image analysis, using word embeddings and convolutional layers	Collected from PhishiTank and OpenPhish benign: 8316 phishing: 7848	Accuracy: 95.35% Precision: 94.39% Recall: 94.26% F1 Score: 94.34%
URL Classification	Character-level, Word-level	Uses parallel convolutional layers to analyze character and word-level URL information	Collected 1.7 M samples from Microsoft browsing telemetry data	TPR: 47.95% Error Rate: 0.28%
	URL features	Combines multi-objective evolution optimization with	Five different URL datasets (Kaggle, Mendeley Data)	Accuracy: 99.04% Recall: 99.48%

		Random Forest for phishing detection		
	N-gram embeddings	Utilizes n-gram embeddings with a four-channel architecture, includes CNN, LSTM, attention layers	Alexa, Majestic (benign): 400 K Phishtank, Openphish (phishing): 400 K	Accuracy: 98.27% F1 Score: 98.26%
	Adversarial URL Generation	A method to generate adversarial URL by obfuscating domain, path, and TLD parts of URL to test the robustness of ML-based phishing URL. detectors	Dataset consists of five different sources with a total of 193,386 benign: 96,693 phishing: 96,693	(Performance reduction occurred due to testing on generated adversarial data) Accuracy: -41.62% F1 Score:-59.17%
	Lexical, Web-scrapped	Combines multiple ML techniques to extract and analyze lexical and web-scrapped features	Dataset consists of 12 different sources with a total of 3,980,870	Accuracy: 99.63% Precision: 99.60%
	URL embedding	Analyzes robustness of ML models against adversarial attacks	Alexa (benign): 10,000 Phishtank (phishing): Not specified	Precision: 91% Recall: About 93% F1 Score: About 92.5%
HTML Classification	HTML content	Uses Normalized Compression Distance to compare HTML content with known phishing pages	Common Crawl (benign): 180,302 Phishtank (phishing): 9034	AUC: 98.68% TPR: About 90% FPR: 0.58%

	HTML content	Integrates MLP for structured data with NLP models for HTML content, fuses embeddings	Alexa (benign): 2000 Openphish (phishing): 2000	Accuracy: 97.18% F1 Score: 96.80%
	URL, HTML	Proposes PhiUSIIL, a phishing URL detection framework combining URL similarity indexing and incremental learning for real-time threat adaptation	Open PageRank Initiative Anon (benign): 134,850 Phishtank, OpenPhish, Malware World (phishing): 100,945	Accuracy: 99.97% Precision: 99.97% Recall: 99.98% F1 Score: 99.98%