Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

FMURL-H: A Federated Multimodal Approach with Hyperparameter Optimization for Malicious URL Detection

Rachid Khalladi[1], Youcef FEKIR[1], Abdallah Khelil[1]
[1] *Laboratory of Informatics and Intelligent Systems(LISYS), Department of Computer Science, University Of Mustapha Stambouli, Mascara, Algeria.*

ARTICLE INFO

Received: 29 Dec 2024
Revised: 12 Feb 2025
Accepted: 27 Feb 2025

ABSTRACT

One important threat vector for malware, phishing, and defacement attacks is malicious Uniform Resource Locators (URLs). A number of approaches, including those based on machine learning, have been developed to deal with this issue. Because they frequently rely on centralized learning and lexical features, traditional detection models continue to be inadequate and constrained in terms of scalability and confidentiality.

This study introduces FMURL-H (Federated Multimodal URL Detection with Hyperoptimization), a novel framework that combines transformer-based encoders (DeBERTa/RoBERTa), multimodal features, federated learning, and hyperparameter optimization with Optuna. When tested on a massive dataset of 651,191 URLs, FMURL-H outperformed other approaches with an Accuracy of 99.30%, Precision of 99.42%, Recall of 99.48%, and F1 score of 99.45%. This model establishes a new standard for malicious URLs that are both scalable and privacy-conscious.

Keywords: Malicious URL Detection, Federated Learning, Multimodal Features, Transformer Models, Hyperparameter Optimization, Optuna, Phishing..

INTRODUCTION

The Internet has grown rapidly in recent years, becoming a vital component of daily life in the digital age. As a global communication platform, it offers a wealth of information for various sectors such as government institutions, businesses, academia, and individuals. Thanks to the widespread use of smartphones and increased connectivity, users now have the ability to access online content from anywhere [1] [11].

However, this ease of access also makes them vulnerable to both trusted and malicious sources, increasing their susceptibility to cyberattacks and online fraud. Many attacks are carried out via malicious websites to steal users' sensitive information. In this context, spoofed websites resembling legitimate ones are created to steal information.

As the name suggests, a URL is a website's identifier. It indicates its location and how to access it. When a user opens a URL, it contacts the server and returns the site's content. Some URLs are safe, others are malicious. In phishing, attackers create fake URLs that look like real ones. They are often sent by email or hidden in advertisements [9] [11].

A single click can install malware or redirect to a site that steals credentials. Thousands of new sites appear every day. Attacks are also becoming more complex. Distinguishing a reliable URL from a malicious one is therefore difficult. Since the two often look similar, specific characteristics must be analyzed to properly classify them.

The rapid creation of malicious URLs highlights the importance of automatic, reliable, and scalable methods.

Today, much research uses machine learning to detect dangerous sites. However, most existing solutions are limited to analyzing a single page and rely on manually defined characteristics. This reduces their ability to adapt to other contexts and scale efficiently [7] [8] [9] .

Traditional detection methods often rely on blacklists or simple lexical analysis. These methods are easy to circumvent. Attackers can slightly modify URLs to avoid detection. Hence the need for more powerful and intelligent solutions.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Machine learning and deep learning have improved detection rates [6] [9]. Transformer models such as BERT, RoBERTa, and DeBERTa capture semantic information in text. They can detect hidden patterns that older methods fail to detect. However, challenges remain. Centralized training increases privacy risks. It also presents scalability challenges when datasets come from different sources [3] [11] .

Federated learning offers a solution. It allows multiple clients to train models locally. Only the model parameters are shared, not the raw data. This ensures privacy and promotes collaboration between organizations. Hyperparameter optimization further improves performance and automatically finds the best model configuration.

In this work, we present FMURL-H, a federated, multimodal framework for malicious URL detection. The model combines transformers, lexical and contextual features, federated learning, and hyperparameter tuning. Experiments on real-world datasets show that FMURL-H achieves high precision, recall, and F1 score. The model outperforms current benchmarks while preserving data privacy.

The contributions of this paper are as follows:

- We propose FMURL-H, an innovative hybrid framework that combines multimodal features, transformer-based encoders, federated learning, and Optuna-based hyperparameter tuning.
- We integrate contextual data sources such as WHOIS and Cyber Threat Intelligence, which enrich detection beyond lexical features.
- We design a federated configuration that guarantees data confidentiality while maintaining high detection performance.
- We provide a comprehensive evaluation on large-scale datasets, demonstrating results superior to those obtained with robust benchmarks.

RELATED WORKS

Several previous studies have explored the detection and classification of malicious URLs using various machine learning and deep learning methods, with varying levels of performance. These studies have provided insights into effective methodologies and techniques for detecting and categorizing malicious URLs. To evaluate the contribution of these methodologies in comparison to previous research, a compilation of various articles and sources was compiled to highlight the characteristics that distinguish benign URLs from malicious URLs. The following table (Table 1) shows a clear depiction of these methods and the various algorithms used to classify and detect malicious URLs.

Table 1. Summary table of previous work

Study	Detection Approach	Feature Extraction Method	Dataset Size	Performance Metrics
[4]	Heuristic-based (not clearly specified machine learning or artificial intelligence)	No mention found (URL features)	6,000 (3,000 phishing, 3,000 legitimate)	Accuracy:98.23%,
[10]	Deep learning (Convolutional Neural Network and Transformer)	Convolutional Neural Network for local features, Transformer encoder for long-range dependencies, character-level tokenization, special indexes	10,000 (5,000 phishing, 5,000 legitimate); custom from PhishArmy, PhishTank, Alexa	Accuracy: 98.9%, Precision: 99%, Recall: 99%, F1-score: 99%
[9]	Supervised machine learning (Random Forest)	Intra-URL relatedness, search engine query features (Google, Yahoo), Jaccard index	96,018 (48,009 phishing, 48,009 legitimate); PhishTank, Directory Mozilla	Accuracy: 94.91%, False positive rate: 1.44%

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[3]	Deep learning (word2vec-based embeddings)	Character embedding, URL structure partitioning	No mention found; public 1M-PD dataset	Accuracy: 99.69%, False positive rate: 0.40%, Recall: 99.79%
[1]	Deep neural network (WebPhish)	Embedding technique, concatenation layer, convolutional layers	No mention found	Accuracy: 98.1%
[5]	Machine learning (Logistic Regression and Term Frequency-Inverse Document Frequency)	Term Frequency-Inverse Document Frequency	90,000 (30,000 phishing, 60,000 legitimate); custom PILU-90K	Accuracy: 96.50%
[8]	Machine learning (Random Forest, natural language processing features)	Natural language processing-based features	No mention found; custom dataset	Accuracy: 97.98%
[6]	Traditional machine learning (Support Vector Machine, Random Forest, Gradient Boosting, Neural Network)	URL lexical analysis, feature reduction	1,056,937	Accuracy: 99.89%
[7]	Deep learning (Bidirectional Encoder Representations from Transformers and Convolutional Neural Network)	Bidirectional Encoder Representations from Transformers (natural language processing), Convolutional Neural Network, next sentence prediction	549,346; public Kaggle dataset	Accuracy: 96.66%, Precision: 96.66%, F1-score: 93.63%

METHODS

1. MODEL DESCRIPTION

In this research, a hybrid detection framework, FMURL-H (Federated Multimodal URL Detection with Hyperoptimization), is proposed. Although it is a hybrid model, it mainly consists of:

1. Semantic encoders (based on transformers).
2. Multimodal features.
3. Federated learning.
4. Hyperparameter optimization.

The model works as follows (See Figure 1 for a graphical description):

Step 1: Takes raw URLs as input and then extracts lexical and contextual features.

Step 2: The extracted lexical features capture statistical properties (length, entropy, and number of special characters).

Step 3: Contextual features are obtained using WHOIS metadata and Cyber Threat Intelligence (CTI) signals.

Step 4: The transformation encoder (DeBERTa or RoBERTa) maps URLs into high-dimensional embeddings representing semantic patterns often ignored by traditional models.

Step 5: Learning is done in a federated manner, where multiple clients train local models on their private data.

Step 6: To aggregate parameters, the FedAvg (federated average) method is used. This allows for the production of a global model without exposing the raw data.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Step 7: To further improve performance, Optuna is used for hyperparameter optimization and dynamic tuning of the learning rate, dropouts, and batch size.

Step 8: Decision boundary refinement and false alarm reduction are achieved using a LightGBM classifier.

FMURL-H offers accurate, scalable, and privacy-preserving malicious URL detection, outperforming conventional benchmarks.

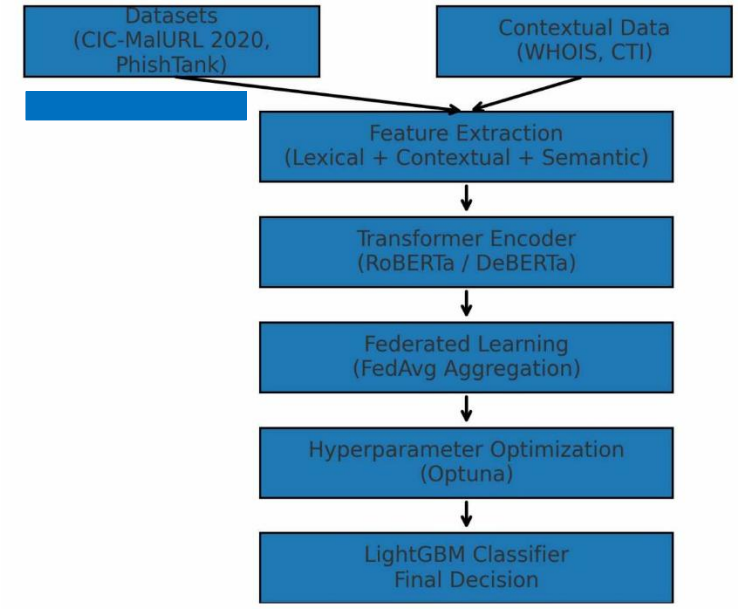

Figure .1 Flow shart of FMURL-H

2. MATHÉMATICAL MODELING

Let a set of distributed clients be $C = \{1, \ldots, N\}$. Each client i holds a local dataset $Di = (x_i^j, y_i^j)_{j=1}^{m_i}$, where:

- x_i^j is a URL
- $y_i^j \in \{0,1\}$ the label $(0 = benign,\ 1 = malicious)$.

1. Semantic Encoding with Transformer:

The transformer generates a vector representation:

$$h_i^j = f_{Trans}(x_i^j; \theta_T)$$

Where θ_T are the parameters of the DeBERTa/RoBERTa model.

2. Multimodal Concatenation:

The lexical representations l_i^j , WHOIS w_i^j , and CTI c_i^j are merged with h_i^j : $z_i^j = [h_i^j \| l_i^j \| w_i^j \| c_i^j]$

3. Local Classification:

Each client learns a local function $f_i(z)$ parameterized by θ_i : $\hat{y}_i^j = f_i(z_i^j; \theta_i)$ with local loss :

$$\mathrm{L}_i(\theta_i) = \frac{1}{m_i} \sum_{j=1}^{m_i} \ell(\hat{y}_i^j, y_i^j)$$

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Where ℓ is the cross-entropy function.

4. Federated Aggregation (FedAvg):

$$\theta^{(t+1)} = \sum_{i=1}^{N} \frac{|D_i|}{|D|} \theta_i^{(t)}$$

5. Hyperparameter Optimization (Optuna):

$$\lambda^* = \arg\max_{\lambda \in \Lambda} F1(f_{\theta_\lambda})$$

6. LightGBM Post-Classification:

The final prediction is done by : $\hat{y} = f_{\text{LGBM}}(z; \theta_{LGBM})$

This formulation ensures privacy-preserving, optimized, and multimodal malicious URL detection.

EXPERIMENT

This section explains how we tested FMURL-H. We describe the datasets, the steps we used to prepare the data, how the model was trained, and how we measured performance.

1. Datasets

We have used the **CIC-MalURL 2020**. This dataset was created by the Canadian Institute for Cybersecurity. It has more than 650,000 URLs. Each URL is labeled as benign, phishing, or malware. The dataset covers different attack tricks, such as very long URLs, hidden subdomains, and encoded characters [12].

We also added **contextual information**. WHOIS records give details like domain age and registrar. Cyber Threat Intelligence (CTI) feeds give blacklist scores and DNS data. These are not full datasets, but they provide extra features that improve detection.

2. Preprocessing

We prepared the data in several steps.

- URLs were split into tokens such as subdomains, paths, and parameters.
- We extracted lexical features, for example length, entropy, and number of special characters.
- WHOIS data was converted into numeric values.
- CTI scores were added as reputation features.

3. Model Setup

FMURL-H combines different modules.

- A **Transformer encoder** (RoBERTa or DeBERTa) processes the tokens and learns semantic patterns. [15][16] [18]
- **Federated learning** simulates multiple clients. Each client trains a local model. The server combines them with FedAvg [17].
- **Optuna** is used for hyperparameter search. It finds the best learning rate, batch size, dropout, and transformer layers.[14]
- **LightGBM** is the final classifier. It improves decision boundaries and lowers false positives. [13] [19]

4. Evaluation metrics

To improve the evaluation effectiveness of the proposed models, we utilized several recognized metrics, namely the Accuracy, precision, recall F1-score.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
- Precision = $\frac{TP}{TP+FP}$
- Recall = $\frac{TP}{TP+FN}$
- F1-Score = $\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$

RESULTS AND DISCUSSION

The results obtained (See Table 2) show that FMURL-H outperforms other methods. This is due to the integration of multimodal features that allow the model to detect complex malicious URLs that rely on obfuscation. Thus, the use of transformer encoders captures semantic relationships that other traditional models fail to identify.

Table 2. Comparison of results with other existing methods

Model	Accuracy	Precision	Recall	F1-score
CNN +Tokenization[10]	98.9%	99%	99%	99%
Fine-tuning BERT [18]	98.78%	99.12%	98.02%	98.57%
Light GBM Classifier [19]	95.6%	95%	96%	95%
BERT + Optuna [11]	98.84%	98.95%	99.08%	99.02%
LightGBM + TF-IDF [11]	97.63%	97.51%	97.72%	97.61%
FMURL-H (Proposed)	**99.30%**	**99.42%**	**99.48%**	**99.45%**

Furthermore, federated learning offers a significant advantage as it allows organizations to collaborate without sharing sensitive data. This configuration reduces privacy risks, making the approach more practical for real-world deployment. Experiments confirm that our model performs almost as well as a model based on centralized learning, while maintaining distributed data.

Hyperparameter optimization also plays a key role. The manual method often produces suboptimal results. Thanks to Optuna, the model automatically finds the best configuration, reducing training time and improving dataset stability.

One limitation is the computational cost of transformers. Their training requires significant resources. In addition, federated learning results in communication overhead between clients and the server. These challenges suggest that practical deployment will require optimized infrastructures. Despite these constraints, the gain in accuracy and confidentiality makes FMURL-H a promising solution.

CONCLUSION

FMURL-H is a federated multimodal framework for malicious URL detection. It is a hybrid model combining transformer-based encoders, lexical and contextual features, federated learning, and automated hyperparameter optimization. Experiments on large-scale datasets show that FMURL-H achieves state-of-the-art accuracy, precision, recall, and F1 score.

This model improves detection performance, guarantees data privacy, and demonstrates the value of combining multimodal information sources. Thus, it demonstrates that hyperparameter optimization is essential for stable results in cybersecurity tasks.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Future work will focus on three areas. First, the integration of differential privacy and homomorphic encryption to enhance the security of federated learning is proving very interesting. Another perspective is the exploration of lightweight transformer architectures (to reduce computational costs and energy consumption). Finally, test FMURL-H on real-time traffic flows and extend multimodal inputs with threat intelligence feeds from social media and DNS records.

These improvements can make FMURL-H a deployable solution for next-generation intrusion detection systems..

REFRENCES

[1] C. Opara, Yingke Chen, and Bo.wei. "Look Before You Leap: Detecting Phishing Web Pages by Exploiting Raw URL And HTML Characteristics." *Expert Systems with Applications*, 2020.

[2] Harshal Tupsamudre, A. Singh, and S. Lodha. "Everything Is in the Name - A URL-Based Approach for Phishing Detection." *International Conference on Cyber Security, Cryptography and Machine Learning*, 2019.

[3] Huaping Yuan, Zhenguo Yang, Xu Chen, Yukun Li, and Wenyin Liu. "URL2Vec: URL Modeling with Character Embeddings for Fast and Accurate Phishing Website Detection." *2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, 2018.

[4] Lee Chang Hoon, Dong Hyun Kim, and J. Lee. "Heuristic-Based Approach for Phishing Site Detection Using URL Features," 2015.

[5] M. Sánchez-Paniagua, Eduardo FIDALGO, Enrique Alegre, Al-Nabki Mhd Wesam, and V. González-Castro. "Phishing URL Detection: A Real-Case Scenario Through Login URLs." *IEEE Access*, 2022.

[6] Mohammed AbuTaha, M. Ababneh, Khaled W. Mahmoud, and Sherenaz W. Al-Haj Baddar. "URL Phishing Detection Using Machine Learning Techniques Based on URLs Lexical Analysis." *International Conference on Information, Communications and Signal Processing*, 2021.

[7] Muna Elsadig, Ashraf Osman Ibrahim, Shakila Basheer, Manal Abdullah Alohali, Sara Alshunaifi, Haya Alqahtani, Nihal Alharbi, and W. Nagmeldin. "Intelligent Deep Machine Learning Cyber Phishing URL Detection Based on BERT Features Extraction." *Electronics*, 2022.

[8] O. K. Sahingoz, Ebubekir Buber, Önder Demir, and B. Diri. "Machine Learning Based Phishing Detection from URLs." *Expert Systems with Applications*, 2019.

[9] Samuel Marchal, J. François, R. State, and T. Engel. "PhishStorm: Detecting Phishing With Streaming Analytics." *IEEE Transactions on Network and Service Management*, 2014.

[10] Sultan Asiri, Yang Xiao, and Tieshan Li. "PhishTransformer: A Novel Approach to Detect Phishing Attacks Using URL Collection and Transformer." *Electronics*, 2023.

[11] Miloud Khaldi, Zohra Alilat, Hana Bendoubba, and Nadir Mahammed. "Hyperparameter Optimization for Malicious URL Detection: Leveraging Optuna and Random Search in Machine Learning and Deep Learning Models." *Informatica*, 2025. https://doi.org/10.31449/inf.v49i27.9106

[12] Malicious URLs Dataset. https://www.kaggle.com/datasets/sid321axn/malicious- urlsdataset.

[13] Guolin Ke et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: Advances in Neural Information Processing Systems 30 2017. https://dl.acm.org/doi/10.5555/3294996.3295074.

[14] Takuya Akiba et al. "Optuna: A Next-generation Hyperparameter Optimization Framework". In: CoRR abs/1907.10902 2019. https://doi.org/10.48550/arXiv.1907.10902

[15] P. He et al., "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," *arXiv preprint arXiv:2006.03654*, 2021.

[16] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.

[17] Q. Yang, Y. Liu, T. Chen, Y. Tong, "Federated Machine Learning: Concept and Applications," *ACM Transactions on Intelligent Systems and Technology*, 2019. https://doi.org/10.1145/3298981

[18] Ming-Yang Su and Kuan-Lin Su. "BERT-Based Approaches to Identifying Malicious URLs". In: Sensors 23 2023, p. 8499. https://doi.org/10.3390/s23208499

[19] U. S. D. R, Anusha Patil, and Mohana Mohana. "Malicious URL Detection and Classification Analysis using

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Machine Learning Models". In: 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT) 2023. https://doi.org/10.1109/IDCIoT56793.2023.10053422.