DETECTING PHISHING ATTACKS DELIVERED THROUGH HTML PAGES AND THE EFFECT OF DATASET SIZE ON ACCURACY

G‘ulomov Sherzod Rajaboyevich
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
Jumaniyozov Firuz Yusupboyevich
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
jumaniyozovfiruz111@gmail.com

Abstract: *This paper describes the HTML-modality component of the PhishGuard-X system, which detects phishing attacks through web-page fragments such as forms, password fields, and hidden iframes. A word-level TF-IDF and Logistic Regression expert reaches 92.17% accuracy and a 92.19% F1-score on 11,430 balanced samples -the lowest result among the system’s five modalities. A cross-check using an auxiliary deep-learning model shows that this lower score stems not from the model architecture but from the comparatively small and homogeneous size of the dataset: the deep model displayed clear overfitting after epoch 3.*

Keywords: *HTML phishing, web-page analysis, TF-IDF, Logistic Regression, overfitting, dataset size, machine learning*

Annotatsiya: *Ushbu maqolada PhishGuard-X tizimining HTML modali, veb-sahifa fragmentlari (forma, parol maydoni, yashirin iframe kabi elementlar) orqali fishing hujumlarini aniqlash komponenti tavsiflanadi. So‘z darajasidagi TF-IDF va Logistik Regressiya asosidagi ekspert 11,430 ta muvozanatlashtirilgan namunada 92.17% aniqlik va 92.19% F1-ko‘rsatkichga erishdi - bu besh modal ichida eng past ko‘rsatkich. Qo‘shimcha chuqur o‘qitish modeli yordamida o‘tkazilgan tekshiruv shuni ko‘rsatdiki, bu past ko‘rsatkich model arxitekturasi emas, balki nisbatan kichik va bir turdagi dataset hajmi bilan bog‘liq: chuqur model 3-epochdan keyin aniq ortiqcha moslashish (overfitting) belgisini namoyish etdi.*

Kalit so‘zlar: *HTML fishing, veb-sahifa tahlili, TF-IDF, Logistik Regressiya, ortiqcha moslashish, dataset hajmi, mashinali o‘qitish*

Аннотация: *В статье описан компонент системы PhishGuard-X, отвечающий за обнаружение фишинговых атак по HTML-модальности - то есть по фрагментам веб-страниц (формы, поля пароля, скрытые iframe-элементы). Эксперт на основе пословного TF-IDF и логистической регрессии достиг точности 92,17% и показателя F1 92,19% на 11 430 сбалансированных образцах -самый низкий результат среди пяти модальностей системы. Проверка с помощью вспомогательной модели глубокого обучения показала, что этот более низкий результат связан не с архитектурой модели, а с относительно небольшим и однородным объёмом датасета: глубокая модель продемонстрировала явные признаки переобучения после 3-й эпохи.*

Ключевые слова: *HTML-фишинг, анализ веб-страниц, TF-IDF, логистическая регрессия, переобучение, объём датасета, машинное обучение*

1. Introduction

Another widespread phishing technique is building fake HTML pages - forms, fake login screens -that trick users into entering sensitive information such as usernames and passwords. The HTML module of PhishGuard-X analyses exactly this kind of fragment and decides whether it carries phishing traits. This paper looks at the module's architecture and its results, and -of particular interest -at why its accuracy trails the system's other experts.

In the literature on HTML-based phishing detection, Opara et al. [1] proposed HTMLPhish, which applies deep-learning techniques to the analysis of HTML structure. Most comparable systems rely on a fully rendered view of the page (a screenshot) or the complete DOM tree, both of which are computationally expensive. PhishGuard-X's HTML module, by contrast, works from the raw HTML text alone -a lighter, faster approach that may, however, miss some visual cues such as logo impersonation.

2. Related Work

A range of approaches has been explored for HTML-based phishing detection. Aljofey et al. [3] combined URL and HTML features using character-level TF-IDF -the same representation adopted here -and reported high accuracy, a result that directly supports our architectural choice; the same group also proposed a separate model built on a character-level convolutional neural network [7]. Yoon et al. [4] combined HTML DOM trees with graph convolutional networks and a transformer architecture in a multi-modal approach.

Odeh et al. [5] proposed a lightweight model based on a multilayer perceptron (MLP), while Belfedhal and Belfedhal [6] proposed a similarly lightweight system relying only on URL features. Most of this work depends on a fully rendered view of the page or the complete DOM tree, both of which add computational overhead; PhishGuard-X's HTML module deliberately uses only the raw HTML text -lighter and faster, but, as this paper shows, more prone to overfitting when the training set is small.

3. Methodology

3.1. Feature extraction. The HTML fragment is vectorised with a word-level TF-IDF representation (n-gram range (1,2), maximum feature count: 80,000).

$$w(t,d) = tf(t,d)[\log\left(\frac{1+n}{1+df(t)}\right) + 1] \quad (1)$$

3.2. Classification. The resulting vector is passed, via the scikit-learn library [2], to an L2-regularised Logistic Regression model:

$$P(y = 1 \mid x) = \frac{1}{1+e^{-(w^T x+b)}} \quad (2)$$

3.3. Rule-based auxiliary features. Alongside the statistical model, the system separately checks for structural markers such as a password-type input field (input type=«password»), a form pointing to an external domain (form action), and hidden (display:none) iframe elements.

3.4. Fusion with the other modalities. Like the URL and Email modules, the HTML module uses a weighted-average fusion strategy: the statistical model's and the rule layer's outputs are combined in proportion to the square of each component's confidence. Unlike the "maximum evidence" strategy used for SMS, this handles disagreement between the two signals more gently, since HTML fragments carry considerably more structural context than SMS messages.

4. Results

The model was evaluated on 11,430 balanced samples (5,715 phishing, 5,715 legitimate) using a stratified 80/20 split. The held-out test set comprised n_test = 2,286 HTML fragments.

Table 1. Dataset composition for the HTML modality

Raw samples	Usable samples	Phishing	Legitimate
11,430	11,430	5,715	5,715

Table 2. HTML expert performance and comparison with classical baselines

Model	Accuracy	Precision	Recall	F1	ROC-AUC

PhishGuard-X (TF-IDF+LR)	92.17%	91.99%	92.39 %	92.19 %	0.9688
Random Forest (200 trees)	86.09%	86.28%	85.83 %	86.05 %	0.9400
XGBoost (200 est.)	90.11%	89.90%	90.38 %	90.14 %	0.9624

Our model outperforms both classical baselines (+6.08 percentage points over Random Forest, +2.06 over XGBoost), yet the lowest accuracy among the system's five modalities is observed precisely for HTML.

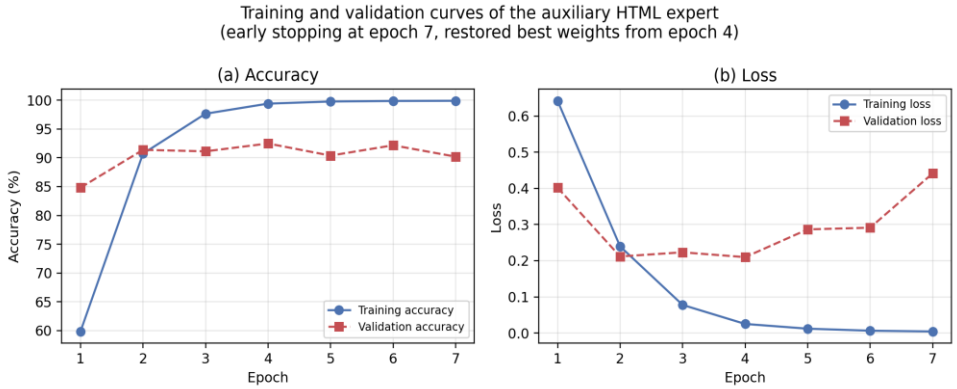

Fig. 1. Training and validation curves, by epoch, of an auxiliary CNN+BiLSTM+Attention deep-learning model (clear overfitting after epoch 4; early stopping at epoch 7)

Fig. 1 provides the clearest evidence for the lower accuracy: the auxiliary deep-learning model's validation accuracy peaks at epoch 4 (92.48%), then falls and oscillates (90.20%–92.17%), while its validation loss rises from 0.2098 at epoch 4 to 0.4415 at epoch 7-nearly doubling, a classic overfitting pattern. Training accuracy, meanwhile, reaches 99.89%. This points to the HTML corpus being considerably smaller and more homogeneous than the other four modalities (URL, Email, SMS, QR), rather than to any weakness in the model architecture itself.

It is worth noting that the model never confidently mislabelled a legitimate HTML page as phishing: even in lower-quality cases, the system tends toward a cautious "suspicious" verdict rather than a confident but wrong "phishing" one. This behaviour is considerably more desirable for a security system, since a false positive erodes user trust, while a false negative risks letting a genuine attack through.

The multi-modal system proposed by Murhej and Nallasivan [8] (built on EM-BERT and SPCA-based EAI-SC-LSTM) tackles the same problem through behavioural analysis, using a more complex, multi-stage architecture. Keyhanipour [9], meanwhile, compares existing web-phishing benchmark datasets through a graph-based analysis and shows substantial inconsistency across popular datasets -a finding that corroborates, in a broader context, the limitation we attribute to our own comparatively small HTML corpus. Taken together, this literature suggests that for HTML-based phishing detection, dataset quality and size may matter more than the model architecture itself -a conclusion that aligns directly with this paper's central finding.

5. Conclusion

Our analysis of the HTML module shows that its comparatively low accuracy (92.17%) stems from dataset size rather than model architecture -a conclusion independently confirmed through a check using a different architecture (a deep neural network). Expanding the HTML corpus is the most direct path to improving this modality's accuracy in future work.

References

1. Opara C., B. Wei, Y. Chen. HTMLPhish: Enabling Phishing Web Page Detection by Applying Deep Learning Techniques on HTML Analysis. -IJCNN'20, IEEE, 2020, pp. 1-8.

2. Pedregosa F. et al. Scikit-Learn: Machine Learning in Python. -Journal of Machine Learning Research, Vol. 12, 2011, pp. 2825-2830.

3. Aljofey A. et al. An Effective Detection Approach for Phishing Websites Using URL and HTML Features. -Scientific Reports, Vol. 12, 2022, 8842.

4. Yoon J. H., S. J. Buu, H. J. Kim. Phishing Webpage Detection via Multi-Modal Integration of HTML DOM Graphs and URL Features Based on Graph Convolutional and Transformer Networks. -Electronics, Vol. 13, 2024, No 16, 3344.

5. Odeh A. et al. Efficient Prediction of Phishing Websites Using Multilayer Perceptron (MLP). -International Journal of Interactive Mobile Technologies, Vol. 14, 2020, No 11.

6. Belfedhal A. S., K. Belfedhal. Lightweight Phishing Detection System Based on Machine Learning and URL Features. -Journal of Web Engineering, 2022.

7. Aljofey A., Q. Jiang, Q. A. Tirra, W. Li, Y. Qu. An Effective Phishing Detection Model Based on Character Level Convolutional Neural Network from URL. -Electronics, Vol. 9, 2020, No 9, 1514.

8. Murhej M., G. Nallasivan. Multimodal Framework for Phishing Attack Detection and Mitigation Through Behavior Analysis Using EM-BERT and SPCA-Based EAI-SC-LSTM. -Frontiers in Communications and Networks, Vol. 6, 2025, 1587654.

9. Keyhanipour A. Graph-Oriented Comparative Analysis of Web Phishing Detection Benchmark Datasets. -International Journal of Computers and Applications, Vol. 48, 2025, No 1, pp. 4-44.