An Explainable AI Study of Phishing Detection Model Degradation Across Real-World Events

Michael Ivanicki
Dept. of Computer Science and Software Engineering
Monmouth University
West Long Branch, NJ, USA
0009-0003-9858-9377

Brian Robert Callahan
Dept. of Computer Science and Software Engineering
Monmouth University
West Long Branch, NJ, USA
0000-0002-1797-8633

***Abstract*—Phishing websites continue to evolve rapidly, driven by technological change and large-scale global events, challenging the long-term reliability of machine learning–based detection systems. While prior research has demonstrated strong classification performance on static datasets, limited work has examined how external real-world factors influence model behavior and feature relevance over time. We investigate the degradation of phishing website detection models and connect observed changes in feature importance to real-world phishing trends and events.**

Using a dataset of 300,000 benign and malicious webpages collected between 2018 and 2020, Random Forest classifiers were trained on year-specific data and evaluated across different temporal test sets. Model behavior was analyzed using global feature importance measures from Random Forests and local explanations generated via LIME. Explainability results were contextualized using real-world phishing statistics.

Results indicate that while core structural features such as the number of links and input fields remain consistently important across years, other features exhibit notable temporal drift. Changes in locally influential features align with periods of increased phishing activity associated with external factors including the expansion of 5G infrastructure and the COVID-19 pandemic. These findings suggest that shifts in attacker behavior during major technological and societal events can directly affect the decision-making patterns of phishing detection models, highlighting the importance of incorporating contextual and temporal analysis when deploying machine learning systems in dynamic and adversarial cybersecurity environments.

***Index Terms*—phishing detection, machine learning, Random Forest, SHAP, LIME**

This is a preprint of the accepted version; the version of record can be found at https://ieeexplore.ieee.org/document/11459104.

Full citation:
M. Ivanicki and B. R. Callahan, "An explainable AI study of phishing detection model degradation across real-world events," in *Proc. 14th Int. Symp. Digit. Forensics Secur. (ISDFS)*, Boston, MA, USA, 2026, pp. 1-5, doi: https://doi.org/10.1109/ISDFS69419.2026.11459104.

I. Introduction

Phishing attacks have become one of the most common cybercrimes, and as a result they are one of the most dangerous. Looking at the Internet Crime Complaint Center's (IC3) "Internet Crime Report" for 2018 through 2020 shows a steep increase in phishing attack numbers. In 2018, phishing attacks were the fifth most popular attack type with 26,379 reported victims [1]. By 2019, that number skyrocketed, boosting phishing to the top attack type with 114,702 reported victims, a 335% increase from 2018 [2]. Subsequently in 2020, the numbers continued to swiftly increase with a reported 241,342 victims for that year, a 110% increase from 2019 [3]. Meta-analysis of phishing trends from industry sources corroborate the explosion of phishing attacks during this three-year span [4].

In order to better understand the reasons behind this immense increase, this paper aims to answer why the number of phishing attacks increased so drastically in such a short period of time. To do so, we explore real world events of the time that give insight to the uptick in attacks, and utilize multiple explainable artificial intelligence methods to see how the attacks themselves have evolved.

II. Background

Previous work in the literature explores both how certain real-world events affected phishing attacks, and how the important features in these attacks have changed throughout the years [5, 6, 7, 8, 9], but none have pieced all of these aspects together to construct a cohesive narrative about the evolution of phishing attacks.

For some articles, the focus is on how the pandemic affected phishing attacks [6, 7]. They discuss statistics on how much phishing attacks increased during that time, and specific ways attackers exploited the situation. We add to this analysis in this paper by using explainable AI to look at what features phishing sites from 2020 were using, and how those differ from phishing sites in 2018 and 2019.

Other works utilize explainable AI methodologies for phishing detection and feature selection, but scarcely touch on how or why the attacks have changed [5, 9, 10]. Others study phishing model degradation but do not use explainable AI or try to tie these changes to real world events [8].

While each individual aspect of this paper has been looked at before, none have combined these ideas to gain a full picture perspective as to the evolution of phishing attacks.

III. Methodology

We used the dataset created by Ejaz et al. [8], a collection of over 300,000 websites spanning 2018 to 2020 and uploaded

to Kaggle [11]. We then extracted the features from each site in the dataset to train the models. Twenty-three features encompassing website content, URL content, and entropy were extracted for each site. The website content features included number of links, scripts, iframes, forms, images, inputs, and meta tags, external ratio, keyword count, and JavaScript event count. Features include the lengths of the title, HTML, URL, hostname, and path. It also included the number of dots, digits, hyphens, and subdirectories. Finally, boolean features, such as if the site was served over https, has an at symbol, or if the URL is the IP address, were included in the dataset for each site. Extraction was accomplished through a Python script that iterated over each site, extracted the defined features, and saved it in CSV format. Our CSVs are available on GitHub at https://github.com/mikeivanicki/AI-Phishing-Websites-CSVs.

After data extraction, we trained three separate AI models using phishing data from each year and tested them on the other years. We used Random Forest, since it balances a strong predictive performance with interpretability and robustness. It is particularly robust to noisy and irrelevant features, and distributional changes over time, which is perfect for studying model degradation and concept drift. Although other training methods such as XGBoost and neural networks may offer improved performance and accuracy over Random Forest, their interpretability is not as strong as what Random Forest offers. Given the nature of this study, a good mix of interpretability and performance is needed, and Random Forest provides just that, making it the best choice to train our models. For the training itself, we split the data into 70% for the training and 30% for the testing. These percentages were chosen because it struck a good balance of having a large enough training set to make a good model, and having a sufficiently large test set to better evaluate temporal degradation. Once trained and tested on all the data, an accuracy graph was created to visualize the efficiency of each model for each year.

To get the full picture of what features were important for each year, we used multiple explainable AI (XAI) methods spanning global and local importance. We utilized Shapley Additive Explanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and Random Forest to evaluate feature importance for each year, and see how they changed over time. Using this set of methods gives us global insight into the overall decision-making process of the model across the entire dataset, and local insight on specific predictions for individual data points.

IV. Results

The results of training the models on all years displayed a strong disparity between the years. As shown in Figs. 1-3, we can see that each model performs significantly worse on the other years, save for the 2020 model on 2019 data. The same trend is also seen in the F1 scores of the models as well. For the 2018 model, the F1 score goes from 0.99 in 2018, to 0.81 in 2019, down to 0.73 in 2020. The 2019 model had an F1 score of 0.99 in the 2019 data, 0.79 in 2018, and 0.89 in 2020. Lastly, the 2020 model had a 0.97 F1 score in 2020 data, 0.96 in 2019, and 0.86 in 2018. This suggests that something has changed over the course of these three years that would cause such significant model degradation.

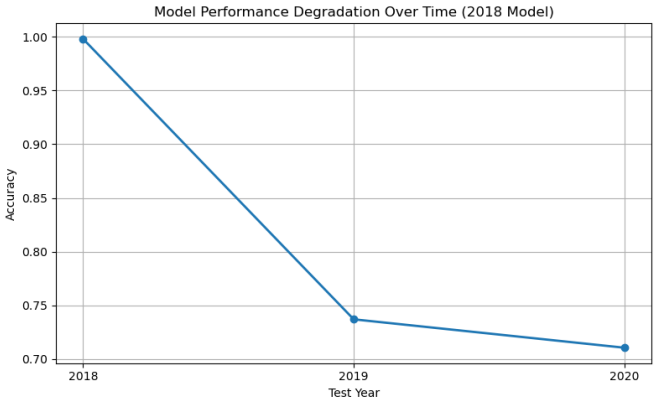

Fig. 1. Model performance degradation over time using the 2018 model.

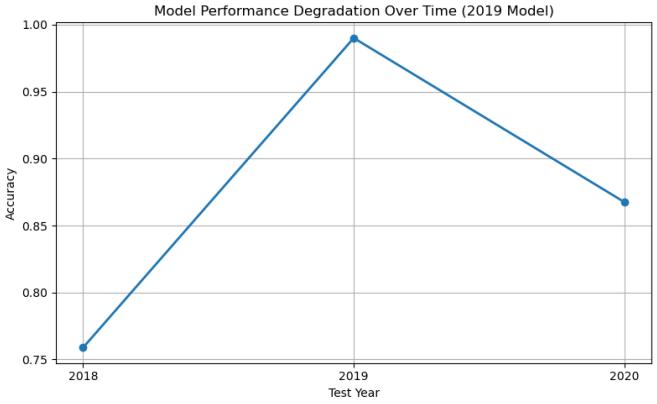

Fig. 2. Model performance degradation over time using the 2019 model.

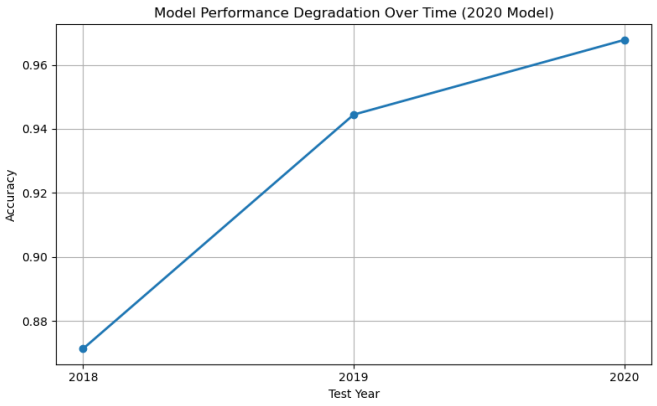

Fig. 3. Model performance degradation over time using the 2020 model.

To get a better understanding of what might have changed between 2018 and 2020, SHAP, LIME, and Random Forest were used for model explainability. Figs. 4-6 show LIME

results. Figs. 7-9 show SHAP results. Finally, Figs. 10-12 show Random Forest results.

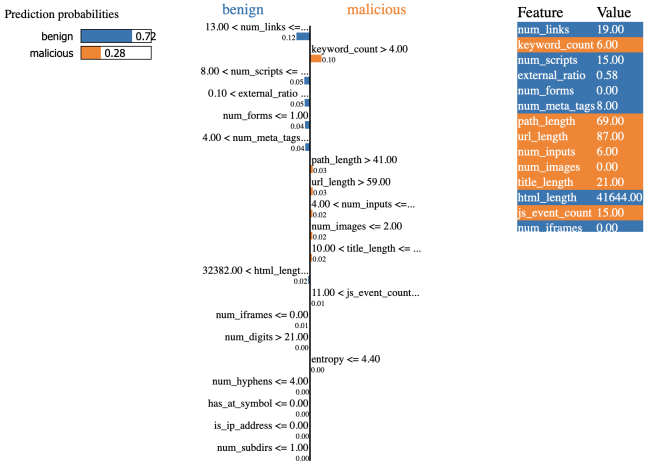

Fig. 4. LIME explanation of a 2018 model prediction on a 2020 example.

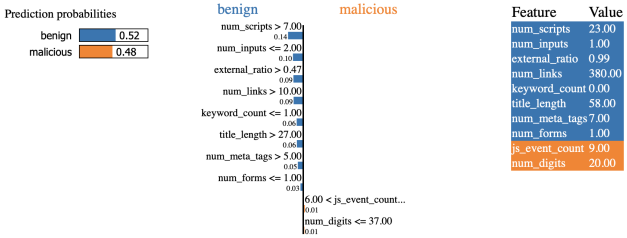

Fig. 5. LIME explanation of a 2019 model prediction on a 2018 example.

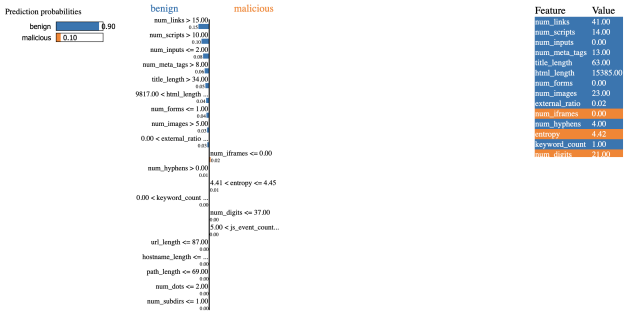

Fig. 6. LIME explanation of a 2020 model prediction on a 2019 example.

Since the results of these three methods did not always agree on which features were most important for each year, they needed to be compared to make a more comprehensive list on which features most or all of them agreed were important. Fig. 13 shows the overall importance of each feature for all three years, deduced from the comparisons of the LIME, SHAP, and Random Forest results.

V. Discussion and Conclusion

Comparing the feature importance results for each year displayed a clear shift, which would explain the performance

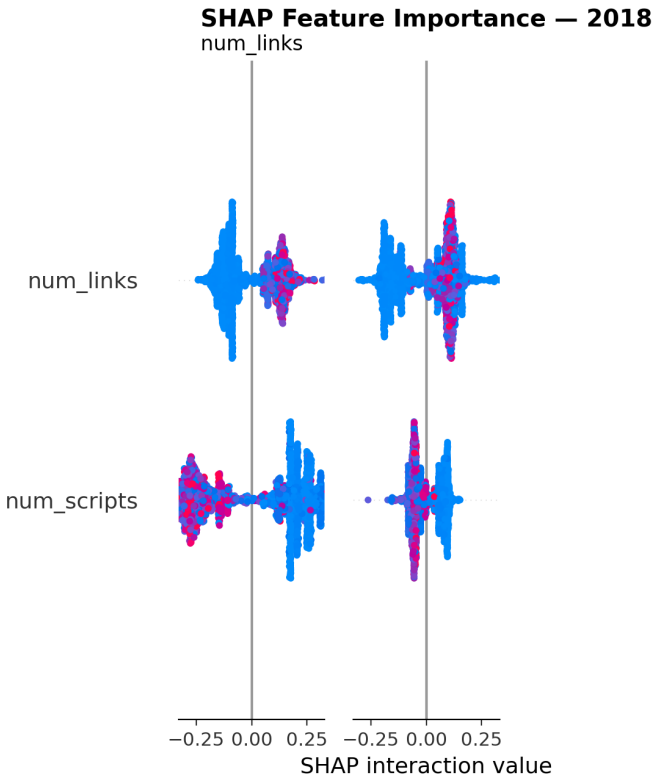

Fig. 7. SHAP feature importance 2018.

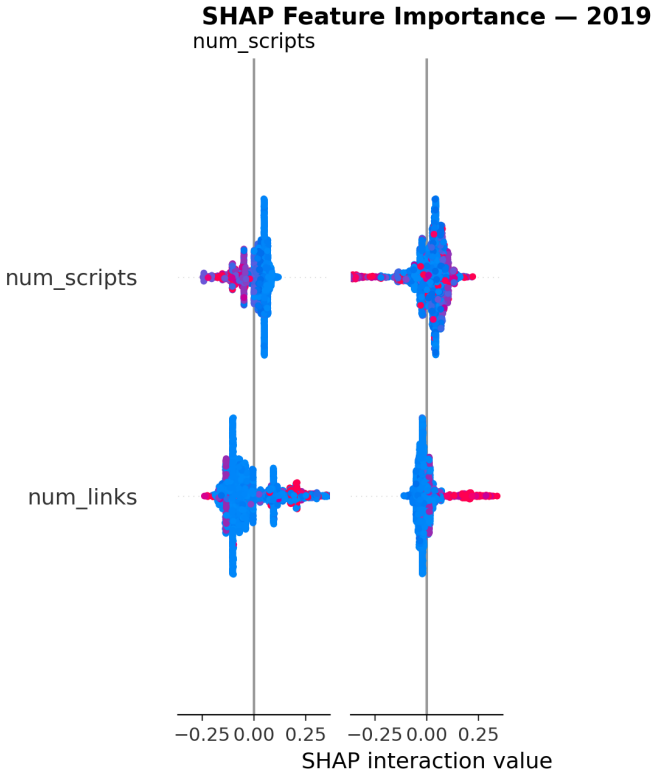

Fig. 8. SHAP feature importance 2019.

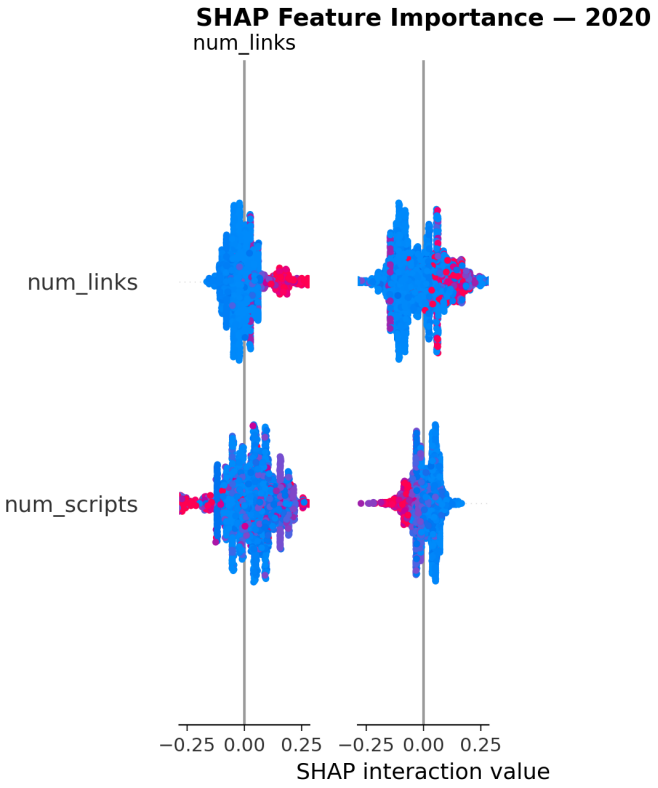

Fig. 9. SHAP feature importance 2020.

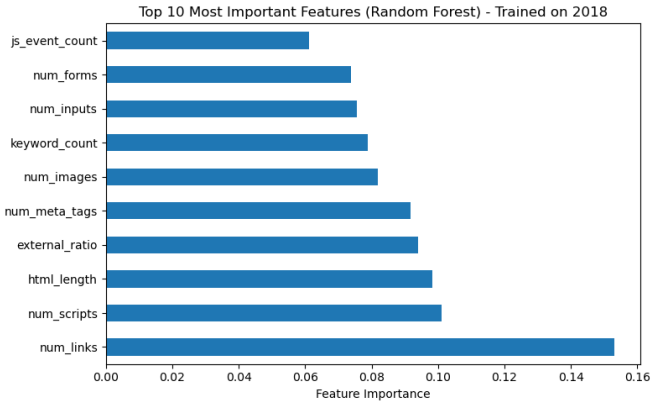

Fig. 10. Top 10 most important features for 2018 using Random Forest.

degradation observed in Figs. 1, 2, and 3. Interestingly, 2020 had seven features that all three explainability methods agreed were important, and is the most accurate predictor for all years. Compared to 2018's four features and 2019's five, it can be inferred that this increase in highly important features explains 2020's increased performance, but also demonstrates the evolution of the attacks.

Another question left unanswered is what was occurring in the world during this time period that fostered this swift evolution of phishing attacks? For 2019, the most likely event that spurred phishing attack growth is the introduction of 5G networks. In the Cybersecurity and Infrastructure Security Agency's (CISA) "Overview of Risks Introduced By 5G Adoption in the United States," they cite 5G infrastructure granting malicious actors more attack vectors as a key risk [12]. Moreover, a study done by Penn State found that attackers could take advantage of certain vulnerabilities in the 5G infrastructure and more easily execute phishing attacks [13].

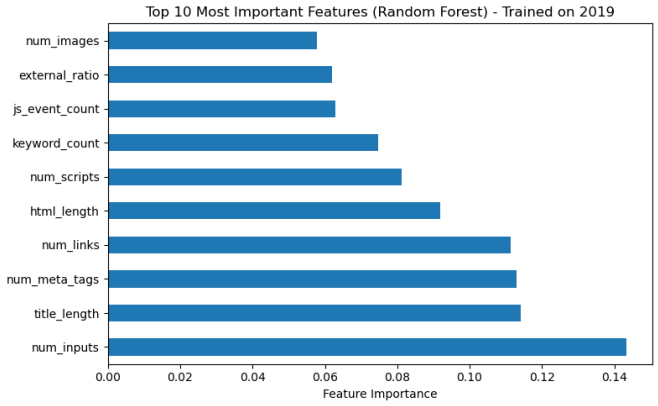

Fig. 11. Top 10 most important features for 2019 using Random Forest.

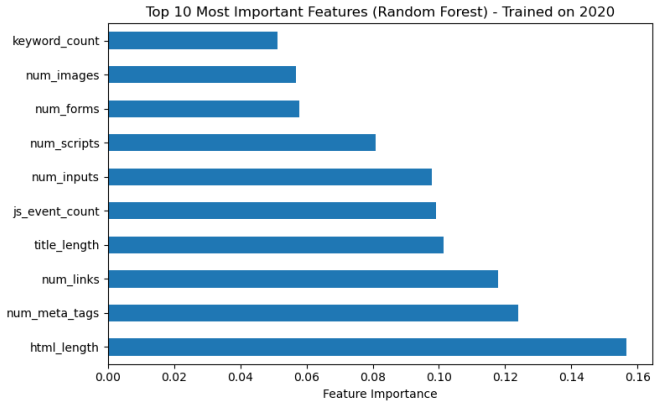

Fig. 12. Top 10 most important features for 2020 using Random Forest.

For 2020, the event that most likely caused the increased phishing attack numbers was the COVID 19 pandemic [6, 14]. In addition to more people being on their devices for longer times than ever before, many attacks impersonated the Centers for Disease Control and Prevention [7]. The pandemic gave attackers a guise to easily exploit people seeking more information during a confusing time.

With all of this information, a narrative emerges that explains not only how phishing attacks evolved from 2018 to 2020, but also why they did as well. Further research will look into the phishing statistics from the years following 2020 to see how trends changed as the world emerged from the pandemic, check for feature importance using data as we approach the current day, and provide context into how real-world events contribute to ever-shifting trends.

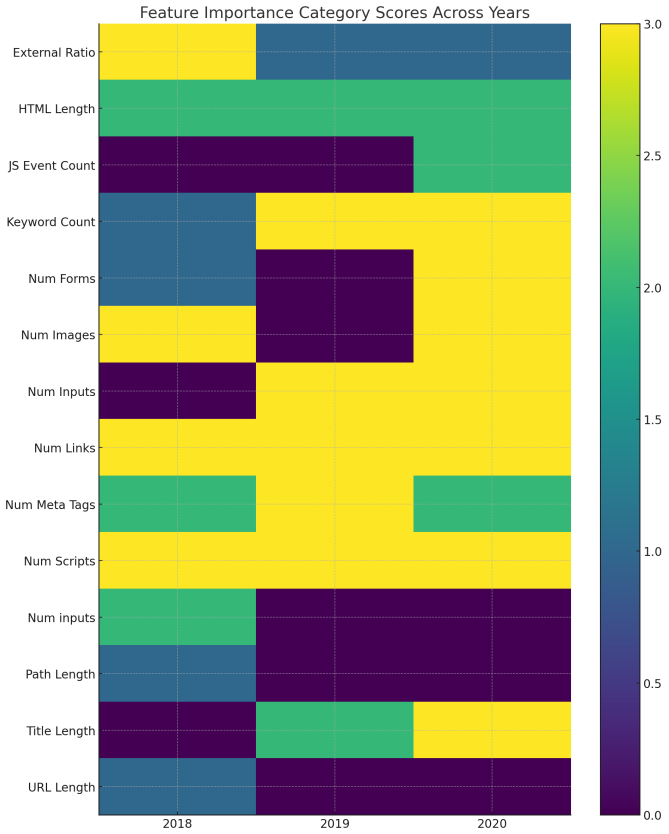

Fig. 13. Feature importance category scores across years.

REFERENCES

[1] FBI, "2018 Internet Crime Report," 2018. [Online]. Available: "https://www.ic3.gov/AnnualReport/Reports/2018_ic3report.pdf.

[2] FBI, "2019 Internet Crime Report," 2019. [Online]. Available: https://www.ic3.gov/AnnualReport/Reports/2019_ic3Report.pdf.

[3] FBI, "2020 Internet Crime Report," 2020. [Online]. Available: https://www.ic3.gov/AnnualReport/Reports/2020_ic3report.pdf

[4] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," Frontiers Comput. Sci., vol. 3, no. 563060, 2021. DOI: https://doi.org/10.3389/fcomp.2021.563060.

[5] H. M. U. Akhtar, M. Nauman, N. Akhtar, M. Hameed, S. Hameed, and M. Z. Tareen, "Mitigating Cyber Threats: Machine Learning and Explainable AI for Phishing Detection," VFAST trans. softw. eng., vol. 13, no. 2, pp. 170–195, Jun. 2025. DOI: https://doi.org/10.21015/vtse.v13i2.2129.

[6] A. F. Al-Qahtani and S. Cresci, "The COVID-19 scamdemic: A survey of phishing attacks and their countermeasures during COVID-19," IET Information Security, vol. 16, no. 5, pp. 324-345, 2022. DOI: https://doi.org/10.1049/ise2.12073.

[7] M. Bitaab, H. Cho, A. Oest, P. Zhang, Z. Sun, R. Pourmohamad, D. Kim, T. Bao, R. Wang, Y. Shoshitaishvili, A. Doupé, and G-J. Ahn, "Scam Pandemic: How Attackers Exploit Public Fear through Phishing," 2020 APWG Symposium on Electronic Crime Research (eCrime), Boston, MA, USA, 2020, pp. 1-10. DOI: https://doi.org/10.1109/eCrime51433.2020.9493260.

[8] A. Ejaz, A. N, Mian, and S. Manzoor, "Life-long phishing attack detection using continual learning," Sci. Rep., vol. 13 no. 11488, 2023. DOI: https://doi.org/10.1038/s41598-023-37552-9.

[9] B. D. Shendkar, P. R. Chandre, S. S. Madachane, N. Kulkarni, and S. Deshmukh, "Enhancing Phishing Attack Detection Using Explainable AI: Trends and Innovations," ASEAN J. Sci. and Technol. Development, vol. 42, no. 1, article 8, 2024, DOI: https://doi.org/10.61931/2224-9028.1604.

[10] S. M. Alasmari, H. Sakly, N. Kraiem, and A. Algaarni, "Phishing detection in IoT: an integrated CNN-LSTM framework with explainable AI and LLM-enhanced analysis." Discov Internet Things, vol. 5, article 102, 2025. DOI: https://doi.org/10.1007/s43926-025-00202-9.

[11] AsifEjaz@ITU, "Phishing_dataset," [Online]. Available: https://www.kaggle.com/datasets/asifejazitu/phishing-dataset/data.

[12] CISA, "Overview of Risks Introduced by 5G Adoption in the United States," 2020, [Online]. Available: https://www.cisa.gov/sites/default/files/publications/19_0731_cisa_5th-generation-mobile-networks-overview_0.pdf.

[13] K. Tu, A. Al Ishtiaq, S. M. M. Rashid, Y. Dong, W. Wang, T. Wu, and S. R. Hussain, "Logic Gone Astray: A Security Analysis Framework for the Control Plane Protocols of 5G Basebands," 33rd USENIX Security Symposium (USENIX Security 24), pp. 3063-3080, 2024.

[14] J. Szurdi, Z. Chen, O. Starov, A. McCabe, and R. Duan, "Studying How Cybercriminals Prey on the COVID-19 Pandemic," Unit42, 2020, [Online]. Available: https://unit42.paloaltonetworks.com/how-cybercriminals-prey-on-the-covid-19-pandemic/.