AI-Powered Phishing and Spam Detection: The Cyber Shield AI Approach

Conference Paper · September 2024

CITATIONS
0

READS
7

5 authors, including:

Naina Nimisha
B. N. M. Institute of Technology
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Siddhant Priyadarshi
B. N. M. Institute of Technology
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

International Journal of Advances in Electronics and Computer Science, ISSN (p): 2394-2835 Volume-11, Issue-9, Sept.-2024
http://iraj.in

AI-POWERED PHISHING AND SPAM DETECTION: THE CYBER SHIELD AI APPROACH

[1]NAINA NIMISHA, [2]SUBRAT PANDEY, [3]SIDDHANT PRIYADARSHI, [4]TEJASWINI R MURGOD, [5]SUSANNA JOHN

[1,2,3,5] Dept. of Artificial Intelligence andMachine Learning, BNM Institute of Technology(Affiliated to VTU) Bangalore, India
[4]Professor in Dept. of Artificial Intelligence and Machine LearningBNM Institute of Technology (Affiliated to VTU) Bangalore, India
E-mail: [1]aiml051@bnmit.in, [2]aiml093@bnmit.in, [3]aim112@bnmit.in, [4]tejaswinirm@bnmit.in, [5]aiml062@bnmit.in

Abstract - The CyberShield AI project is an advanced anti- phishing system that enhances cybersecurity using machine learning classifiers, neural networks, and deep learning. Developed with Python, it features a URL phishing detector and a spam email detector. The URL phishing detector analyzes patterns in URLs, repositories, and IP addresses, employing classifiers like Random Forest and SVM, and leveraging CNN and RNN for high accuracy. The spam email detector uses NLP for content analysis, examining email bodies, attachments, and headers for spam indicators. Both modules store results in MongoDB, and an intuitive Streamlit-based frontend facilitates user interaction, providing actionable insights to mitigate cyberthreats effectively.

Keywords - URL Phishing, Email spam detection, Neural Network, NLP, MongoDB, Streamlit

I. INTRODUCTION

In the contemporary digital landscape, cyber threats have become increasingly innovativ0065, necessitating advanced defences to protect users and organizations. This paper introduces CyberShield AI, a state-of-the-art anti-phishing system designed to combat these evolving threats using machine learning classifiers, neural networks, and deep learning techniques. Developed primarily with Python, CyberShield AI integrates multiple technologies, including the Streamlit library for the frontend, Scrapy for web crawling, and MongoDB for database management.

The system comprises two key modules: a URL phishing detector and a spam email detector. The URL phishing detector evaluates URLs, repositories, and IP addresses, identifying deceptive patterns through feature extraction and web scraping. It employs classifiers such as Random Forest and SVM, combined with CNN and RNN models, to achieve high accuracy in being able to distinguish between legitimate and phishing URLs. The spam email detector utilizes Natural Language Processing (NLP) for comprehensive email analysis, examining content, attachments, and headers for common spam indicators and potential malware.

Both modules store their results in MongoDB, facilitating efficient data management and enabling continuous improvement through iterative refinement. The user-friendly frontend, powered by Streamlit, provides an interactive interface for users to check the degree of risk associated with URLs and emails, thereby empowering them to take informed decisions and effectively mitigate cyber threats. This paper details the design, implementation, and evaluation of CyberShield AI, highlighting its potential to enhance cybersecurity in an increasingly digital world.

II. RELATED WORK

Using a variety of machine learning approaches, several research have been carried out to address the identification of fraudulent URLs and phishing websites. Malak Aljabri and Hanan S. Altamimi's project is noteworthy for its comprehensive approach, employing a spectrum of models including Linear Regression, SVM, Naïve Bayes, Random Forests, and DNN. Their research achieved a remarkable accuracy of 98.82%, highlighting the efficacy of the models in detecting spam, malware, and phishing URLs. However, they pointed out the significant challenge posed by the scarcity of available datasets, which limits further advancements in this domain[1].

Likewise, other research has concentrated on identifying phishing websites through the examination of their shared characteristics. One such project used Random Forest and Decision Tree classifiers to achieve a 97.73% accuracy. This study emphasized the difficulties in interpreting complex models, managing imbalanced datasets, and ensuring the scalability and robustness of detection systems in real-world scenarios[2]. Another notable work in this field employed cross-validation, boosting, and stacking techniques with models such as Decision Trees, SVM, Random Forests, and ANN to detect phishing domains, achieving an accuracy of 97.4%. Despite its success, the study heavily relied on user observation, which was identified as a significant limitation[4].

For the purpose of phishing detection, deep learning algorithms have also been thoroughly investigated. For instance, a project leveraging NLP techniques like BERT, LSTM, CNN, and DNN achieved a 95.47% accuracy indetecting social semantic attacks, although scalability remained a concern[5]. Another project used a hybrid LSD with Canopy feature selection, achieving a 98.12% accuracy with models like Gradient Boosted Trees, Naïve Bayes, and SVM. This study noted underfitting as a performance issue[6].

Furthermore, a highly accurate intelligent phishing detection system using logistic regression, neural networks, locally- deep SVM, boosted decision trees, and decision forests was able to detect phishing attempts even though it recognized theneed for better feature selection strategies in order to combatphishing tactics that would change over time. [8].

Further research includes a study that implemented an Expandable Random Gradient Stacked Voting Classifier (ERG-SVC) for filtering phishing URLs, extracting features like pagerank, IP, redirecting, domain age, and URL length, and achieving an accuracy of 98.118%. This approach, however, relied on third-party attributes for URL feature extraction and presented a complex architecture with potential high cloud deployment costs[9]. Another project employed methods like Gradient Boosted Trees, SMOTE, ADASYN, MWMOTE, and ROSE for fraud detection, achieving a 96.2% accuracy rate but noted challenges related to scalability and real-world implementation hurdled[10]. Together, these studies show how cutting-edge machine learning and deep learning methods might improve cybersecurity while also emphasizing the issues that must be resolved before they canbe applied in the real world.

III. DATASET

The dataset for URL analysis in cybersecurity encompasses a wide range of features critical for evaluating the safety of web links. It includes URL characteristics such as length, the presence of IP addresses or '@' symbols, and the number of subdomains, all of which can indicate potential malicious intent. Domain-related attributes like domain age and expiration provide insights into the legitimacy of websites. Content analysis features, such as title and content size, are useful for detecting spam or phishing attempts. Interaction flags like 'Mouse_Over' and 'Right_Click' can identify suspicious scripts or deceptive web elements. Security indicators, including the presence of HTTPS and 'Tiny_URL,' signal secure or potentially obfuscated links. Behavioral metrics, such as 'Num_Third_Party_Clicks' and 'Num_Popups,' help assess user interactions and exposure to threats. The dataset's final evaluation metrics, 'Final_Val' and 'Result,' determine the safety of URLs, making it required for developing predictive models to protect users from online threats. Additionally, the spam mail detection dataset comprises emails labeled by category (e.g., "spam" or "not spam") and the textual content of the emails, which is crucial for tasks like spam filtering, punctuation, content, header and attachment features for email classification. These all- inclusive characteristics support strong algorithms that can identify emails and reliably forecast URL safety, improving overall cybersecurity measures.

IV. PROPOSED METHODOLGY

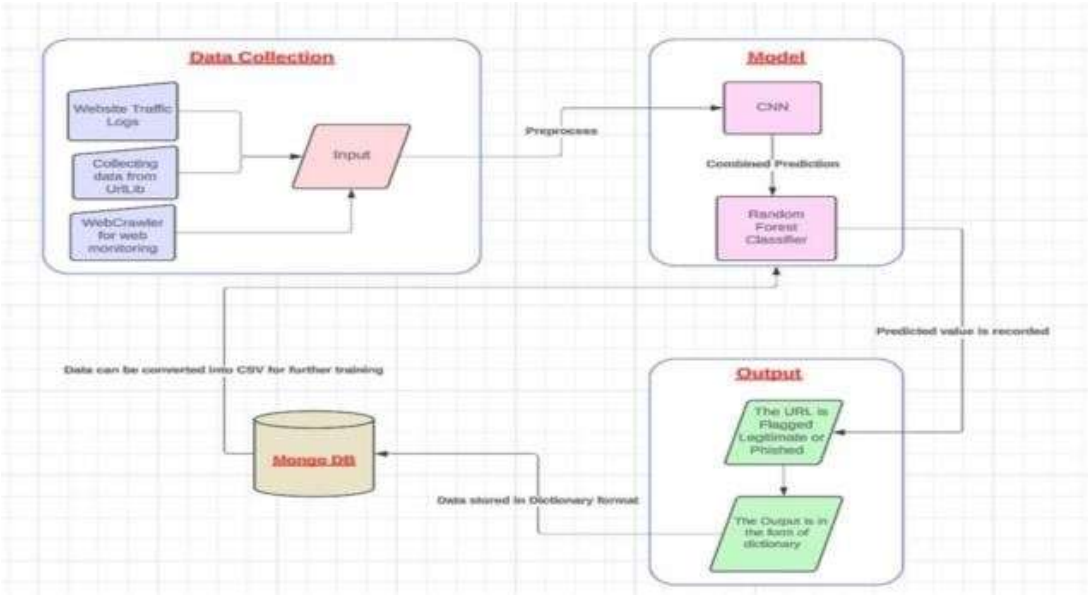

Fig. 1. Architecture of proposed solution

A. Phishing Detector

1) **Feature Extractor:** This module is essential to data preprocessing since it provides the framework for feature extraction from email content and URLs. It carefully breaks down URLs, noting important details such as length, redirection count, and domain name. It also looks into email bodies to identify urgency cues, mood, and frequently used spam phrases. Rich, context-aware data that is necessary for strong classification and analysis is provided by these extracted features, enabling later modules to detect threats effectively.

2) **Model:** This module contains a variety of machine learning algorithms that have been carefully designed for the purpose of detecting cyber threats. The accuracy of the model was improved by using RandomForest stacked on convolutional neural network (CNN). After undergoing thorough training on a variety of datasets and neural network topologies, these models demonstrate superior performance in differentiating between reputable and harmful entities, hence providing effective protection for consumers against phishingand spam attacks.

3) **Scrapping:** This module takes a proactive approach to cybersecurity by using the Scrapy library to scan websites and navigate the digital environment. It is the preferred webcrawler since it carefully examines webpages and extracts insightful data that helps with URL classification and analysis. It improves the system's comprehension of contextual subtleties by breaking down entire websites, which strengthens defenses against phishing attempts and improvesthreat assessment accuracy.

B. Email Spam Detector

1) **Feature Extractor:** This module is a fundamental component of data preparation, making it easier to extract features from email text that are important for identifying cyber threats. It carefully examines email contents, collecting vital information including urgency cues, emotion, and frequently used spam keywords. In order to identify potentialdangers, it also carefully examines email headers, attachment types, and sender information. By giving the system the extensive, context-aware data required for reliable classification and analysis, these extracted features enable later modules to effectively detect and mitigate spam.

2) **Model:** This module contains a plethora of machine learning algorithms that have been painstakingly developed for threat detection particular to emails. This module containsa wide variety of classifiers, ranging from conventional classifiers like Support Vector Machines and Random Forests to sophisticated neural network architectures like Convolutional Neural Networks (CNN) and RecurrentNeural Networks (RNN). These models have been refined through the use of diverse neural network designs and large datasets. As a result, they are able to discern between spam and legitimate emails with greater accuracy, thereby protecting consumers from fraudulent cyberattacks.

3) **Tokenizer:** The module that is responsible for separating textual data from email content into structured units that can be analyzed is at the center of natural language processing(NLP). It divides unprocessed text into meaningful chunks using tokenization techniques, making sentiment analysis, keyword extraction, and other NLP-based activities possible. Tokenizer.py improves the system's comprehension of language subtleties, which enhances threat detection capabilities and strengthens defenses against malicious emailcommunications.

C. Database

After that, the emails and URLs provided for verification are preserved in a MongoDB database for later use in trainingand testing.

V. RESULTS AND DISCUSSION

Figure 2 illustrates how the dataset was trained using a variety of classifiers, including SVM, RandomForest, and neural networks, in order to determine which provided the best accuracy. As can be seen in the confusion matrix in Figure 3, the CNN stacked with RandomForest model that produced the best accuracy of about 87% was selected for additional testing and validation. The training and validation loss and accuracy across the dataset are displayed as epochs are increased in the graphs on Figure 4 and Figure 5.

MODEL	Accuracy
SVM	0.8266
Naive Bayes	0.75
Random Forest	0.857
Simple ANN	0.841
Simple RNN	0.811
Simple CNN	0.848
Simple FNN	0.848
Simple RCNN	0.512
ANN with SVM	0.841
RNN with SVM	0.814
FNN with SVM	0.855
CNN with SVM	0.836
ANN with Random Forest	0.851
RNN with Random Forest	0.866
FNN with Random Forest	0.83
CNN with Random Forest	0.869

Fig. 2. Training of different classifiers and networks

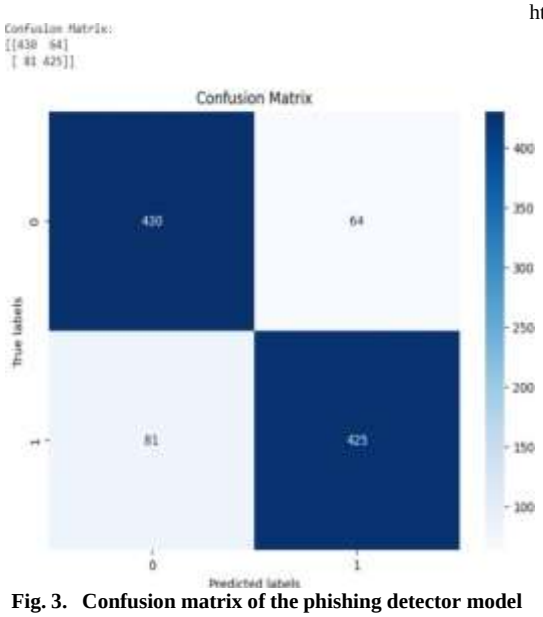

Fig. 3. Confusion matrix of the phishing detector model

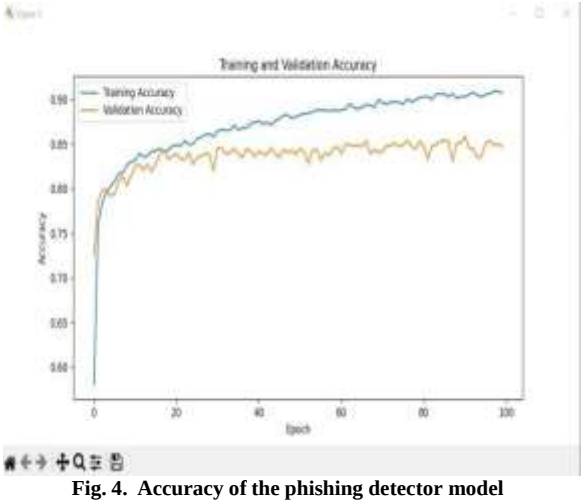

Fig. 4. Accuracy of the phishing detector model

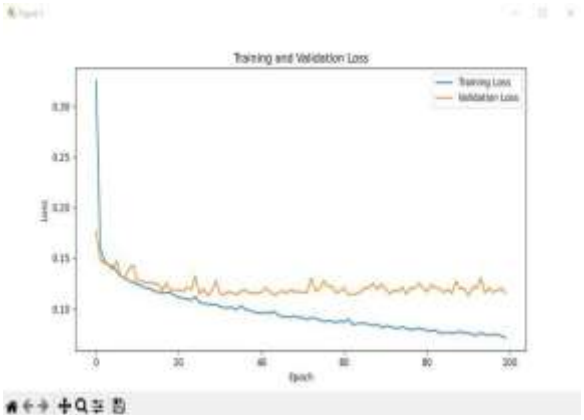

Fig. 5. Loss of the phishing detector model

Similarly, the accuracy and confusion matrix was computed for the email spam detection module, as indicated in figure 6, to assess the model's performance in comparison to other available solutions. The module has a 95% accuracy rate. The validation and training scores for the entire dataset, together with the accuracy as the size grows, are depicted in the two graphs in figures 7 and 8.

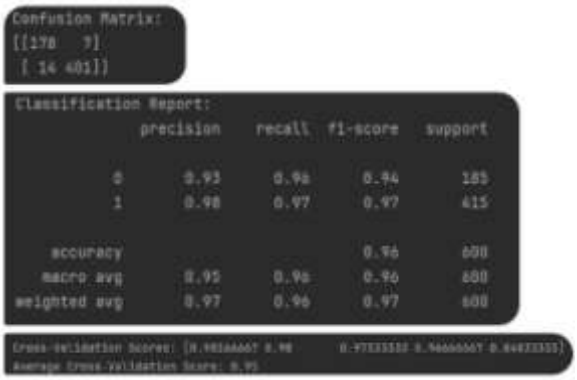

Fig. 6. Report on email spam detection module

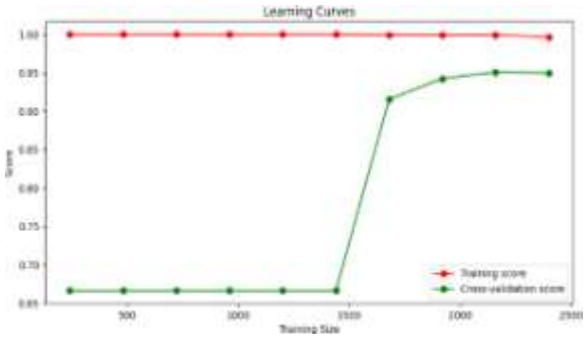

Fig. 7. Score of email spam detection module

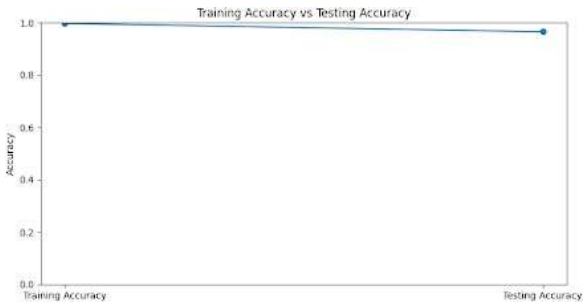

Fig. 8. Accuracy fo email spam detection module

VI. CONCLUSION

In conclusion, CyberShield AI signifies a notable leap forward in cybersecurity by leveraging machine learning and deep learning methodologies to combat phishing and spam attacks. The URL phishing detector module adeptly employs advanced feature extraction and classification techniques, achieving remarkable accuracy in discerning between legitimate and malicious URLs. Similarly, the spam email detector module utilizes Natural Language Processing (NLP) for feature extraction and analysis, effectively identifying spam emails through meticulous scrutiny of content and attachments. Integral to the project's success is its seamless integration with MongoDB, ensuring efficient data storage and retrieval. The user-friendly frontend interface empowersusers to navigate tasks effortlessly, providing actionable insights based on risk assessment and exemplifying the potential of AI in safeguarding digital ecosystems against evolving threats. Looking ahead, CyberShield AI could be enhanced by implementing real-time threat

detection capabilities, employing advanced feature extraction techniques such as semantic analysis and deep learning-based embeddings, integrating with external threat intelligence feeds, and promoting user education and awareness initiatives. Additionally, optimizing scalability and performance through distributed computing and cloud-based infrastructure will ensure the system's readiness to handle growing data volumesand user demand without compromising performance.

ACKNOWLEDGMENT

We would like to express our heartfelt gratitude to all the people and organizations whose assistance made this research endeavor possible. Their tremendous assistance and contributions made a major difference in helping us achieve our goals. Our deepest gratitude goes to our advisors and mentors for their unwavering guidance, support, and invaluable insights throughout the entire research process. We also thank the participants who helped us build and assess the proposed approach by graciously offering their data and feedback. We also acknowledge researchers and developers that made contributions to the creation of the libraries, tools, and datasets used in this investigation. Finally, we would like to express our sincere gratitude to our friends and acquaintances for their understanding, inspiration, and unwavering support, all of which were crucial to the successful completion of this research project.

REFERENCE

[1] M. Aljabri et al., "Detecting Malicious URLs Using Machine Learning Techniques: Review and Research Directions," in IEEE Access, vol. 10, pp. 121395-121417, 2022, doi: 10.1109/ACCESS.2022.3222307.

[2] A. Chawla, “Phishing website analysis and detection using Machine Learning”, Int J Intell Syst Appl Eng, vol. 10, no. 1, pp. 10–16, Mar. 2022.

[3] Z. Azam, M. M. Islam and M. N. Huda, "Comparative Analysis of Intrusion Detection Systems and Machine Learning-Based Model Analysis Through Decision Tree," in IEEE Access, vol. 11, pp. 80348- 80391, 2023, doi: 10.1109/ACCESS.2023.3296444.

[4] Shouq Alnemari and Majid Alshammari, "Detecting Phishing Domains Using Machine Learning" in Appl. Sci., 2023, 13(8), 4649, doi: 10.3390/app13084649.

[5] M. Almousa and M. Anwar, "A URL-Based Social Semantic Attacks Detection With Character-Aware Language Model," in IEEE Access, vol. 11, pp. 10654-10663, 2023, doi: 10.1109/ACCESS.2023.3241121.

[6] A.A. Orunsolu, A.S. Sodiya, A.T. Akinwale, A predictive model for phishing detection, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 2, 2022, Pages 232-247, ISSN 1319-1578.

[7] M. N. Ashtiani and B. Raahemi, "Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review," in IEEE Access, vol. 10, pp. 72504-72525, 2022, doi: 10.1109/ACCESS.2021.3096799.

[8] Mughaid, A., AlZu’bi, S., Hnaif, A. et al. An intelligent cyber security phishing detection system using deep learning techniques. Cluster Comput 25, 3819–3828 (2022), doi: 10.1007/s10586-022-03604-4.

[9] P.D.P.L. Indrasiri, Malka N. Halgamuge, Azeem Mohammad, "Robust Ensemble Machine Learning Model for Filtering Phishing URLs: Expandable Random Gradient Stacked Voting Classifier (ERG-SVC)" in IEEE Access, vol. 9, pp. 150142-150161, doi:10.1109/ACCESS.2021.3124628.

[10] Sandeep Rangineni, Divya Marupaka "Analysis of Data Engineering for Fraud Detection Using Machine Learning and Artificial Intelligence Technologies" in IEEE Access, pp. 2582-5208, doi: 10.56726/IRJMETS43408.

[11] Tanimu, Jibrilla & Shiaeles, Stavros. (2022). Phishing Detection Using MachineLearningAlgorithm.317-322. 10.1109/CSR54599.2022.9850316.

[12] L. Tang and Q. H. Mahmoud, "A Deep Learning-Based Framework forPhishing Website Detection," in IEEE Access, vol. 10, pp. 1509-1521, 2022, doi: 10.1109/ACCESS.2021.3137636.

[13] K. Mridha, J. Hasan, S. D and A. Ghosh, "Phishing URL Classification Analysis Using ANN Algorithm," 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), Kuala Lumpur, Malaysia, 2021, pp. 1-7, doi: 10.1109/GUCON50781.2021.9573797.

[14] Vahid Shahrivari, Mohammad Mahdi Darabi and Mohammad Izadi, “ Phishing Detection Using Machine Learning Techniques”, CoRR - 2020, vol. abs/2009.11116, doi: 10.48550/arXiv.2009.11116.

[15] Mughaid, A., AlZu’bi, S., Hnaif, A. et al. An intelligent cyber security phishing detection system using deep learning techniques. Cluster Comput 25, 3819–3828 (2022), https://doi.org/10.1007/s10586-022- 03604-4.

[16] Sahingoz, Ozgur & Buber, Ebubekir & Demir, Onder & Diri, Banu. (2019). Machine learning based phishing detection from URLs. ExpertSystems with Applications. 117. 345-357.

[17] Vahid Shahrivari , Mohammad Mahdi Darabi and Mohammad Izadi, “ Phishing Detection Using Machine Learning Techniques”, CoRR - 2020, vol. abs/2009.11116, doi: 10.48550/arXiv.2009.11116.

[18] D, Naresh. (2020). Detection of Phishing Websites using an Efficient Machine Learning Framework. International Journal of Engineering Research and. V9. 10.17577/IJERTV9IS050888.

[19] I. Saha, D. Sarma, R. J. Chakma, M. N. Alam, A. Sultana and S. Hossain, "Phishing Attacks Detection using Deep Learning Approach," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 1180-1185, doi: 10.1109/ICSSIT48917.2020.9214132.

[20] M. N. Alam, D. Sarma, F. F. Lima, I. Saha, R. -E. -. Ulfath and S. Hossain, "Phishing Attacks Detection using Machine Learning Approach," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 1173- 1179, doi: 10.1109/ICSSIT48917.2020.9214225.

[21] S. Anwar et al., "Countering Malicious URLs in Internet of Things Using a Knowledge-Based Approach and a Simulated Expert," in IEEE Internet of Things Journal, vol. 7, no. 5, pp. 4497-4504, May 2020, doi: 10.1109/JIOT.2019.2954919.

★★★