[1, *]Dr. G. B. Sambare

[1]Shailesh B. Galande

[1]Sharvari Kale

[1]Pragati Nehete

[1]Vaishnavi Jadhav

[1]Nihal Tadavi

Towards Enhanced Security: An improved approach to Phishing Email Detection

Abstract: - Email phishing assaults remain a widespread catastrophe to individuals and businesses, using human weaknesses to obtain unauthorised access to sensitive information. This research study presents an upgraded email phishing detection system that uses machine learning approaches. To construct a robust and adaptable model, the system includes a complete feature set such as email content analysis, sender reputation, and behavioural patterns. The one that has been proposed uses a new combination of supervised and unsupervised machine learning algorithms to analyse email properties and user behaviour, allowing for the detection of subtle phishing signs. The term of body content analysis, URL analysis, and QR code information are among the features retrieved and analysed utilising advanced natural language processing and pattern recognition algorithms. Novelty lies in the integration of supervised and unsupervised machine learning algorithms to identify subtle phishing indicators, along with advanced natural language processing and pattern recognition algorithms for analyzing email properties. A benchmark dataset is used for the proposed system's validation, and a comparison with other phishing detection techniques shows improved, lower false-positive rates. The results indicate the system's ability to effectively discern phishing emails while minimizing the impact on legitimate communication. The system demonstrates a notable improvement by including a diverse range of features and state-of-the-art machine learning algorithms, which makes it a valuable asset to the cybersecurity toolkit for safeguarding email correspondence.

Keywords: Phishing attack, cybersecurity, Machine Learning, Malicious QR code, Email Phishing, URL phishing, Deep Learning.

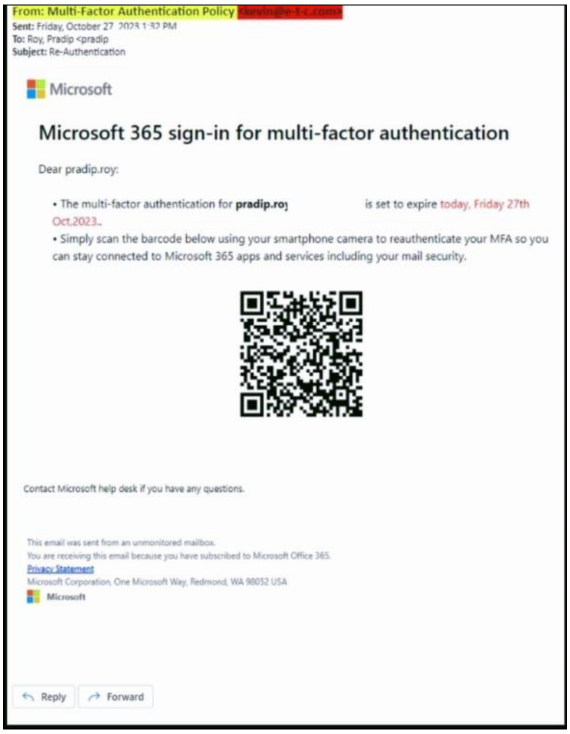
I. INTRODUCTION

The digital era has seen a rise in the prevalence and continual evolution of phishing assaults, which impact both people and organisations. These false emails, which frequently impersonate trustworthy agencies, attempt to trick recipients into disclosing critical information or clicking on dangerous links, making them an effective weapon for cybercriminals. The enormity and sophistication of these assaults highlight the critical need for more effective and comprehensive phishing email detection methods. Statistics show the startling scale of the phishing problem. In 2020, the Anti-Phishing Working Group (APWG) recorded over 266,000 distinct phishing websites, a 46% increase over the previous year. In addition, phishing attempts were identified as the primary cause in 96% of cyber espionage cases and 95% of cyber-espionage-related data breaches in Verizon's 2021 Data Breach Investigations. Emails and chat apps are the most common communication routes used by phishing attacks. Because email attacks are more difficult to detect than other tactics, phishers prefer them [1]; as a result, this study focuses on [2] [3].

Phishing attacks start with emails addressed to online customers. The email contains a false link that sends users to a cloned website that seems identical to the original. This convinces the email recipient that the email and webpage are legitimate. Figure 1 displays an email used for phishing and its main components. This material, obtained from an institution at the University of Massachusetts Amherst, demonstrates how to safeguard internet users from deception. The primary goal of this work is to conduct a comprehensive evaluation of phishing email detection studies that use NLP approaches. The examined 100 research publications published from 2006 to 2022 using predetermined criteria. This research is focused on essential elements of phishing email detection, including NLP, ML algorithms, text characteristics, datasets, and assessment criteria. The survey found no comprehensive review of NLP approaches for detecting phishing emails.

[1] Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India.

*Corresponding author e-mail: santosh.sambare@pccoepune.org



From: Multi-Factor Authentication Policy <kevin@e-1-c.com>
Sent: Friday, October 27, 2023 1:32 PM
To: Roy, Pradip <pradip
Subject: Re-Authentication

Microsoft

Microsoft 365 sign-in for multi-factor authentication

Dear pradip.roy:

- The multi-factor authentication for **pradip.ro** is set to expire today, Friday 27th Oct.2023..
- Simply scan the barcode below using your smartphone camera to reauthenticate your MFA so you can stay connected to Microsoft 365 apps and services including your mail security.

Contact Microsoft help desk if you have any questions.

This email was sent from an unmonitored mailbox.
You are receiving this email because you have subscribed to Microsoft Office 365.
Privacy Statement
Microsoft Corporation, One Microsoft Way, Redmond, WA 98052 USA
Microsoft

Figure 1. Original Phishing Email

The four tiers of the suggested approach are specifically designed to examine different parts of an email for indications of phishing, Figure 2 The first step, titled "Account Verification," is intended to mitigate the danger caused by fake accounts. It attempts to verify the sender's identity by methodically analyzing numerous aspects. In the second stage, "Body and Subject Text Scanning," the textual content of the email's body and topic is scanned using an ensemble model with hybrid feature selection. The third level, "Malicious QR Code Detection," addresses the rising problem of malicious QR codes in email. As QR codes grow increasingly prevalent in numerous facets of daily life, fraudsters have grabbed the chance to conduct assaults. This level is focused on reading and analyzing QR codes to detect possible dangers. The fourth level, "Embedded Link and Image URL Analysis," analyses embedded links, URLs, and image URLs in the email body. The use of this strategy is intended to grow in response to the ever-changing nature of phishing threats, providing individuals and organisations with a strong defence against these harmful attacks.

II. LITERATURE REVIEW

The Recently, a study on research associated with phishing URL detection was carried out [4]. This survey focuses on the features of several ML as well as phishing URL detection strategies, such as batch, online, and representation. Furthermore, [5]-[6] examined several works on phishing URL detection, while [7] and [8] discussed the literature on the area of phishing URL detection along with significant concerns. Mohammad et al. [8] proposed a unique multidimensional technique for identifying phishing attempts in their survey, categorising activities into five categories: machine learning, text mining, people using the service, profile matching, and others. Additionally, the researchers suggested classifying the last group as client-server authentication, honeypot, search engines, and ontology defenses.

[10] In the realm of phishing email detection, HELPHED is introduced, emphasizing the fusion of Ensemble Learning methods with hybrid features for enhanced accuracy. These hybrid features, synthesizing email content and textual traits, offer a precise representation of emails. Two HELPHED methods are proposed: one employing Stacking Ensemble Learning and the other utilizing Soft Voting Ensemble Learning. Both methods leverage distinct Machine Learning algorithms like decision tree and KNN to handle hybrid features concurrently, reducing complexity and elevating model performance. Numerical results with F1-score of 0.9942 is given by model.

[11] A Machine Learning model designed for phishing attack detection within user mailboxes utilizes content-based methods to scrutinize email content for spam identification. The project's objective involves the exploration of machine learning algorithms, with Naïve Bayes chosen for training and Logistic Regression for detection. The conventional ML modelling cycle phases are employed to develop the model, yielding a remarkable result. Naive Bayes and Logistic Regression together achieve an accuracy of approximately 99% in predicting phishing attacks. [12] The proposed advanced model, abbreviated as THEMIS (Recurrence-Convolution-Neural-Networks), integrates attention mechanisms and multilevel vectors. Experimental results reveal an impressive 99.848% accuracy, with a minimal false positive rate of 0.043%. THEMIS demonstrates superior efficacy in identifying phishing emails compared to existing approaches. TensorFlow and Keras were utilized for implementation. The training-validation dataset comprised 5,447 legitimate, 699 phishing, and 6,146 total emails, while the testing set included 2,334 legitimate, 300 phishing, and 2,634 total emails. In THEMIS, the use of RCNN with LSTM addresses long-term dependency issues, where LSTM (Long Short-Term Memory) employs gates at input to manage outputs over various time frames. The attention mechanism, akin to human cognition, selectively focuses on relevant elements for ongoing tasks.

[13] QR codes, while offering various benefits, pose significant security risks. Intruders exploit them to target smartphones, compromising user privacy and accessing sensitive data like login credentials, contacts, photos, videos, and banking information. Such attacks can grant attackers control over mobile devices, enabling unauthorized use of microphone, camera, GPS, and potential involvement in botnet or DDOS attacks. Examples of QR-based threats include phishing, fraud, malware propagation, command injection, and SQL injection attacks. Notably, QR codes lack human readability and can only be interpreted by specific scanning devices. The first documented malicious use occurred in September 2011, involving a QR code directing users to a webpage for the stealthy download of malicious files. [14] In this study, a comprehensive examination of QR code threats in real-world scenarios is conducted. A specialized web crawler is designed to assess the prevalence of QR codes, extract URLs, and discern malicious codes by cross-referencing with blacklists. The research identifies five primary attack strategies, encompassing malware delivery for both Windows and mobile platforms, phishing redirections, exploitation of intermediate sites with known vulnerabilities housing malicious scripts, and targeting vulnerable applications through exploit sites. Throughout the analysis, it becomes evident that attackers consistently exploit QR codes to engage users and capitalize on trust in well-known brands for nefarious purposes. In instances of phishing, QR codes guide victims to counterfeit websites, prompting the input of sensitive information. Noteworthy observations include the utilization of fake business websites, particularly those mimicking Google Play, to distribute malicious Android apps via QR codes. This strategic approach aims to target less sophisticated users by leveraging QR codes to obscure deceptive URLs.

III. PROPOSED METHODOLOGY

Finding phishing emails is a crucial cybersecurity task. Because attackers craft sophisticated Email body content, QR code, URLs, and phishing QR codes and URLs frequently appear to users as authentic. Attackers may then utilize this access to obtain users' personal information for their own purposes. Various challenges have been faced by the existing system that is Dynamic content loading, phishing emails frequently use dynamic content loading techniques, necessitating real-time execution and analysis to detect fraudulent behaviour; URL Obfuscation Techniques, phishers use various obfuscation techniques such as URL encoding, URL shortening, and homograph attacks to make malicious URLs appear legitimate, challenging traditional pattern matching; Redirect Chains, phishers often employ redirect chains to disguise the final destination URL, complicating the analysis process as the malicious content may not be immediately apparent. To properly adopt this approach, organizations and individuals should consider its subsequent stages. The figure 2 shows the flowchart of our proposed model. To begin, monitor and analyze login activity at the email account level for strange patterns or unauthorized access. Second, use powerful machine learning as well as natural language processing algorithms to look for phishing signs in email subject lines and content, such as suspicious terminology and demands for personal information. Third, include QR code and URL scanning technologies to validate the legitimacy of links within emails, and use machine learning to evaluate the legitimacy of URLs based on a variety of parameters. By combining the findings of these several layers of research, a full phishing detection system may be developed, considerably improving email security. It is critical to sensitize email users about the dangers of phishing and urge them to report questionable communications. Organizations and individuals may dramatically improve their capacity to detect and prevent phishing attempts, protect sensitive information, and improve email communication security by using this complete methodology.

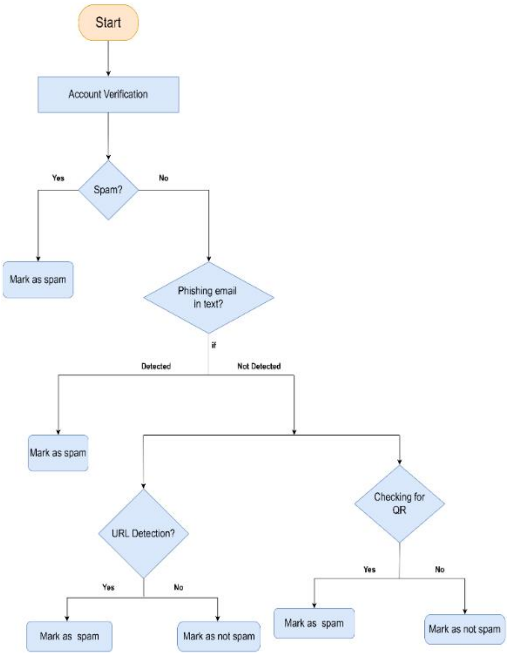


Figure 2. Flowchart of Methodology

A. Pseudo code of methodology

```
START
// Account verification
IF account_verified THEN
  // Check for spam
  IF spam_detected THEN
    Mark as spam
  ELSE
    // Check for phishing text
    IF phishing_text_detected THEN
      Mark as spam
    ELSE
      // Check for QR code
      IF qr_code_present THEN
        // Check for URL
        IF url_detected THEN
          Mark as spam
        ELSE
```

```
          Mark as not spam
      ELSE
        // No QR code, proceed
        Mark as not spam
    ENDIF
  ENDIF
ELSE
  // Account not verified, mark as spam
  Mark as spam
ENDIF
END
```

B. Levels of Proposed Model

The account verification level involves checking of fake account by considering various feature shortlisted. Body and subject text scanning using ensemble model with hybrid feature selection method at level 2. Checking for malicious QR code if present. Checking for embedded links or URLs or image URLs within the body section of email. This system is used to detect the phishing email, so in order to protect them from spam we are using some various levels for detection purposes. For each level we are using various algorithms such as:

1. Account Verification: DT / Naive Bayes

2. Email body and subject part: Ensemble Learning methods with hybrid features

3. URL Detection: CNN

4. Malicious QR detection: CNN

1) Algorithm for level 1 (Account Verification)

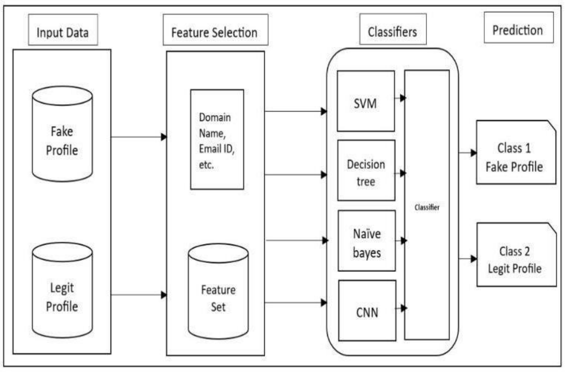


Figure 3. Flowchart for Level 1

Feature extracted:

- Username length: The decision tree can split the data based on the length of the username portion of the email address, allowing you to identify patterns in username length.

- Domain length: The decision tree can split the data based on the length of the domain name portion of the email address, allowing you to identify patterns in domain name length.

- Number of special characters: The decision tree can split the data based on the number of special characters in the email address, allowing you to identify patterns in the usage of special characters.

- TLD (top-level domain): The decision tree can split the data based on the TLD portion of the domain name (e.g., ".com", ".edu", ".gov"), allowing you to identify patterns in TLD usage.

- Username format: The decision tree can split the data based on the format of the username portion of the email address, allowing you to identify patterns in username formats.

- Domain format: The decision tree can split the data based on the format of the domain name portion of the email address, allowing you to identify patterns in domain name formats.

2) Algorithm for level 2 (Subject and Body part)

Process of implementation: For detection of malicious text in body or subject or semantics of an email.

6 stages:

1. Email Parsing stage

2.Content Feature Extraction stage

3.Pre-processing stage

4.Textual Feature Extraction stage

5.Feature Selection stage

6. Ensemble Classification stage

Features shortlisted are:

Body features

1.HTML Code: If there is HTML code in the body of the email

2. HTML forms

3. Scripts- Scripting code in email body

4. Attachments - Number of attachments in email

5. Image link: When a hyperlink is concealed behind an image in the body of an email.

6. Bad words in body- No verify, bunk, debit, payment, suspend, etc.

7. Bad words in subject- No verify, bunk, debit, payment, suspend, etc.

8. Absence of 'RE' (reply to another email) in subject

9. Number of characters in email body

10. Number of words in email body

11. Richness = the number of words count and character count of the email body.

12. The total count of unique terms inside the body of the email

13. The quantity of email sections

Features of the header: -

14. Email encoding- Emails encoding type Base64, Quoted - Printable, 8Bit, 7Bit, Binary, Xtoken, & None.

15. Number of email recipients- often phishing campaigns simultaneously target multiple users.

URL FEATURES: -

16. Number of hyperlinks

17. Text hyperlink- calculates the presence of human-readable text to hide a hyperlink.

18. Number of different HREF

19. Number of dots- calculates the dots of each URL. More than or equal to 4 dots indicate a malicious URL.

20. Check domain

21. @ symbol in URL in email

22. the presence of http:// instead of https:// in URL.

3) Algorithm for level 3 (Checking for URL if present)

To categorise a URL, the CNN first retrieves the labelled training data. Next, the data is randomly divided into sets for training and testing. After preparing training and test data, the CNN architecture, including input, output, and layers, was built for final training. To enable for extracting the important components and create feature vectors, a max-pool layer is being added after each convolution as shown in figure 4. Dropout regularisation was used to prevent the model from overfitting.

- URL Length: if url_length > threshold: malicious else benign
- Domain Reputation: if domain_reputation == poor: malicious else benign
- Use of HTTPS: if uses_https: benign else malicious
- Presence of Subdomains: if subdomain_count > threshold or complex_subdomain_structure: malicious else benign
- Keyword Analysis: if contains_phishing_keywords: malicious else benign

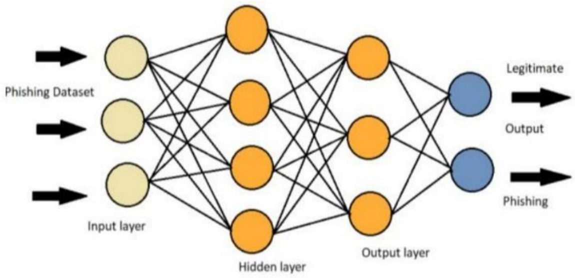


Figure 4 Deep Learning Model

4) Algorithm for level 1 (Account Verification)

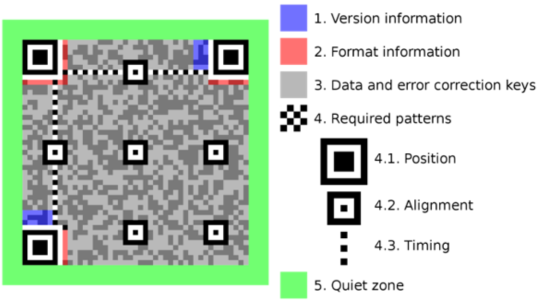


Figure 5 QR Code Structure

A matrix code, or two-dimensional barcode symbol, is what's known as a QR code. Modules Figure 5 is made up of black and white squares, where the black squares stand for a value of 1 and the white squares for a value of 0 [15]. There are forty different types of QR codes; Version 1 is the smallest, with 21 x 21 code pieces. With each new version, there are four more code elements in the side length. Version 40 comprises 31,329 code pieces spread across 177 × 177 modules. In every session, the greatest quantities of data, character types, and error correction levels are defined independently. The Reed-Solomon error correction technique, which has four error correction levels, is implemented via the QR code technology. In addition, the algorithm splits QR codes into separate chunks and fixes errors in each block. Codewords make up each block, and each codeword has eight data modules. The storage capacity decreases as the error correction level increases. The approximate error correction capabilities at each of the four levels are as follows:

- Level L: Low 7 percent of codewords are recoverable;
- 15% of codewords at Level M (Medium) can be recovered;
- Q (Quartile) Level 25% of codewords are recoverable; Level: H (High)
- You can recover 30% of the codewords.

Features that can help distinguish between malicious and benign QR codes. Here are some possible features you can consider:

- Code content: Check the content of the QR code to see if it leads to a known malicious website or application.
- Code size: Malicious QR codes may be smaller or larger than legitimate ones.
- Code complexity: Malicious QR codes may be more complex than legitimate ones.
- Code color: Malicious QR codes may have unusual color combinations or patterns.
- Code source: Malicious QR codes may be distributed through untrusted sources, such as unsolicited emails or social media messages.

C. Dataset Preparation

Guidelines for datasets as mentioned in [10] say that previous studies often use outdated phishing email samples predating 2015, which fail to capture current attack trends, potentially inflating numerical results. Additionally, while many evaluations employ balanced datasets, reflecting an equal distribution of benign and phishing emails, real-world scenarios present imbalanced class ratios, with benign emails outnumbering phishing ones. Therefore, using balanced dataset results in inaccurate results of proposed methodologies. The solution to this can be, using curated datasets which are imbalanced and latest. The Enron email dataset comprises around 500,000 emails by 150 users approximately exchanged among Enron Corporation employees, it stands as one of the few publicly accessible large-scale collections of authentic emails [17]. The Spamassassin Corpus includes 6,047 real emails from Spamassassin developers [20], with 1,897 identified as spam and the remaining 4,150 classified as benign. [16] here the curated dataset from 1998 to 2022 can be found. There are many datasets available publicly for classification of emails [18][19].

D. Performance Metrics

As mentioned in the guidelines of [10] the evaluation of ML-based detection methods should utilize suitable metrics. There are several metrics like precision, recall, Accuracy, receiver F1 score, Matthews correlation coefficient, operating characteristics, training time and receiver operating characteristics, confusion matrix, micro scores, macro scores, weighted scores.

IV. Conclusion and future work

Combining these strategies can offer a robust defense against email phishing overall. As technology evolves, it's crucial to keep current and adjust cybersecurity tactics accordingly. Improving our phishing protection methods protects online transactions from fraud and cyber dangers. The research underscores the existing loopholes in current detection systems, particularly the lack of mechanisms for detecting phishing links and QR codes. It has overcome this challenge by developing a combined system that analyzes email content, URLs, and QR codes. The combination of these approaches employs static as well as dynamic analysis, allowing the system to detect minor phishing signs while adapting the developing strategies used by hostile actors. In the future, there is potential to further enhance email security by developing automated systems for detecting phishing emails. This holistic method strengthens email security while also aligning with a proactively cybersecurity strategy, giving organisations and people a versatile weapon to combat the persistent and advanced nature of email phishing assaults.

REFERENCES

[1] D. V. Lindberg and H. K. H. Lee, “Optimization under constraints by applying an asymmetric entropy measure,” J. Comput. Graph. Statist., vol. 24, no. 2, pp. 379–393, Jun. 2015, doi: 10.1080/10618600.2014.901225.

[2] B. Rieder, Engines of Order: A Mechanology of Algorithmic Techniques. Amsterdam, Netherlands: Amsterdam Univ. Press, 2020.

[3] I. Boglaev, “A numerical method for solving nonlinear integro-differential equations of Fredholm type,” J. Comput. Math., vol. 34, no. 3, pp. 262–284, May 2016, doi: 10.4208/jcm.1512-m2015-0241.

[4] A. Vadariya and N. K. Jadav, ‘‘A survey on phishing URL detection using artificial intelligence,’’ in Proc. Int. Conf. Recent Trends Mach. Learn., IoT, Smart Cities Appl., 2021, pp. 9–20, doi: 10.1007/978-981-15-7234-0_2.

[5] D. Sahoo, C. Liu, and S. C. H. Hoi, ‘‘Malicious URL detection using machine learning: A survey,’’ 2017, arXiv:1701.07179.

[6] C. M. R. D. Silva, E. L. Feitosa, and V. C. Garcia, ‘‘Heuristicbased strategy for phishing prediction: A survey of URL-based approach,’’ Comput. Secur., vol. 88, Jan. 2020, Art. no. 101613, doi: 10.1016/j.cose.2019.101613.

[7] V. V. Satane and A. Dasgupta, ‘‘Survey paper on phishing detection: Identification of malicious URL using Bayesian classification on social network sites,’’ Int. J. Sci. Res., vol. 4, no. 4, pp. 1998–2001, 2013.

[8] A. Aleroud and L. Zhou, ‘‘Phishing environments, techniques, and countermeasures: A survey,’’ Comput. Secur., vol. 68, pp. 160–196, Jul. 2017. 10.1016/j.cose.2017.04.006.

[9] Qi, Q.; Wang, Z.; Xu, Y.; Fang, Y.; Wang, C. Enhancing Phishing Email Detection through Ensemble Learning and Undersampling. Appl. Sci. 2023, 13, 8756. https://doi.org/10.3390/app13158756

[10] Bountakas, Panagiotis and Xenakis, Christos, Helphed: Hybrid Ensemble Learning Phishing Email Detection. Available at SSRN: https://ssrn.com/abstract=4147334 or http://dx.doi.org/10.2139/ssrn.4147334

[11] Sasirekha C, Nandhini R, Karthiga Mai N L, Bhuvaneshwari R S, Chandra V S, 2023, Email Phishing Detection Using Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 11, Issue 03,

[12] YONG FANG, CHENG ZHANG, CHENG HUANG, LIANG LIU and YUE YANG, “Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism,” IEEE, vol. 7, 2019.

[13] R. M. Bani-Hani, Y. A. Wahsheh and M. B. Al-Sarhan, "Secure QR code system," 2014 10th International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, 2014, pp. 1-6, doi: 10.1109/INNOVATIONS.2014.6985772.

[14] A. Kharraz, E. Kirda, W. Robertson, D. Balzarotti and A. Francillon, "Optical Delusions: A Study of Malicious QR Codes in the Wild," 2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, Atlanta, GA, USA, 2014, pp. 192-203, doi: 10.1109/DSN.2014.103.

[15] QR Code. Available online: https://en.wikipedia.org/wiki/QR_code#/media/File:QR_Code_Structure_Example_3.svg (accessed on 6 January 2022).

[16] “Phishing email curated datasets,” Zenodo, Sep. 2023, doi: 10.5281/zenodo.8339691. Available: https://doi.org/10.5281/zenodo.8339691

[17] “CMU School of Computer Science,” Feb. 13, 2024. Available: https://www.cs.cmu.edu/

[18] UCI Machine Learning Repository. Available online: https://archive.ics.uci.edu/ml/datasets.php (accessed on 28 October 2020).

[19] Nazario Dataset. Available online: https://www.monkey.org/~{}jose/phishing/ (accessed on 23 October 2020).

[20] SpamAssassin Dataset. Available online: https://spamassassin.apache.org/ (accessed on 22 October 2020).