

A Hybrid Machine Learning and Reinforcement Learning Approach for Adaptive Phishing Email Detection

Mrs. Kodali Lakshmi Padmavathi¹, Bhedam Pavan Kumar², Dangatla Bhanu Sri³ and Kathula Jaithra⁴

¹Assistant Professor, Department of Artificial Intelligence and Data Science

²⁻⁴Students, Department of Artificial Intelligence and Data Science, Lakireddy Balireddy College of Engineering, Mylavaram, Andhra Pradesh, India

Email: lakshmik@lbrce.ac.in, pavankumarbhedam@gmail.com, dangatlabhanusri@gmail.com, jaithrakathula3@gmail.com

Abstract—Phishing emails are the fake or fraudulent emails that are designed to trick people into giving personal information which may include passwords or credit card details or any other sensitive data. The goal of our project is to create a system that can automatically identify these fake emails and filter them out, and making email communication more safer. The system will analyze the content and patterns in emails to differentiate between the legitimate and the fraudulent emails. This project aims to develop a robust, automated system for identifying and filtering phishing emails, enhancing the security of email communication. Our approach combines machine learning techniques, specifically Random Forest classification, and reinforcement learning via Deep Q-Networks (DQN), to analyze the metadata, textual content, and embedded URLs of emails, discerning legitimate messages from phishing attempts. Through a feedback driven mechanism, the model can learn from incorrect predictions and adapt over time, providing an evolving and proactive defense against phishing threats. This system holds significant potential for deployment in both corporate and personal email environments, contributing to safer email practices across diverse use cases.

Index Terms—Phishing emails, Fraudulent emails, Sensitive data, Email security, Data analysis, Random Forest, Reinforcement Learning, Deep Q-Network (DQN), Feedback driven learning, Metadata analysis, Textual content analysis, Adaptability.

I. INTRODUCTION

A. Background Information

Phishing emails have become one of the most common forms of cyber attacks, targeting individuals and organizations with the intent of stealing sensitive information such as usernames, passwords, and financial data. As digital communication continues to expand, phishing attacks have grown more sophisticated, exploiting unsuspecting email users by mimicking trusted sources. This therefore calls for the need for dependable, automatic systems that can accurately identify and prevent phishing attacks with precision, especially in a scenario where large volumes of email traffic make manual detection infeasible. Traditional approaches to detecting phishing normally employ rule-based systems or blacklists. Yet, both fail to maintain pace with the ever-changing tactics of phishers. Usually, the lists need updating, a task that proves to be very resource

intensive and rather inefficient. Lately, techniques of machine learning seem promising in automatically defining phishing characters in models, based on patterns in the data. Although effective, ML methods alone are limited by their static nature and make it hard to update the systems in real-time in accordance with new phishing strategies. There is a need for a system that achieves high accuracy in detecting phishing emails but at the same time dynamically adapts to emerging threats in real-time. Existing ML techniques are limited in addressing this adaptability, thus a hybrid approach is required which combines static classification with dynamic learning mechanisms to handle evolving phishing tactics.

B. Significance of the Research

1. This paper describes a novel hybrid phishing detection model, bridging traditional machine learning and adaptive reinforcement learning, thereby providing a comprehensive approach to solving the challenges of phishing attacks. It uses a Random Forest classifier for an accurate and robust primary detection while integrating a DQN that enables adaptability through feedback-driven
2. A feedback mechanism that enables continuous improvement by learning from incorrect predictions and identifying new phishing patterns
3. Evaluation of the model across diverse datasets, demonstrating its capability to maintain high accuracy while adapting to novel phishing attempts.

By treating phishing detection as a sequential decision making problem, the DQN component ensures that the system learns optimal detection policies and iteratively refines its criteria over time. This adaptability is critical to handling the diverse and rapidly changing nature of phishing techniques, thereby significantly reducing the risk of misclassification.

II. LITERATURE SURVEY

A. Overview of Relevant Literature

Lingampally Shalini, Sunil Kumar S. Manvi, Naveen Chandra Gowda, Manasa K N [1], This paper addresses phishing email detection, a critical cybersecurity challenge that can cause significant financial and data losses. It examines the usage of ML and DL algorithms to classify phishing emails from non-phishing emails, making distinctions based on textual features as well as metadata. It splits the dataset into two sets: one is 70-30% and the other is 80-20% of training and testing. It compares the performance of different processes: first, a sample expansion stage by using KNN and K-Means algorithms in order to expand the dataset, and second, the testing stage in which the LSTM model is trained using pre-processed data, which includes word segmentation and vector generation. The model's performance is measured by accuracy metrics that indicate a 95% detection rate.

Melad Mohamed Al-Daeef et al. [2], The research paper focuses on identifying the most effective detection features, particularly those extracted from the email body, such as Keywords and URLs. The authors submit a novel feature a novel feature selection framework based on sender information, email content, and recipient interaction to accurately distinguish phishing emails from legitimate ones. Through an analysis of over 8,000 emails, the study finds that URL-based features are more reliable than Keywords, showing a stronger correlation with phishing activity. Additionally, the research emphasizes the importance of sender and content based features, highlighting the crucial role of URL analysis in phishing detection. The paper concludes that optimizing feature selection using these criteria significantly improves detection model performance, providing a more effective approach to combating phishing threats in practical scenarios.

Rohit Valecha, PranaliMandaokar, H. Raghav Rao[3], This paper deals with phishing email detection, a severe challenge in the domain of cyber security that threatens not only financial security but also data integrity. It tries to understand whether machine learning algorithms can detect phishing emails through the analysis of persuasion cues present within the content of an email. The study applies different ML algorithms, including SVM, Random Forest, Naive Bayes, Decision Trees, and Logistic Regression, for email classification based on such cues. Evaluation has been done using training and test sets to test the performances of the models. Performances of the algorithms are estimated using accuracy, precision, recall, and F1-score as metrics. The results indicated a critical role of persuasion cues to detect the phishes; there has also been a better performance of models like Random Forest and SVM in the correct identification of phishing attempts.

Ayman El Aassal et al. [4], This paper addresses the assessment and comparison of phishing detection methods, a crucial aspect of cyber security aimed at preventing significant financial and data breaches. It introduces the "Phish Bench" framework, designed to systematically compare and assess various phishing detection techniques across different datasets. The study evaluates the Performance of machine learning classifiers, including supervised and Neural network-based models, in detecting phishing attacks based on URL, website, and email

features. This paper further discusses the challenges of imbalanced datasets, which are common in phishing detection, and shows the significance of feature selection in improving the accuracy of detection. The results have emphasized the need for innovative approaches to improve the robustness of phishing detection systems against evolving threats.

B. Controversies in the Literature or Gaps

Despite advancements, several controversies persist in phishing detection research. One key debate revolves around feature selection. While Melad Mohamed Al-Daeef et al. [2], argue that URL-based features are the most reliable, Lingampally Shalini et al. [1] and Eder Souza Gualberto et al. [5] highlight the importance of textual and metadata features, indicating that feature effectiveness may depend on the dataset and phishing attack characteristics. Another area of debate is the performance trade-offs between traditional models and advanced neural networks. Yong Fang et al. [6] claim superiority of RCNN models with attention mechanisms, while Nguyet Quang Do et al. [7] warn that deep learning models often suffer from problems such as manual parameter tuning and long training times, thus raising doubts about their feasibility compared to simpler ML methods. Another topic of debate includes dataset augmentation. Lingampally Shalini et al. [1] support augmenting with KNN and K-Means, whereas Ayman El Aassal et al. [4], mention that the risk of using imbalanced datasets is that these might influence the evaluation process. Adaptability in detection systems to changing phishing strategies remains a significant issue. Relatedly, in this regard are papers by Rohit Valecha et al. [3], while Nguyet Quang Do et al. [7] focus a lot on changing attack patterns with dynamic counter mechanisms. Other similar solutions still provide no one general solution over the other: attention mechanism and feature optimization. Moreover, evaluation metrics have inconsistencies in their focus, from accuracy-based evaluation (Lingampally Shalini et al. [1], Eder Souza Gualberto et al. [5]) to precision, recall, and F1 score-based evaluation (Rohit Valecha et al. [3], Yong Fang et al. [6]), which makes direct comparison difficult and obscures a clear understanding of model performance.

III. PROPOSED MODEL

The phishing email detection model we propose identifies phishing attempts by analysing both text and metadata features in emails. The workflow is structured as follows:

A. Dataset Information

The dataset for this phishing email detection model was constructed by merging eight different datasets, as outlined in the paper “Several Curated Datasets and Feature Analysis for Phishing Email Detection with Machine Learning” [8]. This combined dataset consists of over 100,000 email samples. Each sample includes various enriched features such as email metadata (e.g., sender and receiver details), textual content (subject and body), providing a multidimensional view of the data.

B. Data Preprocessing

To prepare the email text, we clean the subject and body, removing noise elements such as punctuation, stop-words, and case variations. The text is converted to lowercase, and lemmatization is applied to reduce words to their root forms, ensuring focus on essential meanings. Metadata features (e.g., sender, receiver, presence of URLs) are also extracted for further analysis.

C. Text Vectorization with TF-IDF

It converts text into numerical vectors through the application of TF-IDF (Term Frequency-Inverse Document Frequency) after preprocessing. This transformation tends to weigh more important terms which best denote phishing activity; words uncommon, unusual, or suspect in nature will denote malicious behaviour most likely. Therefore, these critical terms highlight a way that can separate the phisher's content from regular ones.

D. Feature Engineering and Model Training (Random Forest)

Using the insights from EDA, we engineer key features, including textual details (e.g., word frequency, sentiment analysis) and URL characteristics (e.g., length, presence of suspicious keywords). We then apply a Random Forest classifier, chosen for its robustness and interpretability, as the initial model.

E. Random Forest as Initial Policy

The trained Random Forest model serves as the initial policy for the Deep Q-Network (DQN) agent. It provides a baseline classification of emails as phishing or legitimate, leveraging its knowledge to enhance the DQN

agent's learning efficiency.

F. DQN Agent for Continuous Learning

To improve adaptability, we integrate a Deep Q-Network (DQN), an RL approach that enables the model to refine its phishing classification policy iteratively. In this setup, the DQN agent interacts with the environment (an email classification task), making decisions and receiving feedback based on the correctness of each classification. Positive rewards reinforce correct phishing classifications, while penalties discourage false positives, allowing the model to adapt to new patterns in phishing attacks

G. Integration of DQN and Random Forest

To build more efficient and accurate Phishing email detection we combined the strength of reinforcement learning and supervised learning. The DQN agent uses the Random Forest model to calculate the initial observation for the environment. This observation represents the phishing probability, which helps the agent make its first classification decision. The agent then refines its decision making process by continuously interacting with the environment, receiving rewards or penalties based on the accuracy of its classifications. As it learns, the DQN model gradually improves its performance, potentially outperforming the Random Forest model by adapting to new phishing techniques.

H. Model Architecture Diagram

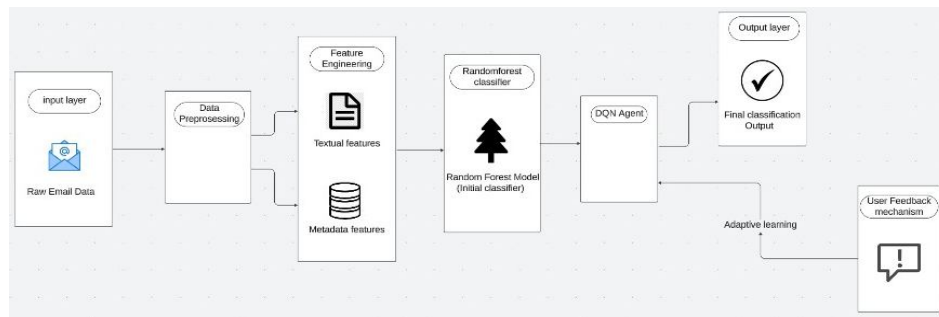


Figure 1 : Architecture of the Hybrid Phishing Detection Model, Illustrating the Integration of Random Forest and Deep Q-Network for Initial Classification and Iterative Learning.

I. Model Flow Chart

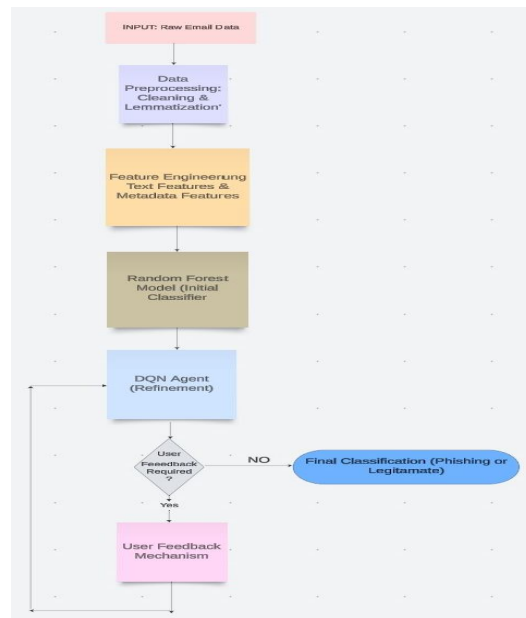


Figure 2: Flowchart of The Proposed Model, Illustrating the Sequence From Data Pre processing to Classification

J. Evaluation Metrics

Accuracy

This metric measures the overall correctness of the model by calculating the proportion of correct predictions out of all predictions made.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Predictions}}$$

Precision

Precision indicates how many of the emails predicted as phishing were actually phishing. It reflects the proportion of true positive phishing detections among all instances labelled as positive..

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall (Sensitivity)

Recall shows the proportion of actual phishing emails that were correctly identified by the model

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

IV. RESULTS

Our phishing detection model attained remarkable accuracy using a hybrid approach combining the Random Forest classification with the Deep Q-Network. The Random Forest classifier reached an accuracy of 98% in independent classification between phishing and legitimate emails, using textual and metadata features. This is comparable or better than the traditional machine learning approaches like Bayesian classifier and Support Vector Machines (SVM), which achieve an accuracy of 90%-95% in similar scenarios. When DQN was incorporated into the system, the model achieved an accuracy of 99%, thus showing its adaptability and ability to learn new phishing strategies dynamically.

A. Data Analysis and Interpretation

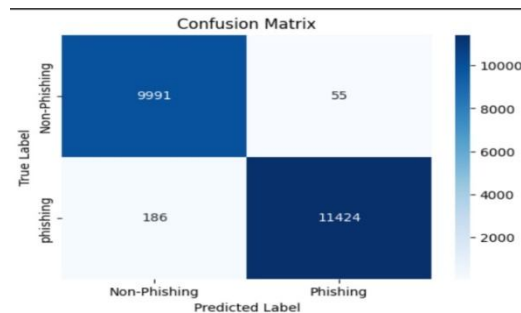


Figure 3: Confusion Matrix Illustrating the Classification Performance of the phishing Detection Model

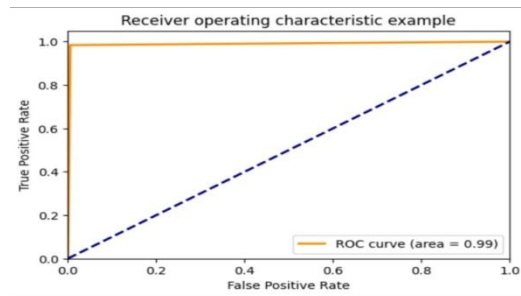


Figure 4: ROC Curve Illustrating the Relationship Between the TP and FP rate of the Model

TABLE I: PERFORMANCE ANALYSIS TABLE FOR RANDOM FOREST

	precision	Recall	f1-score	Support
Non Phishing(0)	0.98	0.99	0.99	10046
Phishing(1)	0.98	0.98	0.99	11610
Overall Accuracy			0.99	21656
Macro Average	0.99	0.99	0.99	21656
Weighted Average	0.99	0.99	0.99	21656

V. DISCUSSION

The proposed phishing email detection model efficiently detect phishing emails with high accuracy. The AutoPhishGuard interface, as presented in Figure 5.1, allows users to easily input email details and immediately obtain real-time predictions whether an email is phishing or legitimate. If an email is labelled as safe, the system provides a "Safe Email" message as in Figure 5.2. Moreover, the feedback mechanism, illustrated in Figure 5.4, further improves the system's effectiveness by refining predictions through user-provided corrections, thereby ensuring continuous enhancement in combating new phishing schemes.

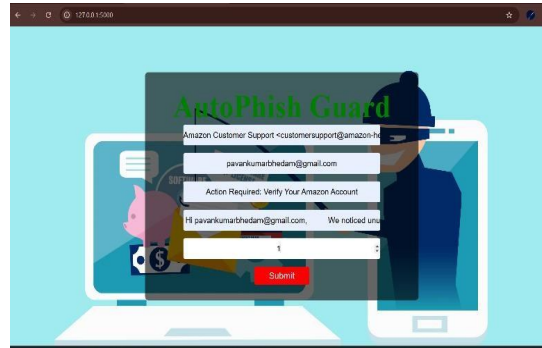


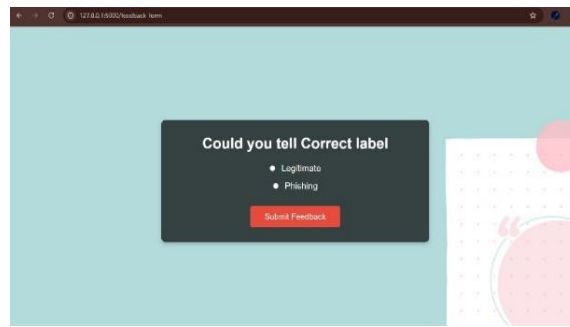
Figure 5.1: The Auto Phish Guard interface enables users to input email details



Figure 5.2: A "Safe Email" message get When an Email is determined to be legitimate



Figure 5.3: A phishing alert notifies users of potential threats for emails classified as phishing



5.4: The Feedback Mechanism Allows users to Provide Input for Incorrect Prediction

V. CONCLUSION

The proposed phishing email detection model demonstrates a robust and systematic approach to identifying phishing emails by combining a Random Forest classifier with DQN. The hybrid model achieves high accuracy, outperforming traditional machine learning methods and adapting dynamically to new phishing techniques. The integration of the user feedback mechanism enhances the ability of the system to refine its predictions, which provides a proactive and reliable solution to real-world challenges in email security. With the user-friendly interface, further simplifying interaction, it makes the model accessible and effective for non-technical users.

FUTURE SCOPE

This phishing detection model can be further developed to improve the feedback system for better adaptability, integrate with real time email platforms to mitigate threats instantly, and enhance multilingual support. Incorporating advanced features like behavioural analysis, testing on diverse datasets, and improving robustness will strengthen the model.

REFERENCES

- [1] L. Shalini, S. S. Manvi, N. C. Gowda and K. N. Manasa, "Detection of Phishing Emails using Machine Learning and Deep Learning," 2022 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2022, pp. 1237-1243, doi:10.1109/ICCES54183.2022.9835846.
- [2] M. M. Al-Daeef, N. Basir and M. M. Saudi, "A Method to Measure the Efficiency of Phishing Emails Detection Features," 2014 International Conference on Information Science & Applications (ICISA), Seoul, Korea (South), 2019, pp. 1-5, doi: 10.1109/ICISA.2014.6847332.
- [3] R. Valecha, P. Mandaokar and H. R. Rao, "Phishing Email Detection Using Persuasion Cues," in IEEE Transactions on Dependable and Secure Computing, vol. 19, no. 2, pp. 747-756, 1 March-April 2022, doi: 10.1109/TDSC.2021.3118931.
- [4] A. El Aassal, S. Baki, A. Das and R. M. Verma, "An In-Depth Benchmarking and Evaluation of Phishing Detection Research for Security Needs," in IEEE Access, vol. 8, pp. 22170-22192, 2020, doi: 10.1109/ACCESS.2020.2969780
- [5] E. S. Gualberto, R. T. De Sousa, T. P. De Brito Vieira, J. P. C. L. Da Costa and C. G. Duque, "The Answer is in the Text: Multi-Stage Methods for Phishing Detection Based on Feature Engineering," in IEEE Access, vol. 8, pp. 223529-223547, 2020, doi: 10.1109/ACCESS.2020.3043396.
- [6] Y. Fang, C. Zhang, C. Huang, L. Liu and Y. Yang, "Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism," in IEEE Access, vol. 7, pp. 56329-56340, 2019, doi: 10.1109/ACCESS.2019.2913705.
- [7] N. Q. Do, A. Selamat, O. Krejcar, E. Herrera-Viedma and H. Fujita, "Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions," in IEEE Access, vol. 10, pp. 36429-36463, 2022, doi: 10.1109/ACCESS.2022.3151903.
- [8] A. I. Champa, M. F. Rabbi and M. F. Zibran, "Curated Datasets and Feature Analysis for Phishing Email Detection with Machine Learning," 2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI), Mt Pleasant, MI, USA, 2024, pp. 1-7, doi: 10.1109/ICMI60790.2024.10585821.
- [9] Serfettin Senurk, Elif Yerli, Ibrahim Sogukpinar, "Email Phishing Detection Prevention by Using Data Mining Techniques", International Conference on Computing Technologies (ICCT), 2017 IEEE.
- [10] G. Sonowal, "A Model for Detecting Sounds-alike Phishing Email Contents for Persons with Visual Impairments," 2020 Sixth International Conference on e-Learning (econf), Sakheer, Bahrain, 2020, pp. 17-21, doi: 10.1109/econf51404.2020.9385451.

- [11] Y. Jin, M. Tomoishi and N. Yamai, "Trigger-based Blocking Mechanism for Access to Email-derived Phishing URLs with User Alert,"2023 InternationalConference on Electronics, Information, and Communication (ICEIC), Singapore, 2023,
- [12] Chien and P. Khethavath, "Email Feature Classification and Analysis of Phishing Email Detection Using Machine Learning Techniques,"2023 IEEEAsia-Pacific Conference on Computer Science and Data Engineering (CSDE), Nadi, Fiji, 2023, pp. 1-8, doi: 10.1109/CSDE59766.2023.10487729.
- [13] O. Abahussain and Y. Harrath, "Detection of Malicious Emails through Regular Expressions and Databases,"2019 International Conference on Innovationand Intelligence for Informatics, Computing, and Technologies (3ICT), Sakhier, Bahrain, 2019