A Gamified Phishing Simulator Using Reinforcement Learning

Keana Leong and Noluntu Mpekoa
University of Johannesburg, South Africa

leongkeana@gmail.com
noluntum@uj.ac.za

Abstract: An organisation's security is fundamentally reliant on its people. Regardless of the sophistication of its cybersecurity infrastructure, the absence of comprehensive training and awareness can lead to vulnerabilities. Traditional phishing awareness training typically involves sending simulated phishing emails to employees, allowing organisations to monitor actions such as link clicks, email reporting, and responses. While this method offers valuable insights into employee behaviour, it often struggles to engage users effectively. This conventional approach may not create a dynamic learning environment conducive to better retention of vital security practices. Furthermore, users generally do not receive immediate feedback regarding their interactions with phishing links, leaving organisations more susceptible to social engineering attacks. This research seeks to address the issue by developing an interactive gamified phishing simulator that employs reinforcement learning (RL). The methodology for this study consists of two key components. First, a literature review was conducted to assess existing phishing awareness techniques and explore how RL can be applied effectively. This review examined the integration of RL within cybersecurity education and explored the impact of gamification on user behaviour. For the RL agent, a dataset comprising both phishing and legitimate emails was compiled. The agent was then trained to discern phishing emails from legitimate ones based on various email features. Then the agent presents users with email challenges and delivers real-time feedback on their selections. The simulator incorporates a reward and badge system that promotes active participation and ongoing learning. This approach aims to overcome the limitations traditionally associated with static phishing training by fostering continuous learning, ultimately reducing user susceptibility. The effectiveness of the proposed simulator was evaluated based on its classification accuracy of phishing and legitimate emails.

Keywords: Reinforcement learning, Gamified phishing, Phishing simulator, Artificial intelligence

1. Introduction

Phishing is executed through various channels, including email, SMS, and voice calls. Email phishing is the most common method, where the victim receives an email that looks like it is from a legitimate source (Alkhalil et al., 2021). These emails often contain manipulative language to encourage users to click on links, download attachments, or provide personal information under the threat of consequences. Smishing involves sending fraudulent text messages that appear legitimate, asking for personal details or prompting the download of an app. Vishing occurs when attackers make phone calls, posing as representatives of legitimate organisations, like banks, to extract sensitive information from victims (Salahdine, Mrabet and Kaabouch, 2021). From an organisational perspective, a single mistake, such as opening a phishing link, can expose an organisation to serious risks such as data breaches, ransomware infections, and Distributed Denial-of-Service (DDoS) attacks. Likewise, at an individual level, attackers can use sensitive information for malicious intent such as identity theft, financial fraud, or gaining access to personal applications (Ravi et al., 2021; Alkhalil et al., 2021). Many users assess the legitimacy of emails based on their overall look and feel, such as formatting, branding, logos, and language used. A poorly designed email with spelling errors or missing elements may raise suspicion, but a well-thought-out phishing email convinces users that it is from a legitimate source. As a result, relying solely on visual cues is an unreliable strategy to detect phishing and thus the reason users are still susceptible to phishing. People are amongst the weakest links for being victims of phishing attacks; organisations must invest resources and time to train staff to identify phishing attacks (Alkhalil et al., 2021; Alsharnouby, Alaca & Chiasson, 2015).

In South Africa (SA), the average data breach is estimated to cost more than R50 million per incident as cybercrimes become more frequent and complex; this cost will increase (Ngejane et al., 2024). SA has experienced substantial financial and data losses due to phishing incidents. Traditional phishing awareness training typically involves sending simulated phishing emails to employees and tracking their responses, such as clicks, reports, and ignored emails. While this provides some insights into behavior, it often lacks engagement and effectiveness. Employees' perception towards simulated phishing training is linked to lower productivity, increased feelings of boredom, anxiety and stress, thus resulting in a lack of user participation and employees not giving the training their full attention (Wannenburg, 2023). This approach fails to create an engaging learning experience, which could lead to better retention of critical security practices. Additionally, users do not receive immediate feedback that they have interacted with a phishing link, increasing the organisation’s vulnerability to social engineering attacks. This delay in feedback prevents employees from learning from their mistakes in real-time, making it harder to improve their ability to recognize phishing attempts in the future (Ngejane et al., 2024). Atemkeng and Wabo (2020) stated that 46% of cybersecurity cases were due to careless or uninformed staff.

Several studies have suggested that awareness training is the most effective approach in preventing attacks, as users are the final line of defence against many cyber-threats.

Although organisations strive to improve user awareness and training in phishing emails, many users still struggle to recognise phishing emails (Alsharnouby, Alaca & Chiasson, 2015). The same phishing simulations are used within the organisation, thus not incorporating dynamic and updated phishing tactics to simulate sophisticated phishing attacks, making the training less effective. Most research focuses on users' responses to internet phishing attempts; however, there is a lack of research on how employees within an organisational space handle phishing scams (Yeoh, 2022; Alkhalil et al., 2021). This study seeks to close this gap by creating an interactive gamification phishing simulator using intelligent agents to improve cybersecurity awareness through an engaging gamified environment. The simulator challenges users to identify phishing threats while providing real-time feedback. To ensure that users remain engaged, it includes a reward and badge system that encourages user participation and continuous learning. Unlike traditional static training methods, this approach offers a dynamic, direct learning experience tailored to real-world scenarios. The primary beneficiaries will be workplace industries aiming to strengthen phishing awareness and reduce user susceptibility. By including a gamified approach into realistic phishing simulations, the application aims to provide users with hands-on experience that better prepares them to recognise real-world phishing attacks for personal and organisational security thus mitigating user susceptibility to being phished (Ariyadasa & Fernando, 2024; Karim, 2019). Increased awareness directly contributes to reducing the various phishing techniques (as mentioned previously) as well as credential and identity theft and financial fraud. For the organisation, this means fewer successful social engineering attacks, thus mitigating against potential data breaches and hackers gaining unauthorised access to the organisation's system. As a result, social engineering becomes more difficult for hackers to exploit the weakest link in cybersecurity defence - the human element.

This research asserts the necessity of developing and implementing a gamified awareness solution to address the limitations associated with traditional static phishing training by fostering continuous learning, ultimately reducing user susceptibility. The remainder of the paper is as follows: Section 2 presents a literature review that was conducted to assess existing phishing awareness techniques and explore how RL can be applied effectively. This review examined the integration of RL within cybersecurity education and explored the impact of gamification on user behaviour. Section 3 presents the design and implementation of the proposed solution. For the RL agent, a dataset comprising both phishing and legitimate emails was compiled. The agent was then trained to discern phishing emails from legitimate ones based on various email features. Section 4 offers an evaluation of the solution, discussions, and future work. While Section 5 concludes the paper.

2. Literature Review

This section delves into the evolution of email phishing detection techniques, beginning with traditional rule-based filters and progressing to machine learning approaches. It is included to establish a basic understanding of how email threats have been managed, the limitations of traditional methods, and the need for adaptive intelligent systems as phishing techniques become more sophisticated. This gap has prompted the integration of machine learning, which brings flexibility, pattern recognition, and improved detection accuracy.

2.1 Phishing Awareness Techniques

Ho et al. (2025) from UC San Diego and the University of Chicago conducted a study for 8 months using a randomised controlled trial to assess two forms of phishing training: Annual security awareness training and embedded phishing training at UC San Diego Health involving 19 500 real world users to determine if employee training through simulated phishing exercises and cybersecurity awareness programs assists organisations as a defence mechanism. Employees received one of several simulated phishing emails covering themes like: account security warnings, shared documents, benefits policy changes and social protocol updates; these simulated phishing emails mimicked real-world scams. Each phishing simulation tracked the following: Click rate of the user, time spent on training if the employee failed and whether training was completed. These researchers found that the annual cybersecurity awareness training that organisations rely on has no impact on phishing defences. Employees who completed the training more recently were just as likely to fall for phishing attempts as those who completed the training over 12 months ago. Additionally, the practice of sending fake phishing emails and training employees who fall for them also showed very little benefit, with training reducing failure rates by only 1.7%. Upon further investigation, they found that engagement with training content was low, with more than 50% of employees closing the training page within 10 seconds of opening it. The employees who completed their interactive training were 19% less susceptible to falling for phishing scams in the future

Khairallah and Abu-Naseer (2024) conducted a study and developed a tool called *EmailAware* that was used to enhance email phishing awareness among university students through gamification. Their goal was to see which training method (gamified learning or traditional video-based education) better improves phishing detection skills. The study consisted of two groups of 30 participants, one group watched video-based phishing training, and the other group used *EmailAware,* which is a gamified platform. To measure the effectiveness of their study, they used Jigsaw Online Phishing Quiz to assess phishing awareness and detection skills and a qualitative survey to gather participant feedback on the gamified approach. Based on their findings, it was revealed that the gamified group outperformed the traditional group by 59%, with more engagement and motivation from this group. The traditional group improved by 41% thus the study suggests the use of gamification to improve cybersecurity training.

Traditional cybersecurity awareness training programs have several weaknesses that limit their effectiveness (Shakya, Pillai, & Chakrabarty, 2023). A major issue is the lack of personalisation; many programs adopt a uniform approach that fails to account for individual learning styles and existing knowledge levels. This generic design reduces learner engagement and comprehension. Studies also highlight that traditional training often lacks interactivity and dynamic elements, making it difficult to maintain participant interest or promote active learning (Salloum et al., 2022). Another significant gap is the absence of standardised metrics to assess training effectiveness or track long-term behavioural change. To address these deficiencies, future training must integrate personalised learning paths, real-time threat simulations, and measurable outcomes. Programs should move toward more interactive, adaptive models that engage users meaningfully and reflect the realities of the cybersecurity landscape (Shakya, Pillai, & Chakrabarty, 2023).

2.2 Machine Learning Techniques

Machine learning (ML) techniques applied to cybersecurity awareness phishing detection are categorised into supervised and unsupervised learning. Supervised learning trains a model on a labelled dataset, where each input is paired with an output label. The goal is for the model to accurately predict the label of new, unseen data once it has learned the mapping between inputs and outputs. Conversely, in unsupervised learning, the model is given a dataset with only input features and no labels, and from this data, the model is to determine the best outcome (Salloum et al., 2022; Dada et al., 2019).

Ahmed et al. (2022) conducted a rigorous investigation aimed at developing a machine learning-based model for detecting phishing emails, addressing both technical limitations and user-specific vulnerabilities. The objective of the study was to propose an optimised ML framework for phishing detection. Data preprocessing included deduplication, removal of missing values, and a standard 70:30 train-test split. Performance metrics used for evaluation included accuracy, precision, recall, and F1-score. The results revealed three major findings. First, a strong correlation between feature richness and detection accuracy was observed, with the 50-feature dataset achieving 100% accuracy using Boosted Decision Trees. Weiss (1999) conducted a comprehensive review of machine learning (ML) and deep learning (DL) techniques applied to spam detection in both email and Internet of Things (IoT) environments. The study aimed to systematically survey existing detection methods, categorising them into supervised, unsupervised, and reinforcement learning approaches, while comparing their performance based on accuracy, precision, recall, and other key metrics. The findings underscored the effectiveness of supervised learning models, with SVM and Naïve Bayes achieving high accuracy rates, up to 99.5% and approximately 89%, respectively, while Random Forests demonstrated performance levels reaching 95.2%.

Taherdoost (2024) emphasises that although machine learning techniques are effective in detecting phishing emails, they face several limitations. For example, Decision Trees (DT) can overfit the training data if not pruned correctly, which reduces the classification accuracy. Misclassifying spam as legitimate is usually harmless, but incorrectly labelling a legitimate email as spam can lead to the loss of valuable information. Additional challenges include the need for large volumes of training data and the constant evolution of attacker strategies, such as using images, synonyms, and intentional misspellings to bypass filters. Furthermore, obtaining high-quality email datasets is difficult, and machine learning models often rely on many features to perform well. These systems also tend to work best when tailored to individual users, and their accuracy drops significantly when applied in general scenarios. Lastly, they require high processing power, which may limit their practicality in resource-constrained environments (Taherdoost, 2024; Thakur et al., 2023).

2.3 Reinforcement Learning (RL)

Reinforcement Learning (RL) is a learning technique within the field of machine learning. It consists of an Artificial Intelligence (AI) agent that interacts with its environment through trial and error, perfecting its choices based on rewards received from past actions (Shakya, Pillai and Chakrabarty, 2023). RL is a new area of study in the literature and its significance within cybersecurity is increasing, it is being widely studied in various disciplines such as robotics, control systems, advertising, video games, autonomous vehicles and autonomous surgeries (Cengiz and Gök, 2023). In the cybersecurity space, it plays a vital role as it can adapt to dynamic and unpredictable environments without the use of predefined models. The use of RL to combat cybersecurity threats can provide security solutions to intrusion detection systems, penetration testing, cyberattacks through Denial of Service and Distributed Denial of Service attacks (DoS and DDoS) and assist with mitigating spoofing and phishing attacks (Oh et al., 2023).

Salahdine, Mrabet and Kaabouch (2021) proposed an innovative intrusion detection algorithm known as Adversarial Environment using Reinforcement Learning. The approach uses Double Deep Q-Network (DDQN), integrating RL with supervised learning. MAGPIE was proposed by Heartfield, which is an anomaly classification architecture that combines unsupervised anomaly detection in real-time with a Multi-Armed Bandit RL algorithm. This approach allows a system to dynamically select models based on human presence inference within a smart home space. Oh et al. (2023) implemented an adversarial cyber-attack simulator used to strengthen cybersecurity using the Deep Reinforcement Learning (DRL) framework. Their approach used an agent-based model that continuously learnt and adapted, making it suitable for an unpredictable nature as network security. Through RL, cybersecurity allows various domains to become more adaptive and automated in their threat detection. In phishing detection, RL agents interact with an email-based environment, refining their ability to classify emails by strategically balancing learning from past experiences and exploring new patterns.

3. Design and Implementation

The initial implementation involved a reflex agent that was developed to classify emails based on a set of predefined rules and patterns, such as the presence of suspicious keywords, URLs, or sender domains. While this approach was straightforward and provided quick results, it quickly became obvious that the reflex agent was limited in its ability to adapt (Salahdine, Mrabet and Kaabouch, 2021). This limitation led to the exploration and eventual adoption of a hybrid approach using a supervised Machine Learning (Random Forest) classifier and Reinforcement Learning agent for personalised awareness training.

This paper focuses on supervised ML algorithms, which have shown high accuracy in classifying emails as legitimate or phishing such as, decision tree, which is used for both classification and regression tasks and is able to manage categorical and numerical inputs. It splits data by checking distinctive features based on how well they separate the categories, by using measures like Gini Index (measuring misclassification risk) and Entropy (measuring information gain) (Salahdine, Mrabet and Kaabouch, 2021). This helps the model determine which features to use at each node in the tree. Random forest (RF) uses multiple decision trees built from random subsets of the data and features. Each tree makes a prediction by looking at a different part of the data and the result is based on the most voted class (phishing or legitimate). This approach improves accuracy and manages noisy or imbalanced data better than a single decision tree.

3.1 Random Forest Design

The primary responsibility of the Random Forest Agent is to accurately classify incoming emails, determining whether they are phishing attempts or legitimate correspondence. Utilising a complex ensemble learning technique, the agent analyses various features of each email, such as sender information, subject lines, and the content within the message. By aggregating the predictions made by multiple decision trees, the Random Forest Agent enhances accuracy and reduces the likelihood of misclassification. This sophisticated approach not only helps in identifying potential threats to user security but also minimises false positives, ensuring that legitimate emails are not mistakenly flagged as harmful.

Dataset: A dataset obtained from Rokibulroni (2023) included 3,650 phishing emails and 2,183 legitimate emails, which comprises 333 LLM-generated spear-phishing emails, 3,317 traditional phishing emails collected from public datasets spanning 1998–2022. There are 1,781 emails from the Enron email corpus, and 402 SpamAssassin marketing emails.

Feature Extraction: The system uses the extract_features function (from email_features.py) to convert each email into a set of numerical features. The features used are presented in Table 1 below:

Table 1: Feature Extractions

has_url	Whether the email body contains a URL (One if yes, zero if no).
sender_has_suspicious_domain	Whether the sender's domain looks suspicious (e.g., too many dashes, numbers, or unusual patterns).
has_urgency	Whether the subject or body contains urgent language (e.g., "urgent", "immediately").
subject_length	The length of the email subject. body_length: The length of the email body.
has_phishing_keyword	Whether the email contains common phishing keywords (e.g., "verify", "password", "security").
has_malicious_attachment	Whether the body mentions a file with a suspicious extension (e.g., .exe, .zip).

Data Preparation: Emails and their labels (*1* for phishing, *0* for legitimate) were loaded from a CSV file. Each email was processed to extract the above features, resulting in a feature matrix (X) and a label vector (y).

Model Training: The data was split into training and test sets (70% training, 30% testing). The Random Forest was then used to train a model. The forest contained twenty-five decision trees. Each tree was trained on a random bootstrap sample of data. At each split in a tree, a random subset of features was considered (to increase diversity among trees). The best split was chosen based on the Gini impurity, which measured how well a feature separates the classes. Trees grew until a maximum depth or until further splits were not meaningful.

Prediction: To classify a new email, its features were extracted and passed to each tree in the forest. Each tree made a prediction, and the final class was determined by majority vote across all trees.

Model Storage and Use: The trained Random Forest model was saved as a pickle file (random_forest_model.pkl). When needed, it was loaded and used to classify new emails in real time.

Integration with RL Agent: When the RL agent presents an email scenario to the user, the Random Forest model was used to determine the correct label (phishing or legitimate) for that email. The user's response was compared to the model's prediction: If the user's answer matches the model, the RL agent receives a positive reward. If not, the agent receives a negative reward. This feedback is used to update the RL agent's Q-table and policy.

3.2 Reinforcement Learning Design

Reinforcement Learning was chosen to enhance adaptability and personalise the training experience, beyond traditional and static ML approaches. The RL agent operates in a gamified phishing simulator, selecting email scenarios and adjusting difficulty dynamically using a Q-table that tracks the effectiveness of past interactions. To choose the next email, the agent uses an epsilon-greedy strategy: it sometimes explores by picking a random difficulty but usually exploits by selecting the difficulty with the highest Q-value for the current state. After the user responds to an email, the agent receives a reward (+5 for a correct answer, -1 for an incorrect one) and updates the Q-table. Table 2 below presents an overview of the agent's environment that was deployed:

Table 2: Reinforcement Learning Agent Environment

Inaccessible	The RL agent does not have access to all information about the environment (e.g., it cannot see future emails or the true intent behind an email). It learns from observed features and past experiences stored in the Q-table.
Non-deterministic	The outcome of the agent's actions (classifying an email) can vary due to the probabilistic nature of user behaviour and the randomness in email content. The same action in the same state may not always yield the same reward.
Episodic	Each email scenario is treated as a separate episode. The agent receives a reward or penalty after each classification, and the Q-table is updated accordingly. The outcome of one episode does not directly affect the next.
Static	The environment (the email being classified) does not change while the agent is deciding. The agent's policy (derived from the Q-table) is updated after each episode, but the email content remains fixed during classification.
Discrete	The state space (features of emails), action space (classify as phishing or legitimate), and rewards are all discrete.

The RL agent was implemented as a simple tabular Q-learning model. An epsilon-greedy policy balanced exploration with exploitation, while updates followed the standard temporal difference learning rule. The use of tabular Q-learning, although basic, provided a fast and functional baseline that made it possible to observe improvements without heavy computational requirements. In the gamified phishing simulator, the environment is a simulated interface where the agent interacts with phishing and legitimate emails. This feedback loop enables the agent to identify which email types and difficulty levels best support user learning, thus improving phishing detection and optimising personalised training outcomes.

4. Evaluation, Discussion and Future Work

4.1 Evaluation and Discussion

The web application supports registration, login, profile management, a leaderboard, and the simulation module. In the proposed system, each user has their personalised training, logging into their profile, able to track their progress. Through gamified elements such as a point system and badges, users feel encouraged to learn and improve their learning in an environment that keeps them engaged. If a user is not able to pass basic levels, which consist of obvious phishing clues, then the system will provide other tasks that are basic till the user can progress to the next level. By doing this, the user is supported on their learning journey and engaged with the content without feeling overwhelmed if a mistake is made.

In comparison to similar systems, the proposed gamified phishing simulator incorporates several novel features that enhance the user learning experience of phishing awareness training in an organisation. The system differs by including gamification with RL; this combination allows a dynamic, engaging, and adaptive training experience. This system gives instant feedback if users fail to spot phishing; through providing immediate feedback, users learn on the spot and this assists with recognising phishing attacks.

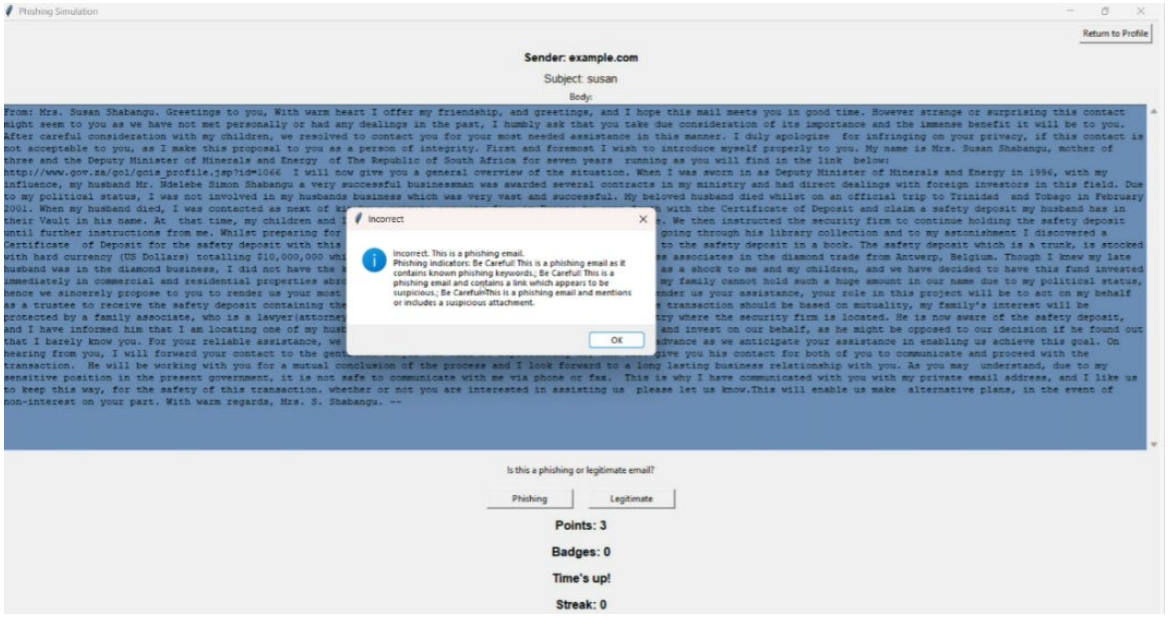

Figure 1: Phishing Simulator

When the RL agent presents an email scenario to the user, the Random Forest model is used to determine the correct label (phishing or legitimate) for that email. The user's response is compared to the model's prediction: If the user's answer matches the model, the RL agent receives a positive reward. If not, the agent receives a negative reward. This feedback is used to update the RL agent's Q-table and policy. The proposed phishing simulator solves problems found in current systems by making learning more interactive, engaging, and adaptable. By using reinforcement learning and gamification, the system keeps users interested and helps them improve their ability to spot and handle phishing attacks. This approach ensures that users are better prepared to deal with real phishing threats, improving the overall cybersecurity of the organisation.

4.2 Future Work

The proposed solution already integrates cutting-edge techniques like Reinforcement Learning and Random Forests for adaptive phishing awareness training. From here, several future research directions could deepen its

impact, broaden its scope, and enhance its effectiveness. Future research could focus on using Explainable AI (XAI) for user feedback and extend training beyond text-based phishing to include image-based, voice-based, and video-based phishing scenarios. Also, studies focusing on Human-Computer Interaction could be conducted to validate the hypothesis that gamification improves awareness and engagement.

5. Conclusions

This study addresses a critical gap by developing an interactive gamification phishing simulator powered by intelligent agents to enhance cybersecurity awareness in a gamified environment. The simulator empowers users to effectively identify phishing threats while delivering real-time feedback. To maximise engagement, it incorporates a robust reward and badge system that motivates active participation and continuous learning. Unlike outdated static training methods, this innovative approach provides a dynamic, immersive learning experience that mirrors real-world scenarios. The primary beneficiaries of this initiative are workplace industries focused on bolstering phishing awareness and lowering user susceptibility. By integrating a gamified strategy into realistic phishing simulations, the application equips users with invaluable hands-on experience that prepares them to recognise and combat actual phishing attacks, ultimately safeguarding both personal and organisational security. This heightened awareness plays a vital role in diminishing various phishing techniques, as well as preventing credential and identity theft and financial fraud. For organisations, this translates to a significant reduction in successful social engineering attacks, which mitigates potential data breaches and prevents unauthorised access to critical systems.

Ethics Declaration: Ethical clearance was not necessary for this study.

AI Declaration: Grammarly has been utilised for editing purposes and for enhancing overall writing quality.

References

Ahmed, N., Amin, R., Aldabbas, H., Koundal, D., Alouffi, B. and Shah, T. (2022) "Machine learning techniques for spam detection in email and IoT platforms: Analysis and research challenges", *Security and Communication Networks*, 2022, p.1862888. Available at: https://doi.org/10.1155/2022/1862888.

Alkhalil, Z., Hewage, C., Nawaf, L. and Khan, I. (2021) "Phishing attacks: A recent comprehensive study and a new anatomy", *Frontiers in Computer Science*, 3. Available at: https://doi.org/10.3389/fcomp.2021.563060.

Alsharnouby, M., Alaca, F. and Chiasson, S. (2015) "Why phishing still works: User strategies for combating phishing attacks", *International Journal of Human-Computer Studies*, 82, pp.69–82. Available at: https://doi.org/10.1016/j.ijhcs.2015.05.005.

Ariyadasa, S. and Fernando, S. (2024) "SmartiPhish: A reinforcement learning-based intelligent antiphishing solution to detect spoofed website attacks", *International Journal of Information Security*, 23, pp.1055–1076. Available at: https://doi.org/10.1007/s10207-023-00778-9.

Atemkeng, M. and Wabo, L.K. (2020) "A review of gamification applied to phishing", *Preprints*. Available at: https://www.preprints.org/manuscript/202003.0139/v1. https://doi.org/10.20944/preprints202003.0139.v1.

Cengiz, E. and Gök, M. (2023) "Reinforcement learning applications in cyber security: A review" *SAUJS*, 27, pp.481–503. Available at: https://doi.org/10.16984/saufenbilder.1237742.

Dada, E.G., Bassi, J.S., Chiroma, H., Abdulhamid, S.M., Adetunmbi, A.O. and Ajibuwa, O.E. (2019) "Machine learning for email spam filtering: Review, approaches, and open research problems", *Heliyon*, 5. Available at: https://doi.org/10.1016/j.heliyon.2019.e01802.

Hale, M.L., Gamble, R.F. and Gamble, P. (2015) "CyberPhishing: A game-based platform for phishing awareness testing", In: *2015 48th Hawaii International Conference on System Sciences*. pp.5260–5269. Available at: https://doi.org/10.1109/HICSS.2015.670.

Ho, G., Mirian, A., Luo, E., Tong, K., Lee, E., Liu, L.,& Voelker, G. M. (2025). Understanding the efficacy of phishing training in practice. In *2025 IEEE Symposium on Security and Privacy (SP)* (pp. 37-54). IEEE.

Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K. and Alazab, M. (2019) "A comprehensive survey for intelligent spam email detection", *IEEE Access*, 7, pp.168261–168295. Available at: https://doi.org/10.1109/ACCESS.2019.2954791.

Khairallah, O., & Abu-Naseer, M. M. (2024). The effectiveness of gamification teaching method in raising awareness on Email Phishing: Controlled Experiment. Thesis, Linnaeus University, Faculty of Technology, Department of computer science and media technology.

Ngejane, C.H., Chishiri, S., Shwayimba, S., Moyakhe, S., Miya, S. and Lwana, G. (2024) "Cyberattack incidents in South Africa: A survey", In: *2024 International Conference on Intelligent Cybernetics Technology and Applications (ICICyTA)*. pp.584–588. Available at: https://doi.org/10.1109/ICICYTA64807.2024.10912850.

Oh, S.H., Jeong, M.K., Kim, H.C. and Park, J. (2023) "Applying reinforcement learning for enhanced cybersecurity against adversarial simulation", *Sensors*, 23, p.3000. Available at: https://doi.org/10.3390/s23063000.

Ravi, R., Shillare, A.A., Bhoir, P.P. and Charumathi, K.S. (2021) "URL based email phishing detection application", *International Research Journal of Engineering and Technology*, 8(4). Available at: https://www.irjet.net/archives/V8/i4/IRJETV8I466.pdf.

Rokibulroni. (2023). Phishing Email Dataset [Data set]. GitHub. https://github.com/rokibulroni/Phishing-Email-Dataset

Salahdine, F., El Mrabet, Z. and Kaabouch, N. (2021) "Phishing attacks detection: A machine learning-based approach", In: *2021 IEEE 12th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*. pp.0250–0255. Available at: https://doi.org/10.1109/UEMCON53757.2021.9666627.

Salloum, S., Gaber, T., Vadera, S. and Shaalan, K. (2022) "A systematic literature review on phishing email detection using natural language processing techniques", *IEEE Access*, 10, pp.65703–65727. Available at: https://doi.org/10.1109/ACCESS.2022.3183083.

Shakya, A.K., Pillai, G. and Chakrabarty, S. (2023) "Reinforcement learning algorithms: A brief survey", *Expert Systems with Applications*, 231, p.120495. Available at: https://doi.org/10.1016/j.eswa.2023.120495.

Taherdoost, H. (2024) "Towards an innovative model for cybersecurity awareness training", *Information*, 15, p.512. Available at: https://doi.org/10.3390/info15090512.

Thakur, K., Ali, M.L., Obaidat, M.A. and Kamruzzaman, A. (2023) "A systematic review on deep learning-based phishing email detection", *Electronics*, 12, p.4545. Available at: https://doi.org/10.3390/electronics12214545.

Wannenburg, M.C., Nieman, A., Steyn, B. and Wannenburg, D.G. (2023) "South Africans' susceptibility to phishing attacks", *Southern African Journal of Accountability and Auditing Research*, 25, pp.53–72. Available at: https://doi.org/10.54483/sajaar.2023.25.1.4.

Weiss, G. (1999) *Multiagent systems: A modern approach to distributed artificial intelligence*. Cambridge, MA: MIT Press.

Yeoh, W., Huang, H., Lee, W.-S., Al Jafari, F. and Mansson, R. (2022) "Simulated phishing attack and embedded training campaign", *Journal of Computer Information Systems*, 62, pp.802–821. Available at: https://doi.org/10.1080/08874417.2021.1919941.