engineering proceedings

MDPI

Proceeding Paper

Adaptive Phishing Detection and Mitigation System Using Huawei Mind Reinforcement Learning with Human Feedback †

Jesher Immanuel B. Hael *, Mark Daniel S. Ortiz and Dionis A. Padilla

School of Electrical, Electronics and Computer Engineering, Mapua University, Manila 1002, Philippines; mdsortiz@mymail.mapua.edu.ph (M.D.S.O.); dapadilla@mapua.edu.ph (D.A.P.)

* Correspondence: jibhael@mymail.mapua.edu.ph

† Presented at the 7th Eurasia Conference on IoT, Communication and Engineering 2025 (ECICE 2025), Yunlin, Taiwan, 14–16 November 2025.

Abstract

Phishing remains a persistent cybersecurity threat, exploiting social engineering to bypass traditional defenses. We developed a phishing detection system that integrates baseline supervised learning with Reinforcement Learning through human feedback (RLHF) to improve adaptability against evolving attack strategies. Implemented using the Huawei MindRLHF framework and deployed on Raspberry Pi hardware, the system was evaluated using a dataset of 135,325 email samples consisting of both phishing and legitimate messages. The baseline supervised model achieved 94.3% accuracy, while the RLHF-enhanced model, through 74 iterations, achieved improved adaptability, reaching a 96.8% accuracy with balanced precision and recall. A multi-component reward function was designed to incorporate correct classification, human agreement, confidence matching, and consistency, enabling the model to refine its decision boundaries beyond automated optimization. Real-time monitoring and feedback were facilitated through a hardware-integrated LCD interface. The results confirm enhanced detection accuracy and reduced error rates, demonstrating its viability for deployment. The findings highlight the potential of human-centered RLHF the resilience and scalability of phishing mitigation systems against emerging cyber threats.

Keywords: phishing detection; reinforcement learning with human feedback (RLHF); Huawei MindSpore; cybersecurity; email security

1. Introduction

Phishing remains a widespread and evolving cyber threat, requiring effective detection and mitigation strategies. Traditional cybersecurity measures often fall short against increasingly sophisticated attacks [1,2]. To address this, AI and machine learning (ML) are increasingly integrated into cybersecurity systems to enhance threat detection and response [1,3], while encryption remains essential for safeguarding sensitive data and ensuring integrity [4,5]. Advanced adaptive AI systems, such as Reinforcement Learning (RL) meet these needs by learning from dynamic threats and incorporating human feedback to enable real-time responses, thereby overcoming the limitations of traditional approaches [1].

Recent advances in phishing detection demonstrate the growing role of AI and ML in cybersecurity [6]. Models such as distilled Bidirectional Encoder Representations from Transformers (DistilBERT) effectively classify spam, while distributed training optimizes

Check for updates

Academic Editors: Teen-Hang Meen, Chi-Ting Ho and Cheng-Fu Yang

Published: 30 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

computational performance [7,8]. Convolutional Neural Networks (CNNs) and Support Vector Machines (SVMs) have proven effective in various classification tasks [9–11], including e-signature verification [12], defect detection [13], and social media text analysis [14]. Beyond these, RL and Deep RL (DRL) introduce adaptive and autonomous defense mechanisms for evolving threats [15,16], such as mitigating DDoS attacks through dynamic resource allocation and adaptive intrusion detection [17]. However, limited research explores RL frameworks that integrate human feedback, leaving a gap that Huawei's MindRLHF could address in phishing mitigation.

Despite progress in AI and RL, most existing approaches rely on static datasets and overlook human factors, limiting adaptability to real-world cyber threats. We investigated Huawei's MindRLHF as a feedback-driven RL framework for phishing detection and mitigation. The objectives are to (1) train an RL model on an annotated email dataset using human feedback, (2) implement a hardware-integrated interface for real-time monitoring and response, and (3) evaluate phishing detection performance through key metrics and convergence analysis.

Huawei's MindRLHF is a practical and adaptive framework for phishing detection and mitigation. By incorporating human feedback into the reward process, the model enhances its learning precision, adaptability, and overall decision-making performance in dynamic threat environments [18]. We focused on email-based phishing, excluding other attack vectors such as malware and network intrusions. Overall, the results of this study can contribute to establishing a foundation for adaptive, feedback-driven cybersecurity systems capable of evolving with emerging threats.

2. Methodology

2.1. Conceptual Framework

Figure 1 illustrates the workflow of the phishing detection system using Huawei MindRLHF v0.3.0. The system processes a shared dataset of phishing and legitimate emails as input, with a baseline model first trained to establish initial performance. Human feedback reinforces correct classifications and refines the model over time. The MindRLHF model, implemented in Huawei MindSpore v2.5.0, analyzes incoming emails to recognize evolving phishing techniques and generate adaptive responses. Continuous learning enhances accuracy and resilience, while results are displayed on an LCD for real-time monitoring. This framework combines automated detection with human feedback to support adaptive and reliable phishing mitigation.

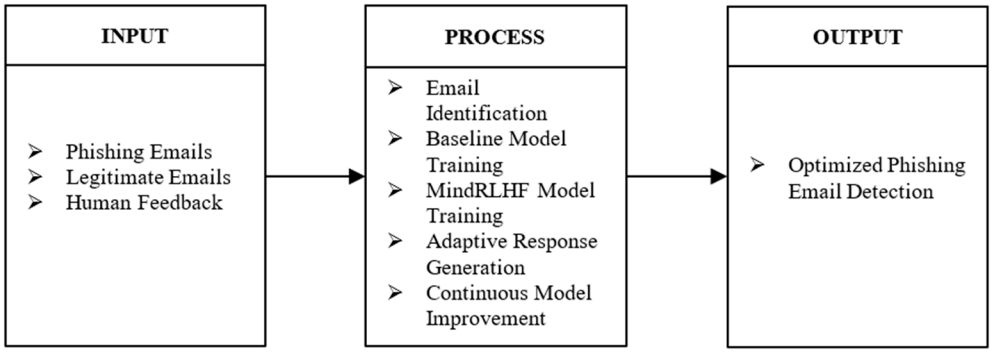


Figure 1. Conceptual framework for the phishing detection and mitigation system.

2.2. Hardware Block Diagram

Figure 2 shows the system component comprising two Raspberry Pi microcomputers connected in a peer-to-peer (P2P) Ethernet topology. The Raspberry Pi 4 serves as the attacker node, generating and transmitting phishing or legitimate emails, while the

Raspberry Pi 5 acts as the defender node, receiving and classifying messages. An LCD touchscreen connected to the defender displays real-time classification results for user verification and feedback. This wired P2P configuration provides a secure and controlled environment for implementing and testing the phishing detection system.

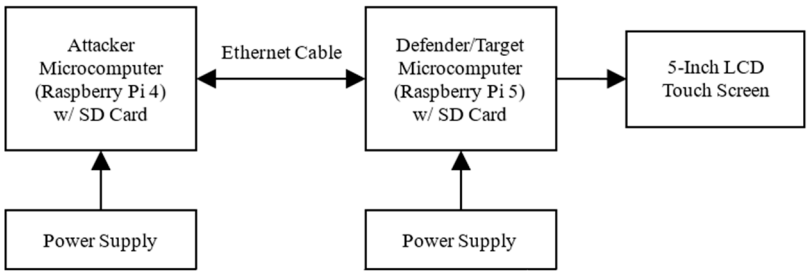


Figure 2. Hardware block diagram of the phishing detection and mitigation system.

2.3. Software

Figure 3 illustrates the software process of the system that begins with data ingestion and preprocessing, where phishing and legitimate emails are collected and cleaned into feature representations. A baseline model is trained for initial classification and evaluated for reliability. The MindRLHF v0.3.0 module then refines the model through reward-based proximal policy optimization over multiple iterations, integrating human feedback to improve decision-making. The trained RLHF model is deployed in the real-time inference module, where incoming emails are classified as phishing, legitimate, or uncertain, with confidence scores guiding decisions. When human feedback is provided, the model updates to reinforce learning. The system continuously monitors performance, retraining if thresholds are not met, and applies security measures to separate phishing from legitimate emails, ensuring adaptive, high-accuracy detection over time.

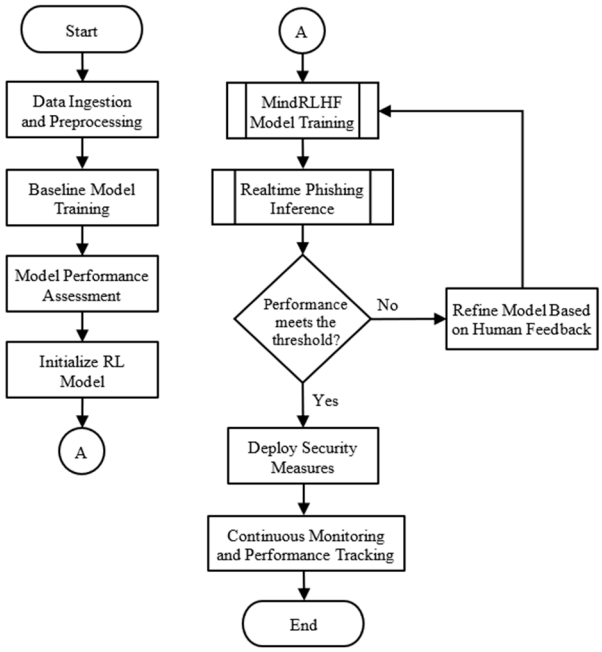
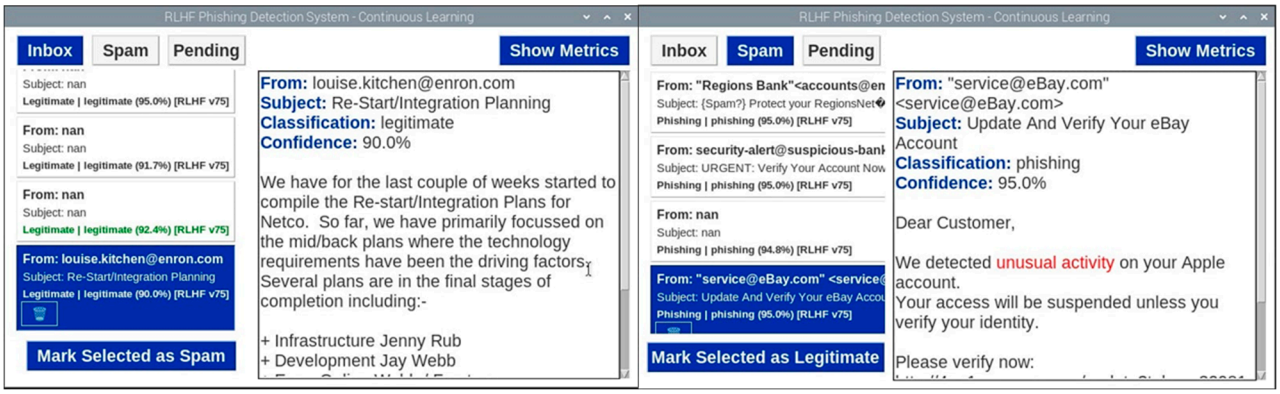


Figure 3. Software process.

2.4. Experimental Setup

Figure 4 illustrates the system setup, specifically the user interface in the casing that shows where phishing emails can be detected and mitigated. Inside the setup, two Raspberry Pi units are interconnected via Ethernet within the casing. The attacker node

(Pi 4) automatically sent email samples, while the defender node (Pi 5) classified them and displayed the results on the touchscreen interface. Users verified or corrected the outputs directly through the display, providing feedback to the RLHF model, which continuously adapted and improved detection accuracy through iterative learning.



(**a**) (**b**)

Figure 4. Experimental setup of the developed system: (**a**) display of legitimate email classification; (**b**) display of phishing email classification.

2.5. Data Gathering and Analysis

A total of 135,325 email samples were collected from multiple sources, including PhishTank, benign uniform resource locators, and the University of California Irvine datasets, comprising both legitimate and phishing instances. Key features such as HTML tags, phishing-related keywords, and suspicious domains were extracted for model training. During the RLHF phase, low-confidence samples were reintroduced and refined through human feedback, allowing iterative improvement in the model's classification accuracy and adaptability. System performance was evaluated using the following standard metrics to measure correctness.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Accuracy (1) measures the overall correctness of email classification, while precision (2) and recall (3) evaluate the system's ability to minimize false positives and false negatives, respectively. The F1 score (4) balances precision and recall, which is important for the imbalanced dataset. Together, these metrics establish benchmarks to ensure high detection rates and effective phishing mitigation.

For RLHF enhancement, the reward function (5) integrated classification correctness (C), human agreement (H), confidence matching (M), and consistency (S) with respective weights: $W_h = 3.0$, $W_c = 2.0$, $W_m = 1.0$, $W_s = 0.5$, to ensure comprehensive evaluation and balanced learning between automated metrics and human feedback.

$$R_{\text{Total}} = W_h H + W_c C + W_m M + W_s S \quad (5)$$

3. Results and Discussion

The baseline model achieved an overall accuracy of 94.3%, correctly classifying 12,223 legitimate and 13,298 phishing emails with minimal false classifications. The baseline model performance is illustrated in the confusion matrix in Table 1. After applying RLHF, the system achieved an overall accuracy of 96.8%, demonstrating enhanced reliability and adaptability in phishing detection. This improvement results from the RLHF reward function, which guided model updates by weighting correctness, human agreement, and prediction consistency.

Table 1. Confusion matrix for baseline model detection.

Predicted	Actual	
	Phishing	**Legitimate**
Phishing	13,298	769
Legitimate	775	12,223

Figure 5 shows the performance of the baseline supervised model and the RLHF-enhanced system. The baseline achieved a 94.3% accuracy, a 92.0% precision, and a 91.7% F1-score, on a training set (81,195 samples) and validated on the validation set (27,065 samples), forming a strong foundation but showing limitations in detecting subtle phishing patterns. After integrating human feedback samples, performance improved to show a 96.8% accuracy, a 95.0% precision, and a 94.7% F1-score, confirming that RLHF effectively refines classification boundaries and enhances adaptability.

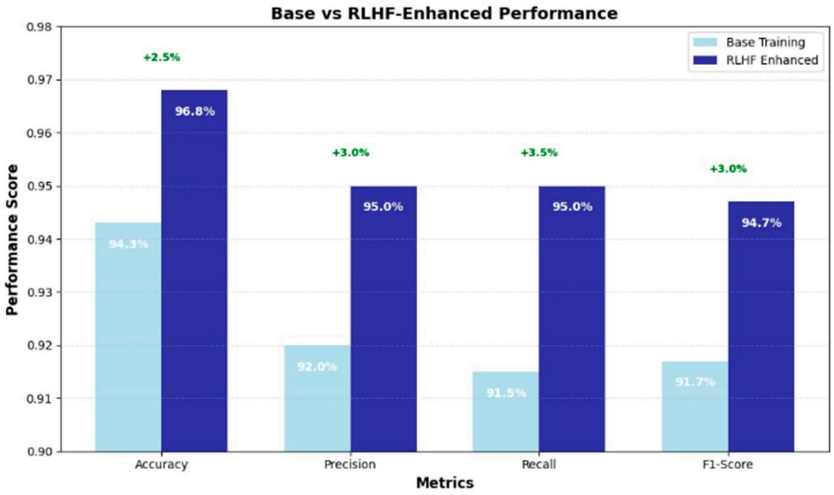


Figure 5. Comparison of model performance metrics.

Figure 6 shows the RLHF reward progression over multiple update iterations, improving from 5.984 to 6.561, a total gain of 0.577. The reward function comprised four components: correct classification (2.0), human agreement (3.0), confidence matching (1.0), and a consistency bonus (0.5), with the highest weight emphasizing human feedback. The progression exhibited rapid early gains, steady mid-phase improvement, and final fine-tuning, indicating effective alignment with feedback.

The results demonstrate that integrating RLHF significantly enhances phishing detection performance compared with conventional supervised models. The reward-driven feedback mechanism not only improved classification accuracy but also strengthened the model's adaptability to new and evolving phishing patterns. These outcomes validate the effectiveness of RLHF in developing intelligent, human-aligned cybersecurity systems capable of maintaining high reliability in real-world deployment.

https://doi.org/10.3390/engproc2026134013

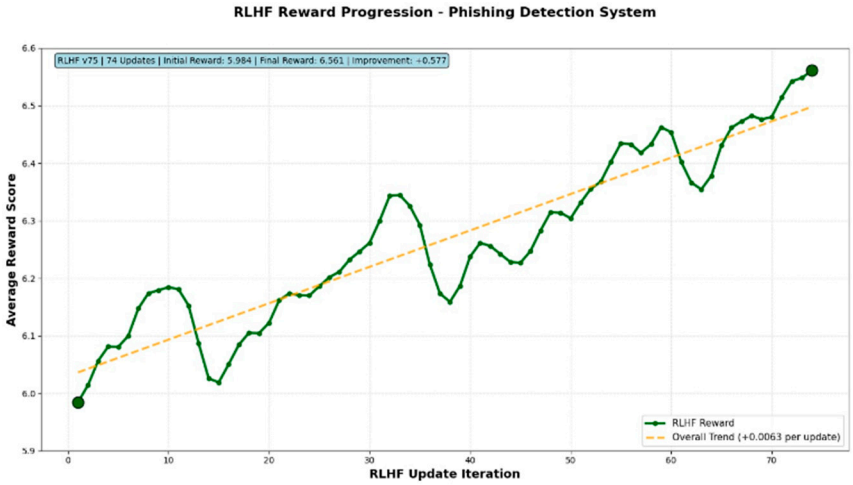


Figure 6. RLHF reward progression through training.

4. Conclusions and Recommendation

We developed a phishing detection system integrating supervised learning with RLHF, implemented on Huawei MindSpore and deployed on Raspberry Pi hardware. The system was trained with an RL model with human feedback, developing a real-time hardware interface. The RLHF-enhanced model improved adaptability and accuracy, reaching 96.8% with balanced precision and recall, while effectively reducing false positives and negatives. The datasets need to include diverse phishing types, languages, and attack patterns. Advanced RL methods, such as actor–critic or hybrid deep models, also must be integrated to enhance scalability and resilience against evolving cyber threats.

Author Contributions: Conceptualization, J.I.B.H. and M.D.S.O.; methodology, M.D.S.O.; software, M.D.S.O.; validation, J.I.B.H., M.D.S.O. and D.A.P.; investigation, J.I.B.H.; data curation, M.D.S.O.; writing—original draft preparation, J.I.B.H.; writing—review and editing, J.I.B.H., M.D.S.O. and D.A.P.; visualization, J.I.B.H. and M.D.S.O.; supervision, D.A.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset utilized in the study is not publicly available but is available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mohamed, N. Current trends in AI and ML for cybersecurity: A state-of-the-art survey. *Cogent Eng.* **2023**, *10*, 2272358. [CrossRef]
2. Karki, S.; Hasan, A.B.M.M.; Sanin, C. Use of ML and AI in Cybersecurity-A survey. In *Procedia Computer Science*; Elsevier B.V.: Amsterdam, The Netherlands, 2024; pp. 1260–1270. [CrossRef]
3. Aiyanyo, I.D.; Samuel, H.; Lim, H. A systematic review of defensive and offensive cybersecurity with machine learning. *Appl. Sci.* **2020**, *10*, 5811. [CrossRef]
4. Yumang, A.N.; Dimaunahan, E.D.; Lazaro, J.B.; Marinas, J.L.T.; Logatoc, J.E.G. Encryption and decryption of vital signs information through a symmetric based cryptography algorithm. In *ACM International Conference Proceeding Series, Association for Computing Machinery, Tokyo, Japan, 15–18 September 2020*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 218–223. [CrossRef]
5. Yumang, A.N.; Dimaunahan, E.D.; Centino, C.K.M.; Doroteo, A.R.J. IoT-based fire mitigation and detection system with AES-256 encryption and android application. In *2023 2nd International Symposium on Sensor Technology and Control, ISSTC 2023*; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2023; pp. 201–206. [CrossRef]

6. Paracha, M.A.; Jamil, S.U.; Shahzad, K.; Khan, M.A.; Rasheed, A. Leveraging AI for network threat detection—A conceptual overview. *Electronics* **2024**, *13*, 4611. [CrossRef]
7. Padilla, D.A.; Fernandez, B.D.P.; Del Rosario, V.I. A distributed training approach on email spam classification using DistilBERT. In *Proceedings of the 2024 7th International Conference on Information and Computer Technologies, ICICT 2024, Honolulu, HI, USA, 15–17 March 2024*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2024; pp. 139–144. [CrossRef]
8. Del Rosario, V.I.; Fernandez, B.D.P.; Padilla, D.A. Email spam classification using DistilBERT. In *Proceedings of the 2023 IEEE 15th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management, HNICEM, Coron, Palawan, Philippines, 19–23 November 2023*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2023. [CrossRef]
9. Ang, M.C.; Taguibao, K.R.C.; Manlises, C.O. Hand gesture recognition for Filipino sign language under different backgrounds. In *Proceedings of the 4th IEEE International Conference on Artificial Intelligence in Engineering and Technology, IICAIET, Kota Kinabalu, Malaysia, 13–15 September 2022*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2022. [CrossRef]
10. Yumang, A.N.; Bautista, F.P.F.; Labarentos, M.S.; Villegas, J.M.J.; Linsangan, N.B.; Pellegrino, R.V. Human detection system using image differencing with email notification system. In *Proceedings of the ACM International Conference Proceeding Series*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 221–225. [CrossRef]
11. Becina, M.D.; Padilla, D.A. Abnormal behavior detection using object detection and tracking. In *Proceedings of the 2022 IEEE 14th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management, (HNICEM), Boracay Island, Philippines, 1–4 December 2022*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2022. [CrossRef]
12. Hernandez, J.D.R.; Mendoza, C.R.O.; Linsangan, N.B. Identifying forged E-signatures using convolutional neural network. In *Proceedings of the 2024 IEEE International Conference on Automatic Control and Intelligent Systems, I2CACIS 2024—Proceedings, Shah Alam, Malaysia, 29 June 2024*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2024; pp. 216–221. [CrossRef]
13. Bascara, F.C.A.; Yumang, A.N. Defect detection and classification of soybean using convolutional neural network. In *Proceedings of the 2024 7th International Conference on Information and Computer Technologies, ICICT, Honolulu, HI, USA, 15–17 March 2024*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2024; pp. 265–270. [CrossRef]
14. Ebora, J.G.O.; Español, J.C.N.; Padilla, D.A. Text classification of Facebook messages using multiclass support vector machine. In *Proceedings of the 2022 13th International Conference on Computing Communication and Networking Technologies, ICCCNT, Kharagpur, India, 3–5 October 2022*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2022. [CrossRef]
15. Kheddar, H.; Dawoud, D.W.; Awad, A.I.; Himeur, Y.; Khan, M.K. Reinforcement-learning-based intrusion detection in communication networks: A review. *IEEE Commun. Surv. Tutor.* **2024**, *27*, 2420–2469. [CrossRef]
16. Sujatha, V.; Prasanna, K.L.; Niharika, K.; Charishma, V.; Sai, K.B. Network intrusion detection using deep reinforcement learning. In *Proceedings of the 7th International Conference on Computing Methodologies and Communication, ICCMC, Erode, India, 23–25 February 2023*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2023; pp. 1146–1150. [CrossRef]
17. James, G.; Abraham, C.; Dipte, S.; Gaat, A.; Siddiqui, A. Distributed denial of service attack mitigation using reinforcement learning. In *Proceedings of the 1st International Conference on Electronics, Communication and Signal Processing, ICECSP, New Delhi, India, 8–10 August 2024*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2024. [CrossRef]
18. Zhang, Q.; Zhao, Y.B.; Kang, Y. Autonomous boundary of human-machine collaboration system based on reinforcement learning. In *Proceedings of the 2020 Australian and New Zealand Control Conference, ANZCC, Gold Coast, QLD, Australia, 26–27 November 2020*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2020; pp. 160–165. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.