Intelligent Automated Penetration Testing Using Reinforcement Learning to Improve the Efficiency and Effectiveness of Penetration Testing

Mohammed Y. A. Aqra(✉) and Xiaoqiang Di

Changchun University of Science and Technology, Changchun, China
mohammed.abokhadeje@gmail.com, dixiaoqiang@cust.edu.cn

Abstract. A penetration test is a process that involves planning, generating, and evaluating attacks that are designed to find and exploit vulnerabilities in digital assets. It can be used in large networks to evaluate the security of their infrastructure. Despite the use of automated tools, it can still be very time consuming and repetitive. The goal of this paper is to develop an intelligent automated penetration testing framework that uses reinforcement learning to improve the efficiency and effectiveness of penetration testing. It utilizes a model-based approach to automate the sequential decision-making process. The framework's main component is a partial observed Markov decision process that is solved using an external algorithm.

One of the biggest challenges in performing penetration tests on large networks is finding and evaluating clusters of vulnerabilities. This paper presents a method that combines a hierarchical network model with a cluster-based approach. It allows for faster and more accurate testing compared to previous methods. The results of the study show that the IAPTF method outperforms other approaches in terms of time, accuracy, and human performance. One of the main advantages of IAPTF is its ability to perform repetitive tests, which is typically not possible with traditional methods. This method could potentially replace manual pen testing.

Keywords: Machin learning · Deep Reinforcement Learning · IAPTS · HRL-GIP

1 Introduction

The rapid emergence and evolution of new wireless communication systems has caused a change in the way they work. The 5G and 6G technologies have already been deployed in various countries. With the increasing number of people using smartphones, the need for more effective and efficient wireless communication has also become more prevalent [1]. The 6G wireless networks are designed to provide high-speed and reliable connectivity. They need advanced network devices and techniques to meet their various requirements. In recent years, the literature has suggested that the main ideas for 6G are the use of

R. Li et al. (Eds.): MOBILWARE 2022, LNICST 507, pp. 30–42, 2023.
https://doi.org/10.1007/978-3-031-34497-8_3

terahertz communication, intelligent reflecting surfaces, and reconfigurable intelligent surfaces [2].

The IRS is a new type of wireless communication technology that is expected to be used in the 6G era. It is a thin metasurface that can reflect electromagnetic waves in an organized manner. This technology can be used to mitigate multi-path problems and is ideal for terahertz and millimeter wave communication [3]. The increasing popularity of IRS devices has led to the development of new technologies that allow them to provide high-speed wireless communication. Some of these include the ability to perform multiple tasks at the same time, such as power transfer and mobile edge computing. In addition, the use of mmWave technology has allowed for the development of unmanned aerial vehicles (UAVs) and smart cities [4].

The increasing popularity of IRS devices has led to the development of new technologies that allow them to provide high-speed wireless communication. Some of these include the ability to perform multiple tasks at the same time, such as power transfer and mobile edge computing [5]. In addition, the use of mmWave technology has allowed for the development of unmanned aerial vehicles (UAVs) and smart cities. Several studies have been conducted on the use of IRS in the design of wireless communication systems that are designed for different performance matrices. These studies are categorized into four main areas: cascade channel estimation, multiuser communication, phase shift optimization, and beamforming optimization [6].

Machine learning techniques are a revolutionary technology that can be used in the development of artificial intelligence (AI) systems. They are capable of learning from vast amounts of data and making predictions for the future [7]. Due to their capabilities, various ML techniques have been widely adopted in the design of wireless communication systems. The various types of deep learning techniques that are commonly used in IRS are labeled as deep learning, reinforcement learning, unsupervised learning, and federated learning. According to the studies conducted on the use of IRS, the advantages of using ML techniques are compared with the conventional methods [8].

The ability to use deep learning techniques in the design of wireless communication systems has been greatly improved by the use of ML. This technology can help improve the accuracy of the channel estimation process by extracting the relationship between the input and output signals. In particular, the use of DL-based IRS in the design of Massive MIMO and OFDM systems has been shown to be more reliable. In [9] study, the authors proposed a DL-based channel estimation method that can be used in the design of an IRS-based system. However, this method could cause issues due to the error propagation when estimating the direct channel. In another study, the authors proposed a method that can be used to estimate the cascaded channel using a convolutional neural network. The literature also covered the various techniques and technologies that are used in the design of IRS-based systems. In [10] study, the authors analyzed the performance of a multi-antenna assisted IRS system. They then conducted a performance analysis of the various techniques involved in the design of the system.

ML techniques are widely used in the design of wireless communication systems due to their capabilities. They can help solve various problems related to the design of wireless communication systems. A study conducted on the use of IRS-assisted technology in a wireless communication system revealed the state-of-the-art survey of the

technology. The authors used microstrip patch antennas and a variety of antennas to study the operation principle of an IRS-based system. They also covered the properties of metasurfaces and their reflections [11].

The literature additionally covered the various techniques and technologies that are used in the design of IRS-based systems. In addition, the authors conducted a survey to learn more about the various aspects of the technology. They also discussed the multiple concepts and features of the channel model. The study additionally conducted a survey to learn more about the implementation structure of the IRS. The literature also covered the various techniques and technologies that are used in the design of IRS-based systems. One of the most important factors that can be considered when it comes to the development of an IRS-based system is the integration of multiple technologies.

1.1 Problem Statement

One of the most challenging issues in the development of machine learning systems is selecting the best method to automate the Agent Pen process. In Pozdniakov, et al., 2020, they present a method that avoids the use of pre-defined models by making the learning algorithm completely model-free. This approach significantly improves the performance of the system while still addressing the quality of attack sequences. Despite the increasing number of attackers and the complexity of the threat landscape, more automation is needed to effectively pentest systems. One of the most challenging factors that can be addressed is the accuracy of the results. To achieve this, they perform a variety of attacks against different targets, including Linux and Windows systems.

One of the most important factors that can be considered when it comes to developing an approximator is the number of features that will represent the state-action space of a given attack. This is because, while building an approximator, it is important to choose the most appropriate features for the specific attack. Another important factor that can be considered is the size of the feature set that will represent the state-action space. To achieve this, they use a recurrent neural network (RNN) that is designed to perform a variety of tasks, such as identifying and validating vulnerabilities. Unfortunately, too many features can negatively affect the performance of the learning algorithm and lead to time-consuming attacks. To address this issue, they have created an autoencoder that can reduce the number of features in the learning algorithm.

1.2 Aim of the Study

This study aims to focus on the use of machine learning (ML) in the field of PT practice to make it more intelligent and efficient. It can be extended to include the use of web and application testing. The study aims to introduce the importance of machine learning in the field of PT practice. It reviews various studies and surveys related to the subject. The last section of the study will introduce the proposed RL model and its various components. The proposed system is called IAPTS and it is mainly used for testing the effectiveness of various techniques related to physical exertion testing. This study also describes the various steps involved in the testing and the results obtained.

2 Literature Review

2.1 Overview of Deep Reinforcement Learning

The goal of the policy is to define the agent's plan of action and how it relates to the environment. The reward values that the environment gives to the agent are the numerical representations of the state's intrinsic desirability. The value function is a long-term function that calculates a discounted return from a specific state after a policy has been followed. The environment model is a representation of the behavior that helps boost the performance of the algorithm [12].

In a supervised manner, DeepMind Alpha Go applied Monte Carlo tree search techniques to a dataset. However, instead of imitating human strategies, it did not see data from human games. Instead, it played a variety of games against itself to learn how to win a game that was not known to humans. An agent is a type of entity that is involved in reinforcement learning. The goal of a reinforcement learning algorithm is to enable an agent to learn quickly and accurately the optimal policy. This process can be done through the representation of the symbol p, which indicates that the goal is to achieve the highest reward value. The goal of a given action is to make the agent closer to its target. The reward RT 2- R determines the reward if the agent is closer or if it is farther away [11, 13].

Deep reinforcement learning is a subfield of machine learning that uses the concepts of deep learning to provide an optimal solution for an agent's experience. This process is carried out through iterations and evaluation of a reward function to determine the ideal behavior for an agent. There are three main approaches to deep reinforcement learning: the value-based approach, the policy-based approach, and the model-based approach.

The goal of value-based reinforcement learning is to enable an agent to find the optimal policy by calculating its value function in the long run. In contrast, policy-based reinforcement learning is focused on finding the optimal policy by implementing the rules that are related to the objective function [14]. The former approach refers to the same action regardless of the state. The latter, on the other hand, involves the use of probabilistic evaluation methods to evaluate the actions taken in a given state. For model-based reinforcement learning, the agent is required to provide a model of the environment that describes the tasks that it can perform in that environment.

It is very challenging to directly compare the performance of model-based and non-model-based reinforcement learning algorithms. In their book, Mackworth and Poole stated that model-based learners are more efficient than their counterparts. They also claimed that fewer experiences are required to learn well with model-free methods. On its own, an agent might face errors and inaccuracies when learning the environment model. This can affect its performance and prevent it from achieving the required tasks. There are various approaches that are designed to integrate the model-free methods with the model-based ones. These include the Monte Carlo method, the value-based method, the policy-based method, and the temporal-difference method [15].

Although the different approaches are implemented in different ways, they share some of the same characteristics. The control that a reinforcement learning algorithm provides is a closed-loop approach. The reward is represented by the system's feedback, and this is delayed by the algorithm. The algorithms used in deep learning are designed to

make decisions based on the sequence of actions taken. They also have long-term rewards that are dependent on the actions' duration. A concept known as credit assignment problem states that the system's dependence on time is due to the various actions that it performs [16, 17].

DRL is a framework that helps in the development of UAV control systems by providing a variety of tools that allow them to work with model-free algorithms. These include the ability to learn online how to perform the target without being trained, and to work in environments that are not familiar to the UAV. With the development of deep learning algorithms, there are now many new tasks that can be performed in the control area. For instance, controlling a swarm of UAVs with minimal resources can be performed safely [18].

2.2 IRS Hardware Architecture and Its Working Principle

The IRS hardware is based on the meta-surface, which is a two-dimensional meta-material. It can be used in the subwavelength frequency range with a large number of atomic layers. This is done through the use of a large number of meta-atoms. The signal response of an IRS Meta surface atom element can be changed by designing it according to the specific requirements of the wireless communication system. For instance, the size, shape, and orientation of the element can be changed to accommodate the dynamic wireless channels generated by the user's mobility [19]. The three main approaches to controlling IRS mechanical reflection are presented in the literature. These include the use of functional materials, electronic devices, and mechanical translation. Examples of electronic devices that are commonly used for this process include field-effect transistor (FETs), PIN diodes, and micro-electromechanical systems (MEMS) [19].

Due to their low energy cost and fast response time, electronic devices are widely used in today's society. In addition, they are also very advantageous for reducing the cost of designing and implementing IRS elements. To achieve the best possible performance, the IRS elements should be linked to a network to acquire knowledge about the environment [20].

This layer is composed of a copper plate, which is used to prevent the signal energy leakage during the IRS's reflection. A control circuit board then excites the element, which then changes the reflection amplitude and phase shift. The IRS smart controller then activates the reflection adaption. The IRS architecture allows the controllers to communicate with each other and with the various network parameters of a network, such as access points and base stations. Through a field-programmable gate array, the controllers can also control the flow of information between the multiple devices [21].

An example of this process is the reduction of the incident energy of a signal by the variation in the resistance of the IRS elements. This allows a scaling of the reflection amplitude to an effective level. This is similar to the work of a radio frequency identification tag, which can regulate the power strength of a signal by changing its load impedance. One of the most important factors that the IRS should consider when it comes to optimizing its reflection design is the availability of independent control over the phase and amplitude of the element. This can be done by implementing a more intelligent hardware design [22].

2.3 IRS Reflection

The channel model for wireless communication based on the IRS is composed of three components. These are the channel link between the IRS elements and the BS, the reflection of the IRS elements, and the channel between the two entities. In this type of channel, the characteristics of the signal are different from those of a traditional direct channel. The IRS utilizes a BS-based approach to receive and distribute superposed multi-path signals. Each element of the device receives and distributes the associated signal with an amplitude and phase that are ideal for a specific channel model [21].

2.4 IAPTS Memory, Expertise Management and Pre-Processing

The goal of this research is to develop a framework that will allow a large and medium-sized LAN to be equipped with a reliable and secure RL learning system. This will require dealing with a huge amount of intimations, which is very complex and can be difficult to solve given the limited computing power and time available. One of the most challenging aspects of this project is the modeling and simulation of the PT as a POMDP environment [1].

Resource management is a must in order to ensure that the resources are used efficiently. The system memory is used to store the data that the system processes. These include the environment attributes, which are used to define the policies and actions that an agent can perform within the environment, and the agent's memory, which is used to acquire knowledge and experiences. The first part of this study will focus on searching a policy as an agent acts in a certain environment [3]. The reward value will then be determined by a human expert. This method will allow researchers to conduct a research facilitation. This will allow IAPTS to adapt and adjust the tests that it deploys after a successful exploitation. It will also allow it to perform post-exploitation tasks such as privilege escalation and pivot [6].

This module will also allow researchers to develop a deeper understanding of the system's operations and procedures in order to provide it with the necessary "expert knowledge." In practical terms, this will allow IAPTS to perform various tasks such as extracting PGs and executing the testing plan. It will additionally allow it to keep track of the results of the tests and update its status in real time [7].A parallel knowledge-based system will also be implemented as part of the framework. This system will capture the details of the actions performed by the human expert. It will additionally extract knowledge from the data that the security system collects. This method will allow the system to perform post-exploitation tasks such as privilege escalation and pivot [1].

Despite the potential of artificial intelligence, it is still not able to model the intuition of humans. This issue will be solved by allowing the human expert to interact with the system. A mechanism that will allow security analysts to provide feedback will also be implemented to solve this issue[2].The feedback collected during the test will be stored in the system memory, which will then be used for future use. This method will allow the system to perform post-exploitation tasks such as privilege escalation and pivot. In addition to the human expertise, the memory will also feature a variety of features and workflows, such as the implementation of direct learning [5].

One of the most effective ways to improve the learning performance of the RL algorithms is by implementing a prioritized experience replay method. This method was initially implemented to enhance the learning capabilities of the system. However, in order to accommodate the technical requirements of the system, we modified the method. In addition to choosing the most relevant and plausible policies, we also injected other transition sequences that were validated by a human expert. These sequences allow the system to retrieve value function information from the previous tests and improve its efficiency. The resulting sequences can be replayed to reduce the number of POMDP observations and transitions, which improves the algorithm's efficiency [6].

In addition to the human expertise, the system will also feature a variety of features and workflows, such as the implementation of direct learning. One of the most important factors that will be introduced in this project is the GIP LP Solve algorithm, which was designed to improve the performance of IAPTS. This method will allow the system to capture the necessary expertise in form of a decision policy. Another effective way to demonstrate the importance of learning is by implementing a test scenario that is inspired by the real-world experience of re-testing a network after some upgrades or updates. This scenario will allow the system to perform multiple tasks such as extracting PGs. After the tests are performed, the system will use the previous generation of PGs as an initial belief in the output [7].

3 Testing the Model

The decision to embed RL within the practice of PT was based on its relevance and suitability. Its implementation also took into account the various decision-making problems that are involved in the process. The second challenge was to choose between two models: one that is model-free and another that is model-based. The concept of PT practice is that it involves choosing a decision-making task that is related to the relevant decision policy. This strategy helps minimize the number of iterations that the user has to perform in order to achieve the best possible outcome. In addition, the quality of the solution is related to the relevance of the decision policy. The first step in the process of solving is to identify the appropriate approach. In this paper, the researchers discussed the various advantages of using a model-free RL and the two options that are available for implementing this strategy: policy search and value iteration.

The next step was to choose between the two models. The decision to use policy search was made after considering the various advantages of the strategy. This strategy is supported by the initial goal of optimizing the process and fully automating the decision-making tasks within the PT practice. We then build a decision tree from the POMDP solution's output. After the selection of the appropriate model and the corresponding technique, the next step was to choose the appropriate solution modes. This process is carried out to enhance the efficiency and accuracy of the decision-making process. The goal of this work is to provide an intelligent automation of the tasks within the PT practice.

The paper also discussed the various advantages of using a comprehensive framework for developing and testing both the model-free and the model-based solutions. This approach allows the developers to start with the approximate mode and test both the proposed and the actual solution. After selecting the best solution, PERSEUS was the first algorithm to be implemented during the early stages of the project. Due to the importance of accuracy and efficiency, many exact solving algorithms were considered for inclusion in IAPTF. However, after assessing the various advantages of GIP, it was concluded that it was the most efficient candidate. To ensure that the solution is flexible and can handle different input types, multiple modifications were made to the GIP and PERSEUS versions.

The researchers tested and implemented the exact and approximate methods of solving the POMDP problem. The former method involves optimizing the value function of the model over all possible belief's states, while the latter method is more computational. For most applications, exact solutions are very challenging to perform due to their complexity. The exact solution that is used for calculating the value function of POMDPs is very cost-effective as it allows the user to perform the optimal action on any belief state. However, the computational power required to implement this algorithm is very high due to the exponential growth of the belief space. The approximate method is more inexact because it relies on a single choice of belief states. In this paper, the researchers will use the method known as Point-Based Value Iteration to perform a comparison and guidance exercise. It takes into account the initial belief and then picks the belief points according to a forward simulation.

The exact solution provided by the PBVI algorithm is an approximate representation of the value of the model that takes into account the various belief points in the model. It then ensures that the value function increases with every iteration. In practice, this method is very useful for solving large problems that involve multiple belief points. The PBVI algorithm also maintains a single value per point, which helps it to successfully solve large problems. In IAPTS, we implemented the PERSEUS algorithm, which performs a randomized point-based value optimization process on each of the POMDPs. It ensures that the value of the points in the belief set is improved in each backup stage. The Perseus algorithm also performs backup stages depending on the convergence criteria and the number of policy changes.

The exact solution provided by the PBVI algorithm is an approximate representation of the value of the model that takes into account the various belief points in the model. It then ensures that the value function increases with every subsequent belief. However, the computational power required to implement this algorithm is very high due to the complexity of the problem. To minimize the time required to implement this algorithm, the researchers decided to use a variety of algorithms that use a combination of linear programs and a prune-dominated vector.

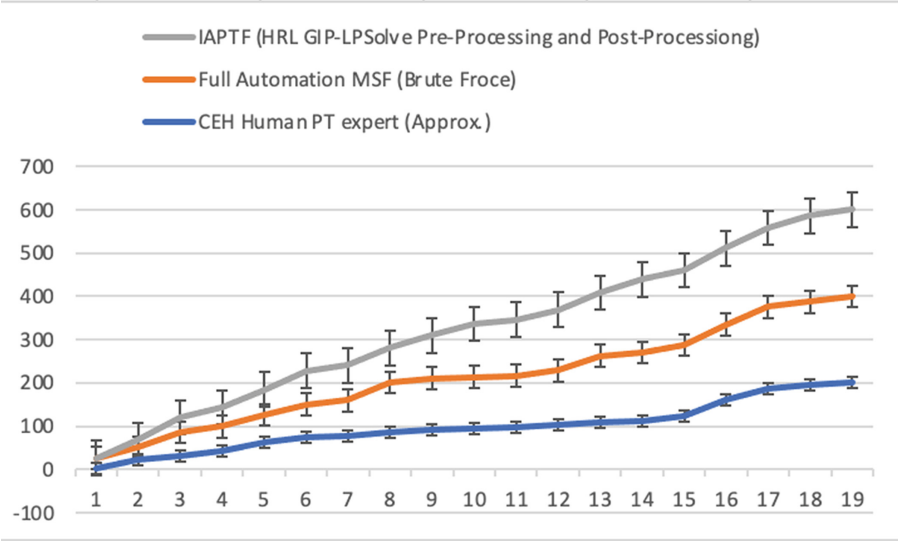

Fig. 1. Using different algorithm and belief handling techniques, this can solve different size POMDP problems.

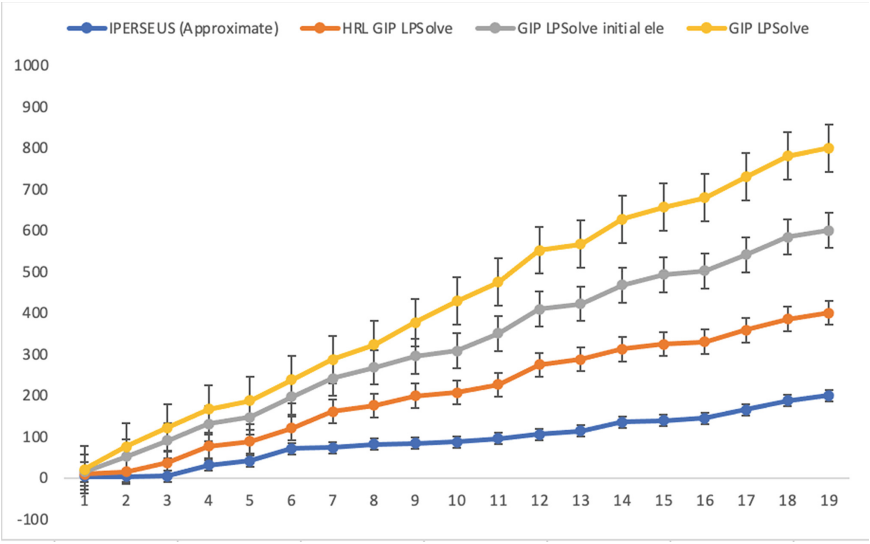

Fig. 2. Re-test the same network. After introducing a significant number of changes to the network configurations

4 Results and Discussion

The HRL approach requires around two days to perform compared with the regular RL-GIP. Going beyond the 100-machine size, HRL is more efficient and can reach 200 machine size. It performed almost as well as the PERSEUS algorithm. In addition, HRL-GIP can also be significantly faster than the previous generation. The HRL-GIP effect is widely used in large networks to perform effective value optimization. It can reach a good rate in 100-machine networks with over 20 clusters.

Security clustering often results in a large number of security clusters. This means that many small POMDPs are placed on top of these security clusters. This can result in IAPTF having to perform a lot of data manipulation and optimization to solve the problem. Using a model-free RL method is very advantageous for reducing the time required to implement the algorithm.

The goal of evaluating IAPTS is to provide a comprehensive analysis of the various aspects of the process, starting with validating the approach and examining the results in real-world situations. After that, the evaluation process continues with analyzing the accuracy and relevance of the results. The output of the program is then converted into a more understandable format by implementing policies graphs.

In addition to the time, it takes to solve the POMDP problem, other factors such as the number of tasks that need to be performed by the Metasploit MSF will also be taken into account to arrive at the overall time that IAPTS will spend testing on the test-bed networks. The results of the study presented in Fig. 1 and Fig. 2 show that IAPTS outperforms both the manual and automated versions of PT when it comes to solving algorithms on different LANs. The authors' experience as a consultant for PT also contributed to the findings. It is additionally clear that IAPTS is more cost-effective than both manual and automated versions of PT. In addition, the various discount rates that were considered during the development of IAPTS were aimed at preserving the realistic nature of the system.

Following the multiple simulations and testing, the discount rate for "0.5" was selected. We then decided to introduce some changes in the algorithm to improve its performance. One of these is the use of a more short-term approach to improve the performance of IAPTS. The researchers also decided to prioritize the observations and transitions through the use of the associated probabilities.

The results of the evaluation process were very good, with the new GIP-LP Solve achieving better performance than the previous generation. In addition, the improved performance of this variant was also evidenced by the significant time savings that it made in both the PG accuracy and time consumed. To further improve the performance of the program, we additionally re-tested the same network with or without changing the machine configurations.

The results of the tests were very impressive, especially when re-testing the same network that was shown in Fig. 2 The quality of the decisions that were produced by IAPTS was also beyond human expertise. In particular, when using the GIP LP Solve initial belief algorithm, the program was able to produce decisions that were more accurate than those that were produced by a human expert.

4.1 Simulation

Each performance was repeated five times with random seeds to ensure that the results were reproducible. The shading regions on the graph represent the variability in the runs. After training the agents, we are looking at two indicators: the number of steps that the agents have to take to reach their target and the maximum score that they can achieve. The HA-DRL algorithm can give an agent around 20 points depending on the scenario. It can also reduce the score by a small amount due to invalid actions it takes to reach the assets. We have tested the algorithm on various scenarios with different action space configurations.

The complexity of a scenario can increase the number of hosts and the number of actions in the action space. However, with only 2 additional agents to train, the number of actions in the action space can only increase from 49 to 4646. In all the tested scenarios, the HA-DRL convergence was superior to the IAPTS agent. In these scenarios, the complexity of the algorithm was significant enough to show its superiority, while the cost of training was not too high. Compared to the previous generation, the HA-DRL algorithm performed better in learning the optimal policy.

The performance of the IRS algorithm on different scenarios was unstable, especially when it was training with 20 and 30 hosts. It only managed to learn the optimal policy in one out of four runs. In the 50 hosts' scenario, the HA-DRL performed well, as it was able to train the agents successfully in one out of four runs. However, it was not able to get the most out of the training experience, as it took around 2,000 episodes to get the agents up to speed. The RIS algorithm was able to achieve the optimal number of actions in each scenario. This is the first demonstration of how deep learning can be used in an automatic testing system to handle large action pools.

5 Conclusions

This paper presents an approach to embedding RL techniques into the cyber security domain. We use a hierarchical RL representation to address the complexity of the PT domain. This method overcomes the scaling-up challenges that were encountered in addressing large POMDPs on networks with many nodes. The proposed approach involves segregating the network into small clusters and then processing the network's various attacking vectors in a way that is more effective than that of certified ethical hackers. This method can be used to deal with different types of networks and their complexity. The proposed IAPTF framework is a versatile and comprehensive approach that enables human experts to perform more complex tasks without requiring them to spend a lot of time. It can also reveal unexpected combinations that are typically ignored in manual testing.

References

1. Nomikos, N., Zoupanos, S., Charalambous, T., Krikidis, I.: A survey on reinforcement learning-aided caching in heterogeneous mobile edge networks. IEEE Access **10**, 4380–4413 (2022). https://doi.org/10.1109/ACCESS.2022.3140719

2. Rasheed, F., Yau, K.L.A., Noor, R.M., Wu, C., Low, Y.C.: Deep reinforcement learning for traffic signal control: a review. IEEE Access **8**, 208016–208044 (2020). https://doi.org/10.1109/ACCESS.2020.3034141
3. Park, D.Y., Lee, K.H.: Practical algorithmic trading using state representation learning and imitative reinforcement learning. IEEE Access **9**, 152310–152321 (2021). https://doi.org/10.1109/ACCESS.2021.3127209
4. Liu, D., Wang, Z., Lu, B., Cong, M., Yu, H., Zou, Q.: A reinforcement learning-based framework for robot manipulation skill acquisition. IEEE Access **8**, 108429–108437 (2020). https://doi.org/10.1109/ACCESS.2020.3001130
5. Liu, C.L., Chang, C.C., Tseng, C.J.: Actor-critic deep reinforcement learning for solving job shop scheduling problems. IEEE Access **8**, 71752–71762 (2020). https://doi.org/10.1109/ACCESS.2020.2987820
6. Li, M.L., Chen, S., Chen, J.: Adaptive learning: a new decentralized reinforcement learning approach for cooperative multiagent systems. IEEE Access **8**, 99404–99421 (2020). https://doi.org/10.1109/ACCESS.2020.2997899
7. Rothmann, M., Porrmann, M.: A survey of domain-specific architectures for reinforcement learning. IEEE Access **10**, 13753–13767 (2022). https://doi.org/10.1109/ACCESS.2022.3146518
8. Gasperov, B., Kostanjcar, Z.: Market making with signals through deep reinforcement learning. IEEE Access **9**, 61611–61622 (2021). https://doi.org/10.1109/ACCESS.2021.3074782
9. Lee, H., Cha, S.W.: Reinforcement learning based on equivalent consumption minimization strategy for optimal control of hybrid electric vehicles. IEEE Access **9**, 860–871 (2021). https://doi.org/10.1109/ACCESS.2020.3047497
10. Zhang, Q., Lin, J., Sha, Q., He, B., Li, G.: Deep interactive reinforcement learning for path following of autonomous underwater vehicle. IEEE Access **8**, 24258–24268 (2020). https://doi.org/10.1109/ACCESS.2020.2970433
11. Lee, H., Song, C., Kim, N., Cha, S.W.: Comparative analysis of energy management strategies for HEV: dynamic programming and reinforcement learning. IEEE Access **8**, 67112–67123 (2020). https://doi.org/10.1109/ACCESS.2020.2986373
12. Alharin, A., Doan, T.N., Sartipi, M.: Reinforcement learning interpretation methods: a survey. IEEE Access **8**, 171058–171077 (2020). https://doi.org/10.1109/ACCESS.2020.3023394
13. Hu, Y., Hua, Y., Liu, W., Zhu, J.: Reward shaping based federated reinforcement learning. IEEE Access **9**, 67259–67267 (2021). https://doi.org/10.1109/ACCESS.2021.3074221
14. Kubalik, J., Derner, E., Zegklitz, J., Babuska, R.: Symbolic regression methods for reinforcement learning. IEEE Access **9**, 139697–139711 (2021). https://doi.org/10.1109/ACCESS.2021.3119000
15. Chen, S.Y.C., Yang, C.H.H., Qi, J., Chen, P.Y., Ma, X., Goan, H.S.: Variational quantum circuits for deep reinforcement learning. IEEE Access **8**, 141007–141024 (2020). https://doi.org/10.1109/ACCESS.2020.3010470
16. Elavarasan, D., Durairaj Vincent, P.M.: Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. IEEE Access **8**, 86886–86901 (2020). https://doi.org/10.1109/ACCESS.2020.2992480
17. Mohammed, M.Q., Chung, K.L., Chyi, C.S.: Review of deep reinforcement learning-based object grasping: techniques, open challenges, and recommendations. IEEE Access **8**, 178450–178481 (2020). https://doi.org/10.1109/ACCESS.2020.3027923
18. Jin, M., Lavaei, J.: Stability-certified reinforcement learning: a control-theoretic perspective. IEEE Access **8**, 229086–229100 (2020). https://doi.org/10.1109/ACCESS.2020.3045114
19. Green, S.A., et al.: Mapping mental health service access: achieving equity through quality improvement. J. Public Health **35**(2), 286–292 (2013)

20. Thomson, L.J., Camic, P.M., Chaterjee, H.J.: Social Prescribing: A Review of Community Referral Schemes. University College London, London (2015)
21. Shepherd, M., Butler, L.: The underuse of couple therapy for depression in improving access to psychological therapies services (IAPTS): a service evaluation exploring its effectiveness and discussion of systemic barriers to its implementation. J. Family Therapy **43**(4), 493–515 (2021). https://doi.org/10.1111/1467-6427.12323
22. Ghanem, M.C., Chen, T.M.: Reinforcement learning for efficient network penetration testing. Information **11**(1), 6 (2020). https://doi.org/10.3390/info11010006