# AI in Penetration Testing: A Systematic Mapping Study

Sulaiman O. Alwabisi
*College of Computer & Info Sciences*
*King Saud Univerisity*
Riyadh, KSA
446100616@student.ksu.edu.sa

***Abstract*—The integration of Artificial Intelligence (AI) into penetration testing presents transformative opportunities for enhancing cybersecurity practices. This systematic mapping study investigates the current state of AI-driven penetration testing by reviewing 57 primary studies published between 2015 and 2025. Through a rigorous quality assessment process based on ten criteria, high-quality papers were selected for detailed analysis. The study is guided by four research questions: (RQ1) What are the key AI techniques currently utilized in penetration testing? (RQ2) What are the major challenges and limitations associated with integrating AI into penetration testing? (RQ3) How effective are AI-driven techniques in enhancing the penetration testing process? (RQ4) What are the current research trends, gaps, and future directions? Findings reveal that reinforcement learning (RL), deep learning (DL), generative AI models, and supervised learning are the most frequently adopted techniques. These approaches contribute significantly to automation, improved vulnerability detection, scalability, and optimization of penetration testing workflows. However, several challenges persist, including scalability limitations, training inefficiencies, integration difficulties, and ethical concerns related to AI bias and misuse. Emerging trends indicate growing interest in large language models (LLMs) and hierarchical reinforcement learning to address current limitations. This study highlights the evolving landscape of AI-assisted penetration testing and provides a roadmap for future research aimed at enhancing model effectiveness, real-world applicability, and ethical deployment. The insights gathered from this mapping study offer valuable guidance for researchers and practitioners seeking to advance automated and intelligent cybersecurity solutions.**

***Index Terms*—Artificial intelligence, Penetration Testing, Machine learning**

## I. Introduction

Penetration testing is a critical component of cybersecurity, serving as a proactive measure to identify and exploit vulnerabilities within systems, thereby enhancing their security posture. Its importance is underscored by the increasing frequency and sophistication of cyberattacks, which necessitate robust defenses to protect sensitive data and maintain operational integrity. Penetration testing simulates real-world attacks to uncover weaknesses that could be exploited by malicious actors, providing organizations with a clear understanding of their security gaps and the potential impact of these vulnerabilities if left unaddressed [1], [2]. This process not only identifies vulnerabilities but also tests the effectiveness of existing security measures, offering insights into how far an attacker could penetrate the system and what data could be compromised [2]. The results of penetration tests are typically compiled into detailed reports that include recommendations for mitigating identified risks, such as updating systems, correcting misconfigurations, and applying missing patches, thereby enabling organizations to prioritize and address security issues effectively [2], [3]. Furthermore, penetration testing is essential for compliance with various industry standards and regulations, which often mandate regular security assessments to ensure data protection and privacy [1]. The practice also plays a vital role in the Software Development Life Cycle (SDLC), ensuring that security is integrated from the early stages of application development, thus reducing the likelihood of security flaws in deployed applications [2]. Additionally, the advent of AI and machine learning has introduced new dimensions to penetration testing, allowing for more efficient and automated testing processes that can adapt to evolving threats and reduce the resource-intensive nature of traditional methods [3], [4] . By leveraging these technologies, penetration testing can become more frequent and comprehensive, addressing the challenges of complex and large-scale network environments [3]. Overall, penetration testing is indispensable for maintaining robust cybersecurity defenses, providing organizations with the necessary insights and tools to protect against potential breaches and ensure the resilience of their digital assets.

The importance of AI in penetration testing is underscored by its potential to enhance the efficiency and effectiveness of vulnerability assessments in increasingly complex systems. AI techniques, such as machine learning (ML), are being integrated into penetration testing to automate and improve the identification of vulnerabilities, which is crucial as systems become more sophisticated and adversarial techniques evolve [5]. The use of AI in this domain is not only about automating tasks but also about providing a more intelligent approach to vulnerability assessment, which can lead to better-informed mitigation strategies [5]. Moreover, AI can serve as a sparring partner for human testers, augmenting their capabilities and compensating for the shortage of skilled security professionals. This collaboration between AI and human testers can lead to new capabilities, enhancing the overall penetration testing process [6]. Large language models (LLMs), for instance, can assist in both high-level task planning and low-level vulnerability hunting, providing a closed-feedback loop that allows

for continuous improvement and adaptation to new threats [6]. Additionally, AI-driven approaches, such as the use of attack graphs and genetic algorithms, offer scalable solutions that can evolve over time to better fit the testing environment, thus improving the robustness of security assessments [5]. The integration of AI in penetration testing also opens up opportunities for creating standardized testbeds that can provide a common platform for evaluating and benchmarking different security approaches, further advancing the field [5]. Overall, AI's role in penetration testing is pivotal in addressing the challenges posed by modern cybersecurity threats, offering innovative solutions that enhance both the efficiency and effectiveness of security testing processes. The objectives of this study are to:

- Identify the key AI techniques utilized in penetration testing.
- Examine the major challenges and limitations in applying AI to penetration testing.
- Evaluate the effectiveness of AI-driven techniques in improving the penetration testing process.
- Highlight current research trends, gaps, and propose future directions for AI-enhanced penetration testing.

To achieve these objectives, the following research questions have been formulated:

- RQ1: What are the key artificial intelligence (AI) techniques currently utilized in penetration testing?
- RQ2: What are the major challenges and limitations associated with integrating AI into penetration testing?
- RQ3: How effective are AI-driven techniques in enhancing the penetration testing process?
- RQ4: What are the current research trends, gaps, and future directions for AI-enhanced penetration testing?

The remainder of this study is organized as follows: Section II provides background information on AI and penetration testing. Section III describes the methodology for selecting and analyzing relevant research papers. Section IV presents the results and discussion organized around research questions. Finally, Section V concludes the study and suggests directions for future work.

## II. Background About AI and Pentesting

This section provides the foundational knowledge necessary to understand the context and significance of applying Artificial Intelligence (AI) in penetration testing. It begins by introducing key concepts in cybersecurity and the role of penetration testing as a proactive security measure. The section then explores the fundamentals of AI, including major techniques and their relevance to security domains. By establishing a clear understanding of both penetration testing and AI, this background sets the stage for the systematic mapping and analysis conducted in the later sections of the study.

### A. Artificial Intelligence (AI)

Artificial Intelligence (AI) is a broad and multifaceted field that encompasses various technologies and methodologies aimed at enabling machines to perform tasks that typically require human intelligence. According to the European Commission JCR report, AI is defined as any machine or algorithm capable of observing its environment, learning from it, and taking intelligent actions or proposing decisions based on the knowledge and experience gained. This definition highlights the adaptability and learning capabilities of AI systems, which are central to their functionality [7] . AI is divided into several core scientific domains, including Reasoning, Planning, Learning, Communication, and Perception, each contributing to the overall capabilities of AI systems. For instance, the Learning domain involves systems that can automatically learn, decide, predict, and adapt without explicit programming, often utilizing machine learning techniques such as neural networks and reinforcement learning. Communication within AI primarily involves natural language processing (NLP), which enables machines to understand and generate human language, facilitating tasks like information extraction and text mining [7]. The application of AI in cybersecurity, particularly offensive AI, demonstrates its potential to enhance cyber-attacks by automating and conducting penetration testing, which traditionally requires significant manual effort and expertise [8]. AI's role in cybersecurity is not limited to offensive tactics; it also includes defensive strategies, such as using AI to automate penetration testing processes, thereby reducing time and costs while increasing efficiency [9]. Furthermore, AI's integration into software testing (ST) showcases its versatility, where it supports various testing activities, including test case generation and optimization, through techniques like genetic algorithms and NLP [7]. This integration underscores AI's potential to transform traditional engineering practices by providing innovative solutions across diverse fields, from cybersecurity to software testing, thereby enhancing efficiency and effectiveness in these domains.

Figure 1 illustrates the hierarchical relationship between Artificial Intelligence (AI), Machine Learning (ML), Deep Learning, and Generative AI. Each layer represents a specialization of the previous one, with AI being the broadest concept and Generative AI and Deep Learning representing more specific advancements within Machine Learning.

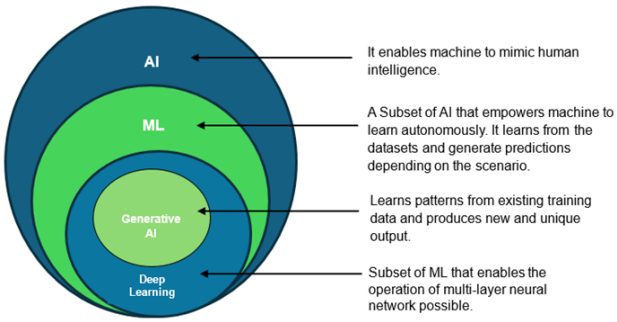


Fig. 1. Relationship between AI, ML, DL, and Generative AI [10]

### B. Penetration Testing

Penetration testing, often referred to as pentesting, is a critical process in cybersecurity aimed at evaluating the security of systems and networks by simulating attacks. This process helps identify vulnerabilities that could be exploited by malicious actors, thereby allowing organizations to strengthen their defenses. The pentesting process is typically structured into several key phases, each with specific objectives and methodologies.

**Penetration testing follows a structured process divided into key phases** that simulate real world cyberattacks. These phases include information gathering, exploitation, post exploitation, and reporting. Each phase plays a critical role in identifying vulnerabilities and assessing potential risks. The following outlines each phase and its role in the overall testing lifecycle.

- Information Gathering
  1) The initial phase of pentesting involves collecting data about the target system to identify potential vulnerabilities.
  2) Techniques such as traffic monitoring, port scanning, and operating system detection are commonly used to gather relevant information [11].
  3) This phase is crucial for understanding the system's architecture and identifying weak points that could be exploited in subsequent phases.
- Attack and Penetration
  1) Once vulnerabilities are identified, the next step is to exploit these weaknesses to gain unauthorized access to the system.
  2) This involves executing known exploits, which can be programs or specific data designed to take advantage of the discovered vulnerabilities [11].
  3) The goal is to compromise the target system and gain privileged access, which can then be used to further explore the network or system.
- Post-Exploitation
  1) After gaining access, the pentester may continue to explore the system to identify additional vulnerabilities or to establish a persistent presence.
  2) This phase can involve setting up backdoors or other means of maintaining access to the system [5].
  3) The focus is on understanding the full extent of the system's vulnerabilities and the potential impact of an attack.
- Reporting
  1) The final phase involves documenting all findings, including identified vulnerabilities, the methods used to exploit them, and the potential impact on the organization.
  2) The report also includes recommendations for mitigating the identified risks and strengthening the system's security posture [3].
  3) This documentation is essential for informing system administrators and developers about necessary security improvements.

The structured approach to penetration testing, from information gathering to reporting, is essential for identifying and mitigating security vulnerabilities. As systems become more complex, the integration of AI and automation in pentesting processes offers promising avenues for enhancing cybersecurity defenses. Future research and development in this area could lead to more sophisticated and efficient pentesting methodologies, ultimately contributing to more robust security frameworks.

## III. Paper Selection and Review Schema

To comprehensively address the research questions outlined in Section I, a systematic approach was adopted to select, evaluate, and analyze relevant research papers. This section details the methods used for paper collection, the criteria for inclusion and exclusion, the quality assessment process, and the analysis framework applied to the selected studies.

### A. Paper Collection

Relevant papers were collected through a structured search process across major academic databases, including IEEE Xplore, ACM Digital Library, SpringerLink, Elsevier, MDPI, Wiley, and Scopus. The search was performed using carefully selected keywords related to Artificial Intelligence (AI) and penetration testing. An example of a typical search string used is as follows: "Penetration testing" AND ("software engineering" OR "cybersecurity") AND ("AI" OR "machine learning") AND ("2015-2025") Additional search strings included combinations such as:

- "Artificial intelligence" AND "penetration testing"
- "Machine learning" AND "penetration testing"
- "Automated penetration testing" AND ("AI" OR "deep learning")
- "Ethical hacking" AND ("artificial intelligence" OR "machine learning")

In total, 57 papers were selected. The distribution of papers by publisher or source is presented in Figure 2.

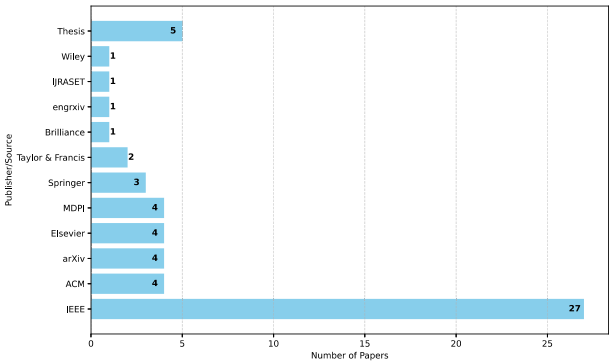


Fig. 2. Distribution of Papers by Publisher/source.

Figure 2 shows that the majority of selected papers were published by IEEE, followed by ACM, Elsevier, and MDPI. This highlights the strong presence of AI and cybersecurity research in well-established venues.

### B. Inclusion and Exclusion Criteria

To ensure the relevance and quality of the selected studies, specific inclusion and exclusion criteria were applied during the paper selection process.

*1) Inclusion Criteria:*

- The paper must contribute to answering at least one of the defined research questions (RQ1–RQ4).
- The paper must be published after 2015.
- The paper must match the selected keywords defined in Section III-A.

### C. Exclusion Criteria

- Papers that do not contribute to answering any research question.
- Papers published before 2015.
- Papers that scored less than 10 points in the Quality Assessment (QA) criteria.
- Non-English articles.

Only papers meeting all inclusion criteria and none of the exclusion criteria were selected for further analysis.

### D. Quality Assessment

To ensure the rigor and relevance of the selected studies, the quality assessment involved ten predefined criteria as detailed in Table I, designed to evaluate each study's relevance, methodology, and contributions in relation to the research questions.

TABLE I
EVALUATION CRITERIA AND THEIR RELATION TO RESEARCH QUESTIONS

| Criterion | Description | Related to |
|---|---|---|
| Q1 | Relevance to AI-driven penetration testing | RQ1 |
| Q2 | Explicit discussion of AI techniques in penetration testing | RQ1 |
| Q3 | Highlights challenges or solutions in AI penetration testing | RQ2 |
| Q4 | Well-defined methodology for AI model or penetration testing | RQ2 |
| Q5 | Comparative analysis with traditional penetration testing | RQ3 |
| Q6 | Evaluation metrics clearly reported | RQ3 |
| Q7 | Contribution of new insights or improvements | RQ4 |
| Q8 | Comparative analysis with related work | General |
| Q9 | Rank of the paper | General |
| Q10 | Availability of tools, datasets, or source code | General |

Each paper was evaluated against the ten criteria using a three-point scale:

- Score 2: Fully satisfies the criterion (well-explained and highly relevant).
- Score 1: Partially satisfies the criterion (mentioned but lacks sufficient depth or clarity).
- Score 0: Does not satisfy the criterion (criterion not addressed or unclear).

The total quality score for each paper was calculated by summing the individual scores across all criteria. Papers scoring 10 points or above were considered acceptable for detailed analysis, while papers scoring below 10 were excluded. The distribution of quality scores among the 57 papers is illustrated in Figure 3.

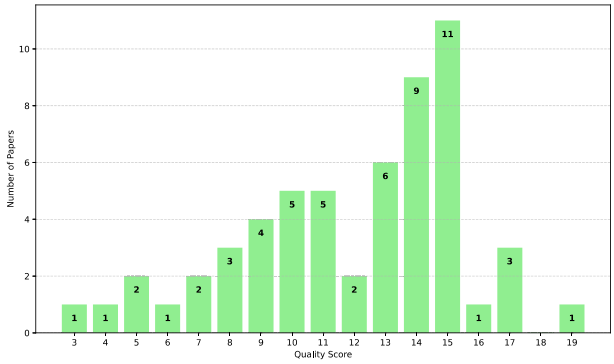


Fig. 3. Distribution of Papers by Quality Score.

As shown in Figure 3, most papers scored between 10 and 15, indicating a generally good level of quality and relevance. Papers with scores below 10 were excluded, ensuring that only studies with sufficient rigor contributed to the final analysis.

### E. Distribution of papers by year

Analyzing the distribution of the selected papers over time provides valuable insights into the evolution of research focus in AI-driven penetration testing. Tracking publication trends helps to identify periods of increased academic and industry attention, as well as emerging areas of innovation.

Additionally, the distribution of papers by year of publication is presented in Figure 4.

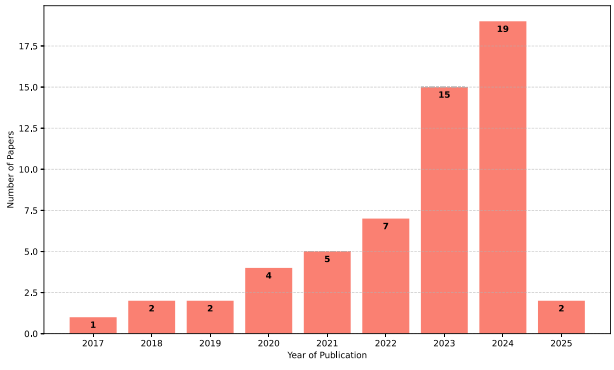


Fig. 4. Distribution of Papers by Year of Publication.

The publication trend reveals a significant increase in research output in recent years, with over 60% of the selected papers published between 2023 and 2025. This reflects the

growing research interest and rapid technological advancements in the field of AI-driven penetration testing. The selected papers were analyzed based on various aspects, including the AI techniques employed, application domains within penetration testing (such as vulnerability discovery, attack simulation, and reporting automation), reported challenges and limitations, evaluation metrics, and suggested future research directions. These findings will be discussed in detail in the next section, where each aspect is organized to answer the corresponding research questions.

## IV. Results and Discussion

This section presents and discusses the findings derived from the final set of selected studies, which were filtered based on rigorous quality assessment criteria. The discussion is structured around the four research questions guiding this study, covering the current artificial intelligence (AI) techniques utilized in penetration testing, the benefits of integrating AI into the process, the key challenges and limitations faced, and the future research directions in this evolving field. Each subsection synthesizes evidence from the selected literature to provide a comprehensive understanding of how AI is transforming penetration testing practices.

### A. AI Techniques Currently Used in Penetration Testing

This subsection addresses (RQ1), which explores the primary AI techniques currently employed in the field of penetration testing. The literature reveals a diverse range of approaches that enhance various stages of the testing process, from vulnerability discovery to attack path optimization. These techniques are categorized into four main groups based on their underlying learning strategies: Reinforcement Learning (RL), Deep Learning (DL), Generative AI models, and Supervised Machine Learning. The following subsections detail how each of these methods contributes to the advancement of automated and intelligent penetration testing systems.

*1) Reinforcement Learning (RL):* Reinforcement Learning (RL) plays a transformative role in penetration testing by automating and optimizing the process, which traditionally relies heavily on human expertise and is time-consuming. RL approaches, such as those using Markov Decision Processes (MDPs) and Partially Observable Markov Decision Processes (POMDPs), allow for the modeling of penetration testing as a sequence of decision-making tasks where the RL agent learns optimal attack strategies through interaction with the environment, without requiring a predefined model of exploit outcomes [11], [12]. This capability is particularly advantageous given the dynamic nature of cyber environments, where new vulnerabilities continuously emerge, making it challenging to maintain up-to-date models [11]. RL agents can explore a broader range of attack vectors than human testers, adapting to changes in the system's state and learning from both successes and failures to improve performance over time [13].

*2) Deep Learning Models (DL):* Deep Learning Models (DL) play a significant role in enhancing penetration testing by automating and optimizing the process of identifying vulnerabilities in complex network environments. The integration of deep learning with reinforcement learning, known as Deep Reinforcement Learning (DRL), is particularly effective in this domain. DRL combines the decision-making capabilities of reinforcement learning with the pattern recognition strengths of deep learning, allowing for the development of sophisticated models that can navigate and exploit network vulnerabilities more efficiently than traditional methods. For instance, the Deep Q-Learning Network (DQN) is a prominent DRL approach that has been used to automate penetration testing by identifying optimal attack paths through a network, leveraging a combination of attack trees and vulnerability data to train the model [14]. This approach is beneficial in dynamic network scenarios where traditional methods may struggle due to the complexity and variability of the environment [15].

*3) Generative AI Models:* Generative AI models, such as Large Language Models (LLMs) and Generative Adversarial Networks (GANs), significantly enhance the efficiency of penetration testing by automating and optimizing various aspects of the process. LLMs can rapidly identify vulnerabilities by simulating a wide range of potential attack scenarios, allowing security teams to focus on the most critical vulnerabilities and implement necessary countermeasures more swiftly [16]. These models can automate the generation of test scenarios, reducing the need for manual intervention and enabling a more extensive evaluation of potential vulnerabilities, which not only saves time but also ensures a more thorough and comprehensive testing process [16]. Additionally, GANs can be used to create diverse and realistic attack vectors, such as malicious payloads and SQL injection attempts, which target potential vulnerabilities in web applications. This iterative process of generating and refining attack vectors until they bypass application defenses effectively discovers vulnerabilities [17].

*4) Supervised Machine Learning:* The integration of supervised machine learning algorithms into penetration testing has shown potential to significantly enhance the efficiency of the process. Supervised learning models, which are trained on labeled datasets, can identify likely points of access with greater accuracy by analyzing system behavior and network configurations, thereby saving time and resources for security professionals [18]. These models are particularly effective in automating tasks such as vulnerability identification, which traditionally require extensive manual effort and expertise. For instance, decision tree algorithms have been employed to select the most effective exploits during penetration tests, although they are noted for their sensitivity to changes in training data, which can affect prediction accuracy [5], [19], Despite this sensitivity, the use of decision trees and other supervised learning techniques can streamline the penetration testing lifecycle by automating the selection of test cases and reducing the need for human intervention in repetitive tasks [19]. Moreover, the application of machine learning in penetration testing is not limited to decision trees; other

algorithms like Long Short-Term Memory (LSTM) and Belief-Desire-Intention (BDI) models have also been explored for their ability to detect vulnerabilities more effectively [19]. The efficiency gains from these algorithms are evident in their ability to process large volumes of data quickly and accurately, which is crucial given the increasing complexity and scale of modern networks.

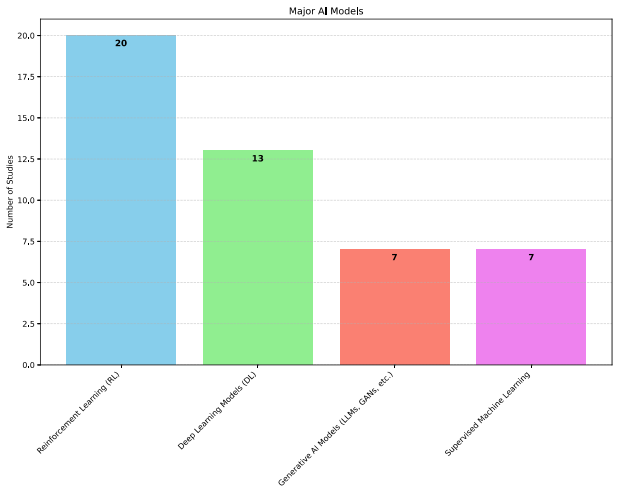


Fig. 5. AI Techniques Used in Penetration Testing.

However, the effectiveness of these models can be limited by the quality and comprehensiveness of the training data, as well as the computational resources required to train and deploy them [18]. Overall, while supervised machine learning algorithms offer promising improvements in the efficiency of penetration testing, their implementation must be carefully managed to address challenges related to data sensitivity and computational demands. Figure 5 provides a visual summary of the main AI techniques applied in penetration testing, including RL, DL, generative models, and supervised learning

### *B. Benefits of Using AI in Penetration Testing*

This section addresses RQ3 by highlighting the key benefits that artificial intelligence brings to penetration testing. The integration of AI technologies has significantly transformed traditional testing methods by enhancing automation, accuracy, scalability, and adaptability. Based on the reviewed studies, AI contributes to more efficient vulnerability detection, reduces manual effort, supports decision-making, and enables continuous learning from dynamic environments. The following subsections elaborate on these advantages, grouped into six thematic areas reflecting the practical impact of AI in penetration testing workflows.

*1) Automation of Penetration Testing:* Artificial Intelligence (AI) significantly enhances the automation of penetration testing by leveraging techniques such as Reinforcement Learning (RL) and Machine Learning (ML) to simulate and execute cyber-attacks more efficiently and effectively. AI-driven frameworks like Shennina and PenBox utilize RL to automate various phases of penetration testing, including network and service enumeration, vulnerability assessment, and attack path generation, which traditionally require extensive manual effort and expertise [9], [20]. For instance, Shennina employs a RL approach to train models that can detect and exploit vulnerabilities, optimizing the process by selecting reliable remote exploits and eliminating false positives [9]. Similarly, PenBox, developed by the European Space Agency, integrates AI to automate the execution of penetration tests, using Deep Q-Learning to optimize attack paths without human intervention, thus reducing the time and cost associated with manual testing [20]. The use of AI in penetration testing also involves the application of Partially Observable Markov Decision Processes (POMDP) and Planning Domain Definition Language (PDDL) to model and solve complex attack scenarios, although these methods often require prior knowledge of network configurations [20]. Overall, AI's ability to learn from data, adapt to new threats, and execute complex attack strategies autonomously makes it a powerful tool for automating penetration testing, thereby improving the resilience of systems against evolving cyber threats.

*2) Efficiency Improvement and Time Saving:* AI can significantly enhance the efficiency and time-saving aspects of penetration testing by automating various phases of the process and optimizing decision-making. Reinforcement Learning (RL), a subset of AI, is particularly effective in this domain, as it can automate the exploration of attack paths and adapt strategies based on real-time feedback from the environment. For instance, the Shennina framework utilizes RL to automate network and service enumeration, vulnerability assessment, and attack path generation, which reduces the need for manual intervention and accelerates the penetration testing process [9]. Similarly, the PenBox framework developed by the European Space Agency integrates AI techniques, including Deep Q-Learning, to automate penetration testing in space mission operations, thereby reducing human interaction and optimizing the process in terms of performance and cost-effectiveness [20]. Overall, AI's ability to automate complex tasks, learn from interactions, and optimize decision-making processes makes it a powerful tool for improving the efficiency and effectiveness of penetration testing.

*3) Improved Vulnerability and Threat Detection:* AI can significantly enhance vulnerability and threat detection in penetration testing by automating and optimizing various aspects of the process. Reinforcement learning (RL), a subset of AI, is particularly effective in this domain as it allows AI models to learn optimal strategies for penetration testing through trial and error, thereby improving efficiency and reducing the need for human intervention [20], [21]. For instance, the use of Q-Learning, a type of RL, has been shown to automate the penetration testing process effectively, enabling the AI to identify vulnerabilities and suggest optimal attack paths without prior knowledge of the network configuration [20]. This approach not only accelerates the testing process but also enhances the accuracy of threat detection by continuously learning from the environment and adapting to new scenarios. Additionally, large

language models (LLMs) like CIPHER have been developed to assist in penetration testing by providing expert-level guidance and reasoning, which is particularly beneficial for beginners in the field [22]. These models are trained on extensive datasets of penetration testing scenarios, allowing them to offer detailed explanations and recommendations, thus improving the overall effectiveness of vulnerability detection [22].

*4) Scalability and Learning Improvement:* AI techniques, particularly Reinforcement Learning (RL), offer promising avenues for enhancing the scalability and learning capabilities of penetration testing. The application of RL in penetration testing allows for the automation of attack strategies without requiring a pre-defined model of the environment, which is crucial given the dynamic nature of cyber threats and network configurations. This adaptability is achieved through the interaction of RL agents with simulated environments, where they learn optimal attack paths by maximizing rewards associated with successful exploits [11]. The use of network attack simulators, such as those developed by J. Schwarz, provides a flexible and computationally efficient platform for training these RL agents, avoiding the high costs associated with virtual machine-based simulations [20]. Moreover, the integration of Deep Q-Learning (DQN) within these frameworks ensures a level of randomization and adaptability, which is essential for handling the vast state spaces typical of large networks [20].

*5) Optimization and Planning Improvements:* AI can significantly enhance the optimization and planning of penetration testing practices by automating complex tasks and improving decision-making processes. Reinforcement Learning (RL), a subset of AI, has been particularly effective in this domain. For instance, the Shennina framework utilizes RL to automate network and service enumeration, vulnerability assessment, and attack path generation, thereby optimizing the penetration testing process by reducing false positives and ensuring efficient exploitation of vulnerabilities [9]. Similarly, the integration of Deep Q-Learning Networks (DQN) in penetration testing frameworks allows for the discovery of optimal attack paths with high accuracy, as demonstrated by the ability of DQN to find the easiest paths to exploit vulnerabilities in simulated environments [17]. The use of AI in penetration testing also extends to the development of automated frameworks like PenBox, which employs Q-learning to maximize rewards through optimal policy learning, thus enhancing the efficiency and cost-effectiveness of penetration tests [20]. These AI-driven approaches not only streamline the planning phase by generating attack trees and simulating various attack scenarios but also improve the overall effectiveness of penetration testing by enabling the identification of the most efficient attack strategies.

*6) Enhancement of Cybersecurity Training and Audits:* Artificial Intelligence (AI) significantly enhances cybersecurity training and audits of penetration testing practices by automating complex tasks and improving efficiency. AI, particularly through reinforcement learning (RL), can simulate realistic cyber threats and automate penetration testing, which traditionally requires extensive manual effort and expertise. This automation not only reduces the need for skilled professionals but also allows for more frequent and comprehensive testing, thereby improving the overall security posture of organizations [9]. AI models, such as the Intelligent Automated Penetration Testing System (IAPTS), utilize RL to learn from interactions within a network environment, enabling them to autonomously discover and exploit vulnerabilities. This capability allows AI systems to perform tests that human experts might overlook due to time constraints or complexity, thus broadening the scope of penetration testing and enhancing its effectiveness. Moreover, AI can optimize the use of resources by prioritizing relevant tests and reducing network congestion, which is often a challenge in traditional penetration testing [23]. Overall, AI's ability to automate and enhance penetration testing practices not only improves the efficiency and effectiveness of cybersecurity audits but also contributes to the development of more robust and adaptive security strategies. Figure 6 highlights the six key benefits of using AI in penetration testing, such as automation, efficiency, and scalability.

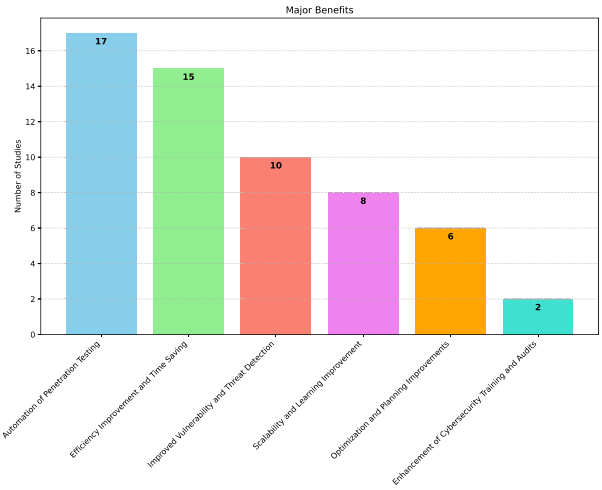


Fig. 6. Benefits of AI Integration in Penetration Testing

### C. Challenges in Integrating AI with Penetration Testing

This section responds to RQ2 by discussing the major challenges and limitations associated with integrating AI into penetration testing practices. While AI offers significant improvements in efficiency and capability, its adoption also introduces technical, operational, and ethical hurdles. Issues such as scalability, training complexity, limited data availability, model transparency, and deployment difficulties are commonly reported in the literature. This section organizes these challenges into four thematic areas to provide a structured overview of the obstacles that must be addressed for broader and more effective use of AI in this domain.

*1) Scalability and Real-World Applicability Issues:* Integrating AI with penetration testing presents several scalability issues, primarily due to the complexity and resource demands of simulating realistic network environments. One significant challenge is the high computational cost associated with

running virtual machines (VMs), which are often used to simulate network environments for AI training and testing. This can slow down the processing of AI algorithms and require substantial computational assets, making it difficult to scale up to larger networks or more complex scenarios [20]. Additionally, the need for extensive data collection and logging to train AI models, such as those used in the Shennina framework, can be resource-intensive and may not scale efficiently across diverse network configurations [9]. The variability in network topologies and the need for customized configurations further complicate scalability, as AI models must be adaptable to different environments without extensive manual intervention [20]. Another issue is the integration of AI with existing penetration testing tools, which often have heterogeneous interfaces and require significant effort to automate and chain together effectively [24]. The reliance on human intervention in AI-driven penetration testing, as seen with tools like PentestGPT, highlights the difficulty in achieving fully automated solutions that can scale without human oversight [24]. Moreover, the challenge of maintaining a balance between exploration and exploitation in reinforcement learning models can lead to inefficiencies, particularly when scaling to larger state and action spaces [20]. The need for continuous updates and adaptations to AI models to keep pace with evolving cyber threats also poses a scalability challenge, as it requires ongoing resource investment and expertise [9]. Furthermore, the potential for AI models to be trained on outdated or incomplete data sets can limit their effectiveness and scalability in real-world applications [24]. The complexity of managing and optimizing hyperparameters in deep reinforcement learning models adds another layer of difficulty, as it requires extensive experimentation and tuning to achieve optimal performance across different scenarios [20]. Finally, the ethical and security concerns associated with deploying AI-driven penetration testing tools at scale, such as the risk of misuse by malicious actors, necessitate robust oversight and governance frameworks, which can be challenging to implement and maintain at scale [24].

*2) Limitations of Language Models and AI Bias Risks:* Language models (LMs) and AI-based penetration testing face several limitations and risks, particularly concerning AI bias and the need for human intervention. One significant limitation is the reliance on human input during penetration testing, as current AI models like PentestGPT require human assistance to perform tasks such as navigating websites and interpreting outputs, which indicates that full automation is not yet feasible [24]. Additionally, the performance of LMs in penetration testing is inconsistent across different tasks, with models like Llama 3.1-405B and GPT-4o struggling with complex tasks such as privilege escalation and exploitation, especially on medium to hard-level machines [24]. This inconsistency highlights the models' limitations in handling complex cybersecurity scenarios without human oversight. These limitations and risks underscore the importance of responsible development and ethical oversight in the deployment of AI-driven cybersecurity solutions, as well as the need for ongoing research to address these challenges and improve the reliability and security of AI-based penetration testing tools.

*3) Training, Convergence, and Data Efficiency Challenges:* AI-based penetration testing faces several challenges related to training, convergence, and data efficiency. One significant issue is the overfitting of expert knowledge when using imitation learning, which can hinder the balance between exploration and exploitation, leading to slower convergence and reduced efficiency in intelligent penetration testing algorithms [25]. The complexity of algorithms, such as those used in reinforcement learning (RL), often results in prolonged training times and difficulties in achieving convergence, particularly in poorly simulated environments [25]. Additionally, the integration of expert knowledge, while beneficial, can be scenario-dependent and poorly interpretable, further complicating the training process and potentially leading to overfitting [25]. The use of RL methods, such as the Markov decision process (MDP), is challenged by the computational complexity that arises as network scenarios expand, making it difficult to apply these methods to large-scale networks efficiently [25].

*4) Tool and Model Integration and Deployment Challenges:* AI-based penetration testing faces several challenges related to tool and model integration and deployment. One significant challenge is the high computational cost associated with running virtual machines (VMs) for network simulations, which can slow down AI algorithms and require substantial computational resources, as seen in the use of network simulators like NS3 and mininet that are not specifically designed for penetration testing [20]. Additionally, the integration of AI models, such as those using reinforcement learning (RL), into existing frameworks like PenBox, requires careful consideration of the environment to ensure fidelity to real-world scenarios while maintaining flexibility and cost-effectiveness [20]. Figure 7 outlines the primary challenges in integrating AI with penetration testing, including scalability, training, and ethical concerns.

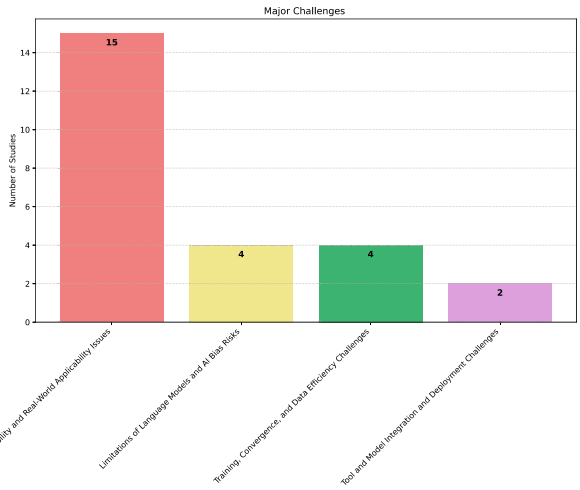


Fig. 7. Challenges of AI Integration in Penetration Testing

### D. Future Directions and Research Gaps

This final section addresses RQ4, focusing on the current research trends, existing gaps, and prospective directions for AI-enhanced penetration testing. The reviewed studies emphasize the importance of advancing algorithm design, improving scalability, fostering ethical AI practices, and developing robust methodologies. In addition, the need for real-world validation, standardized datasets, and cross-domain collaboration is highlighted. The subsections that follow group these insights into four key areas, aiming to guide future research efforts and innovation in this field.

*1) Scalability and Real-World Deployment Improvements:* Future directions for scalability and real-world deployment improvements in AI-based penetration testing are multifaceted, focusing on enhancing the adaptability and efficiency of these systems. One significant area is the integration of advanced reinforcement learning (RL) techniques, such as combining RL with genetic algorithms and fuzzy techniques, which could optimize security evaluation methodologies and address the complexity of training observed in RL systems [17]. Additionally, the development of more sophisticated network simulators, like the Network Attack Simulator, offers a flexible and scalable environment for testing AI models, which is crucial for real-world applicability [20]. The use of model-based approaches in RL could further enhance scalability by allowing agents to learn approximate environment transition functions, enabling offline learning and reducing the need for extensive real-world data collection [20].

*2) Model and Algorithm Enhancements:* Future directions in AI-based penetration testing, particularly concerning model and algorithm enhancements, are diverse and promising. One significant area of focus is the integration of reinforcement learning (RL) techniques, such as Q-learning and Deep Q-Networks (DQN), to automate and optimize penetration testing processes. These methods are being explored to improve the efficiency and effectiveness of penetration tests by enabling agents to learn optimal attack strategies through trial and error in simulated environments, thereby reducing human intervention and resource consumption [4], [20], [21]. The use of RL in penetration testing is particularly promising for its ability to handle complex decision-making processes in dynamic environments, although challenges such as the sparse reward problem and large action spaces remain [3].

*3) Automation and Methodology Development:* Future directions in AI-based penetration testing focus on enhancing automation and developing methodologies that leverage advanced AI techniques to improve efficiency and effectiveness. One promising area is the integration of Reinforcement Learning (RL) to optimize attack path definitions and reduce human intervention, as demonstrated by the European Space Agency's PenBox project, which aims to automate penetration testing in space systems using RL to maximize rewards and minimize costs [20]. The CIPHER model exemplifies efforts to create domain-specific LLMs that can guide penetration testing processes, emphasizing the importance of specialized training and benchmarking to enhance AI's practical utility in cybersecurity [22]. Moreover, the use of network simulators and the development of dynamic attack trees are being investigated to improve the representativeness and scalability of AI-driven penetration testing environments, which could lead to more realistic and comprehensive security assessments [20]. These advancements suggest a future where AI not only automates routine tasks but also adapts to evolving threats, providing a robust framework for continuous improvement in cybersecurity practices.

*4) Security, Ethics, and Bias Mitigation:* Designing AI-based penetration testing to minimize potential security risks and ensure ethical considerations involves several strategic approaches. Firstly, the integration of Reinforcement Learning (RL) techniques, such as Q-Learning, can automate penetration testing processes, enhancing efficiency and reducing human error, which is crucial for maintaining security integrity [9], [21]. The use of RL allows the AI to learn from real-time observations and adapt its strategies, which can help in identifying vulnerabilities without causing unintended harm to the systems being tested [21]. Moreover, employing frameworks like MITRE &CK provides a structured method to analyze AI-powered cyberattacks, ensuring that the tactics used are well-documented and understood, which is essential for ethical transparency and accountability [9]. Additionally, the development of realistic testbeds, such as the AI4SIM framework, allows for the simulation of advanced attacks in a controlled environment, minimizing the risk of real-world impact while providing valuable data for refining AI models [9]. Ethical considerations are further addressed by ensuring that AI models are trained and tested in environments that mimic real-world conditions without exposing actual systems to risk. This involves using simulations like the Vehicular Ad Hoc Network (VANET) to test AI models in a safe and controlled manner, which helps in understanding the potential impacts of AI-driven attacks and refining the models to prevent misuse [21]. Figure 8 illustrates the main future research directions and existing gaps in AI-enhanced penetration testing.

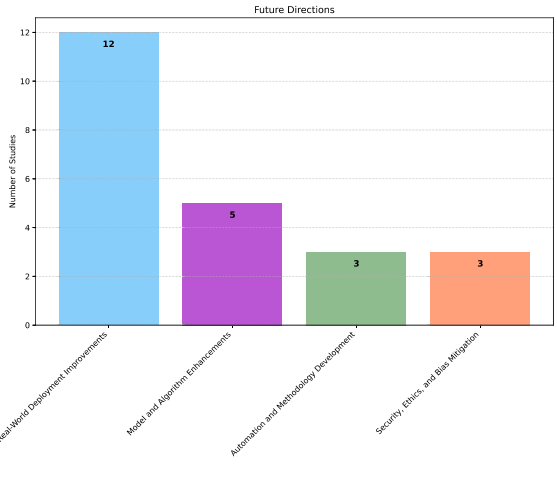


Fig. 8. Future Directions and Research Gaps in AI-based Penetration Testing

This section highlighted the key AI techniques, benefits, challenges, and future directions in penetration testing based on selected studies. While AI offers significant advantages, several limitations remain. These findings form the basis for the next section, which provides the overall conclusion of the study on the integration of AI into penetration testing.

## V. Conclusion

This systematic mapping study investigated the growing role of Artificial Intelligence (AI) in the field of penetration testing, with the aim of identifying current techniques, understanding the associated challenges, evaluating the effectiveness of AI-driven approaches, and outlining future research directions. A total of 57 research papers published between 2015 and 2025 were initially identified using a structured search strategy across major academic databases. Following a rigorous quality assessment process based on ten predefined criteria, selected papers were selected for in-depth analysis. This assessment ensured that only studies of sufficient relevance and methodological rigor contributed to the final synthesis. The study addressed four research questions, each exploring a critical aspect of AI-driven penetration testing. In response to the first research question, the analysis revealed that the most commonly applied AI techniques include Reinforcement Learning, Deep Learning, Generative AI models such as large language models and generative adversarial networks, and Supervised Machine Learning. These techniques are employed to automate various stages of the penetration testing process, including reconnaissance, attack path planning, exploitation, and reporting. Notably, reinforcement learning and deep Q-learning networks have shown strong potential for automating complex decision-making and adapting to dynamic network environments. Regarding the second research question, the study identified several key challenges that hinder the seamless integration of AI into penetration testing. These include scalability limitations when applying models to large and complex networks, training inefficiencies and convergence problems, integration difficulties with existing cybersecurity tools, and ethical concerns related to AI bias, misuse, and lack of transparency. The current limitations of language models in handling advanced penetration testing tasks and their dependence on human interaction were also emphasized. In answering the third research question, the findings demonstrate that AI significantly enhances the efficiency, speed, scalability, and accuracy of penetration testing. AI enables automated vulnerability detection, supports intelligent decision-making, and facilitates real-time adaptation to changing environments. Frameworks such as Shennina and PenBox were highlighted as practical examples of AI implementation, demonstrating the power of reinforcement learning and deep learning to improve penetration testing effectiveness while reducing manual effort. For the fourth research question, the study identified key research gaps and proposed future directions. These include the development of more scalable and adaptable AI models, the creation of standardized testbeds for benchmarking, interdisciplinary collaboration between cybersecurity and AI communities, and the advancement of ethical and explainable AI systems. The increasing use of large language models for supporting penetration testing tasks such as command execution and vulnerability assessment also presents opportunities for future improvement, particularly in addressing current limitations in memory retention and autonomous decision-making. In conclusion, this study confirms that AI has the potential to transform penetration testing by making it more intelligent, automated, and adaptive. While traditional penetration testing methods remain valuable, the integration of AI introduces new possibilities for improving speed, precision, and scope. To fully realize this potential, the field must continue to address the technical, practical, and ethical challenges identified in this Study. This includes advancing AI models, refining methodologies, and ensuring that AI-driven systems are deployed responsibly and securely in real-world cybersecurity environments.

## References


[1] D. Garg and N. Bansal, "A systematic review on penetration testing," in *2021 2nd Global Conference for Advancement in Technology, GCAT 2021*. Institute of Electrical and Electronics Engineers Inc., 10 2021.

[2] S. A. Altayaran and W. Elmedany, "Integrating web application security penetration testing into the software development life cycle: A systematic literature review," in *2021 International Conference on Data Analytics for Business and Industry, ICDABI 2021*. Institute of Electrical and Electronics Engineers Inc., 2021, pp. 671–676.

[3] C. Greco, G. Fortino, B. Crispo, and K. K. R. Choo, "Ai-enabled iot penetration testing: state-of-the-art and research challenges," 2023.

[4] S. X. S. C. Austin O'Brien, "Automated post-breach penetration testing through reinforcement learning," in *IEEE*. IEEE, 2020.

[5] D. R. Mckinnel, T. Dargahi, A. Dehghantanha, and K.-K. R. Choo, "A systematic literature review and meta-analysis on artificial intelligence in penetration testing and vulnerability assessment," *sciencedirect*, 2019.

[6] A. Happe and J. Cito, "Getting pwn'd by ai: Penetration testing with large language models," in *ESEC/FSE 2023 - Proceedings of the 31st ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. Association for Computing Machinery, Inc, 11 2023, pp. 2082–2086.

[7] D. Amalfitano, S. Faralli, J. C. R. Hauck, S. Matalonga, and D. Distante, "Artificial intelligence applied to software testing: A tertiary study," *ACM Computing Surveys*, vol. 56, 6 2023.

[8] A. Castagnaro, M. Conti, and L. Pajola, "Offensive ai: Enhancing directory brute-forcing attack with the use of language models," in *Proceedings of the 2024 Workshop on Artificial Intelligence and Security*. ACM, 11 2024, pp. 184–195. [Online]. Available: https://dl.acm.org/doi/10.1145/3689932.3694770

[9] S. Karagiannis, C. Fusco, L. Agathos, W. Mallouli, V. Casola, C. Ntantogian, and E. Magkos, "Ai-powered penetration testing using shennina: From simulation to validation," in *ACM International Conference Proceeding Series*. Association for Computing Machinery, 7 2024.

[10] "Ml and generative ai." [Online]. Available: https://www.tutorialspoint.com/gen-ai/ml-and-generative-ai.htm

[11] J. Schwartz, "Autonomous penetration testing using reinforcement learning," *Thesis ,UOQ*, 2018.

[12] M. C. Ghanem and T. M. Chen, "Reinforcement learning for intelligent penetration testing," *IEEE*, 2018.

[13] Clintswood, D. G. Lie, L. Kuswandana, Nadia, S. Achmad, and D. Suhartono, "The usage of machine learning on penetration testing automation," in *Proceedings - 2023 3rd International Conference on Electronic and Electrical Engineering and Intelligent System: Responsible Technology for Sustainable Humanity, ICE3IS 2023*. Institute of Electrical and Electronics Engineers Inc., 2023, pp. 322–326.

[14] H. Zhenguo, "Autonomous penetration testing using drl," *Thesis ,JAIST*, 2021.

[15] Q. Li, R. Wang, D. Li, F. Shi, M. Zhang, and A. Chattopadhyay, "Dynpen: Automated penetration testing in dynamic network scenarios using deep reinforcement learning," *IEEE Transactions on Information Forensics and Security*, 2024.

[16] E. Hilario, S. Azam, J. Sundaram, K. Imran Mohammed, and B. Shanmugam, "Generative ai for pentesting: the good, the bad, the ugly," *International Journal of Information Security*, vol. 23, pp. 2075–2097, 6 2024.

[17] H. A. Udupa, B. S. Anavi, B. Goyal, S. P. Kasturi, and P. Agarwal, "Advanced reinforcement learning based penetration testing," in *1st International Conference on Electronics, Computing, Communication and Control Technology, ICECCC 2024*. Institute of Electrical and Electronics Engineers Inc., 2024.

[18] M. Patil, D. Thakare, A. Bhure, S. Kaundanyapure, and D. A. Mune, "An ai-based approach for automating penetration testing," *International Journal for Research in Applied Science and Engineering Technology*, vol. 12, pp. 5019–5028, 4 2024.

[19] V. Saber, D. Elsayad, A. M. Bahaa-Eldin, and Z. Fayed, "Automated penetration testing, a systematic review," in *3rd International Mobile, Intelligent, and Ubiquitous Computing Conference, MIUCC 2023*. Institute of Electrical and Electronics Engineers Inc., 2023, pp. 373–380.

[20] A. Confido, E. V. Ntagiou, and M. Wallum, "Reinforcing penetration testing using ai," in *IEEE Aerospace Conference Proceedings*, vol. 2022-March. IEEE Computer Society, 2022.

[21] P. Garrad and S. Unnikrishnan, "Reinforcement learning in vanet penetration testing," *Results in Engineering*, vol. 17, 3 2023.

[22] D. Pratama, N. Suryanto, A. A. Adiputra, T.-T.-H. Le, A. Y. Kadiptya, M. Iqbal, and H. Kim, "Cipher: Cybersecurity intelligent penetration-testing helper for ethical researcher," *MDPI*, 8 2024. [Online]. Available: http://arxiv.org/abs/2408.11650

[23] M. C. Ghanem and T. M. Chen, "Reinforcement learning for efficient network penetration testing," *Information (Switzerland)*, vol. 11, 1 2020.

[24] I. Isozaki, M. Shrestha, R. Console, and E. Kim, "Towards automated penetration testing: Introducing llm benchmark, analysis, and improvements," *arXiv*, 10 2024. [Online]. Available: http://arxiv.org/abs/2410.17141

[25] Y. Wang, Y. Li, X. Xiong, J. Zhang, Q. Yao, and C. Shen, "Dqfd-aipt: An intelligent penetration testing framework incorporating expert demonstration data," *Security and Communication Networks*, vol. 2023, 2023.