Introduction of AI into Penetration Testing

April 23rd, 2024

Utah Valley University

By Logan McKinley

Table of Contents

Abstract

This paper explores the numerous ways in which AI is being introduced into the world of cybersecurity. With that comes the review of AI models that have been created around the world to learn, find, and exploit known vulnerabilities of a chosen network. AI has become somewhat of a controversial topic in today's world and with its recent use in cybersecurity, that controversy and worry has grown exponentially. Research shows that this is only the beginning of these models being used in widespread cybersecurity plans and projects, in the near future it is projected to grow to AI models researching, organizing, and carrying out their own penetration tests without the need for humans to hold its hand the whole way through. You will see many instances of AI models that are already widely used and available to the public (ChatGPT), which will further this work even more due to it being accessible to use in everyday workflows and projects.

Introduction of AI into Penetration testing

The integration of Artificial Intelligence has redefined penetration testing, offering greater efficiency and effectiveness in identifying and mitigating cyber threats. In the dynamic landscape of cybersecurity, the introduction of Artificial Intelligence marks a pivotal moment in the evolution of penetration testing. With its ability to quickly analyze large amounts of data, detect anomalies, and predict potential threats, AI holds the promise of forever changing how organizations protect their digital assets. This paper explores how AI has been introduced into cyber security, and its role in penetration testing, spotlighting its innovative methodologies and its impact on enhancing cyber security.

To know why the introduction of AI into penetration testing is a substantial step forward in the cybersecurity field, we must first understand a few important concepts. What is penetration testing? How is it used in a corporate environment? What are the rules and regulations of a penetration test? How would AI handle those same rules and regulations? This knowledge will help guide you in the process of deciding whether this needs to be pursued and perfected or dropped all together.

Understanding Penetration Testing

The definition of Penetration testing from NIST (National Institute of Standards and Technology) is, “A method of testing where testers target individual binary components or the application as a whole to determine whether intra or intercomponent vulnerabilities can be exploited to compromise the application, its data, or its environment resources.” (2024) Penetration testing is used around the world both internally and by third party companies that are hired to perform them. They are used to identify vulnerable areas in a company’s network,

systems, and even physical security. Once identified, they are documented and given back to the hiring company who can patch those vulnerabilities that risk a potential breach. Refer to Figure 1 ([[Stages of Penetration Testing]], 2024).

Scope

When a third-party company is hired to perform a penetration test, they are given a documented scope that is required to be followed religiously. The scope states what in the company is allowed to be tested and what is off-limits to the pen testing company. Any breach of the scope given can have substantial legal consequences.

Rules of Engagement

In the realm of research, the scope outlines the parameters for permissible testing, while the rules of engagement prescribe the methodologies acceptable within this defined scope. These regulations articulate the approved tools, techniques, and procedures for conducting assessments on the accessible systems, thus ensuring consistency with the established testing framework and objectives.

Reporting Requirements

Reporting requirements for penetration testing play a crucial role in facilitating informed decision-making and enhancing cybersecurity posture. The format of the report, whether it is a comprehensive document or a concise presentation, should be tailored to the needs of executive leadership, IT security teams, and regulatory bodies. The report should encompass detailed findings, remediation recommendations, risk assessments, and supporting evidence to provide a comprehensive overview of the organization's security vulnerabilities and areas for improvement.

Importance

This approach allows security teams to address weaknesses before malicious actors can exploit them, thereby reducing the risk of data breaches, financial losses, and reputational damage. Additionally, penetration testing provides valuable insights into an organization's security posture, helping to prioritize remediation efforts and allocate resources effectively. The rigorous assessment and validation of security measures through penetration testing not only enhances an organization's resilience to cyber threats but also fosters trust and confidence among clients and partners.

The Integration of AI in Penetration Testing

Automation has long been sought after within penetration testing, reflecting a broader trend in technology where humans seek to minimize their direct involvement. This path defines the ongoing evolution of cybersecurity. The absence of the tools we have today would extend the duration of penetration testing procedures from weeks to potentially months, highlighting the need for current technological advancements in expediting testing processes.

Pros

1. **Automation**: AI enables automated scanning and analysis of vulnerabilities, saving time and resources.
2. **Accuracy and Precision**: AI-driven tools can detect complex vulnerabilities with high accuracy and precision.
3. **Efficiency**: AI-based penetration testing can manage large-scale assessments more efficiently than manual methods.

4. **Proficiency**: AI systems can be trained to adapt and learn from new threats, improving their effectiveness over time.
5. **Scalability**: AI technologies can scale to address the growing complexity and volume of cybersecurity threats.
6. **Data-Oriented**: AI-driven tools can analyze vast amounts of data to identify patterns and trends that may indicate potential vulnerabilities.

Cons

1. **Control**: AI systems may lack transparency and control, leading to uncertainty about their decision-making processes.
2. **Human Factor**: Overreliance on AI may diminish the role of human expertise and intuition in penetration testing.
3. **Morals and Ethics**: AI-driven penetration testing raises ethical concerns regarding privacy, consent, and potential misuse of technology.
4. **Unemployment**: The automation of penetration testing tasks by AI may lead to job displacement for human security professionals.
5. **Bias**: AI algorithms may exhibit biases inherent in the data used to train them, leading to inaccurate or unfair results.
6. **Privacy**: AI-driven penetration testing tools may inadvertently compromise user privacy by collecting and analyzing sensitive data without consent.

Considering the cons, many of these challenges can be solved through adequate human education, training, and oversight. However, in the rapidly expanding AI landscape, it becomes

essential to identify reliable sources for acquiring the necessary knowledge and tools to harness AI's full potential in the cybersecurity sphere.

Machine Learning

What is machine learning? How is machine learning used in penetration tests today? “Machine learning are technologies and algorithms that enable systems to identify patterns, make decisions, and improve themselves through experience.” Kuswandana, L. (2021). Machine learning plays a crucial role by augmenting human expertise, enabling systems to emulate human-like thinking and decision-making processes. By analyzing large amounts of data, machine learning algorithms can identify subtle patterns indicative of potential vulnerabilities or threats within complex networks and systems. This capability has significantly advanced the automation of penetration testing, allowing for more efficient and effective security assessments. Moreover, the continuous evolution of machine learning techniques has enabled them to simulate cognitive functions like the human brain, including learning, problem-solving, and reasoning. As a result, machine learning-driven approaches have become indispensable tools for cybersecurity professionals, offering unparalleled insights and capabilities in identifying and mitigating security risks.

AI Learning Models

An AI learning model, also referred to as a machine learning model, is a framework essential for enabling computers to learn from data and make decisions or predictions autonomously. These models are constructed based on mathematical algorithms that analyze vast datasets to identify patterns and relationships within the information. Unlike traditional programming methods, where explicit instructions are provided for each task, AI learning models leverage algorithms to learn from experience and refine their performance over time.

Examples of AI learning models encompass various techniques such as linear regression, decision trees, and sophisticated architectures like convolutional neural networks and recurrent neural networks. Through continuous training and refinement, AI learning models empower systems to adapt to the latest information and solve complex problems, thus driving advancements across numerous domains.

The two learning models that are most common and that you will see in this paper are Linear Regression and Reinforcement Learning. Linear Regression is a simple and widely used machine learning algorithm for predicting a continuous output variable based on one or more input variables. It fits a linear equation to the observed data points, minimizing the difference between the predicted and actual values. Reinforcement learning is a branch of machine learning where an agent learns to make decisions by trial and error, aiming to maximize long-term rewards received from its interactions with the environment.

Real-World Applications

Numerous strides have been made in integrating AI into penetration testing, with companies and individuals developing AI models tailored to specific stages. However, the predominant focus lies on the initial steps outlined in Figure 1: reconnaissance and scanning. These phases are deemed paramount as they lay the groundwork for strategizing and executing successful attacks. Current AI models include Shennina, GyoiThon, and LLM's (large language models) like ChatGPT.

Shennina

Shennina was created by a group of students in the Computer Science department at Bina Nusantara University in Indonesia. It is one of the more in-depth models that has been created to

conduct all steps within the test, “Shennina is an AI based penetration testing tool which provides network and service scanner, vulnerabilities exploitation tester, attack path generation, and Metasploit integration to launch the attack.” Kuswandana, L. (2021).

The students show that Shennina’s process can be broken down into some of the same steps we saw in Figure 1. Step 1 is overlooked for this model, showing that the initial planning and reconnaissance will need to have human intervention which introduces a sort of security so that Shennina is not picking its own victims at random. The students say, “To start off the process, Shennina scans the target network and looks for any open ports with running services that might be exploitable. The scanned results will then be used for the second step which is to look for any vulnerabilities that could be found in the ports or services of the target. The vulnerabilities database was served in a dataset previously for Shennina to learn. Shennina will notify any vulnerabilities found in the target.”

Following this, Shennina can generate an attack path to exploit the identified vulnerabilities. This path is then saved as an h5 file, a format utilized for storing extensive data in multidimensional arrays. Once stored, the model transitions into the exploitation phase, autonomously following its designated path to breach the system and endeavor to gain access.

One of the most crucial functions that Shennina can do is in the last step of its test. “To close off, Shennina will generate an exploitation report of the performed attack. This report will be stored in an md file format. This includes the target IP, attack result, vulnerable service/port, exploit name, details, and references, shell type, and the used payload. This report provides us with a wealth of information regarding the performed attack. In the perspective of cyber security, this will help with the reporting of penetration test practices.” Documentation is key in all things

cybersecurity but especially when running a penetration test as lack thereof can have legal repercussions.

The students opted for the RL (Reinforcement Learning) algorithm for their model, leveraging its capability to identify flaws that might typically be overlooked by human pen testers. This is owed to its capacity to explore broader space and employ attack tree tactics beyond the reach of human capabilities. According to the Bina Nusantara students, “The capability of reinforcement learning to learn from experience and progressively improve performance is also another benefit. In complex systems where it would be challenging for penetration testers to recognize every potential attack vector, this is very helpful.” Refer to Figure 2 ([[Shennina Methodology]], 2021).

GyoiThon

Created by students at Universiti Teknologi Malaysia, “GyoiThon is a machine learning-based penetration testing tool developed by Masuya Masafumi that can use other penetration testing tools like Scrapy to identify the vulnerabilities found on the target websites.” Hoang, L. V. (2022). This tool has nine different modes of usage depending on what kind of test is being pursued. The default mode includes the following processes:

i. Gathering HTTP responses from target URLs by using the Web crawling feature, identifying product or version using string pattern matching.

ii. Identifying vulnerabilities based on the reported vulnerabilities in the National Vulnerability Database (NVD), by their CVE numbers.

iii. Locating unnecessary HTML or JavaScript comments.

iv. Locating unnecessary debug messages.

v. Login pages assessment.

Refer to Figure 3 ([[GyoiThon Procedure]], 2021).

GyoiThon effectively scanned and detected vulnerabilities on both port 80 and port 443 in default and ML modes. Notably, there was a significant disparity in the vulnerabilities discovered between the two modes. However, a consistent trend emerged where port 80 exhibited a higher number of vulnerabilities. This disparity can be attributed to the fact that the tests were conducted on 'localhost,' which might lack the latest version of HTTPS, whereas port 443 enables data transmission over a secure network. Refer to Figure 4 ([[Default Mode vs ML Mode]], 2022). Each mode, as depicted in Figure 4, exhibited flaws concerning "Unidentified Vulnerabilities," as noted by the students, “there are still many vulnerabilities that cannot be identified. This is due to those vulnerabilities are being not reported in the National Vulnerability Database (NVD).”

PenBox

Developed by the European Space Agency (ESA) and the European space industry, The PenBox tool addresses the need for robust security measures within their data systems, which manage assets of significant value. As digital media and internet usage become increasingly prevalent, ensuring the security and strength of these systems becomes top priority. PenBox, short for Penetration testing in a Box, was created to integrate automated security penetration testing seamlessly into the life cycle of these systems. By executing attack patterns, PenBox aims to uncover weaknesses and vulnerabilities within space systems, enhancing security engineering efforts by providing automated testing capabilities. Its user-friendly graphical interface makes

security testing accessible even to non-experts, thereby raising awareness about security risks and their potential impact on mission-critical operations.

PenBox stands out because it does not just identify vulnerabilities in computer systems; it goes a step further by simulating attacks to demonstrate the potential consequences of security breaches in real-world scenarios. PenBox operates within specific attack scenarios, with each action building upon the results of previous steps, ensuring a comprehensive and focused approach to security testing. Its adaptability is spearheaded by its code structure, providing the use of custom attack patterns and future enhancements. However, achieving effectiveness requires seamless integration and interaction with a variety of penetration testing tools, a challenge addressed through its flexible and adaptable design.

Additionally, PenBox is exploring the integration of machine learning techniques to enhance its performance and decision-making capabilities. By leveraging machine learning, PenBox aims to reduce reliance on human intuition, optimize tool selection, and enhance overall efficiency. This integration highlights PenBox's commitment to remaining at the front of security testing innovation, continually evolving to address the ever-growing challenges of safeguarding space systems against new threats. Refer to Figure 5 ([[PenBox architecture]], 2022)

LLMs

Going from the realm of autonomous pen testing like we have seen in the past 2 AI models. LLMs are being utilized for a more manual approach with extra assistance. This is important to note due to how widely accessible LLMs like ChatGPT are available to everyday users on the internet. “Combining human operators with AIs creates new capabilities instead of cloning existing ones. Furthermore, keeping a human in the loop reduces the potential ethical

problems imposed by the use of AIs [6]. Recent research indicates that the efficiency gains provided by the use of AI-based systems are greatest for low-skilled workers [7], augmenting human operators with a generative AI might thus also benefit the training of novice penetration testers." Happe, A., & Cito, J. (2023).

This dilemma presents a significant challenge for technology users across the globe. For most, accessing guidance on creating an .exe file embedded within a PDF—capable of execution upon a click—seems as straightforward as requesting a step-by-step tutorial from ChatGPT. However, it is important to note that ChatGPT has safeguards in place to prevent any form of assistance related to hacking, malware, or unethical activities. Despite these safeguards, individuals often circumvent these restrictions and receive the information they seek from the AI model. "During the development of the script, the ethics filter was infrequently triggered. Adding "do not ask questions or provide judgments" to command prompts seems to significantly reduce denials. The optional "detail additional vulnerabilities" step was more often denied due to ethical reasons, but this had no impact on the overall hacking progress. Slight prompt variations were successful in reducing GPT3.5's ethical concerns, e.g., instead of asking for "exploits for vulnerabilities" we asked for "verification commands for vulnerabilities" Happe, A., & Cito, J. (2023).

Training and Education

Educational institutions and training providers are recognizing the importance of incorporating AI-driven penetration testing courses into their curriculum to equip aspiring cybersecurity professionals with the requisite knowledge and skills to navigate this evolving landscape effectively. These courses delve into the theoretical foundations of AI, machine

learning algorithms, and their applications in penetration testing, providing students with hands-on experience in utilizing AI-powered tools and techniques to identify and eliminate security vulnerabilities. By immersing learners in real-world scenarios and practical exercises, these programs not only enhance their technical proficiency but also foster critical thinking and problem-solving abilities crucial for effective cybersecurity practice. A few of these courses include ones from Pluralsight, the SANS Institute, and Cybrary.

Continuous learning and professional development are encouraged to stay ahead of evolving trends and advancements in AI-driven cybersecurity. Overall, training and education initiatives play a pivotal role in equipping cybersecurity professionals with the knowledge and skills needed to leverage AI technologies effectively in penetration testing practices.

Present-day Obstacles

In AI-driven penetration testing, ethical considerations and trust are essential. The adoption of AI technologies raises concerns regarding privacy, fairness, and accountability. Ensuring ethical AI deployment has transparency in algorithmic decision-making, addressing biases, and establishing clear oversight. Transparency and explainability are essential for building trust in AI systems, allowing users to understand their operations and assess their reliability. Additionally, mitigating bias and promoting fairness in AI algorithms is crucial to prevent unwanted outcomes in vulnerability assessments and threat identification.

Accountability and responsibility are integral aspects of ethical AI usage in penetration testing. Organizations must clarify roles and establish mechanisms for oversight and accountability to address potential adverse consequences of AI-generated actions. Collaboration and transparency among stakeholders are essential for fostering trust in AI-driven penetration

testing. Open dialogue and knowledge-sharing promote a mutual understanding of AI technologies' capabilities and limitations, creating responsible innovation in cybersecurity practices. By prioritizing transparency, fairness, and ethical conduct, organizations can build trust in AI-driven penetration testing and uphold ethical standards while leveraging AI's potential to enhance cybersecurity resilience.

Security and Privacy Implications

While AI is not new in general, its application within cybersecurity is recent and increasingly accessible. The main considerations revolve around privacy concerns, adherence to legal regulations, and the potential for AI tools to be weaponized. In today's cybersecurity world, threat actors possess access to a many automated tools, such as Burp Suite, Metasploit, and Kali Linux, which streamline the process of exploitation. However, defenders also utilize these very tools to protect networks against malicious activities daily. Introducing additional tools that can conduct scans, formulate attack strategies, execute exploits, and document findings with minimal human intervention raises concerns about the potential threat posed to networks, especially when initiated by an actor targeting a victim machine.

Future Directions and Opportunities

The future trajectory of AI remains unclear, given its availability to the public. However, as technology continues to evolve, it holds potential to address several everyday challenges. In the realm of cybersecurity, the landscape is full of possibilities for AI-driven innovations, with many reports in this paper highlighting key points and trends in current pen testing tools, "On the one hand, we plan to expand the training dataset with additional network topologies, so as to improve the versatility and stability of the DQN model. On the other hand, we consider

integrating a network service scanning function into the framework, so that information on real target environments can be automatically provided to the DQN model, leading to more accurate results for actual network topologies." Hu, Z. (2020). Others talk about how AI will serve as a companion along side humans in pen testing to get accurate results, "However, challenges such as inaccurate results, financial constraints, and the lack of human expertise highlight the importance of careful supervision and management. The future of pen testing likely involves a hybrid approach, where AI complements human testers rather than replaces them entirely."

As stated before, the role of AI in cybersecurity remains uncertain, yet promising. With ongoing advancements and the evolving landscape of cyber threats, the future of AI in penetration testing holds immense potential. As technologies continue to mature and adapt, AI's capacity to advance security measures and strengthen defenses against detailed attacks becomes more apparent. However, the journey towards integrating AI seamlessly into cybersecurity will require exploration, innovation, and collaboration among industry experts, researchers, and practitioners. Embracing the transformative power of AI, while addressing potential challenges and ethical considerations, will undoubtedly shape the trajectory of cybersecurity in the years to come.

Conclusion

In conclusion, the integration of Artificial Intelligence into penetration testing represents a significant advancement in the realm of cybersecurity. As explored in this paper, AI's introduction into cyber defense strategies introduces a new era of efficiency and effectiveness in identifying and mitigating cyber threats. By leveraging AI's ability to analyze vast amounts of data, detect anomalies, and predict potential vulnerabilities, organizations can enhance their

security posture and better safeguard their digital assets. However, while AI offers immense promise, its implementation also raises important considerations regarding ethics, privacy, and bias, which must be carefully addressed to ensure responsible and effective use. Moving forward, continued research, innovation, and collaboration will be essential to unlock the full potential of AI in penetration testing and strengthen our collective defense against evolving cyber threats.

References

Kuswandana, L. (2021). The Usage of Machine Learning on Penetration Testing Automation. IEEE Xplore, 2021.10335188. https://ieeexplore-ieee-org.ezproxy.uvu.edu/document/10335188

Hoang, L. V. (2022). Leveraging Deep Reinforcement Learning for Automating Penetration Testing in Reconnaissance and Exploitation Phase. IEEE Xplore https://ieeexplore-ieee-org.ezproxy.uvu.edu/document/10013801

Seelen Jagamogan, R., Ismail, S. A., Hassan, N. H., & abas, H. (2022) Penetration Testing Procedure using Machine Learning. https://ieeexplore-ieee-org.ezproxy.uvu.edu/document/9870951

Happe, A., & Cito, J. (2023) Getting Pwn'd by AI: Penetration testing with large language models. arXiv.org. https://arxiv.org/abs/2308.00121

Hu, Z. (2020, October). Automated Penetration Testing Using Deep Reinforcement Learning. IEEE Xplorehttps://ieeexplore-ieee-org.ezproxy.uvu.edu/document/9870951/authors

Singh, N. (2020, October). Automated versus Manual Approach of Web Application Penetration Testing. IEEE Xplore https://ieeexplore-ieee-org.ezproxy.uvu.edu/document/9225385

Confido, A. (2022, August). Reinforcing Penetration Testing Using AI. IEEE Xplore https://ieeexplore-ieee-org.ezproxy.uvu.edu/document/9843459

Chaudhary, S. (2020, August). Automated Post-Breach Penetration Testing through Reinforcement Learning. IEEE Xplore https://ieeexplore-ieee-org.ezproxy.uvu.edu/document/9162301

Ghanem, M. C. (2019, January). Reinforcement Learning for Intelligent Penetration Testing. IEEE Xplore https://ieeexplore-ieee-org.ezproxy.uvu.edu/document/8611595

Gale Business. (2022, September). Global Penetration Testing Market Report 2022: Integration of AI, Ml, and Other Technologies in Penetration Testing Presents Opportunities. https://go-gale-com.ezproxy.uvu.edu/ps/i.do?p=GBIB&u=utahvalley&id=GALE%7CA719461270&v=2.1&it=r&sid=ebsco&aty=ip

Imperva. (2023, December 20). What is penetration testing: Step-by-step process & methods: Imperva. Learning Center. https://www.imperva.com/learn/application-security/penetration-testing/

NIST. (2024). Penetration testing - glossary*: CSRC*. CSRC Content Editor. https://csrc.nist.gov/glossary/term/penetration_testing

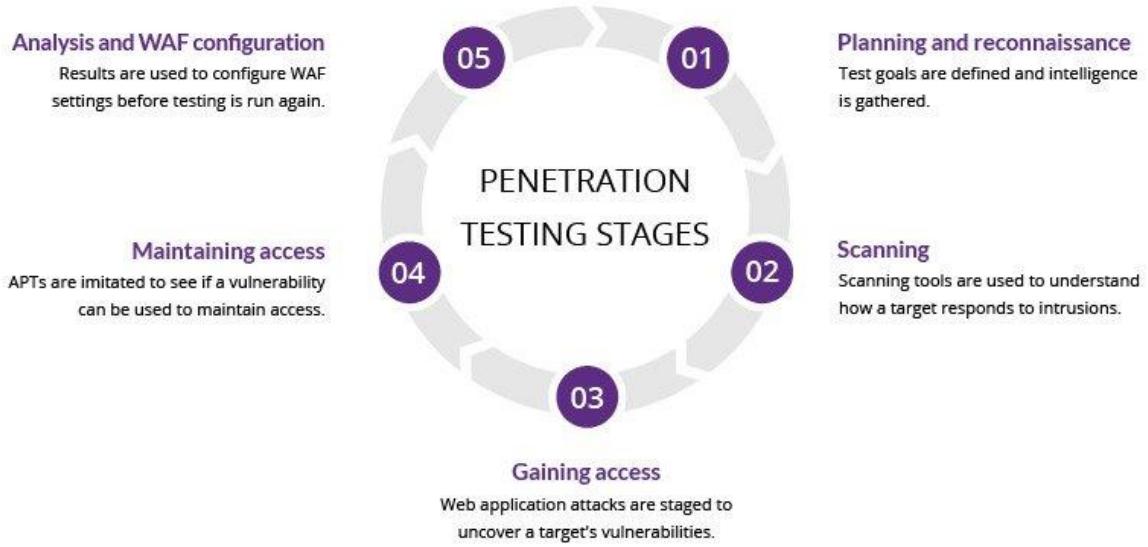
Rai, V. (2023, May 24). Chatgpt for pen testing (pt. 1). Redfox Security. https://redfoxsec.com/blog/chatgpt-for-pen-testing-part-1/#:~:text=The%20application%20of%20ChatGPT%20in,to%20phishing%20or%20impersonation%20attempts.

Mukherjee, A. (2024, March 27). Can ai enhance penetration testing?. Evolve Security Automation and Orchestration by Threat Intelligence. https://www.threatintelligence.com/blog/ai-penetration-testing

Figures

Figure 1

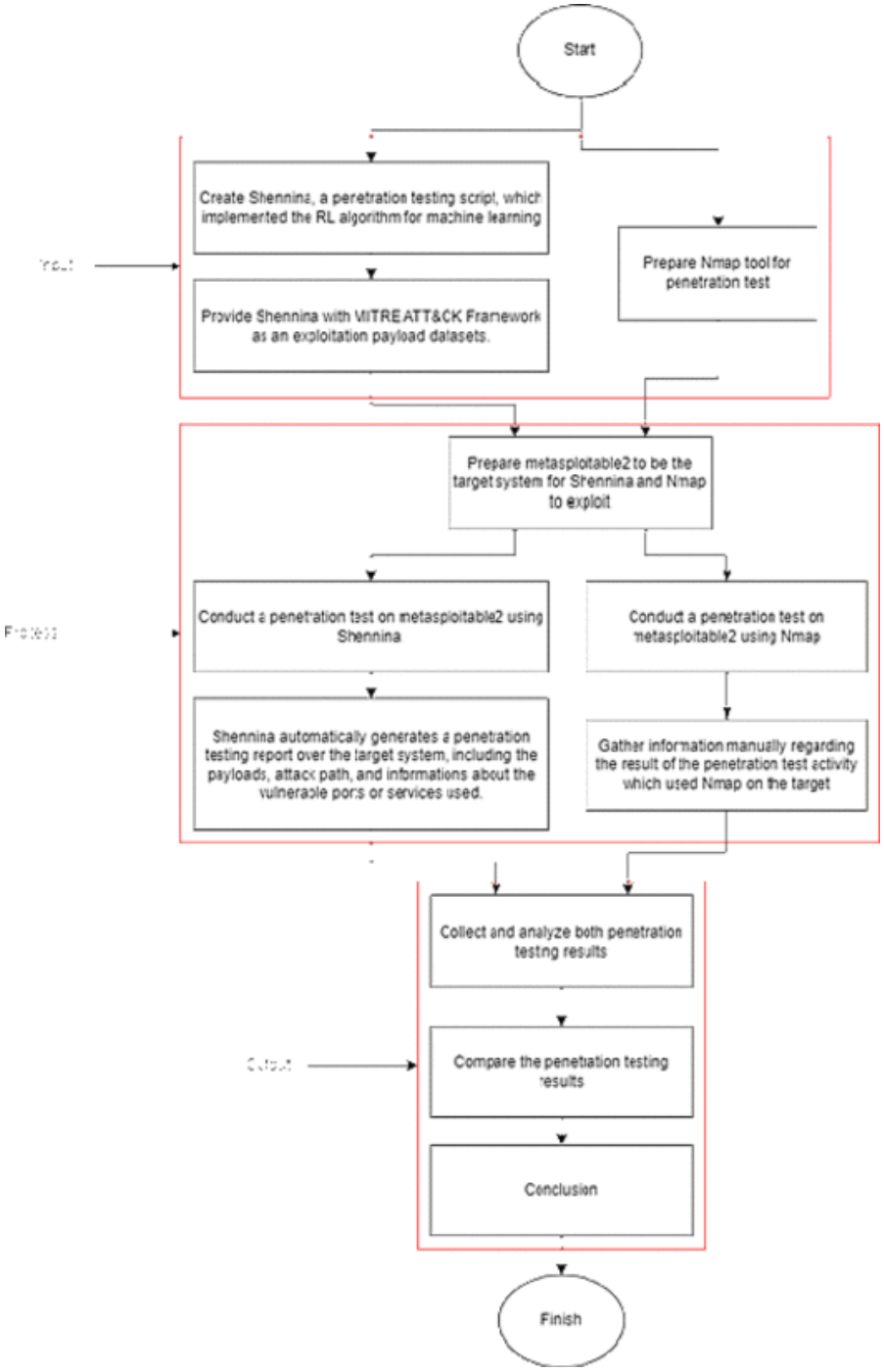
Stages of Penetration Testing



[Stages of Penetration Testing]. (2024). https://www.imperva.com/learn/application-security/penetration-testing/

Figure 2

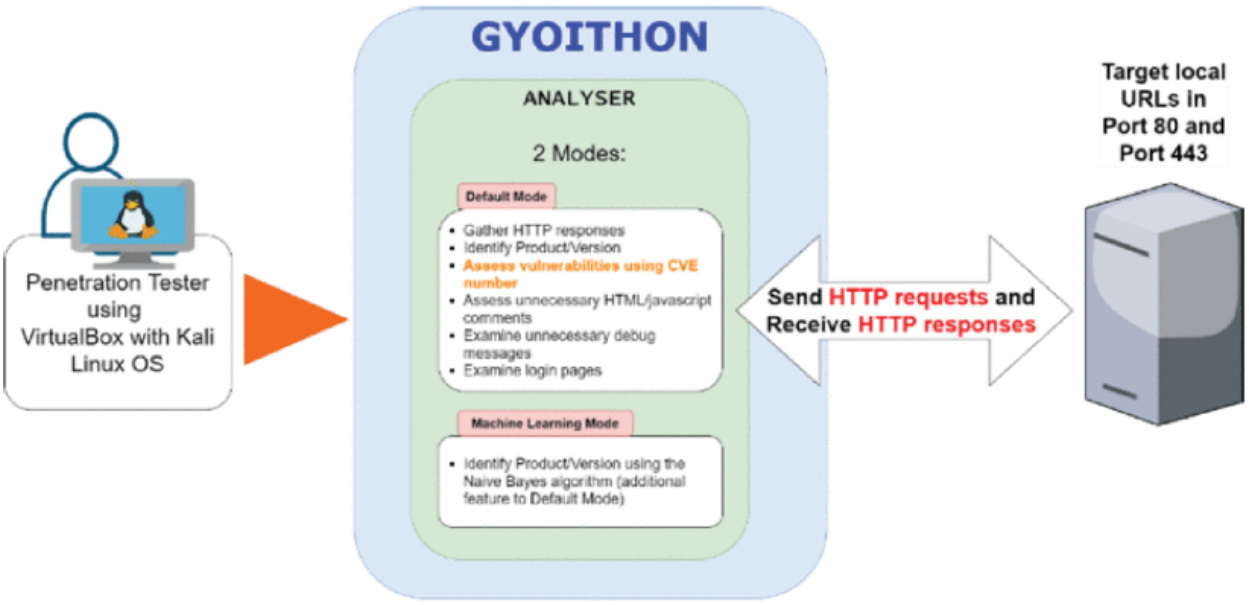
Shennina Methodology



[Shennina Methodology]. (2021) https://ieeexplore-ieee-org.ezproxy.uvu.edu/document/10335188

Figure 3

GyoiThon Procedure



[GyoiThon Procedure]. (2022) https://ieeexplore-ieee-org.ezproxy.uvu.edu/document/9870951

Figure 4

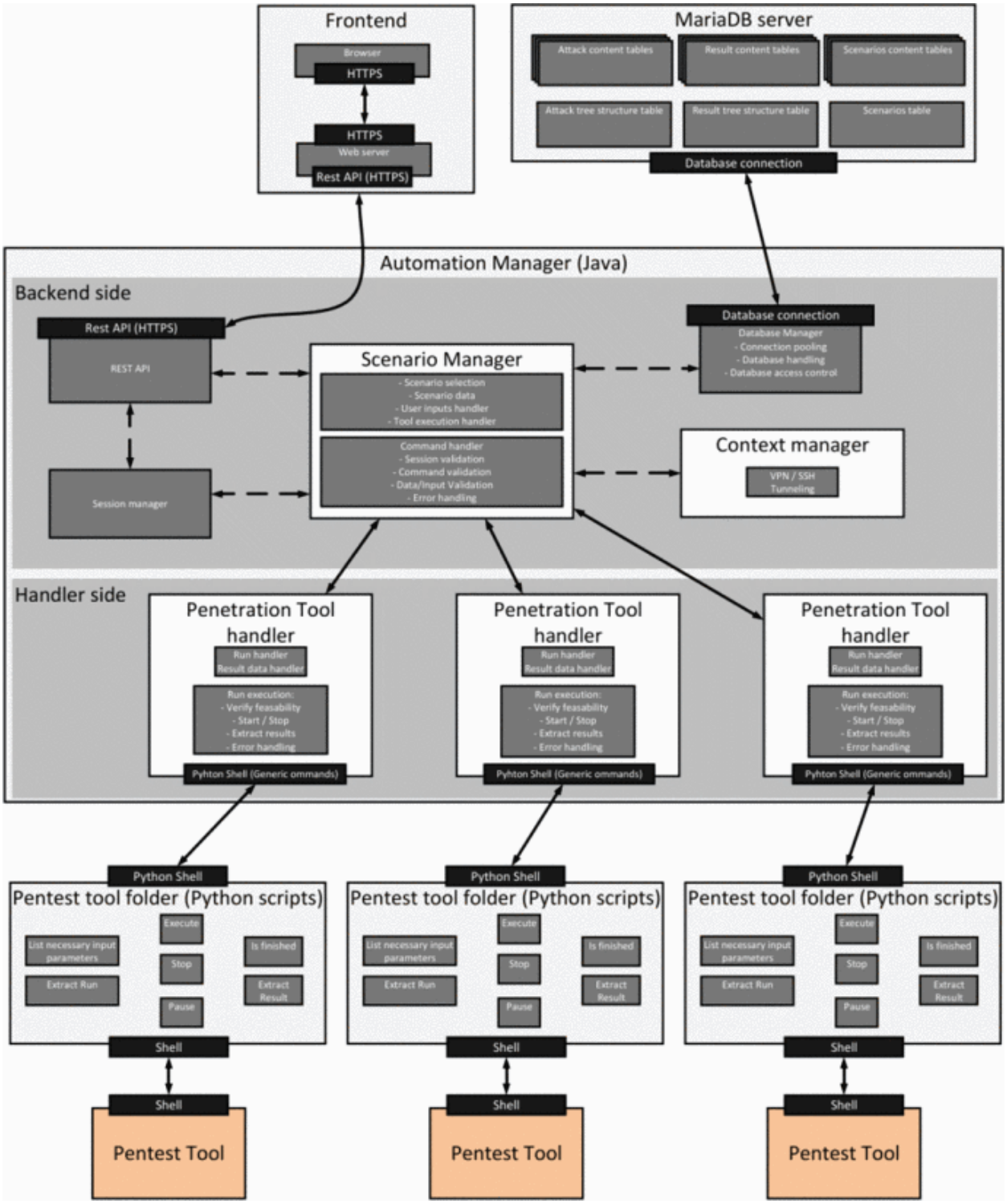
Default Mode vs ML Mode

Vulnerabilities identified by its CVE number	Default Mode				Machine Learning (ML) mode			
	Port 80		Port 443		Port 80		Port 443	
	CMS Websites	*All Websites*	*CMS Websites*	*All Websites*	*CMS Websites*	*All Websites*	*CMS Websites*	*All Websites*
CVE-2006-5877	54	101	7	10	54	101	3	10
CVE-2007-2834	54	101	7	10	54	101	3	10
CVE-2008-1459	-	-	-	-	24	49	-	-
CVE-2008-4668	-	-	-	-	24	49	-	-
CVE-2008-5053	-	-	-	-	24	49	-	-
CVE-2010-0425	54	101	7	10	54	101	3	10
CVE-2010-0834	54	101	7	10	54	101	3	10
CVE-2010-2068	54	101	7	10	54	101	3	10
CVE-2010-4156	54	101	7	10	54	101	3	10
CVE-2011-0754	54	101	7	10	54	101	3	10
CVE-2011-5254	48	48	-	-	72	72	-	-
CVE-2012-2376	54	101	7	10	54	101	3	10
CVE-2012-3575	48	48	-	-	72	72	-	-
CVE-2015-4642	1	26	-	-	1	26	-	-
CVE-2016-7405	1	26	-	-	1	26	-	-
CVE-2016-8670	1	26	-	-	1	26	-	-
CVE-2020-26596	48	48	-	-	72	72	-	-
Unidentified vulnerabilities (*Cannot search*)	187	334	9	30	187	361	9	30
TOTAL	766	1,364	65	150	910	1,610	33	150

[Default Mode vs ML Mode]. (2022) https://ieeexplore-ieee-org.ezproxy.uvu.edu/document/9870951

Figure 5

PenBox architecture



[PenBox architecture]. (2022) https://ieeexplore-ieee-org.ezproxy.uvu.edu/document/9843459