IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, March 2026
Impact Factor: 8.2

Review of AI-Assisted and Automated Penetration Testing Techniques

Aditya Agale, Arpit Kadam, Vedangee Poyeraker, Dr. Renuka Deshpande
Dept. of Artificial Intelligence & Machine Learning
Shivajirao S. Jondhale College of Engineering, Dombivli (E), India

Abstract: *This review surveys recent advances in automated and AI-assisted penetration testing, focusing on the design and evolution of intelligent security assessment frameworks. We examine the transition from traditional manual and tool-driven penetration testing approaches toward automated systems enhanced by machine learning and large language models (LLMs). The paper analyzes how modern AI-based techniques integrate reconnaissance tools, vulnerability scanners, and reasoning models to improve contextual understanding and decision-making during security assessments. A critical review of existing automated and LLM-assisted penetration testing frameworks is presented, comparing their capabilities, limitations, and practical applicability in enterprise environments. The review further discusses challenges related to false positives, scalability, explainability, and ethical considerations in AI-driven security testing. By synthesizing state-of-the-art literature and identifying key research gaps, this review aims to provide researchers and practitioners with a clear understanding of current trends, limitations, and future directions in intelligent penetration testing.*

Keywords: *Artificial intelligence*

I. INTRODUCTION

Penetration Penetration testing is a vital cybersecurity practice. It is used to detect vulnerabilities in systems, networks, and web applications before attackers exploit them. Recent studies have introduced penetration testing frameworks that incorporate large language models, such as PentestGPT, AutoPen, VulnBot, and xOffense. These systems demonstrate how language models assist in reconnaissance, analytical reasoning, and attack planning tasks [1]–[5]. Such developments indicate a shift toward more intelligent, semi-autonomous security evaluation approaches.

Beyond LLM frameworks, several studies have examined how artificial intelligence is being incorporated into modern penetration testing environments. These studies highlight the transition from traditional manual methods to more intelligent and automated approaches [6]. AI-driven vulnerability scanners and automated web assessment tools enhance detection efficiency across different attack surfaces [7]. Formal automation models also contribute to structuring and standardizing penetration testing workflows [8]. Systematic reviews of LLM-based vulnerability detection and AI-powered testing tools provide deeper academic insight into the development and effectiveness of these techniques [9], [10].

More Machine learning and reinforcement learning techniques have been examined to enhance adaptability and decision-making capabilities in autonomous penetration testing systems [11]. Large language models are also utilized for security data analysis and CVE detection tasks [12], while research in AI-assisted software testing supports automated vulnerability identification processes [13]. Conventional web application penetration testing tools continue to serve as baseline methods for evaluating AI-driven approaches [14], [15], and broader studies on generative AI in cybersecurity provide context for these emerging advancements [16].

Systematic mapping studies and analytical reviews explore the evolving role of AI within penetration testing research [17]. Collaborative testing approaches for generative AI systems introduce alternative evaluation methods [18], while industry-focused research on LLM-assisted vulnerability analysis offers practical implementation perspectives [19]. Structured and curriculum-based autonomous testing frameworks seek to enhance staged automation strategies [20].

DOI: 10.48175/IJARSCT-31870

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, March 2026 Impact Factor: 8.2

Despite these developments, multiple challenges remain, including false positives in automated scanners, scalability issues, and deployment constraints within DevSecOps environments [21], [22]. Further reviews of AI-based automated penetration testing tools highlight the importance of stronger comparative evaluation across existing approaches [23]. Conformity with established vulnerability standards such as the OWASP Top 10 enables structured risk categorization [24], while internet-connected device search platforms demonstrate practical reconnaissance capabilities relevant to security assessments [25].

II. REVIEW METHODOLOGY

This review adopts a structured methodology to examine existing studies on automated and AI-assisted penetration testing. The aim is to analyze, compare, and synthesize earlier research to identify prevailing trends and limitations within this domain. Foundational LLM-based penetration testing frameworks, including PentestGPT, AutoPen, VulnBot, and xOffense, illustrate how large language models support autonomous security testing systems [1]–[4].

A Subsequent research expanded agent-based architectures and assessed the broader integration of artificial intelligence in modern penetration testing environments [5], [6]. AI-driven vulnerability scanners and structured automation models reflect the movement toward more intelligent security evaluation processes [7], [8]. Comprehensive reviews of LLM-based vulnerability detection and advancements in AI-powered testing tools further reinforce the academic and practical basis of automated penetration testing [9], [10].

The research also examines reinforcement learning and machine learning approaches to enhance adaptability and decision-making in autonomous penetration testing systems [11]. Large language models are utilized for security data analysis and CVE detection activities [12], while studies on AI-assisted software testing support automated vulnerability identification techniques [13]. Conventional web penetration testing tools are reviewed as baseline references for comparison with AI-driven methods [14], [15], and broader studies on generative AI in cybersecurity place these advancements within a wider technological context [16].

To gain a clearer understanding of research trends, systematic mapping studies and collaborative AI-based penetration testing approaches are considered [17], [18], together with industry-oriented analyses of LLM-assisted vulnerability research [19]. Structured autonomous web testing frameworks further illustrate staged automation methods [20].

Practical concerns such as false positives in automated scanners and continuous security testing within DevSecOps environments are examined to assess real-world deployment factors [21], [22]. Additional reviews of AI-based automated penetration testing tools offer comparative perspectives across different methodologies [23]. The discussion also aligns with established vulnerability standards such as the OWASP Top 10 [24] and incorporates real-world reconnaissance capabilities through internet-connected device search platforms [25].

III. ANALYSIS OF EXISTING TECHNIQUES

The domain of automated and AI-assisted penetration testing has progressed considerably with developments in machine learning, reinforcement learning, and large language models. Initial research largely concentrated on conventional tool-based and rule-driven techniques for vulnerability detection and exploitation. These approaches depended heavily on predefined signatures and manual analysis of outputs, which restricted scalability and limited contextual reasoning capabilities [6], [8].

A. Traditional Penetration Testing Approaches

Initial penetration testing methods mainly depended on manual evaluation and rule-based security tools. Security analysts performed reconnaissance, identified vulnerabilities, and carried out exploitation using predefined scripts along with their domain expertise. Although these techniques offered in-depth insights, they were heavily reliant on human skill, required significant time, and were challenging to scale across large and evolving infrastructures [6], [8]. Additionally, such approaches may lead to inconsistent results due to differences in analyst experience and capability.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, March 2026
Impact Factor: 8.2

B. Automated Vulnerability Scanning Techniques

To enhance efficiency, automated vulnerability scanners were developed to handle activities such as network scanning, service enumeration, and web application assessment. By integrating multiple scanning components, these tools reduce manual workload and expand coverage across target systems [7], [10]. However, research shows that automated scanners frequently produce numerous false positives and lack contextual awareness, making manual verification necessary to assess the true security impact of identified vulnerabilities [21].

C. Machine Learning-Based Security Assessment Techniques

Recent studies have investigated the application of machine learning techniques to improve vulnerability detection and prioritization. These machine learning–driven methods examine historical vulnerability data, system behavior, and application characteristics to identify security weaknesses more precisely than conventional rule-based approaches. Although detection performance has improved, many of these methods remain task-specific and lack the capability to reason across different stages of the penetration testing process [9], [11], [14].

D. LLM-Assisted and Autonomous Penetration Testing Frameworks

With the rise of large language models, researchers have introduced LLM-assisted penetration testing frameworks that can interpret vulnerability descriptions, correlate scanning results, and support attack planning. Platforms such as PentestGPT and AutoPen exhibit enhanced automation and contextual awareness through natural language reasoning and coordinated multi-agent mechanisms [1], [2]. Multi-agent architectures further increase autonomy by allocating responsibilities such as reconnaissance, scanning, and reporting to specialized agents [3]–[5].

E. Challenges and Limitations of Existing Techniques

Despite considerable advancements in automated and AI-assisted penetration testing, current techniques still encounter several challenges that restrict their practical use. Many proposed frameworks have not been validated at large enterprise scale and are mostly tested in controlled or simulated environments, which limits their generalizability and scalability in real-world settings [3], [5]. A widely reported issue in the literature is the high rate of false positives produced by automated vulnerability scanners and AI-driven models, requiring manual confirmation by security analysts [9], [11].

Another significant limitation concerns the explainability and reliability of AI-assisted penetration testing systems. Although large language model–based frameworks offer advanced reasoning and automation capabilities, they may generate inconsistent or hallucinated outputs, creating concerns regarding trust and transparency in security evaluations [12], [17]. These issues, together with ongoing false positive challenges observed in automated systems [21]

Table I. Comparison of AI-Assisted Penetration Testing Frameworks

Ref	Technique	AI Used	Domain	Key Limitation
[1]	LLM-assisted Pentesting	GPT-based LLM	Web & Network	Limited enterprise validation
[2]	Autonomous Pentesting	LLM Agents	Web Apps	High computational overhead
[3]	Multi-Agent Framework	LLM + Agents	Network	Simulated evaluation only
[4]	AI-Driven Pentesting	Multi-Agent AI	Web & Network	Scalability concerns
[5]	LLM Agent Framework	LLM	Web Apps	LLM reasoning dependence
[6]	Survey Study	Machine Learning	General	No experimental validation
[9]	Vulnerability Detection	LLM	Software	Mostly conceptual analysis
[10]	RL-Based Testing	Reinforcement Learning	Network	Training instability

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, March 2026
Impact Factor: 8.2

IV. DISCUSSION AND IDENTIFIED RESEARCH GAPS

The reviewed literature indicates that artificial intelligence has substantially improved the automation of penetration testing activities. AI-assisted and LLM-based frameworks enhance vulnerability detection, contextual reasoning, and automation across multiple phases of the penetration testing process [1], [2], [6]. Nevertheless, existing studies highlight several unresolved issues and research gaps that must be addressed to achieve dependable, scalable, and enterprise-ready automated penetration testing solutions. Moreover, although many frameworks report promising experimental outcomes, their real-world implementation in complex and diverse enterprise environments remains limited.

A. Limited Contextual Understanding and Generalization

Although AI-driven methods enhance detection performance, many systems depend on task-specific models or predefined scenarios, restricting their ability to generalize across varied real-world environments [9], [11]. Differences in system configurations, application architectures, and evolving threat landscapes often diminish the effectiveness of trained models when applied beyond controlled testing conditions [14].

B. Explainability and Trust in AI-Assisted Penetration Testing

Multiple studies emphasize the limited explainability of AI-based penetration testing frameworks, especially those built on large language models [12], [17]. The lack of transparent reasoning and clear justification for vulnerability prioritization lowers confidence in automated results and slows adoption within enterprise security workflows where interpretability is essential [21].

C. Scalability and Real-World Deployment Challenges

Most proposed frameworks are tested in small-scale or simulated settings and lack validation within large, complex enterprise infrastructures [3], [5]. Factors such as computational overhead, integration with existing security systems, and support for continuous assessment are often insufficiently addressed, revealing a gap between experimental prototypes and fully deployable penetration testing solutions [18].

D. Evaluation Metrics and Real-World Validation

A major research gap also involves the absence of standardized evaluation metrics for measuring the effectiveness of automated penetration testing systems [6], [10]. Many studies depend on limited benchmarks or qualitative assessments, which makes cross-comparison and robustness evaluation difficult across diverse environments. Comprehensive real-world testing and unified evaluation frameworks remain largely missing in current research efforts [22].

Summary

In summary, while existing research demonstrates the potential of AI-assisted and LLM-based penetration testing techniques, several critical gaps remain unresolved. Limitations related to generalization, explainability, scalability, and evaluation hinder the transition of current research prototypes into reliable enterprise solutions. Addressing these challenges through standardized benchmarks, explainable AI techniques, and large-scale validation is essential for advancing the practical adoption of intelligent penetration testing systems.

V. CONCLUSION AND FUTURE DIRECTIONS

This review paper presents a comprehensive analysis of existing research on automated and AI-assisted penetration testing techniques. By examining traditional approaches, automated scanning tools, machine learning–based methods, and recent LLM-assisted frameworks, the study highlighted how artificial intelligence has significantly enhanced the

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, March 2026 Impact Factor: 8.2

efficiency and scope of penetration testing processes. The reviewed literature demonstrates a clear shift toward greater automation and contextual reasoning in modern security assessment systems.

However, the analysis also revealed that despite notable advancements, current approaches face persistent challenges related to scalability, explainability, generalization, and real-world deployment. These limitations indicate that while AI-assisted penetration testing shows strong potential, further research is required to bridge the gap between experimental frameworks and practical enterprise-ready solutions. Overall, this review provides a structured understanding of current techniques and identifies key research gaps that can guide future advancements in intelligent penetration testing.

Based on the analysis of existing literature, future research should focus on improving the explainability and transparency of AI-assisted penetration testing systems to enhance trust and adoption in enterprise environments. Developing standardized evaluation metrics and benchmarks would also enable more consistent comparison of different approaches across diverse real-world scenarios. Additionally, integrating continuous assessment capabilities and improving scalability remain critical areas for further investigation.

In addition to the reviewed studies, the proposed project aims to explore the integration of LLM-assisted reasoning with existing penetration testing tools to provide more contextual vulnerability analysis and prioritization. The project will focus on consolidating scan outputs and presenting actionable insights rather than proposing new detection techniques, thereby aligning with the research gaps identified in the literature.

REFERENCES

[1] Deng, J., et al., "PentestGPT: Evaluating and Harnessing Large Language Models for Automated Penetration Testing," Proceedings of the USENIX Security Symposium, 2024.
[2] Xu, S., et al., "AutoPen: Towards Autonomous Penetration Testing Using Large Language Models," arXiv preprint arXiv:2308.06782, 2023.
[3] Zhang, H., et al., "VulnBot: A Multi-Agent Autonomous Penetration Testing Framework," arXiv, 2024.
[4] Li, Y., et al., "xOffense: AI-Driven Multi-Agent Framework for Autonomous Penetration Testing," arXiv, 2024.
[5] Wang, R., et al., "AutoPentester: LLM Agent-Based Framework for Automated Penetration Testing," arXiv, 2024.
[6] Sharma, A., and Gupta, R., "The Role of Artificial Intelligence in Modern Penetration Testing: A Systematic Review," IEEE Access, 2023.
[7] Khan, M., et al., "Smart Web Vulnerability Scanner Using AI-Based Autonomous Agents," International Journal of Engineering Research & Technology (IJERT), 2022.
[8] Verma, P., et al., "Automated Penetration Testing: Formalization and Realization," Computers & Security, Elsevier, 2023.
[9] Zhou, Y., et al., "LLM-Based Vulnerability Detection: A Systematic Literature Review," arXiv, 2024.
[10] Kim, J., et al., "Enhancing Cybersecurity with AI-Powered Penetration Testing Tools," IEEE Access, 2022.
[11] Rahman, F., et al., "Analysis of Autonomous Penetration Testing Through Reinforcement Learning," Sensors, MDPI, 2023.
[12] Patel, N., et al., "Leveraging Large Language Models for Security Data Analysis and CVE Detection," Applied Sciences, MDPI, 2023.
[13] Al-Amin, S., "AI-Assisted Software Testing and Vulnerability Detection," Master's Thesis, LUT University, 2022.
[14] Singh, R., and Kumar, P., "A Survey on Web Application Penetration Testing Tools," Electronics, MDPI, 2022.
[15] Mehta, S., et al., "New Approaches for Web Application Vulnerability Scanning," International Journal for Research in Applied Science & Engineering Technology (IJRASET), 2021.
[16] Deng, L., et al., "Generative AI in Cybersecurity: A Comprehensive Review," Journal of Information Security and Applications, 2023.
[17] Hassan, M., et al., "AI in Penetration Testing: A Systematic Mapping Study," TechRxiv, 2023.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, March 2026
Impact Factor: 8.2

[18] Sommer, M., et al., "Collaborative Penetration Testing for Generative AI Systems," IEEE Security & Privacy, 2024.
[19] Bishop Fox Research, "LLM-Assisted Vulnerability Research and Patch Diffing," Technical Report, 2023.
[20] Lin, Z., et al., "CurriculumPT: LLM-Based Autonomous Web Penetration Testing Framework," Applied Sciences, MDPI, 2024.
[21] Torres, P., et al., "False Positives in Automated Vulnerability Scanning Tools: Causes and Mitigation," IEEE Security & Privacy, 2022.
[22] Sommer, R., et al., "Continuous Security Testing in DevSecOps Environments," IEEE Software, 2023.
[23] Kaur, R., and Singh, S., "AI-Based Automated Penetration Testing Tools: A Review," Journal of Cyber Security Technology, 2023.
[24] OWASP Foundation, "OWASP Top 10 Web Application Security Risks," 2021.
[25] Shodan, "Shodan: Search Engine for Internet-Connected Devices," Technical Documentation, 2023