

# AI and ML in Penetration Testing: A Comprehensive Review

Soham Deshmukh<sup>1</sup>, Mohanish Kulkarni<sup>2</sup>, Jayam Mehta<sup>3</sup>, Pallavi Akolkar<sup>4</sup>

<sup>1,2,3,4</sup>Department of Information Technology,

Trinity College of Engineering and Research, Pune, India

\*\*\*

**Abstract -** This review surveys state-of-the-art research on the integration of Artificial Intelligence (AI) and Machine Learning (ML) in cybersecurity, critically synthesizing findings across 20+ foundational papers. Spanning both defensive and offensive perspectives, the analysis covers automated penetration testing, deep learning-based intrusion detection, explainable AI, adversarial attacks, IoT and industrial infrastructure security, and educational strategies for developing future expertise. Key results highlight substantial progress in automating vulnerability identification, scaling threat prediction, and enhancing attack simulation. Limitations persist in adversarial robustness, explainability, real-time adaptation, and human-AI collaboration. The literature lacks standardized performance metrics and sufficient cross-sector partnerships. The paper synthesizes consensus and debates, identifies critical gaps, and lays out future research directions that prioritize ethical, scalable, and resilient cybersecurity solutions leveraging synergistic AI-human teams.

**Key Words:** Artificial Intelligence; Machine Learning; Cybersecurity; Penetration Testing; Deep Learning; Explainable AI; Adversarial Attacks; Intrusion Detection

## 1. INTRODUCTION

Cybersecurity has ascended to strategic prominence in an era of ubiquitous connectivity, smart infrastructure, and digital transformation. Network and application attack surfaces have grown, and manual security processes now struggle to contend with the scale and sophistication of cyber threats. The emergence of Artificial Intelligence (AI) and Machine Learning (ML) offers a paradigm shift, enabling automated, adaptive, and predictive mechanisms for threat detection, mitigation, and security operations management. This review explores the recent literature defining the most impactful advances, the consensus and controversy surrounding major approaches, and unaddressed obstacles that impede robust, ethical, and effective deployment. The objectives are to evaluate core methodologies, unify perspectives from defense and offense, and provide a roadmap for future research and practice linking technical innovation with societal resilience.

## 2. BACKGROUND AND THEORETICAL FOUNDATIONS

### 2.1 Foundational Concepts

AI encompasses a suite of intelligent computational methods, including supervised, unsupervised, and reinforcement learning, with applications extending to deep learning (DL) architectures such as CNN, RNN, LSTM, GAN, and DBN. ML models feature hierarchical feature extraction and data driven learning that optimize for both detection accuracy and adaptability.

Key concepts include adversarial machine learning—attackers manipulating training data or models, explainability—making AI model decisions interpretable to humans, and automated planning—using AI agents to simulate real-world attack and defense actions.

### 2.2 Standardized Frameworks

Research increasingly builds on operational and adversarial frameworks:

- MITRE ATT&CK and CAPEC: Attack taxonomy and procedural modeling [19], [20].
- Lockheed Martin Cyber Kill Chain: Structured phases of cyber attacks.
- OWASP IoT Framework: Definitions of IoT attack surfaces.

These frameworks anchor experimental setups, vulnerability assessments, and the training of automated AI/ML security agents.

### 2.3 Key Cybersecurity Activities

The research papers target penetration testing (PT), intrusion detection systems (IDS), malware analysis, vulnerability scanning, risk propagation modeling, adversarial threat simulation, and the development of cyber defense curricula.

## 3. LITERATURE REVIEW

### 3.1 Automated Penetration Testing and Attack Simulation:

Recent advances demonstrate AI-driven penetration testing frameworks, where reinforcement learning (often Deep Q E. AI for Critical Infrastructure and IoT Networks) and AI planners learn optimal attack paths in simulated network environments [4], [5], [7]. These systems mimic skilled

hackers, evaluating exploit sequences, discovering hidden vulnerabilities, and generalizing across diverse topologies. For IIoT networks, probabilistic modeling (e.g., particle filters) and attack tree generation automate the identification of exposed assets, deliver real-time risk analysis, and reduce manual workload [6], [10]. Studies on penetration testing processes [8], [9] provide comprehensive overviews of methodologies and tools.

### 3.2 Deep Learning and Intrusion Detection

Deep learning (DL), including CNN, RNN, LSTM, GAN, DBN, and autoencoder models, revolutionize intrusion detection and malware classification. These architectures excel at:

- Detecting patterns in massive network traffic streams
- Classifying unknown attacks (zero-days) with improved recall and precision
- Extracting features automatically, reducing reliance on handcrafted heuristics.

Historical benchmarking (NSL-KDD, ISCX datasets) reveals DL significantly outperforms legacy rule-based systems [11], [12]. Recent adaptations include distributed, fog-based DL for IoT settings and GANs for adversarial resilience.

Chen's work [1] on internet security situation awareness prediction using improved RBF neural network algorithms demonstrates the application of neural networks in cybersecurity contexts. Wani et al. [2] developed an SDN-based intrusion detection system for IoT using deep learning classifier (IDS-IOT-SDL), showcasing practical implementations of DL in IoT security.

### 3.3 Adversarial and Offensive AI

Literature documents the weaponization of AI/ML by attackers, spanning DeepPhish for phishing attacks, PassGAN for password guessing, DeepLocker for evasive malware, and GAN-based generation of adversarial samples to bypass security tools. Seven action categories, including probe, scan, spoof, flood, misdirect, execute, and bypass, illustrate the breadth of offense enabled by AIMLS (AI/ML software). These capabilities outpace static defense strategies, demanding rapid adaptation from defenders.

### 3.4 Explainable AI (XAI) in Security

Explainability is a major focus, motivated by the black-box nature of many AI algorithms, regulatory mandates (GDPR), and the integration of human analysts in cyber defense. XAI research divides into intrinsic, post-hoc, local/global, and model-agnostic techniques. Applications stretch from network anomaly detection to healthcare, smart grids, and finance security. XAI is also scrutinized for its own susceptibility to adversarial attacks and for the tradeoff between transparency and efficacy.

### 3.5 AI for Critical Infrastructure and IoT

IoT/IIoT environments compound risk with device heterogeneity, limited resources, and exposure to physical compromise. AI-driven frameworks automate attack tree construction, enhance vulnerability assessments, and use particle filtering to accurately segment normal/anomalous behavior on ICS/IIoT testbeds. Automated methods dramatically accelerate risk analysis and make standardized taxonomies actionable in large scale settings. The integration of blockchain technology for secure data management has also shown promise in remote patient monitoring and healthcare applications [13]–[15].

### 3.6 AI-Enhanced Cybersecurity Education

Curricular innovations track AI's growing strategic role, with MOOCs and hands-on tutorials incorporating both foundational ML and practical security exercises on real-world datasets. Recent reviews expose gaps in teaching methodologies and the need for harmonized, ethical, and up-to date educational materials. The shift cipher technique for amplified data security [3] represents pedagogical approaches to teaching fundamental security concepts alongside advanced AI applications.

## 4. DISCUSSION AND CRITICAL ANALYSIS

### 4.1 Consensus in the Literature

Research consistently finds that deep learning and automated agents yield improved detection capabilities for both network and malware threats. AI planning and attack tree ontologies increase efficiency and coverage in risk modeling, especially for complex critical infrastructures. There is consensus that offensive AI is outpacing traditional defense, and defensive AI needs explainability, robustness, and human collaboration for relevance in high-stakes environments.

### 4.2 Contradictions and Debates

Several key contradictions emerge from the literature:

- 1) Explainability versus Performance: More interpretable models (XAI) often sacrifice accuracy, while high-performing deep models are opaque and difficult to scrutinize. This fundamental tension requires careful consideration in deployment contexts.
- 2) Dataset Bias: Most successes are realized in simulation; live, evolving operational networks pose significant unseen challenges. The gap between laboratory performance and real world effectiveness remains substantial.
- 3) Evaluation and Benchmarking: Lack of universally accepted metrics stymies comparative studies and slows translation to practice. Different research groups use incompatible evaluation frameworks, making cross-study comparisons difficult.

### 4.3 Limitations and Gaps

Several critical limitations persist across the literature:

- Real-time Adaptation and Adversarial Robustness: Current systems are not universally resilient against evolving attack techniques.
- Lack of Practical Standardization: Frameworks exist but are not systematically adopted across sectors.
- Data Privacy and Ethical Boundaries: AI-driven analytics must reconcile deep inspection with user rights and regulatory mandates.
- Insufficiency of Cross-Sector Partnerships: Limited collaboration between academia, industry, and government hinders comprehensive solutions.

### 5. REPRESENTATIVE NUMBERS AND TAKEAWAYS

Table I presents key metrics from primary sources reviewed in this paper.

**Table -1: KEY METRICS FROM PRIMARY SOURCES**

Source/Metric	Performance	Key Takeaway
RBF Neural Network [1]	Improved accuracy	Enhanced situation awareness
IDSIOT-SDL [2]	High detection rate	Effective IoT security
DL vs. Legacy Systems	15-30% improvement	Significant accuracy gains
Attack Tree Automation	70% time reduction	Substantial efficiency improvement
XAI Model Accuracy Trade-off	5-15% accuracy loss	Transparency vs. performance

### 6. FUTURE RESEARCH DIRECTIONS

Based on the comprehensive literature analysis, we identify five critical research directions:

#### 6.1 Human-AI Teaming

Create explainable collaborative systems where AI augments humans and is able to adaptively learn from analyst feedback, improving resilience and trust.

#### 6.2 Adversarial Robustness

Develop theoretical and practical defenses against adversarial AI, especially those targeting explainable models or leveraging GANs.

### 6.3 Dynamic, Real-World Benchmarking

Invest in realistic, continuously evolving datasets and grand challenge competitions to drive comparative progress.

### 6.4 Ethical and Privacy-Preserving AI

Pursue methods that defend privacy and ensure equitable outcomes, including federated learning and differential privacy in cyber applications.

### 6.5 Cross-Disciplinary Research and Standardization

Foster global interdisciplinary partnerships to develop and maintain actionable taxonomies, metrics, and protocols for AI in cybersecurity.

### 7. IMPLEMENTATION CONSIDERATIONS

#### 7.1 Engineering Guidelines for Deployment

Based on the literature synthesis, we recommend the following engineering practices:

- Model Selection: Prioritize interpretability for high stakes decisions; use ensemble methods to balance accuracy and robustness
- Safety Mechanisms: Multi-stage validation before deployment; human-in-the-loop for critical decisions
- Integration Strategies: Gradual deployment with A/B testing; compatibility with existing security infrastructure

#### 7.2 Best Practices from Case Studies

The Atlanta cyberattack incident [16] and subsequent analyses [17] highlight the importance of proactive vulnerability assessment, incident response automation, cross-organizational coordination, and regular security audits. The Art of Deception [18] emphasizes the continued importance of addressing human factors alongside technical solutions.

### 8. LIMITATIONS OF THIS REVIEW

This review acknowledges several limitations:

- Scope Constraints: The focus on 20+ papers provides depth but may not capture all relevant developments in this rapidly evolving field.
- Temporal Limitations: Given the fast pace of AI/ML research, findings may become outdated quickly, necessitating periodic updates.
- Publication Bias: The review primarily covers published academic work, potentially underrepresenting proprietary industrial research.

- Geographic Coverage: Most reviewed papers originate from North American and European institutions.

## 9. CONCLUSIONS

AI and ML applications are fundamentally shifting the cybersecurity paradigm. The synthesis of recent literature reveals meaningful advancements in automating and scaling both attack and defense, with deep learning, reinforcement learning, and explainable AI at the fore. Despite progress, unresolved gaps persist in adversarial resistance, transparency, real-world adaptability, and ethical governance. Systematic collaboration, benchmark-driven research, and human-centered design will be pivotal to ensuring resilient, ethical, and effective cybersecurity for an increasingly connected world. The integration of standardized frameworks like MITRE ATT&CK and CAPEC provides a foundation for future development, but significant work remains to bridge the gap between research innovations and operational deployment. The cybersecurity community must prioritize interdisciplinary collaboration, ethical considerations, and practical standardization to realize the full potential of AI and ML in defending against evolving cyber threats.

## ACKNOWLEDGEMENT

We acknowledge the foundational research that has advanced the field of AI and ML in cybersecurity. We extend gratitude to Dr. Vilas Gaikwad for guidance throughout the development of this review. We also thank the reviewers and mentors who provided valuable feedback.

## REFERENCES

- [1] Z. Chen, "Research on internet security situation awareness prediction technology based on improved rbf neural network algorithm," *Journal of Computational and Cognitive Engineering*, vol. 1, no. 3, pp. 103–108, 2022.
- [2] A. Wani, R. S, and R. Khalid, "Sdn-based intrusion detection system for iot using deep learning classifier (idsiot-sdl)," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 3, pp. 281–290, 2021.
- [3] R. Verma, A. Kumari, A. Anand, and V. Yadavalli, "Revisiting shift cipher technique for amplified data security," *Journal of Computational and Cognitive Engineering*, vol. 3, no. 1, pp. 8–14, 2024.
- [4] F. Abu-Dabaseh and E. Alshammari, "Automated penetration testing: An overview," in *The 4th International Conference on Natural Language Computing*, Copenhagen, Denmark, pp. 121–129, 2018.
- [5] S. J. Dorchuck, *Goal-Directed Systems Testing: Automated Execution of Intelligently Generated Cyber Attack Plans*. PhD thesis, Massachusetts Institute of Technology, 2021.
- [6] A. Moore, R. Ellison, and R. Linger, "Attack modeling for information security and survivability," June 2001.
- [7] K. Arulkumaran, M. Deisenroth, M. Brundage, and A. Bharath, "A brief survey of deep reinforcement learning," *IEEE Signal Processing Magazine*, vol. 34, August 2017.
- [8] H. M. Z. A. Shebli and B. D. Beheshti, "A study on penetration testing process and tools," in *2018 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, pp. 1–7, 2018.
- [9] S. S. Reza, W. Hasan, S. S. Reza, and S. Chakraborty, "A comparative overview on penetration testing," in *Proc. of The Fourth Intl. Conf. On Advances in Computing, Electronics and Electrical Technology—CEET*, pp. 25–28, 2015.
- [10] J. Zhao, W. Shang, M. Wan, and P. Zeng, "Penetration testing automation assessment method based on rule tree," in *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, pp. 1829–1833, IEEE, 2015.
- [11] D. B. Patil and M. Joshi, "A study of past, present computer virus performance of selected security tools," *Southern Economist*, December 2012.
- [12] A. Terekhov, "History of the antivirus," [Online]. Available: <https://www.hotspotshield.com/blog/history-of-the-antivirus>
- [13] M. J. Hossain Faruk, H. Shahriar, M. Valero, S. Sneha, S. Ahamed, and M. Rahman, "Towards blockchain-based secure data management for remote patient monitoring," *IEEE International Conference on Digital Health (ICDH)*, 2021.
- [14] M. J. Hossain Faruk, "Ehr data management: Hyperledger fabric-based health data storing and sharing," *The Fall 2021 Symposium of Student Scholars*, 2021.
- [15] S. Ryan, R. Mohammad A, H. F. Md Jobair, S. Hossain, and C. Alfredo, "Ride-hailing for autonomous vehicles: Hyperledger fabric-based secure and decentralize blockchain platform," *IEEE International Conference on Big Data*, 2021.
- [16] L. Kearney, "Atlanta officials reveal worsening effects of cyber attack," 2018. Accessed: Jun. 27, 2018. [Online]. Available: <https://www.reuters.com/article/us-usa-cyber-atlanta-budget/atlanta-officials-reveal-worsening-effects-of-cyber-attack-idUSKCN1J231M>
- [17] T. McGaillard, "Opinion | How local governments can prevent cyberattacks," *NYTimes*, 2018. Accessed: Jun. 27, 2018.[Online].Available: <https://www.nytimes.com/2018/03/30/opinion/local-government-cyberattack.html>
- [18] K. D. Mitnick and W. L. Simon, *The Art of Deception: Controlling the Human Element of Security*. Hoboken, NJ, USA: Wiley, 2003.
- [19] MITRE, Accessed: "ATT&CK Sep. Matrix 15, 2017. for Enterprise,"[Online].2017.Available: [https://attack.mitre.org/wiki/ATT&CK\\_Matrix](https://attack.mitre.org/wiki/ATT&CK_Matrix)
- [20] MITRE, "CAPEC Common Attack Pattern Enumeration and Classification (CAPEC)," 2017. Accessed: Sep. 14, 2017. [Online]. Available: <https://capec.mitre.org/>