International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Critical Gaps in Defending Against AI-Powered Cyber Attacks: A Comprehensive Review of Challenges, Vulnerabilities, and Future Directions

1st Shubham Ramtekkar
CSE(Cyber Security)
G H Raisoni College of Engineering&Management, Nagpur
shubham.ramtekkar.cyb@ghrietn.raisoni.net

2nd Srushti Wani
CSE(Cyber Security)
G H Raisoni College of Engineering&Management, Nagpur
srushti.wani.cyb@ghrietn.raisoni.net

3rd Harshal Khobragade
Dr.Priti Bihade(HoD)
CSE(Cyber Security)
G H Raisoni College of Engineering&Management,
harshal.khobragade.cyb@ghrietn.raisoni.net
priti.bihade@ghrua.edu.in

Abstract

Artificial Intelligence (AI) has significantly advanced the field of cybersecurity. Today, AI enables systems to automatically detect threats, predict potential attacks, and respond in real time. These capabilities have made cybersecurity systems faster and more efficient than ever before.
However, the major challenge is that the same technology used by defenders is also being used by attackers. Cybercriminals are now leveraging AI to create attacks that are more intelligent, faster, and much harder to detect.
In this review paper, various research studies have been analyzed to identify the critical gaps in defending against AI-powered cyberattacks. These gaps mainly include lack of explainability in AI systems, issues related to data quality and availability, weaknesses in current defense mechanisms, and the rapid evolution of attack techniques. The paper concludes that traditional cybersecurity approaches are no longer sufficient, and there is a strong need to develop more robust, transparent, and adaptive AI-based defense systems.

Introduction

In today's digital era, Artificial Intelligence (AI) has become a powerful tool in strengthening cybersecurity systems. Organizations are increasingly adopting AI-based technologies such as machine learning (ML) and deep learning (DL) to detect threats, monitor network activity, and respond to security incidents more efficiently. These technologies help analyze large volumes of data, identify unusual patterns, and detect potential attacks in real time. As highlighted in previous research, ML-based systems have proven to be effective in identifying both known and previously unseen cyber threats [1], [2].

However, while AI has significantly improved defensive capabilities, it has also created new opportunities for attackers. Cybercriminals are now using AI to design more advanced and intelligent attacks. For example, AI can be used to automatically generate malware, making it harder for traditional security systems to detect and block such threats. It is also being used to create highly realistic phishing emails by analyzing user behaviour and communication patterns, increasing the chances of successful attacks [3], [4].

Another growing concern is the use of deepfake technology. AI-generated fake identities, voices, and videos are being used for fraud, impersonation, and social engineering attacks. These types of threats are particularly dangerous because they are difficult to distinguish from real content, making them more convincing and harder to detect [5]. This clearly shows that cyberattacks are no longer just manual or simple—they have become automated, adaptive, and highly intelligent.

Despite the progress in AI-based cybersecurity solutions, several important challenges still exist. One major issue is that AI systems themselves can be attacked. Adversarial attacks involve manipulating input data in a way that misleads machine learning models, causing them to make incorrect decisions. Studies have shown that even highly accurate models can be fooled with carefully crafted inputs [6].

Another challenge is the dependency on high-quality data. AI models require large datasets for training, but in cybersecurity, such data is often limited, unbalanced, or sensitive. This makes it difficult to build models that can perform effectively in real-world environments. In addition, many AI systems lack transparency and are often referred to as "black boxes," meaning their decision-making process is not easily understood. This lack of explainability can reduce trust and make it harder for security professionals to take appropriate actions [7].

Furthermore, cyber threats are constantly evolving. Attack techniques change over time, and AI models trained on past data may not always perform well against new types of attacks. This issue, known as concept drift, creates a gap between the capabilities of existing defense systems and the rapidly advancing strategies used by attackers.Overall, AI plays a dual role in cybersecurity. On one hand, it strengthens defense systems and improves threat detection. On the other hand, it enables attackers to develop more sophisticated and scalable attacks. This makes AI a double-edged sword in the field of cybersecurity. Therefore, it is important to identify and understand the weaknesses in current AI-based defense systems. This review paper focuses on these critical gaps and highlights the need for more robust, adaptive, and transparent solutions to effectively defend against AI-powered cyberattacks in the future**.**

Objective

The use of Artificial Intelligence (AI) in cybersecurity has brought both advantages and serious challenges. While AI helps in detecting and preventing attacks, it has also made cyberattacks more intelligent, faster, and harder to detect. As a result, current cybersecurity systems are struggling to keep up with these advanced threats. This section explains the major gaps in existing defense mechanisms in a clear and practical way.

A. Speed Mismatch Between Attackers and Defenders

One of the most important issues is the difference in speed between attackers and defenders. AI-powered attacks can be executed in seconds or minutes because they are automated and do not require human intervention. These systems can scan networks, find vulnerabilities, and launch attacks almost instantly. On the other hand, many cybersecurity defense systems still rely on human monitoring or slower automated responses. This delay creates a serious problem, as by the time a threat is detected, the damage may have already been done. This speed gap makes it very difficult for organizations to respond effectively to modern cyberattacks.

B. Difficulty in Detecting Unknown (Zero-Day) Attacks

Most traditional cybersecurity tools are designed to detect known threats based on previous attack patterns or signatures. However, AI-powered attacks are often new and unpredictable. They can exploit unknown vulnerabilities, known as zero-day attacks, which have not been seen before. Because these attacks do not match existing patterns, many security systems fail to detect them. This allows attackers to bypass defenses easily and remain undetected for a long time, increasing the potential damage.

C. Weaknesses in AI-Based Defense Systems

Although AI is used to strengthen cybersecurity, AI systems themselves are not completely secure. Attackers can exploit weaknesses in these systems using techniques like adversarial attacks, where small changes are made to input data to confuse the AI model. Another issue is data poisoning, where attackers manipulate the training data used by AI systems. This can cause the system to learn incorrect patterns and make wrong decisions. As a result, AI-based security tools may fail to identify real threats or may even allow malicious activities to pass as normal behaviour.

D. Lack of Transparency in AI Decision-Making

Many AI systems work as "black boxes," meaning it is not clear how they arrive at a particular decision. In cybersecurity, this lack of transparency is a serious problem because security analysts need to understand why a threat was flagged or ignored. Without clear explanations, it becomes difficult to trust AI systems or take appropriate action. This also makes it harder to debug errors or improve system performance.

E. Shortage of Skilled Professionals

Another major gap is the lack of professionals who have expertise in both AI and cybersecurity. AI-based security systems are complex and require specialized knowledge to design, implement, and maintain. Many organizations do not have enough skilled personnel to manage these advanced systems. This limits their ability to fully utilize AI technologies and leaves them vulnerable to sophisticated attacks.

F. Overdependence on Traditional Security Methods

Despite the rise of AI, many organizations still rely heavily on traditional security approaches such as firewalls and rule-based systems. These methods are effective against basic threats but are not designed to handle intelligent and adaptive AI-driven attacks.AI-powered malware, for example, can change its behavior to avoid detection, making traditional defenses ineffective. This overdependence on outdated systems creates a major weakness in cybersecurity infrastructure.

G. Increasing Complexity of Digital Systems

Modern IT environments are becoming more complex, with the use of cloud computing, Internet of Things (IoT) devices, and hybrid networks. While these technologies offer many benefits, they also increase the number of entry points for attackers.AI can quickly analyze these complex systems and identify weak points that humans might miss. Securing such a large and interconnected environment is extremely challenging, and even a small vulnerability can lead to a major security breach.

H. Rise of Autonomous and Scalable Attacks

AI has made it possible to create cyberattacks that can operate independently without human control. These autonomous attacks can continuously scan systems, adapt to defenses, and launch multiple attacks at the same time.This scalability means that a single attacker can target thousands of systems simultaneously. Such attacks are not only difficult to detect but also very hard to stop once they begin.

I. Lack of AI-Specific Security Frameworks

Current cybersecurity frameworks are not fully designed to handle AI-related threats. Issues like model theft, data manipulation, and AI misuse are not properly addressed in traditional security policies. There is a need for

new security frameworks that specifically focus on protecting AI systems, including their data, algorithms, and deployment environments. Without these frameworks, organizations remain unprepared for AI-driven risks.

Related Works and Research Findings

In recent years, many researchers have studied how Artificial Intelligence (AI) is being used in cybersecurity. Their work shows that AI has improved threat detection and response, but at the same time, it has also introduced new challenges. This section explains these research findings in a simple and clear way, focusing on the critical gaps in defending against AI-powered cyberattacks.

A. Performance of AI-Based Intrusion Detection Systems

Many studies show that AI-based intrusion detection systems (IDS) are more advanced than traditional security systems. These systems can automatically detect suspicious activities by analyzing network traffic and user behaviour.

Research findings suggest that AI improves detection speed and accuracy. It can identify patterns that are difficult for humans or rule-based systems to detect. However, these systems are not perfect. They often generate false alarms or fail to detect new types of attacks that were not included in their training data.

This shows an important gap: even though AI systems are powerful, they still struggle to handle real-world situations where attacks are constantly changing.

B. Weakness of AI Systems Against Adversarial Attacks

Another important area of research focuses on how attackers can trick AI systems. Studies show that AI models can be easily manipulated using special techniques called adversarial attacks.

For example:

- Small changes in input data can confuse the AI system
- Malicious data can be added during training (data poisoning)
- Attackers can design inputs that bypass detection systems

Research experiments have shown that these techniques can significantly reduce the accuracy of AI systems. In some cases, harmful activities are incorrectly classified as safe.

This highlights a major gap: AI systems are not strong enough to defend themselves against targeted attacks.

C. Difficulty in Detecting Modern AI-Powered Attacks

Recent studies show that cyberattacks are becoming more intelligent due to the use of AI. These attacks can adapt, learn, and change their behavior to avoid detection.

For example:

- Malware can change its code to remain undetected
- Phishing attacks can be personalized using AI
- Attacks can hide within encrypted network traffic

Research findings indicate that even advanced AI-based security systems struggle to detect these types of attacks, especially in complex environments.

This reveals another critical gap: current defense systems are not fully capable of handling dynamic and evolving threats.

D. Issues with Data and Training

AI systems depend heavily on data for training. However, many research studies highlight problems related to data quality and availability.
Some common issues include:

- Datasets are outdated and do not reflect current threats
- Data is often unbalanced, leading to biased models
- Real-world attack data is difficult to access

Because of these issues, AI models may perform well in testing environments but fail in real-world situations.
This creates a significant gap: without proper data, AI systems cannot effectively detect new and unknown cyberattacks.

E. Improvements Through Deep Learning Techniques

Researchers have also explored the use of advanced deep learning models such as neural networks to improve cybersecurity systems. These models can analyze complex patterns and improve detection accuracy.
Studies show that deep learning-based systems:

- Perform better than traditional machine learning methods
- Reduce false positives
- Detect more complex attack patterns

However, even these advanced models are not completely secure. They are still vulnerable to adversarial attacks and require continuous updates.
This shows that although improvements are being made, there is still a gap in achieving fully reliable AI-based security systems.

F. Growing Concern of Adversarial Machine Learning

Adversarial machine learning has become a major topic in recent research. It focuses on how attackers can exploit AI systems and how to defend against such attacks.
Research findings suggest that:

- Attackers can generate inputs that fool multiple systems at once
- Techniques like Generative Adversarial Networks (GANs) are used to create sophisticated attacks
- Existing defense methods are not fully effective

Many researchers agree that there is no complete solution yet to protect AI systems from these threats.
This highlights a critical gap: the security of AI systems themselves is still an open problem.

G. Gap Between Research and Real-World Application

Although many advanced solutions have been proposed in research papers, their practical use in real-world systems is still limited.
Some common challenges include:

- High cost of implementation
- Complexity of AI systems
- Lack of skilled professionals
- Difficulty in integrating with existing systems

Because of these challenges, many organizations continue to rely on traditional security methods, which are not strong enough to handle AI-powered attacks. This creates a major gap between theoretical research and practical cybersecurity implementation.

Area	What Research Shows	Simple Explanation	Critical Gap
AI in Cybersecurity Defense	AI helps in detecting threats faster, responding automatically, and analyzing user behavior.	The system can quickly notice unusual actions like a login from a new location and mark it as suspicious.	These systems are not fully flexible and may fail to detect new or unknown attacks.

Area	What Research Shows	Simple Explanation	Critical Gap
AI-Powered Cyber Attacks	Attackers are using AI to create smarter malware, phishing emails, and deepfake scams.	Cyberattacks are becoming more intelligent and can change their behavior to avoid detection.	Security systems struggle to keep up with fast-changing and adaptive attacks.
Adversarial AI and Model Exploitation	AI models can be tricked using techniques like data poisoning and adversarial inputs.	Attackers can confuse the system so that it makes wrong decisions or ignores real threats.	AI systems are not strong enough to resist such targeted manipulation.
Data and Training Limitations	AI depends on data, but available datasets are often outdated or incomplete.	If the system learns from poor or old data, it cannot detect new types of attacks properly.	Lack of quality and updated data reduces the effectiveness of AI systems.
Emerging Research Trends	Researchers are working on Explainable AI, hybrid models, and better threat intelligence systems.	These ideas aim to make AI more understandable and reliable in cybersecurity.	Most of these solutions are still in development and not widely used in real systems.

Identified Gaps and Research Needs

Even though Artificial Intelligence (AI) has improved cybersecurity a lot, there are still many important gaps that make systems vulnerable to modern cyberattacks. These gaps are mainly due to technical limitations, data issues, and the fast-changing nature of cyber threats. This section explains these problems in a simple and practical way.

A. Lack of Explainability

One major issue with AI systems is that they often work like a "black box." This means we can see the output (decision), but we don't know how the system actually made that decision.

Because of this:

- Security experts may not trust the system's decisions
- It becomes difficult to fix errors or improve the system
- It is hard to justify decisions in sensitive areas

In simple words, if we don't understand how AI thinks, we cannot fully depend on it.

B. Adversarial Vulnerabilities

AI systems can be easily tricked by attackers using special techniques. Even very small changes in input data can confuse the system.

For example:

- A small change in malware can make it look safe
- Attackers can design inputs that bypass detection

This means AI systems can make wrong decisions without even realizing it.

C. Data Issues

AI systems depend completely on data for learning. But in cybersecurity, good quality data is hard to find.

Some common problems are:

- Not enough data
- Data is not balanced (more normal data than attack data)
- Data does not represent real-world situations

Because of this, AI systems may not perform well in real-world conditions.

D. Concept Drift

Cyberattacks keep changing over time. New types of attacks are created regularly, and old patterns become outdated.

This creates a problem called **concept drift**,
where:

- The AI system is trained on old data
- But new attacks behave differently

As a result, the system becomes less effective over time if it is not updated regularly.

E. Automation vs Control

AI allows systems to work automatically without human involvement. This is useful because it saves time and responds quickly.

However, too much automation can be risky:

- The system may block normal users by mistake
- It may fail to understand complex situations
- Wrong decisions can be made without human checking

So, human involvement is still very important for better control and accuracy.

F. Weak Threat Intelligence Integration

Cybersecurity systems need to stay updated with the latest threats. This is done using threat intelligence.
But many AI systems:

- Do not use real-time threat updates
- Do not learn from new attack information

Because of this, they may miss new and advanced cyber threats.

G. Lack of Standardized Frameworks

There are no proper universal rules or frameworks for securing AI-based systems.

This creates problems like:

- Different organizations use different methods
- No clear guidelines for implementation
- Difficulty in measuring security performance

Without proper standards, it becomes hard to build strong and consistent security systems.

H. Ethical and Governance Issues

AI in cybersecurity also raises ethical concerns.

Some important issues are:

- **Bias:** AI may make unfair decisions based on biased data
- **Privacy:** Large amounts of user data are collected and analyzed
- **Misuse:** AI can be used for harmful purposes like surveillance or cyberattacks

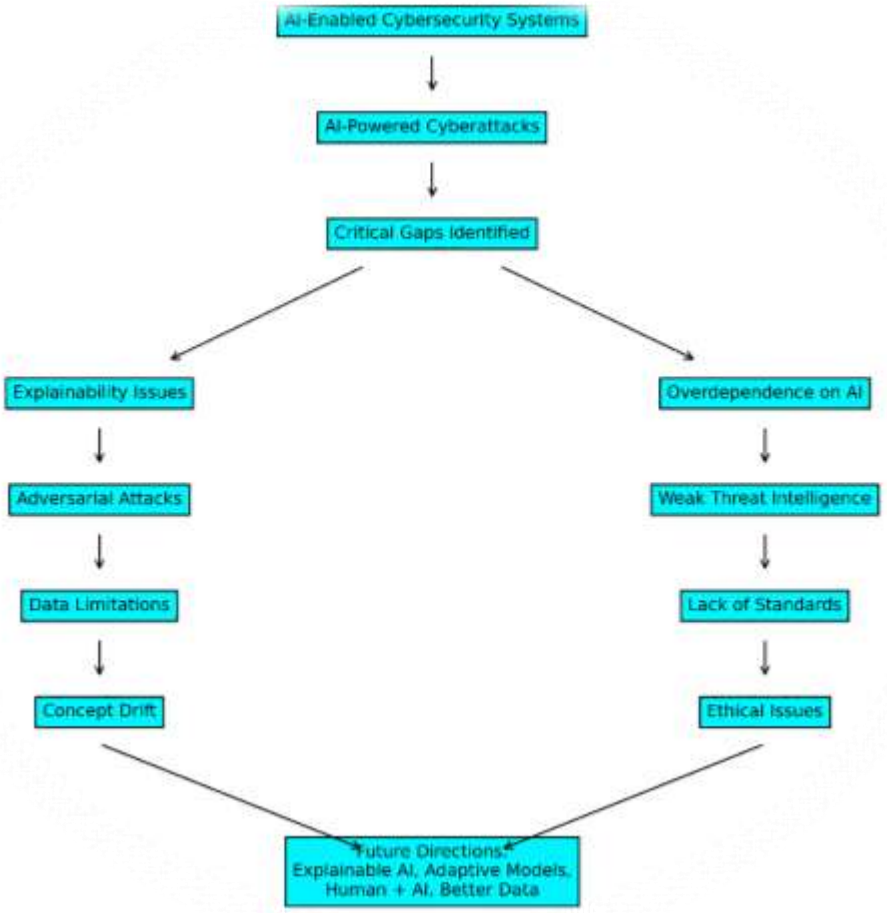
These issues reduce trust in AI systems and create challenges in their adoption.

Conclusion

Artificial Intelligence (AI) has become a very powerful tool in cybersecurity. It helps systems detect threats faster, respond in real time, and analyze user behaviour more effectively. With the help of technologies like machine learning and deep learning, security systems can now handle large amounts of data and identify complex attack patterns that were difficult to detect earlier. Because of this, AI has made modern cybersecurity systems stronger and more efficient.

However, the same technology is also being used by attackers. Cybercriminals are now using AI to create more advanced threats such as self-learning malware, deepfake scams, and highly personalized phishing attacks. These attacks are not only smarter but also adaptive, meaning they can change their behaviour over time. This makes them much harder to detect and stop using traditional security methods.

This study highlights several important gaps in defending against such AI-powered cyberattacks. One major issue is that many AI systems work like a "black box," where we cannot clearly understand how decisions are made. This reduces trust and makes it difficult for security professionals to verify or improve the system. Another



serious problem is that AI models can be easily tricked through adversarial attacks, where even small changes in input data can lead to wrong predictions.

Data is another key challenge. AI systems depend heavily on data, but in many cases, the available data is limited, outdated, or not balanced. Because of this, AI models may not perform well in real-world situations. Also, as cyberattacks keep evolving, older models become less effective over time. This problem, known as concept drift, shows the need for continuous updates and improvements.

Another concern is over-dependence on automation. While automated systems can work quickly, they may make wrong decisions in complex situations if there is no human involvement. In addition, many AI systems are not properly connected with real-time threat intelligence, which means they may fail to detect the latest threats. The lack of proper standards and frameworks for AI security also makes it difficult to build consistent and reliable systems.

There are also ethical issues to consider, such as bias in decision-making, privacy concerns, and the misuse of AI technologies.

In simple terms, traditional cybersecurity methods are no longer enough to deal with modern AI-powered attacks. There is a strong need to develop better systems that are more transparent, adaptive, and reliable. Future improvements should focus on making AI more understandable, improving its resistance to attacks, using better and updated data, and combining AI with human expertise.

References

[1 X. Huang, M. Kantarcioglu, B. Malin, and X. Jiang, "Adversarial machine learning," *ACM Computing Surveys*, vol. 54, no. 5, 2020..

[2] T. Sowmya and E. A. Mary Anita, "A comprehensive review of AI-based intrusion detection systems," *Measurement: Sensors*, 2023.

[3] M. Macas and C. Wu, "Deep learning methods for cybersecurity and intrusion detection systems," 2020.

[4] IBM Security, "Threat Intelligence Index Report," 2023

[5] M. Ring et al., "A Survey of Network-Based Intrusion Detection Data Sets," *Computers & Security*, 2019..

[6] Adversarial ML in NIDS, *Expert Systems with Applications*, 2021..

[7] A. Alotaibi and M. A. Rassam, "Adversarial Machine Learning Attacks against Intrusion Detection Systems," *Future Internet*, 2023.

[8] I. Javadov, "Performance Evaluation of AI-Driven IDS in Encrypted Networks," 2024.

[9] Adversarial Attacks on ML-based IDS, *PMC*, 2022.

[10] T. Heartfield and G. Loukas, "Social engineering attacks and human factors," *IEEE Security & Privacy*, 2022.

[11] ENISA, "Artificial Intelligence Cybersecurity Challenges," European Union Agency for Cybersecurity, 2022..

[12] S. Wang et al., "Concept Drift Detection for Streaming Data," *IEEE Transactions*, 2020..

[13] S. Furnell et al., "Alert fatigue in security systems," *Information Security Journal*, 2024.

[14] M. Brundage et al., "Malicious Use of AI," Oxford, 2018.

[15] P. Puhakainen and M. Siponen, "Role-based security training effectiveness," *MIS Quarterly*, 2022.

[16] A. Kapoor, "Adversarial Machine Learning Attacks and Defenses in AI-Driven Cybersecurity Systems," *SSRN*, 2025.

[17] M. Penmetsa *et al.*, "Adversarial Machine Learning in Cybersecurity: A Review on Defending Against AI-Driven Attacks," *SSRN*, 2025..

[18] M. Musser *et al.*, "Adversarial Machine Learning and Cybersecurity: Risks, Challenges, and Legal Implications," *arXiv*, 2023.

[19] Adversarial Attacks and Defenses in AI-Powered Cybersecurity Systems," *Journal of Computing and Information Technology*, 2021..

[20] H. Ke, J. Xu, Y. Wang, H. Chen, and Z. Shen, "Adversarial Machine Learning in Cybersecurity: Attacks and Defenses," *International Journal of Management Science Research*, 2025

[21] AI Cybersecurity Framework Study, "Artificial Intelligence Cybersecurity Dimensions and Offensive AI Threats," *Springer AI & Ethics*, 2024.

[22] "A Meta-Survey of Adversarial Attacks Against Artificial Intelligence Algorithms," *Neurocomputing*, 2025.

[23] S. T. Erukude, V. C. Marella, and S. R. Veluru, "AI-Driven Cybersecurity Threats: A Survey of Emerging Risks and Defensive Strategies," *arXiv*, 2026.

[24] E. Anthi *et al.*, "Adversarial Attacks on Machine Learning Cybersecurity Defences in Industrial Control Systems," *arXiv*, 2020.

[25] Kumari, Shalini. (2025). Blockchain for Tamper-Proof Digital Evidence Logging and Chain of Custody in Cybercrime Investigations. myresearchgo. 1. 11-20. 10.64448/myresearchgo..vol.1.issue.8.03.