Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 18s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Artificial Intelligence in Offensive Cyber Security: Emerging Techniques, Threat Landscapes, Ethical Implications and Implementation O Fast Gradient Sign

Kaniyarasu[1], Dr. P.Vivekanandan[2], S Saravanakumar[3], Dr. M.Sakthivadivel[4], S. Arunbalaji[5], K. Manikandan[6]

[1]Assistant professor (SS), Dr.Mahalingam College of Engineering and Technology, Pollachi., Kaniyarasu.k@gmail.com
[2]Professor & Head, Dr.Mahalingam College of Engineering and Technology, Pollachi. drpvivekanandan@gmail.com
[3]Assistant professor (SS), Dr.Mahalingam College of Engineering and Technology, Pollachi., Saravanacs84@gmail.com
[4]Assistant professor (SS), Dr.Mahalingam College of Engineering and Technology, Pollachi., Saravanacs84@gmail.com
[5]Part time Research Scholar, Kamban College of Arts and Science, arunbalajithebest@gmail.com
[6]Part time Research Scholar, Kamban College of Arts and Science, kmanikndan@gmail.com

Cite this paper as: Kaniyarasu, Dr.P.Vivekanandan, S Saravanakumar, Dr. M.Sakthivadivel, S.Arunbalaji, K.Manikandan, et.al (2025) Artificial Intelligence in Offensive Cyber Security: Emerging Techniques, Threat Landscapes, Ethical Implications and Implementation O Fast Gradient Sign. *Journal of Neonatal Surgery*, 14 (18s), 274-278.

ABSTRACT

Artificial Intelligence (AI) is rapidly transforming the cybersecurity landscape, empowering not only defensive systems but also enabling highly targeted and automated offensive strategies. This paper explores the evolving role of AI in offensive cybersecurity, where attackers leverage machine learning and deep learning to improve the efficiency, speed, and adaptability of cyberattacks. AI is being used to automate vulnerability discovery, craft convincing phishing messages using natural language processing, and create polymorphic malware capable of evading traditional detection systems. One of the most notable offensive techniques examined in this study is the Fast Gradient Sign Method (FGSM) an adversarial attack algorithm that subtly manipulates input data to fool AI-based detection models. FGSM exemplifies how attackers can use a model's own gradients to generate malicious inputs that remain undetected while causing misclassification, posing a serious risk to AI-driven security tools. Tools powered by deep learning, such as voice and video deepfakes, are further pushing the boundaries of social engineering attacks, making them more believable and harder to detect. These advancements present growing threats to individuals, organizations, and national infrastructure, highlighting the urgent need for ethical guidelines and global regulatory frameworks. By analyzing both real-world incidents and technical strategies like FGSM, this paper aims to shed light on the offensive potential of AI and calls for international collaboration to ensure that its use in cyberspace remains safe, transparent, and accountable.

Keywords: Artificial Intelligence, Offensive Cybersecurity, Machine Learning Attacks, AI-Powered Malware, Cyber Threat Automation

1. INTRODUCTION

In recent years, Artificial Intelligence (AI) has emerged as a transformative force in the field of cyber security. Traditionally, cyber security focused heavily on defense detecting intrusions, preventing data breaches, and mitigating threats. With the evolution of cyber threats, AI began playing a crucial role in identifying patterns, analyzing large volumes of data, and automating responses to attacks faster than any human could. While these developments have significantly improved our defensive capabilities, a more unsettling trend has also emerged AI is increasingly being weaponized for offensive purposes. The background of AI in cybersecurity was rooted in using algorithms for better threat detection and predictive analysis. Early AI applications were focused on anomaly detection, user behavior analytics, and threat intelligence. These tools helped organizations stay one step ahead of attackers by identifying suspicious activity before damage could be done. However, the same technologies that made networks more secure have also opened the door for malicious actors to use AI for launching smarter, faster, and more personalized attacks.

Offensive cybersecurity strategies powered by AI are on the rise, and their sophistication is growing rapidly. Attackers now use AI to automate phishing campaigns, develop intelligent malware, and exploit system vulnerabilities with unprecedented precision. Tools like deepfakes, natural language generation, and autonomous botnets have changed the threat landscape dramatically. These techniques not only increase the success rate of attacks but also make them harder to trace and stop. What was once considered advanced cyber warfare is now becoming accessible to even small-scale threat actors thanks to open-source AI models and tools. This study aims to explore how AI is being integrated into offensive cybersecurity strategies, what techniques are being used, and what the broader implications are for digital safety. The objective is to examine the tools and tactics driven by AI in offensive operations, evaluate their impact on cybersecurity, and shed light on the ethical, legal, and strategic challenges they pose. By understanding the offensive side of AI in cybersecurity, this research hopes to contribute toward the development of more balanced, secure, and forward-looking cyber defense frameworks.

2. AI-DRIVEN OFFENSIVE TECHNIQUES

Artificial Intelligence (AI) is no longer confined to the realm of cybersecurity defense. It is now actively shaping offensive strategies, enabling attackers to execute more precise, automated, and evasive attacks than ever before. AI's ability to analyze massive datasets, adapt to changing environments, and learn from feedback has made it a powerful weapon in the hands of malicious actors. Several AI-driven offensive techniques are emerging as significant threats, each exploiting a unique aspect of AI's capabilities.

One major application is AI in penetration testing and vulnerability scanning. Traditionally, penetration testing required skilled professionals to identify security flaws, but now AI can automate much of this process. Machine learning algorithms can scan large networks, analyze behavior patterns, and uncover hidden vulnerabilities without human intervention. This technology, when misused, allows attackers to simulate extensive reconnaissance and exploit system weaknesses at scale and speed. Another advanced offensive use is machine learning to evade Intrusion Detection Systems (IDS). Attackers train models to understand how IDS mechanisms work and craft traffic or payloads that mimic legitimate behavior. Over time, these systems learn which actions avoid detection, allowing malicious activity to slip through undetected. This cat-and-mouse game becomes even more dangerous as adversarial AI learns from the defenses it faces and adapts accordingly.

The rise of deepfake and AI-generated social engineering attacks is perhaps one of the most socially alarming developments. AI can now generate highly realistic voice and video deepfakes to impersonate trusted individuals CEOs, government officials, or family members convincing targets to reveal sensitive information or transfer funds. Combined with AI-generated text, attackers can conduct large-scale spear-phishing campaigns that are indistinguishable from legitimate communication. Finally, AI-enhanced malware, such as polymorphic and metamorphic malware, has become a significant concern. These types of malware change their code and behavior patterns with each infection, making them extremely difficult to detect using traditional signature-based systems. When paired with AI, they can dynamically adapt in real-time, intelligently responding to system defenses and maximizing damage.

3. CASE STUDIES AND REAL-WORLD INCIDENTS

As artificial intelligence continues to evolve, it has not only empowered defenders but also found its way into the arsenal of attackers. Real-world incidents have started to surface where AI-driven tools have been used to carry out cyberattacks with a level of precision, stealth, and personalization that traditional methods could not achieve. These case studies and emerging tools demonstrate how AI is no longer a futuristic threat—it is already influencing the tactics used in today's cyber landscape. One of the most discussed examples is DeepLocker, a proof-of-concept AI-powered malware developed by IBM Research. DeepLocker was designed not to cause harm, but to show what AI-enhanced malware could be capable of. Unlike traditional malware that targets systems indiscriminately, DeepLocker uses facial recognition, geolocation, and voice matching to ensure that it only activates its payload when it encounters a specific target. For example, it could sit silently on thousands of devices, completely undetectable, and only launch an attack when it recognizes a specific individual's face through a webcam. This type of precision targeting makes detection much harder and significantly increases the potential impact of an attack.

Another notable incident occurred in 2019, when cybercriminals used AI-generated voice deepfakes to impersonate a CEO's voice in a phone call, convincing a company executive to transfer $243,000 to a fraudulent account. The attackers mimicked the executive's accent, speech rhythm, and tone with startling accuracy. This marked one of the earliest publicly known cases of deepfake voice technology being used in financial fraud. In the realm of automated phishing, AI has been employed to scrape personal data from social media and generate convincing spear-phishing emails that are tailored to individual targets. These messages often contain personalized references, making them far more believable and increasing their chances of success. Unlike manual attacks, these can be generated and launched at scale.

These real-world examples show that AI is not only a tool for defense but also a potential weapon when used maliciously. As these techniques become more refined and accessible, there is a pressing need for cybersecurity systems, ethical regulations, and awareness campaigns to stay ahead of such intelligent threats.

4. ADVANTAGES FOR ATTACKERS

The integration of Artificial Intelligence into offensive cybersecurity has tipped the balance in favor of cyber attackers in several unsettling ways. What once required time, effort, and deep technical skills can now be achieved faster, more efficiently, and with greater precision. AI grants attackers significant advantages—making their operations smarter, more adaptive, and increasingly difficult to detect or counter. One of the biggest advantages AI offers is speed, scalability, and adaptability. Unlike traditional attack methods that require manual intervention, AI can rapidly scan thousands of systems for vulnerabilities in a fraction of the time. It can adapt its attack strategy on the fly, based on the behavior of the target environment. For example, AI can alter its code or delivery method if it senses it's being watched by a security system. This adaptability allows attackers to remain one step ahead of conventional defenses, adjusting tactics in real-time to avoid detection.

Another powerful advantage is target profiling. AI can scrape data from social media, websites, and publicly available records to build detailed profiles of individuals or organizations. This information is then used to craft hyper-personalized attacks—whether through phishing emails, fake phone calls, or impersonation attempts. The more convincing the attack, the higher the success rate, and AI makes that level of personalization possible at an unprecedented scale. What used to take hours of research can now be automated, processed, and acted upon within minutes.

Furthermore, AI allows for real-time decision-making in attack vectors. Using predictive analytics and reinforcement learning, attackers can deploy AI models that react dynamically to security protocols as they unfold. If a system flags suspicious activity, the AI can quickly change its behaviour such as slowing down its actions, masking its presence, or switching to a different attack route without needing human input. This real-time responsiveness makes it extremely difficult for security teams to trace or shut down an attack before damage is done. In essence, AI has given attackers the tools to act faster, smarter, and more invisibly. As these technologies become more accessible, the gap between defensive readiness and offensive innovation continues to grow highlighting the urgent need for next-generation security solutions that can match AI’s evolving capabilities.

5. CHALLENGES AND LIMITATIONS

While Artificial Intelligence offers powerful advantages in the realm of offensive cybersecurity, it is not without its challenges and limitations. The same complexity that makes AI effective also brings with it a set of obstacles that can restrict its misuse. From the quality of data it relies on, to the computational resources it demands, and the growing ability of defenders to detect AI-generated threats, the offensive use of AI is not without hurdles.One of the biggest challenges is data dependency and training bias. AI models are only as good as the data they are trained on. If the input data is incomplete, outdated, or biased, the AI’s performance can be flawed and even counterproductive. For instance, if an attacker’s machine learning model is trained on limited or noisy cybersecurity datasets, it may fail to recognize defensive mechanisms accurately, leading to unsuccessful or poorly executed attacks. Additionally, biases in data can result in predictable behavior, which can be exploited by defenders once patterns are discovered.

Another key limitation is the requirement for computational resources and access. Training and running advanced AI models require significant processing power, often supported by expensive GPUs, cloud infrastructure, and technical expertise. While large state-sponsored threat actors may have these resources, small-time hackers or amateur cybercriminals may struggle to access or afford the computing environment necessary to develop and deploy AI-driven attacks at scale. This makes the use of AI in offensive cybersecurity a high-barrier entry point for many. Moreover, as AI becomes more prevalent in attacks, defenders are also leveraging AI to detect and neutralize AI-based threats. Sophisticated defensive systems can now identify anomalies that hint at AI-generated behavior, such as unusual timing patterns, language structures, or rapidly evolving attack techniques. Cybersecurity tools enhanced with AI can learn to detect synthetic activity over time, creating an arms race where both attackers and defenders are constantly evolving. The smarter AI gets on the offense, the more advanced the defenses become.

6. ETHICAL AND LEGAL IMPLICATIONS

The integration of Artificial Intelligence into offensive cybersecurity brings with it not only technical concerns but also deep ethical and legal implications. As the line between innovation and exploitation becomes increasingly blurred, society is faced with a critical challenge: how to embrace the power of AI while preventing its misuse. This dilemma becomes especially urgent when AI tools intended for good are repurposed for harm, exposing individuals, organizations, and even nations to serious threats.At the heart of this issue lies the dual-use dilemma. Many AI technologies, such as machine learning models or automation frameworks, were initially created to solve complex problems and improve digital safety. However, these same tools can also be adapted to perform cyberattacks more efficiently. For example, a natural language processing model designed to detect phishing emails can be inverted to generate convincing phishing content. This dual-use nature of AI creates a gray area where it becomes difficult to regulate development without stifling innovation.

To address these concerns, there is a growing demand for AI governance and responsible use. Developers, researchers, and organizations must be conscious of the ethical impact of the technologies they build. This means implementing internal checks and balances, ensuring transparency in how AI is trained and deployed, and establishing ethical review processes—especially when working on tools that can be exploited for offensive purposes. Responsible AI isn't just about building smarter systems it's about making sure they align with the broader values of safety, privacy, and human rights.

Finally, there is an urgent need for robust policy and regulation frameworks at both national and international levels. As AI-driven cyber threats transcend borders, fragmented regulations leave many loopholes open for exploitation. Governments, tech leaders, and international bodies must collaborate to create standardized policies that define acceptable use, penalize abuse, and promote global cybersecurity norms. Regulatory efforts should also focus on transparency, accountability, and traceability in AI systems to ensure they are not used maliciously.

7. IMPLEMENTATION OF FAST GRADIENT SIGN METHOD

The Fast Gradient Sign Method (FGSM) is a widely recognized adversarial attack technique used in cybersecurity to expose vulnerabilities in AI-based systems. Originally developed to test the robustness of neural networks, FGSM has become a powerful offensive tool for attackers. By making small, calculated changes to input data based on the gradient of the model's loss function FGSM can trick machine learning models into making incorrect predictions. In cybersecurity, this can be used to create adversarial malware samples or manipulate network traffic patterns in ways that evade intrusion detection systems (IDS) and antivirus software. The simplicity and speed of FGSM make it particularly dangerous, as even minimal perturbations that are nearly invisible to humans can completely fool an AI model. As AI becomes increasingly embedded in threat detection systems, understanding and defending against attacks like FGSM is essential to building more secure and resilient cyber defenses.

Fast Gradient Sign Method (FGSM) – Step-by-Step Algorithm
1. Input the original data sample
Let x be the original input (e.g., image, network traffic data, etc.), and y the true label.
2. Define the loss function
Use a standard loss function J(θ, x, y), such as cross-entropy, where θ are the model parameters.
3. Compute the gradient of the loss w.r.t the input
Calculate: $\nabla_x J(\theta, x, y)$
This gives the direction in which to adjust the input to increase the loss.
4. Determine the perturbation
Apply the sign of the gradient to create a small perturbation in the input:
$\delta = \varepsilon \cdot \mathrm{sign}(\nabla_x J(\theta, x, y))$
Where ε is a small scalar value that controls the magnitude of the perturbation.
5. Generate the adversarial input
Add the perturbation to the original input to get the adversarial example:
x_adv = x + δ
6. Feed the adversarial input to the AI model
Evaluate the model with x_adv. The goal is to have the model misclassify x_adv as something other than y.
7. Check if the attack is successful
If the model prediction for x_adv ≠ y, the adversarial attack has succeeded.

The Fast Gradient Sign Method (FGSM) is a well-known technique used to trick AI models by making small, calculated changes to input data. What makes FGSM so effective—and dangerous is how simple and fast it is to use. The core idea is to slightly modify the input (like an image, text, or network data) in a way that pushes the model into making a wrong prediction, even though the changes might be too small for a human to notice.

Here’s how it works: every AI model learns by minimizing a "loss function," which tells it how wrong it is during training. FGSM takes advantage of this by calculating how to increase that loss basically showing the model what would confuse it the most. It does this by finding the direction of change that would most impact the output, using the model’s own internal calculations (called gradients). Then, it adds a very small adjustment in that direction. This adjustment is controlled by a factor called epsilon (ε), which decides how big the change should be.

In cybersecurity, FGSM can be a powerful offensive tool. For instance, attackers can slightly alter malicious code or emails so they slip past AI-powered security systems undetected. They can even use FGSM to create "adversarial malware" programs that look harmless to AI models but act maliciously in reality. Because it’s fast and doesn’t require complex setup, FGSM is often used in large-scale attacks or testing environments to show how fragile AI systems can be. Understanding FGSM helps us see how attackers can misuse AI and highlights why it’s important to build models that aren’t easily fooled. It also serves as a reminder that as smart as AI has become, it’s still vulnerable to clever manipulation—especially when used in critical areas like cybersecurity.

This chart shows how a small, intentional change to input data can fool an AI model using the Fast Gradient Sign Method (FGSM). The green dot represents a normal input that’s correctly classified. The arrow shows a small shift—called a perturbation—pushing the input toward misclassification. The orange dot is the adversarial version, which now tricks the model. This simple change highlights how attackers can outsmart AI systems with minimal effort.

8. CONCLUSION

In today’s fast-paced digital world, Artificial Intelligence has emerged as both a powerful shield and a potential threat in the field of cybersecurity. While it has significantly improved our ability to detect and respond to threats, this paper highlights how AI is also being exploited by cybercriminals to launch more intelligent, faster, and harder-to-detect attacks. From automating system scans to creating convincing deepfakes, attackers are now using AI tools that often surpass traditional security defenses. Real-life examples like AI-driven phishing campaigns and advanced tools such as IBM’s DeepLocker show that these threats are no longer just possibilities they’re already here. One particularly concerning technique is the Fast Gradient Sign Method (FGSM), which manipulates AI systems by making tiny, almost invisible changes to input data, causing those systems to misinterpret what they see. This can be especially dangerous in areas like malware detection or biometric security. These adversarial attacks expose a critical weakness in current AI models and underline the urgent need for more resilient systems. Even though offensive AI has its challenges, such as needing large datasets and high computing power, the pace of development means we can’t afford to wait. We must invest in smarter, more secure AI defenses and ensure ethical practices are built into every stage of development. Just as important is the need for international cooperation—without clear global guidelines, the misuse of AI could spread unchecked across borders. Moving forward, the cybersecurity community must work together to shape a future where AI strengthens our defenses rather than undermines them. If developed responsibly, AI has the potential to be one of our greatest assets in protecting the digital world. But if left unchecked, it could also become one of the most serious threats we’ve ever faced.

REFERENCES

1. Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access, 6, 14410–14430. https://doi.org/10.1109/ACCESS.2018.2807385
2. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023
3. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (pp. 39–57). https://doi.org/10.1109/SP.2017.49
4. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
5. Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. IEEE Transactions on Neural Networks and Learning Systems, 30(9), 2805–2824. https://doi.org/10.1109/TNNLS.2018.2886017
6. Kumar, A., & Garg, P. (2020). Role of artificial intelligence in cybersecurity: A review. Journal of Information and Optimization Sciences, 41(6), 1355–1364. https://doi.org/10.1080/02522667.2020.1801853
7. Brundage, M., Avin, S., Clark, J., Toner, H., & Eckersley, P. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. Future of Humanity Institute. https://arxiv.org/abs/1802.07228
8. Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 23(5), 828–841. https://doi.org/10.1109/TEVC.2019.2890858
9. Liu, Y., Chen, X., Liu, C., & Song, D. (2017). Delving into transferable adversarial examples and black-box attacks. arXiv preprint arXiv:1611.02770.
10. Gilmer, J., Metz, L., Faghri, F., Schoenholz, S. S., Raghu, M., Wattenberg, M., & Goodfellow, I. (2018). Adversarial spheres. arXiv preprint arXiv:1801.02774.
11. Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1310–1321). https://doi.org/10.1145/2810103.2813687
12. Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I., & Tygar, J. D. (2011). Adversarial machine learning. In Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence (pp. 43–58). https://doi.org/10.1145/2046684.2046692
13. Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533.
14. Mirsky, Y., Doitshman, T., Elovici, Y., & Shabtai, A. (2019). Kitsune: An ensemble of autoencoders for online network intrusion detection. Network and Computer Applications, 100, 146–160. https://doi.org/10.1016/j.jnca.2017.09.014
15. Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1903.12261
16. Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. (2016). Towards the science of security and privacy in machine learning. arXiv preprint arXiv:1611.03814.
17. IBM Research. (2018). DeepLocker: How AI can power a stealthy new breed of malware. IBM Security Intelligence. https://securityintelligence.com/posts/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/
18. National Institute of Standards and Technology. (2023). AI Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce. https://www.nist.gov/itl/ai-risk-management-framework
19. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206
20. Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. Science, 361(6404), 751–752. https://doi.org/10.1126/science.aat5991