A review: strategies and challenges for preventing AI cyber attacks

Musab A. M. Ali[1,†], Nessibeli Askarbekova[2,*,†], Damelya Yeskendirova[2,*,†] and Zhadyra Ongenbayeva[2,*,†]

[1] College of Computing and Information Sciences, University of Technology and Applied Sciences, Sultanate of Oman
[2] International Information Technology University, 34/1 Manas St., Almaty, 050000, Kazakhstan

Abstract

The increasing integration of artificial intelligence (AI) into critical sectors such as autonomous vehicles, healthcare, finance, and cybersecurity makes these systems prime targets for cyberattacks. AI systems, especially those based on machine learning (ML) and deep learning (DL), exhibit vulnerabilities due to their reliance on large datasets, complex algorithms, and "black-box" decision-making processes. This paper explores the weaknesses of AI systems, focusing on adversarial attacks, data poisoning, model inversion, and evasion attacks. Specific examples include how adversarial inputs can mislead self-driving cars or cause diagnostic errors in healthcare systems. Defense mechanisms like adversarial training, defensive distillation, feature squeezing, model assembling, and gradient masking are discussed as methods to protect AI systems. However, these defenses face limitations, such as high computational costs and susceptibility to advanced attacks. The study concludes by emphasizing the need for ongoing research into strengthening AI defenses, securing private data, and balancing performance with robust cybersecurity. Ensuring AI security is crucial as cyber threats continue to evolve in complexity and scope.

Keywords

artificial intelligence, machine learning, deep learning, model techniques, cyber attacks

1. Introduction

Artificial intelligence (AI) technologies are built in complex systems that play critical roles and interact with infrastructure or human lives. AI operates in a variety of sectors, like self-driving cars, healthcare, finance, and cyber security which is why using AI systems makes them a target for cyber attackers. Their advanced automation and data processing capabilities were also their weakness in some ways as these systems are targeted by crooks. In this study, the details and features of these threats are studied by example prorate malware, why such strong defending measures need to be built in AI systems, and a summary about how we can protect our AI system from attacks.

AI solutions are susceptible to threats because they depend on large datasets and complicated algorithms, thus it can be difficult for many of them that build upon machine learning (ML) and deep learning (DL) models. These concerns become more relevant when one considers the "black-box" nature of many AI models, where it is not possible to interpret decision-making processes—this complicates their security landscape even further. If not properly addressed, this lack of transparency can facilitate the attack by an adversary that clandestinely manipulates AI outputs undetected and potentially has disastrous consequences in safety-critical applications [1]. These include various successful studies showing that AI systems are all too vulnerable to cyberattacks. A common threat

DTESI 2024: 9th International Conference on Digital Technologies in Education, Science and Industry, October 16–17, 2024, Almaty, Kazakhstan
* Corresponding author.
† These authors contributed equally.
m.alrawi@iitu.edu.kz (M. Ali); n.askarbekova@iitu.edu.kz (N. Askarbekova); d.eskendirova@iitu.edu.kz (D. Yeskendirova); zh.ongenbayeva@iitu.edu.kz (Zh. Ongenbayeva)
0000-0003-0677-5993 (M. Ali); 0009-0006-6230-5063 (N. Askarbekova); 0000-0003-4270-1908 (D. Yeskendirova); 0009-0001-8497-2955 (Zh. Ongenbayeva)

model in AI is the malicious adversary who can feed a trained AI bad data to have it return a false or harmful prediction, known as an adversarial attack [2]. In autonomous driving systems, such perturbations are translated to catastrophic failures where the system misidentifies traffic signs resulting in accidents [1]. In healthcare, adversarial examples can result in diagnostic AI systems mistaking medical images and thus recommending wrong treatments [3].
One of the first steps to develop effective defense systems is to be able to understand what are the types of cyber attacks an AI system can suffer from. Adversarial attacks which include poisoning of data, inversion of models, and evasion attacks are thus possible.

1. Adversarial attacks: Adverse attack approaches creating small, often imperceptible perturbations to input data cause the AI models to create a wrong decision. An AI system for a self-driving car recognizes an image of a stop sign (right) as a picture of associated traffic lights and the same device considers a similar graphic with some added noise to correspond to another illustration presenting yield. In this attack, the training data of an AI model is manipulated so that it can learn false patterns causing biased or incorrect output from trained models running in actual deployment [4].
2. Model inversion: Model inversion attempts to recreate the input data from a model's output and as a result reveals sensitive information. Model inversion attacks constitute privacy leaks in facial recognition [5].
3. Evasion attacks: Evasion takes place when training or deploying an AI system that feeds in inputs designed to thwart detection or purposely misclassify. In cyber security-associated contexts, evasion attacks are prominent for AI models used to see malware or fraudulent events [3].

The practical defenses against AI cyber attacks:

1. Algorithms for adversarial robustness: Adversarial training improves AI accuracy with a failure aspect of production models which are trained on the misclassified usage taught them to recognize attacks and fight back. It involves high computational costs and leads to overfitting certain attacks [6]-[7].
2. Defensive distillation: A related technique called defensie distillation involves training a distilled model on the soft outputs and allows to reduction sensitivity of the final model to input perturbations. Its effectiveness may be evaded by advanced attacks [2].
3. Image feature compression: Image simply means compressing the characteristics in name, that is transforming small-size input data into a larger one to eliminate adversarial attacks. Before data is fed into the model, it filters objections to filter potential threats [8].
4. Ensemble: Ensembling different models' output to make it harder for the adversarial attacker. They might somehow be resource-intrusive and may introduce latency [4].
5. Obfuscating gradients: Obfuscated gradients hide the actual gradient from attackers, causing them to struggle more in producing adversarial examples. It is not a silver bullet, considering that the attacker would find ways to break it [9].

2. Literature review

2.1. Overview of AI vulnerabilities

As we rely on AI for critical infrastructure and day-to-day applications, the increased attack surface leaves these systems vulnerable to all kinds of cyber threats. Due to the complexity and data-dependent nature, AI model techniques, especially ML and DLs, are inherently at risk. These weaknesses are exacerbated by the “black-box” character of several AI approaches, in which decision-making models—operating on an estimated grader—are inaccessible to direct examination and the forecast or detection AI system is vulnerable [9]. Asystems are notoriously brittle against adversarial attacks where tiny, sometimes imperceptible changes to input data can cause the AI

model's output to go astray [10-15]. [2] showed that a small amount of perturbation in input data could lead to an ANN making wrong image classifications, and this had possibly catastrophic implications when considering the use of machine learning models by applications such as autonomous driving or for medical diagnosis. These adversarial attacks rely on the model's sensitivity to imperceptible input changes that have a pronounced effect on the models' outputs. Data poisoning attacks, manipulating the learning patterns through tampering with training data of an AI system. Such attacks, which distort the meaning of some AI model base by creating bias or errors during inferencing [4] are called adversarial. The damage data poisoning can cause is substantial, especially in industries that rely on AI for do-or-die decisions (e.g., finance, healthcare, and security).

2.2. Types of cyber attacks on AI systems

An overview of different types of cyber attacks that AI systems faces to safeguard them. Various common attack vectors such as adversarial attacks, model inversion, data poisoning, and evasion attacks are described in the study:

1. Adversarial attacks: These are the most commonly discussed forms of cyber threats towards AI systems. These techniques are called attacks since they introduce tiny perturbations in the input data and result in incorrect or unexpected outputs from an AI model. [17] in this brought to light the importance of adversarial examples and in doing so showed that state-of-the-art models can be tricked. This is not restricted to digital environments; these situations can occur in the real world as well. They can be effective in the physical world, where very minor perturbations of objects or images cause misclassification by AI [16].
2. Data poisoning: These attacks, which include the training phase of an AI model by injecting the malicious data into a training set, an attacker causes learning of wrong patterns in the model leading to delivery biased/incorrect answers. [3] also underscored the difficulty of making AI systems that are less capable of being detected as they develop more sophisticated, integrated systems with vital infrastructures.
3. Model inversion: It reverse engineer the model's outputs to reconstruct input data and reveal behind privacy. Such an attack can be impactful in settings where AI models are used to deal with private information (e.g., biometrics and medical data). Model inversion attacks can also expose private data [7] as well which highlights the requirement for more privacy-preserving methods to be incorporated in AI model deployment homes.
4. Evasion attacks: This type of attack targets AI systems during the inference phase, with attackers sending crafted inputs that are in turn intended to be either not detected or misclassified. Such attacks are pertinent in cybersecurity use cases, where AI models are used to identify malware or fraudulent activities [3]. [18] presented an exhaustive survey of evasion attacks and existing countermeasures arguably exemplifying the cat-and-mouse game which replicates in the AI domain anytime between offense driven by attack generation strategy refinement that gets countered directly or else through side-channel detection response improve on part of the defender.

2.3. Defensive strategies against AI cyber attacks

Adversarial training: Adversarial training uses adversarial examples during the training process to strengthen AI models and make them more robust against manipulations. This incurs deeper computational costs and may cause overfitting to certain adversarial strategies [7]:

1. Defensive distillation: it decreases the influence of attacks by building a distilled model that is trained on predictions rather than hard labels from a pre-trained and full-precision version. It can be bypassed by more advanced attacks that need investigation [2], [9].
2. Feature squeezing: Feature squeezing protects the model from adversarial examples by reducing input data complexity. Abstract of an activated neuron in a feature map made

through a convolutional layer is available are hard to succeed filtering data before it reaches the Classification Model creates intentional but non-functional approximations of features or samples reduces entropy filters out possible adversaries [16]. The use of multiple models together.

3. Model ensemble techniques: This does help to make the system more secure — it makes it harder for adversarial examples to deceive all three systems. This approach can be very slow and cause delays [4].
4. Gradient masking: Gradient masking masks the gradients used by attackers to generate adversarial examples, increasing attack difficulty albeit not rendering attacks impossible and hence raising a call for research on this subject [9].

Summary and Research Gaps shows in Table 1.

Table 1
Frequency of Special Characters

Title/Topic	Summary	Research Gaps
Adversarial Training	Adversarial training involves training AI models on adversarial examples to improve robustness, but it can lead to increased computational costs and potential overfitting. [5], [19]	Balancing security and performance remains a challenge, with potential issues of overfitting
Defensive Distillation	Defensive distillation reduces a model's sensitivity to small input perturbations by training a distilled model on soft outputs, though it can be bypassed by advanced attacks. [1], [9]	Further research is needed to counteract sophisticated adversarial techniques that bypass distillation
Feature Squeezing	Feature squeezing simplifies input data, reducing the effectiveness of adversarial attacks by filtering out malicious alterations before they reach the model [9]	Exploring the trade-offs between accuracy and robustness, and the impact on real-world applications
Model Ensemble Techniques	Model ensemble techniques involve using multiple models and aggregating their outputs to enhance robustness, though this approach can be resource-intensive [4]	Resource-intensiveness and latency issues in model ensemble techniques need to be addressed
Gradient Masking	Gradient masking obscures the gradients used by attackers to generate adversarial examples, making it harder to craft successful attacks, though it is not foolproof [9]	Research is required to develop more reliable methods that cannot be easily circumvented by attackers

3. Research methodology

The research method using the systematic literature review (SLR) explains existing studies related to AI security that prevent cyber attacks on AI systems. The SLR methodology was selected due to its systematic method for identifying, reviewing, and combining existing research studies so that provide a thorough understanding of the currently known state of knowledge as well as areas requiring further investigation. We analyzed the selected literature review studies — which ensures that key themes are identified, and categorized based on AI security. The study describes attacks and classifies cyber threats against AI systems as adversarial attacks, data poisoning, model inversion,

and evasion. Research gaps to highlight current research gaps, e.g., what is known and not well understood about such attacks (hence requiring more investigation), what are the challenges in deploying existing countermeasures, or when it comes to developing new paradigms.

4. Results

We were interested in looking at some of the main studies on AI security as either affecting vulnerabilities or defensive strategies against cyber attacks.

1. Adversarial Attacks (The Ultimate Adversary of AI systems): A tiny change in input data can parse the wrong output. These attacks work well in both digital and physical settings, showing the importance of strong defensive capabilities [1].
2. Data poisoning attacks: Data poisoning exploits the behavior of AI models when they are constructed to take blocks of additional information which pollutes and corrupts that data structure at training, leading directly or indirectly to biased responses from them. These types of attacks are very dangerous for critical applications including medical and finance [4].
3. Model Inversion: This kind of attack allows the reconstruction or restoration of input data based on the output information available from a model. It creates privacy issues, especially when this type can be carried out in an environment that is very sensitive to various applications Some stronger privacy-preserving techniques may be necessary to address these possible attacks [5].
4. Adversarial training: While adversarial training is beneficial in strengthening models, it comes with its own set of problems like higher computation and risk of overfitting to a particular kind of attack [19].
5. Defensive distillation: Using the defensive distillation method makes the model less sensitive to input perturbations bypassed by advanced attacks, but it leaves room for improvement and research [2].

5. Discussion

A literature review of existing works provides important insights and challenges towards AI-based cybersecurity:

- Adversarial attack: Adversarial training, has a lot of potential destructiveness but is associated with cost and high risk for overfitting (hence it may not be the panacea itself). Consistent with this principle, a resilient network can be used in hybrid approaches of machine learning defense where it combines feature squeezing and model ensembles to collectively improve system-wide robustness [8], [19].
- Data poisoning and model inversion: These attacks demonstrate the necessity of improved defenses, countermeasures, and privacy-preserving mechanisms to mitigate fundamental vulnerabilities in AI systems [4], [5].
- Defensive techniques: Although defensive distillation and gradient masking are effective countermeasures, their susceptibilities to stronger attacks indicate that continuous research is necessary to harden both methods. The practical utility of techniques model ensembles [2] and homomorphic encryption comes with resource requirements as well as performance trade-offs [9].
- Secure deployment and monitoring: Practices of continuous monitoring, and secure deployments are very important for being agile because cyber threats do not stand. These approaches worked well for long-term security, but they are based on static guarantees that will fail against new types of attacks unless the applied techniques get updated too [2], [4].

- Research gaps: Some of the major areas for research include enhancing protection from advanced adversarial attacks, improving homomorphic encryption towards real-time usage, and striking a balance between security, performance, and computational costs.

6. Conclusion

Looking beyond targeted adversarial data perturbations, the above studies highlight secure deployment as a critical layer of defense and monitoring for adversaries. This is not enough, so an innovation model needs to be. For example, future research needs to address vulnerable areas such as the need for better defenses against more sophisticated adversarial attacks and new privacy-preserving learning techniques that maintain security guarantees while supporting applications capable of homomorphic encryption in real-time.

In addition, the need to apply classic cyber hygiene with AI-specific security controls and an evolving threat landscape is what helps in keeping it secure, reliable as well as effective. With cyber threats being at their peak of evolution in a digital world, developing tougher and more scalable solutions seems to be the path to future AI security.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

[2] Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. IEEE Symposium on Security and Privacy (SP), 582-597.

[3] Liu, Q., Chen, P. Y., Liu, S., et al. (2019). A survey on security threats and defensive techniques in machine learning. arXiv preprint arXiv:1810.04065.

[4] Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317-331.

[5] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. IEEE Symposium on Security and Privacy (SP), 39-57.

[6] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

[7] Carlini, N., Katz, G., Barrett, C., & Dill, D. L. (2017). Provably minimally-distorted adversarial examples. arXiv preprint arXiv:1709.10207

[8] Xu, W., Evans, D., & Qi, Y. (2017). Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155.

[9] Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. Proceedings of the 35th International Conference on Machine Learning (ICML), 274-283.

[10] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4700-4708.

[11] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

[12] Chen, J., Su, H., Zhang, Z., & Chen, C. (2019). Robustness verification of tree-based models. Advances in Neural Information Processing Systems (NeurIPS), 12317-12328.

[13] Zhou SZhu TYe DYu XZhou W(2024)Boosting Model Inversion Attacks With Adversarial ExamplesIEEE Transactions on Dependable and Secure Computing10.1109/TDSC.2023.328501521:31451-1468Online publication date: May-2024
[14] Z. Ayan, B. Alimzhan, M. Olga, Z. Timur, and Z.Toktalyk, Quality of service management in telecommunication network using machine learning technique, Indonesian J. of Electr. Eng. and Comput. Sci., vol. 32, no. 2, pp. 1022–1030, 2023. doi: 10.11591/ijeecs.v32.i2.pp1022-1030.
[15] O.A. Manankova, M.Z. Yakubova, M.A. Rakhmatullaev, and A.S.Baikenov, "Simulation of the Rainbow Attack on the SHA-256 Hash function," J. of Theoret. and Appl. Inf. Tech., vol. 101, no. 4, pp. 1594–1603, 2023.
[16] Kurakin, A., Goodfellow, I. J., & Bengio, S. (2017). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.
[17] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
[18] Filip Radlinski, Krisztian Balog, Fernando Diaz, Lucas Dixon, and Ben Wedin. 2022. On Natural Language User Profiles for Transparent and Scrutable Recommendation. In Proceedings of the 45th Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3477495.3531873
[19] Aleksander Madry and Aleksandar Makelov and Ludwig Schmidt and Dimitris Tsipras and Adrian Vladu "Towards Deep Learning Models Resistant to Adversarial Attacks,| International Conference on Learning Representations,2018.