Column: Energy and Security

Cybersecurity Issues in AI

Deepak Puthal
Newcastle University

Saraju P. Mohanty
University of North Texas

Artificial Intelligence (AI) is omnipresent to make current edge technology smarter and automatic without much human intervention. The machine learning (ML) including deep learning (DL) frameworks employ "learn" to power AI system by learning from the data [1][2]. The process of learning and utilizing them for future decisions leads to novel attack patterns in AI [3]. Attacks in the AI are a purposive manipulation of processes based on the underlying AI system's weaknesses that lead to malfunction at the end goal. Nowadays, AI system is running with mission-critical applications such as autonomous driving and smart grids; thus attacks in the AI system may lead to the consequences of loss of life.

Many solutions exist for designing AI algorithms for mitigating the cyber security issues. However, identifying security issues and challenges in AI systems and designing solutions to mitigate them is still an open problem [3][4]. If we look into the working model of an AI system closely, it is nothing but a machine that took input and process to give output, as shown in Figure 1(a). The well-known cyber attacks in an AI system such as "Input Attacks" and "Poisoning Attacks" have been illustrated in Figure 1.

Input Attacks: Attackers alter the input data provided to an AI system to manipulate the system's output desired by the attackers. Several contributions exist in this problem for neural networks. The abstract model is shown in Figure 1(b).

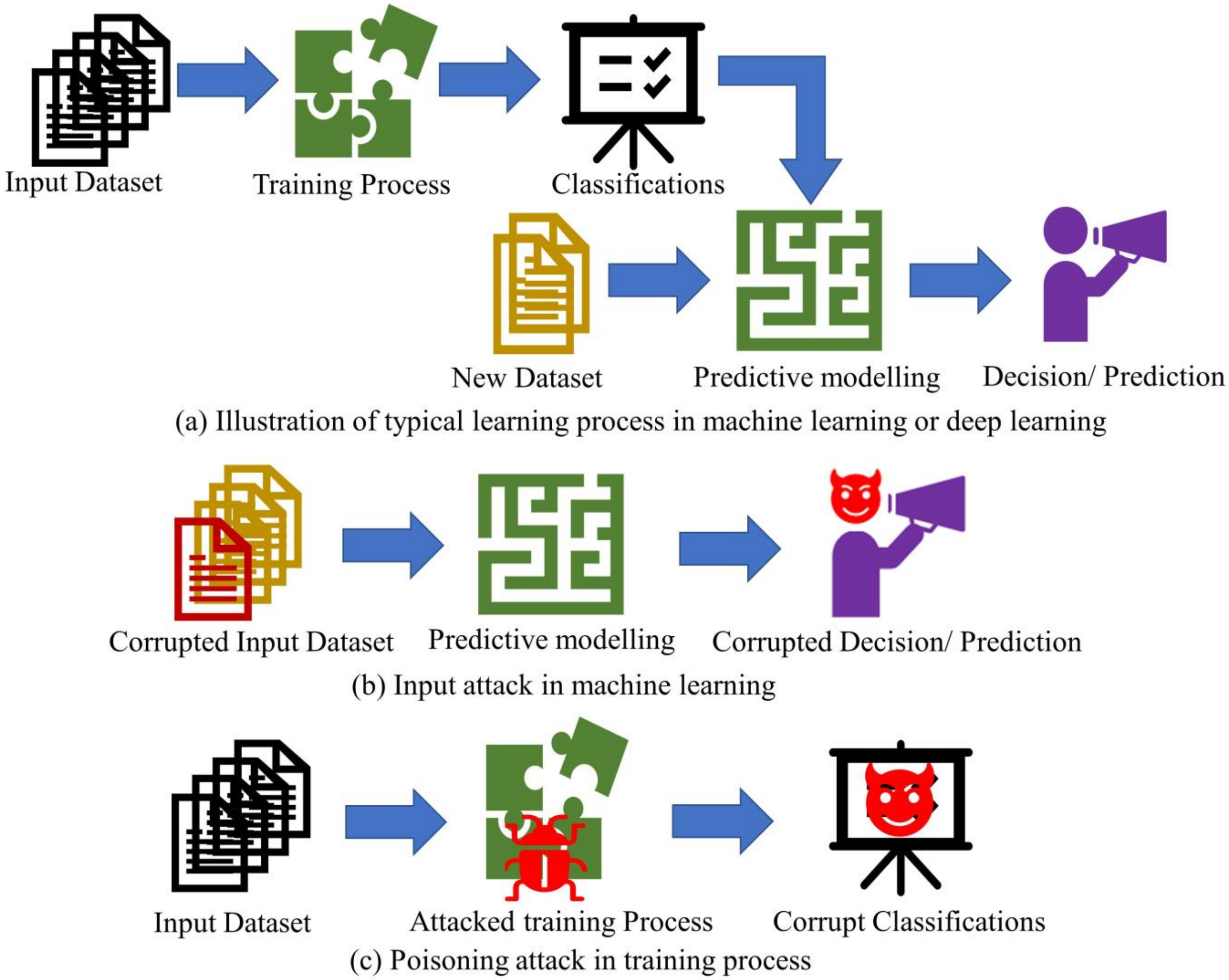
Poisoning Attacks: Attackers either alter the data used for system training or tamper with the training process. These types of attacks majorly appear during the system development and training, i.e., the AI system development's initial process. The abstract model is shown in Figure 1(c).

Along with the above two common attacks, there is another possible type of attack (e.g. Evasion Attacks), where attackers learn offline to discover information for future attacks [4].

USE CASE: SELF-DRIVING CAR

Let us consider the use case of the self-driving car recognizing stop signs to describe the above-specified attacks [5]. Let's assume the attacker little modify or partially cover the stop sign on the road to fool the AI system [5][6]. This type of attack is known as the "Input Attack" in the AI system. It could be mitigated by combining the input data from multiple sources and applying a data aggregation algorithm [6].

In order to detect a stop sign, an AI system learn from the thousands of dataset containing the stop sign. Attackers can spoof the learning process by inputting some fake data sets of stop signs by giving wrong instructions, such as taking the right turn. Finally, even a clear stop sign on the road could not the detected by the AI algorithm. This type of attack is known as the “Poisoning Attacks” in the AI system, where an attacker finds a backdoor to exploit the system. The type of attacks can be mitigated using the data provenance techniques [6].



(a) Illustration of typical learning process in machine learning or deep learning

(b) Input attack in machine learning

(c) Poisoning attack in training process

Figure 1: A selected attack patterns in AI.

FUTURE OF AI SECURITY

Due to the demand, current use, and future aspects of AI, securing the AI system cannot be unbeaten. Unlike software system, we should give best to secure the AI [6]. One of the primary and effective will be the AI assurance processes by extending the software assurance processes. As discussed above, the basics model of AI technique, assurance should be at the training and testing phase while including the AI process software [6]. Privacy and trust are still an open problem in AI, which needs to get attention in the future for secure system building.

The classification of the cyber attacks in the AI system is shown in Figure 2 [7][8]. It listed the attacker's capabilities and goals during the AI system's end-to-end process, such as training, modeling, classifications, and decision or prediction.

CONCLUSION

AI security is vital; however, it should be designed carefully by considering the suitable assurance processes based on system requirements. Data should be secure throughout the process of an AI system.

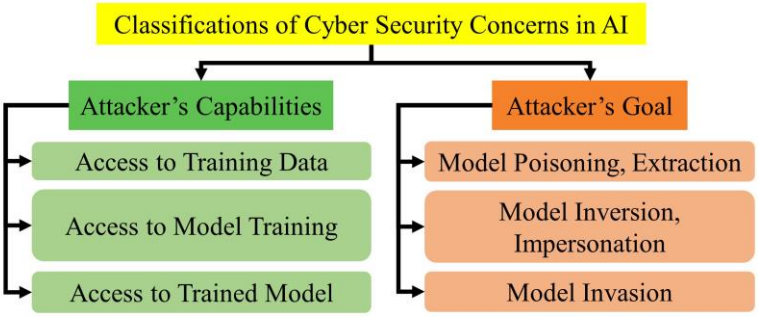


Figure 2: security classifications for AI

REFERENCES

[1] J. Lemley and P. Corcoran, "Deep Learning for Consumer Devices and Services 4 - A Review of Learnable Data Augmentation Strategies for Improved Training of Deep Neural Networks," *IEEE Consumer Electronics Magazine*, Vol. 9, No. 3, pp. 55-63, May 2020.

[2] Z. Li, V. Sharma, and S. P. Mohanty, "Preserving Data Privacy via Federated Learning: Challenges and Solutions", *IEEE Consumer Electronics Magazine*, Vol. 9, No. 3, May 2020, pp. 8--16.

[3] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning", *Machine Learning*, Vol. 81, No. 2, pp. 121-148, 2010.

[4] N. Pitropakis, E. Panaousis, T. Giannetsos, E. Anastasiadis, and G. Loukas, "A taxonomy and survey of attacks against machine learning", *Computer Science Review*, Vol. 34, pp. 100199, 2019.

[5] M. Comiter, "Attacking Artificial Intelligence", Belfer Center Paper, 2019. https://www.belfercenter.org/sites/default/files/2019-08/AttackingAI/AttackingAI.pdf. Accessed: March 12, 2021.

[6] E. Bertino, "Attacks on artificial intelligence", *IEEE Security & Privacy*, Vol. 19, No. 1, pp. 103-104, 2021.

[7] S. Kundu, "Can you trust your machine learning system?", Keynote *IEEE Computer Society Annual Symposium on VLSI*, 2019. http://www.eng.ucy.ac.cy/theocharides/isvlsi19/keynotes_files/Sandip-Kundu_ISVLSI-2019_Keynote.pdf. Accessed: March 12, 2021.

[8] M. Jere, T. Farnan and F. Koushanfar, "A Taxonomy of Attacks on Federated Learning", *IEEE Security & Privacy*, vol. , No. 01, pp. 0-0, 5555.

ABOUT THE AUTHORS

Deepak Puthal is an Assistant Professor in the School of Computing at Newcastle University, Newcastle upon Tyne, UK. Contact him at: deepak.puthal@newcastle.ac.uk.

Saraju P. Mohanty is a Professor in the Department of Computer Science and Engineering, University of North Texas, Denton, USA. Contact him at: smohanty@ieee.org.