CASS

CENTRE for AEROSPACE & SECURITY STUDIES

Adversarial Attacks on Machine Learning – An Appraisal

Shaza Arif

Researcher, National Security

Working Paper

President
AIR MARSHAL FARHAT HUSSAIN KHAN (RETD)

Edited by:
SARAH SIDDIQ ANEEL

Layout
HIRA MUMTAZ

All correspondence pertaining to this publication should be addressed to CASS, through post or email at the following address:

Centre for Aerospace & Security Studies

cass.editor@gmail.com/
cass.thinkers@gmail.com

Centre for Aerospace & Security Studies

+92 051 5405011

cassthinkers

cass.thinkers

@CassThinkers

Old Airport Road, Islamabad, Pakistan
www.casstt.com

CENTRE for AEROSPACE & SECURITY STUDIES

Adversarial Attacks on Machine Learning – An Appraisal

Working Paper

Shaza Arif

Researcher, National Security

TABLE OF CONTENTS

ABSTRACT

Technological evolution continues to define the contours of modern-day society. Amongst the proliferating list of enablers, Machine Learning (ML) has emerged as a driving force for technological advancement at a fast pace. Its efficiency and ability to process data, learn patterns and assess underlying relationships in a short time has accelerated the growth of ML across diverse enterprises. However, as its applications grow, so have the efforts to counter it. Adversarial attacks have emerged as a potent threat to ML that can lead to unforeseen consequences. These attacks can be executed in two ways - by tampering with the training data or the model itself. Adversarial attacks undermine ML's efficiency and potentially threaten societies increasingly dependent on it. The Working Paper explores several types of adversarial attacks to highlight ML's vulnerability. Taking cues from various experiments conducted and conferences convened, the paper discusses implication scenarios of adversarial attacks on the civil and military sectors.

Keywords: Technology, Machine Learning, Adversarial Attacks, Risks, ML Scenarios.

INTRODUCTION

Advances in emerging technologies have accelerated their integration into peoples' daily lives. Increasing access to information is opening new opportunities and benefits. Amongst other technologies spearheading this race, Machine Learning (ML) stands out as an important enabler, unleashing new potential in different fields with each passing day. The positive results manifested by ML have led to its growth and stimulated more in-depth research in the field.

While it is true that new technologies offer novel capabilities and incentives, they also come with new vulnerabilities. Despite the accuracy and complexity, ML systems are attack-prone and can be misled into error. This vulnerability can be exploited by attackers who can disrupt these systems for various objectives. Hence, in a parallel and more nefarious direction, efforts are being made to develop programmes that can deceive and manipulate ML by forcing it to make wrong decisions.

The *Working Paper* briefly discusses ML and its types followed by deliberations on adversarial attacks that threaten ML such as data poisoning and attacking the model programme. Furthermore, it analyses the implications of this phenomenon in the civil and military domains through scenarios using primary and secondary data. The paper ends with recommendations on how to mitigate threats from adversarial attacks.

MACHINE LEARNING (ML)

Before discussing adversarial attacks, it is necessary to discuss the concept of Machine Learning (ML) itself. ML is a subset of Artificial Intelligence (AI). Over the years, AI has made significant breakthroughs in several fields and is tagged as an important asset in the Fourth Industrial Revolution.[1] AI refers to 'computing technologies that exhibit what humans consider intelligent behavior.'[2] It is a simulation in which machines are so highly programmed that their cognitive capability 'equate' or 'simulate' human intelligence. Hence, ML or its employment can be understood simply as the 'brain' of AI which uses data to learn from a given phenomenon automatically using neural networks.

Neural networks comprise a series of algorithms (a computational process used for solving or performing computations[3]) to process underlying relationships and recognise patterns in a dataset by using a procedure that mimics the human brain.[4] It consists of three node layers:

1. An input layer which receives the initial data;
2. Multiple hidden layers for the computation process; and,
3. An output layer which indicates the final result.[5]

ML is divided into four broad categories (Table 1). All the ML categories use the available data and train themselves over time to predict outcomes for situations they have not been trained for. For every training round, the ML system is judged or evaluated on the basis of a data subset in accordance with some measure of the system's quality. The principle is that the availability of more data will lead to the development of a robust model and more accuracy of a given application.

---

1 David Mhlanga, "Artificial Intelligence in the Industry 4.0, and its Impact on Poverty, Innovation, Infrastructure Development, and the Sustainable Development Goals: Lessons from Emerging Economies?" *Sustainability* 13, no. 11 (2021): 5788-5804.

2 Anne Johnson and Emily Grumbling, *Implications of Artificial Intelligence for Cybersecurity: Proceedings of a Workshop* (Washington, D.C.: The National Academies Press), 1.

3 Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein, *Introduction to Algorithms* (Massachusetts: MIT Press, 2009), 5.

4 Martin C Nwadiugwu, "Neural Networks, Artificial Intelligence And The Computational Brain," (MSc diss. University of Ilorin Nigeria, 2020).

5 Gavril Ognjanovsk, "Everything you Need to know about Neural Networks and Backpropagation-Machine Learning Easy and Fun," *Towards Data Science*, January 14, 2019, https://towardsdatascience.com/everything-you-need-to-know-about-neural-networks-and-backpropagation-machine-learning-made-easy-e5285bc2be3a.

Table 1: Machine Learning Categories

Supervised Learning	A human operator collects data, feeds it into the system, and labels it until the system starts to learn the process of labeling by itself. Labelled training data and existing training examples predict a function.
Unsupervised Learning	Unsupervised learning refers to the process in which models use unlabelled datasets for the learning process. Data is entered into the system but labelling and classification occur autonomously by dividing available data into groups and identifying patterns according to given characteristics.
Semi-Supervised Models	This involves models which learn from labelled as well as unlabelled data.
Unsupervised Learning	This is an environment-driven approach where models are trained using rewards or penalty.

Source: Iqbal H. Sarker, "Machine Learning: Algorithms, Real World Applications and Research Directions," SN Computer Science 2:160, (2021), https://link.springer.com/content/pdf/10.1007/s42979-021-00592-x.pdf.

ADVERSARIAL ATTACKS

Despite the super-intelligence ascribed to Machine Learning, it is vulnerable to manipulation and deception. It can be impaired using adversarial attacks, which come in many forms. However, the fundamental objective of such attacks is to cause the ML model to contradict the developer's intent. The episodes or attacks are conducted using adversarial examples. These include inputs with added perturbation that are difficult to recognise by the human eye but stimulate the algorithms to yield wrong decisions. The adversarial attacker attempts to feed malicious input to ML models to have an erroneous output while the working may appear unmodified.[6]

A. Adversarial Attacks & their Execution: Important Factors

To better understand adversarial attacks, it is important to discuss certain related factors as discussed in the subsequent sections.

1. Adversary's Goal

The attacker's main objective is to fool the ML model with misleading information, malfunctioning the algorithm, and subverting the original predictions. This goal may be achieved through a targeted or untargeted attack. The former disrupts the output as per the attacker's exact desire, i.e. predicts a specific class,[7] whereas the latter forces the model to produce anything but the right answer. For example, in a targeted attack, a perturbation may be added in a tree's picture to make it appear human. On the contrary, in an untargeted attack, the perturbation is intended to classify the image as anything but the image of the tree. Targeted attacks are usually more complex vis-à-vis time and resources, while untargeted attacks are more commonly used.

2. Information Level of the Attacker

The goal of an adversarial attack is usually achieved based on the amount of information an attacker has. The level of his/her access is a critical factor in adversarial attacks. It may either be a black box or a white box attack.[8] In a black box setting, the

---

6 Petri Vähäkainu, Marrti Lehto and Annti Kariluoto, "Adversarial Attack's Impact on Machine Learning Model in Cyber-Physical Systems," *Journal of Information Warfare* 19, no. 4 (2020): 57-69, https://www.jstor.org/stable/27033645#metadata_info_tab_contents.

7 Pradeep Rathore, Arghya Basak, Sri Harsha Nistala and Venkataramana Runkana, "Untargeted, Targeted and Universal Adversarial Attacks and Defenses on Time Series" (paper presented at 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, July 19-24, 2020).

8 Yinghua Zhang, Yangqiu Song, Jian Liang, Kun Bai et al., "Two Sides of the Same Coin: White-Box and Black-Box Attacks for Transfer Learning," (paper presented at the 26th ACM SIGKDD

ML attacker does not possess complete access or control regarding the model's working.[9] On the contrary, in a white box setting, one has complete access to the internal working, i.e., information regarding classifiers, algorithms, training parameters, and training data.[10] In a black box attack, large quantities of inputs may be fed into the model to receive outputs, allowing the attacker to study the model and develop a similar model to fool the original one. Given that the attacker has ample knowledge, the manipulation can be done during the data collection phase, training, or classification process in a white box attack. The black box scenarios are more challenging for the attacker due to access limitations.

3. Attacker's Capability

Capability of the attacker refers to his/her potential and abilities to control features or labels of the training data, and injection or modification of the present data. It also includes overseeing which part of the data is being influenced by the attack. Generally, the attacker with more understanding of the training data is considered a strong attacker, given that s/he can devise more informed attacks against the system.

4. Attacker's Strategy

An attacker's strategy refers to techniques employed to alter the training data to maximise the effectiveness of the attack and optimal outputs. It aims to seek ways to manipulate the features of the training data, modify the labels and identify vulnerable points on datasets. Hence, the nature of the adversarial attack and the strategy used to execute the attack remains dependent on the information available regarding the model and the level of ingress to the model. Greater accessibility to the information may ensure better decision-making regarding the execution of the adversarial attack.

B. Effectiveness of Attacks

As earlier mentioned, ML's effectiveness depends on two factors:

1. Kind/ attributes of the data being used, and,
2. Working/performance/efficiency of the concerned algorithms.

International Conference on Knowledge Discovery & Data Mining, Virtual Event California, August 23-27, 2020).

[9] Kshitiz Aryal, Maanak Gupta, and Mahmoud Abdelsalam, "A Survey on Adversarial Attacks for Malware Analysis," (paper, arXiv, 2021), https://arxiv.org/abs/2111.08223.

[10] Aryal, Gupta, and Abdelsalam, "A Survey on Adversarial Attacks for Malware Analysis."

Hence, there are two primary ways to attack ML which are as follows:

1. Attack the data.
2. Attack the ML model.

1. Attack the Data

Data can be impaired using poisoning attacks. Poisoning attacks are conducted on data fed in the ML model or the labels used by models to classify the data. The goal is to render an ML model ineffective to produce the required outcome by interfering with its learning process.

Data Poisoning

Training its systems with abundant data is imperative for ML.[11] Data is collected in vast quantities from different sources such as social media, cellphones, internet, geo-location tools, payments and business archives. Since ML models require abundant data to verify its authenticity before it is used to train the system and is challenging. Such scenarios lead to a situation where the ML developer can encounter lack of front-end control, providing opportunities to potential attackers to deceive and manipulate the system by employing carefully crafted samples in the training sets. 'Data poisoning', hence, refers to attacks in which the attacker impairs the training data used in the ML model.[12] These attacks are becoming a major threat to ML systems/databases[13] since they infiltrate and insert incorrect/misleading information.[14] The errors are interpreted as patterns by the algorithms.

If an ML model learns from the corrupted/ tampered data, it will draw wrong results no matter how advanced the system is. Given that algorithms will use this corrupted data to learn and train itself, it will affect the entire process of analysis which could lead to

---

[11] Laura Verdea, Fiammetta Marullia and Stefano Marronehttps, "Exploring the Impact of Data Poisoning Attacks on Machine Learning Model Reliability," *Procedia Computer Science* 192 (2021): 2624-2633, https://www.sciencedirect.com/science/article/pii/S1877050921017695.

[12] Miguel A. Ramirez, Song-Kyoo Kim, Hussam Al Hamadi, Ernesto Damiani et al., "Poisoning Attacks and Defenses on Artificial Intelligence: A Survey," (paper, arXiv, 2022), https://arxiv.org/pdf/2202.10276.pdf.

[13] Emad Alsuwat, Hatim Alsuwat, Marco Valtorta and Csilla Farkas, "Adversarial Data Poisoning Attacks against the PC Learning Algorithm," *International Journal of General Systems* 49, no.1, (2020) 3-31, https://www.tandfonline.com/doi/abs/10.1080/03081079.2019.1630401.

[14] James Thorpe, "What is Data Poisoning and Why should we be Concerned?" *International Security Journal,* September 13, 2021, https://internationalsecurityjournal.com/what-is-data-poisoning/#:~:text=Data%20poisoning%20involves%20tampering%20with,unintended%20and%20even%20harmful%20conclusions.

unintended consequences. This phenomenon downgrades the concerned model's performance and manipulates how the model classifies available data/information. A simple example of data poisoning attacks dates back to 2016 when Microsoft launched a bot named 'Tay' to interact with other users using ML.[15] In less than 16 hours, Microsoft had to shut off the system because it started using inflammatory language. The reason for Tay's rude behaviour was the deliberate attempt of some Twitter users who fed offensive tweets in the bot's algorithm, impacting its response.[16]

Data poisoning can be done through the following ways (Table 2):

Table 2: Data Poisoning Methods

Data Modification	The attacker can change / add or eliminate the attributes from the training data. However, attacker does not possess the access to the algorithm.
Label Modification	In this kind of attack, the attacker is able to control labels assigned to a certain training points.[17]
Data Manipulation	This refers to addition of an invisible watermark which enables backdoor channels in the model.
Data Injection	In these attacks, the invader may insert poisoned / harmful features (with a label) that may contaminate the data sets. However, the attacker can add only to the pre-existing sets and cannot alter the training data or the ML algorithm.

Source: Charles Hu, Yen-Hung (Frank) Hu, "Data Poisoning on Deep Learning Models," (paper presented at 2020 International Conference on Computational Science and Computational Intelligence (CSCI), New York, December 16, 2020).

[15] Elle Hunt, "Tay, Microsoft's AI Chabot, gets a Crash Course in Racism from Twitter," *Guardian,* March 24, 2016, https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter.

[16] Naveen Joshi, "Countering the Underrated Threat of Data Poisoning Facing Your Organization," *Forbes*, March 17,2022, https://www.forbes.com/sites/naveenjoshi/2022/03/17/countering-the-underrated-threat-of-data-poisoning-facing-your-organization/?sh=52d215abb5d8.

[17] Rahim Tehari, Raza Javidan, Mohammad Shojafar, Zahra Pooranian et al., "On Defending Label Flipping Attacks on Malware Detection Systems," *Neural Computing & Applications* 32, no. 18 (2020), https://link.springer.com/article/10.1007/s00521-020-04831-9.

2. Attack the ML Model

Similar to how data can be corrupted by the attacker for incorrect predictions, it is also possible to achieve the same outcome via attacking the ML model. Given below are some of the ways models can be corrupted:

- Poisoning Attacks on Models

As explained earlier, poisoning attacks are usually conducted on data during the training phase. However, it is also possible to attack respective models after they have been developed. Attackers can disrupt the model by deliberately crafting evaluation-time input for which the model produces incorrect output. These alterations may be so minuscule that they may go unnoticed by a human operator but prove disastrous in terms of the output.

One form of poisoning attacks is 'logic corruption' in which the attacker directly jeopardises the learning process of the algorithm, forcing it to make incorrect judgements. In logic corruption, the poisoning attack is executed on the structure of the model.[18] For example, an attacker may disrupt the classifiers of a particular model[19] forcing it to make wrong judgements. 'Boiling frog attacks' are a prominent example of poisoning attacks. The attacker steadily weakens the model by injecting small amounts of poisoned data in each round of the training cycle of the model. In such a model, the repeated attacks leave a minimal impact with each re-training cycle and go undetected. However, over time, the incremental impact of such attacks can be quite significant.[20] Both supervised and unsupervised ML is vulnerable to such kinds of attacks.

- Evasion Attacks

These are the most prevalent attacks against a ML model. While poisoning attacks occur during the training phase, the evasion attacks are executed during the testing

---

[18] Chen Wang, Jian Chen, Yang Yanga, Xiaoqiang Ma et al., "Poisoning Attacks and Countermeasures in Intelligent Networks: Status quo and Prospects," *Digital Communications and Networks* 8, no.2 (2022): 225-234, https://www.sciencedirect.com/science/article/pii/S235286482100050X#bib32.

[19] Ibid.

[20] Eric Chan Tin, Victor Heorhiadi, Nicholas Hopper and Yongdae Kim, "The Frog-Boiling Attack: Limitations of Secure Network Coordinate Systems," *ACM Transactions on Information and System Security (TISSEC)* 14, no. 3 (2011): 1-23, https://dl.acm.org/doi/abs/10.1145/2043621.2043627?casa_token=mJPvpgJyQf8AAAAA:tkAB8GC0Ah8g5oXb6bM4DEpwT0hYw_Rd_fDPQFc8XAEWwih0flrmn3MQ2on2l0wVz_sVBy9fX736.

phase. These attacks include modifying the testing data in the deployment phase to avert detection by a classifier.[21] The model is fed with an adversarial example which appears as untampered data to the human eye but is misclassified by the model and ultimately shows the wrong results. Misclassification occurs due to decision boundaries. These attacks may be targeted or untargeted. By slightly distorting an image in the testing phase, the ML models can misinterpret it with high confidence. For instance, a slight noise, which may not be detected by a human being can lead to classification of the image of a 'panda' as a 'gibbon' with 99.3% confidence.[22]

- Model Stealing

'Model stealing' refers to reconstructing the original pre-built model or extracting information from the original pre-trained model. The attacker deliberately attacks the adversarial algorithms to fool the ML models into adding some data into the training data for the wrong output. Although, with new advancements, the safeguards against such systems are becoming more complex, it remains vulnerable.

ML models comprise large amounts of data which is used to make predictions. The attack can be aimed for information leakage. In such an attack, the attacker has some level of ingress to the information the developer would not desire to be released. Some of these models are likely to be 'leakier' than others and may include ample information that may be easily accessible.[23] This is particularly concerning for algorithms that tend to retain information. For example, the ability to recover data from a model can reveal critical details meant to be private. Furthermore, high-quality labelled data sets such as ImageNet are expensive.[24] Similarly, sophisticated models such as BERT are challenging to train and optimise.[25] Consequently, developers often use public datasets and prefer transfer learning from existing models. The pertained models pose a security risk given that there is more probability of data being hacked. They also have the potential to be used as backdoors.[26] The attacker can force the system to

---

[21] Ziyi Bao and Luis Muñoz-González, "Mitigating Evasion Attacks against Machine Learning Systems through Dimensionality Reduction and Denoising," (diss., Imperial College London, 2018).

[22] Karen Hao, "How we Might Protect Ourselves from Malicious AI?" *MIT Technology Review*, May 19, 2019, https://www.technologyreview.com/2019/05/19/135299/how-we-might-protect-ourselves-from-malicious-ai/.

[23] David Trott, "Deceiving Machines: Sabotaging Machine Learning," *Chance* 33, no. 2, (2020): 20-24, https://www.tandfonline.com/doi/abs/10.1080/09332480.2020.1754067.

[24] Ibid.

[25] Ibid.

[26] Ibid.

categorise particular inputs as faulty without affecting the model's performance on the normal dataset. The attacks may alternate the performance of such systems, impacting the overall performance and effectiveness of a particular ML system. For example, in a research paper titled 'Pre-trained Models: Past, Present and Future,' the authors observe that pre-trained models such as BERT offer diverse opportunities for ML and advocated for their employment on a larger scale to maximize the benefits.[27] However, after several experimentations, the authors were convinced that there were plenty of vulnerabilities that could be exploited by attackers with regards to the reliability of the data and overconfidence of the ML system. Hence, increased reliance can lead to more incidents of model stealing, paving way for unwanted events.

[27] Xu Han, Zhengyan Zhanga, Ning Dinga, Yuxian Gu et al., "Pre-trained Models: Past, Present and Future, *Science Direct* 2 (2021): 225-250.

IMPLICATIONS OF ADVERSARIAL ATTACKS ON ML SYSTEMS

The growing applications of ML are followed by a corresponding increase in the risk that comes from adversarial attacks. The vulnerability of ML systems can have grave consequences for the civilian and military sectors. Given the growing integration of ML systems with existing technologies, coupled with the potential of adversarial attacks to disrupt them, the consequences can put civilians and soldiers at risk**.**

A. Impacts on the Civilian Sector

ML is extending its reach in diverse civilian sectors. Plenty of examples can be discussed in the civilian sector that may be impacted by adversarial attacks, be it security, health, or law. Banks are using ML for security of the infrastructure as well as networks via fraud pattern detection and risk assessment.[28] The use of ML has considerably raised the performance in diagnostic image analysis and is predicted to have a transformative impact on the health sector.[29] In addition, ML has also been proven as an effective tool in data curation for the health sector.[30] ML is also being used extensively for surveillance on citizens.[31] Given below are some of the scenarios that highlight the implications of adversarial attacks in this area:

Scenario I

The use of facial recognition systems is becoming more prevalent. From mobile unlocks to security systems in banks, their use is growing swiftly. These systems are vulnerable to poisoning attacks and can be exploited for notorious activities.[32] The attacks can range from a trivial disturbance in an image to sophisticated attacks where an ML model's parameters are modified.[33] Such circumstances could lead to circumvention of identification or reaching a false match. The results of several experiments have revealed that even alterations to an eyeglass frame could fool

---

[28] Praveen Kumar Donepudi, "Machine Learning and Artificial Intelligence in Banking," *Engineering International* 5, no. 2 (2017): 83-86, https://www.abc.us.org/ojs/index.php/ei/article/view/490/973.

[29] Trishan Panch, Peter Szolovits and Rifat Atun, "Artificial Intelligence, Machine Learning and Health Systems," *JGlob* 8, no 2, (2018): 020303-020311, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6199467/#R9.

[30] Panch, Szolovits and Atun, "Artificial Intelligence, Machine Learning and Health Systems."

[31] Steven Feldstein, *The Global Expansion of AI Surveillance* (Washington, D.C.: Carnegie Endowment for International Peace, 2019),11.

[32] Ying Xu, Kiran Raja, Raghavendra Ramachandra and Christoph Busch, "Adversarial Attacks on Face Recognition Systems," in *Handbook of Digital Face Manipulation and Detection Advances in Computer Vision and Pattern Recognition,* eds. Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez and Christoph Busch (Cham: Springer, 2022),139.

[33] Ibid.

advanced and sophisticated facial recognition systems.[34] Such attacks can potentially put security systems at greater risk.

Scenario II

Self-driving cars stand out as one of the most discussed example in this regard. Self-driving vehicles will become more common in the future.[35] While it may be normal to assume that such cars will comprise efficient ML systems, experiments reveal that they remain vulnerable to adversarial attacks. Self-driving autonomous vehicles could be targeted using physical adversarial attacks which need not even be overly complex. For example, such cars use road signs for guidance. Experiments confirm that by tampering with road signs, images can be misclassified by the vehicle, while they may appear normal to a human.[36] Guidance systems can be manipulated by simple tricks such as covering the road signs with stickers or paint to create an adversarial stop / speed limit sign that the vehicle would misinterpret leading to accidents.

Scenario III

Adversarial attacks also pose challenges for the health sector.[37] Adversarial techniques can be used to manipulate data to fool medical/healthcare systems. In a series of experiments, it has been revealed that adversarial inputs can interfere with ML systems to misclassify medical images. The perturbations were tiny, yet they were able to mislead the model by classifying malignant tumors as benign and vice versa.[38] Hence, such attacks may pose major challenges in disease detection such as cancer, whose timely detection, if attacked, can have severe consequences.

---

[34] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer and Michael K. Reiter, "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition" (paper presented at the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, October 24-28, 2016).

[35] Jenny Cusack, "How Driverless Cars will Change our World," *BBC,* November 30, 2021, https://www.bbc.com/future/article/20211126-how-driverless-cars-will-change-our-world.

[36] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, et al., "Robust Physical-World Attacks on Deep Learning Visual Classification," (paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Utah, June 18-23, 2018).

[37] Hokuto Hirano, Akinori Minagi and Kazuhiro Takemoto, "Universal Adversarial Attacks on Deep Neural Networks for Medical Image Classification," *BMC Med Imaging* 21, no.9 (2021), https://doi.org/10.1186/s12880-020-00530-y.

[38] Samuel G. Finalayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam, and Isaac S. Kohane, "Adversarial Attacks on Medical Machine Learning," *Science 363*, no. 6433 (2019): 1287-1289.

B. Impacts on the Military Sector

Adversarial attacks on ML systems can bring more ambiguity to already complex, evolving and complicated warfare. They have the potential to impact the battlefield both in conventional as well as hybrid platforms.[39]

Deception is an inherent element of warfare. While the target of deception operations are human enemy combatants, the expanding scope of ML is opening new opportunities to target machines that are becoming more involved in the battlefield.

In conventional warfighting, systems with ML models are more appealing because they are more advanced and mitigate the danger of human casualties. Over time, more autonomy is being given to weapons such as drones.[40] More dependence on ML can lead to more vulnerability to adversarial attacks. To be fair, military systems are heavily secure with various safeguards. However, if attackers can break into an integrated ML system by any means given the pace of advancements of adversarial attacks, it can turn a winning battle into a losing one in the blink of an eye.

Secondly, surging availability of relevant data is strengthening military capabilities worldwide. AI's efficiency is pushing militaries to gather data using this tool,[41] but sifting through this data is equally important. The amount of AI-generated Intelligence, Surveillance and Reconnaissance (ISR) data is likely to increase.[42] Today, various ground, air, and space-based capabilities are used to collect data for situational awareness. While the vision of highly automated data gathering and processing, which may shrink decision-making cycles, looks very appealing, the same can become a grave concern if exploited by potential attackers. It offers security risks through the possibility of data manipulation. If subjected to data poisoning, the outcome can be disastrous. An attacker can hack into an ML model directly to distort situational awareness, which can undermine stability.

---

[39] Christopher Ratto, Michael Pekala, Neil Fendley, Nathan Drenkow et al., "Adversarial Machine Learning and the Future Hybrid Battlespace," (paper presented at IST-190 Research Symposium (RSY) on Artificial Intelligence, Machine Learning and Big Data for Hybrid Military Operations (AI4HMO), Science & Technology Organization, October 5-6, 2021).

[40] John R. Hoehn, Kelley M. Sayler and Michael E. DeVine, *Unmanned Aircraft Systems: Roles, Missions, and Future Concepts*, report (Washington, D.C.: Congressional Research Service, 2022), https://s3.documentcloud.org/documents/22089315/unmanned-aircraft-systems-roles-missions-and-future-concepts-july-18-2022.pdf.

[41] Zachary Davis, "Artificial Intelligence on the Battlefield," *Prism* 8, no. 2 (2019): 114-131.

[42] Davis, "Artificial Intelligence on the Battlefield."

Major powers are integrating ML into their militaries at an expedited pace. The US has heavily invested in AI and sees autonomy as a central feature of future warfare.[43] It sees algorithmic warfare as a defining element of the future and is keen to integrate ML in different domains, including command and control, combat, surveillance, planning, logistics and counter-terrorism.[44] China is also pushing for ML-based capabilities in the military and is looking towards more autonomy in its land, air and sea-based platforms.[45] It has made considerable progress vis-à-vis Unmanned Aerial Vehicles (UAVs), ground-based robotics and is experimenting with autonomous vessels[46]. It is also keen to integrate ML in its military logistics and is also working on refining manned-unmanned teaming using ML. Russia is venturing into integrating ML in its defence forces with a particular focus on information security, cyber warfare, unmanned systems and electronic warfare capabilities.[47]

Scenario I

In a conflict between state X and state Y, numerous autonomous drones might be employed to collect important information regarding each various facilities. State X may deliberately execute a targeted attack on state Y's drone to misclassify the information given to the adversary. For instance, it may deceive the ISR systems into labelling state Y's troops as state X's. The manipulated information can have serious consequences if used for decision-making in such situations. The applications of AI support systems where critical decisions are made in a short time based on quick analysis of enormous data. The increasing reliance on ML in the military sector[48] can

---

[43] David Vergun, "Artificial Intelligence, Autonomy Will Play Crucial Role in Warfare, General Says," *Department of Defence*, February 8, 2022, https://www.defense.gov/News/News-Stories/Article/Article/2928194/artificial-intelligence-autonomy-will-play-crucial-role-in-warfare-general-says/.

[44] Forrest E. Morgan, Benjamin Boudreaux, Andrew J. Lohn, and Mark Ashby et al., *Military Applications of Artificial Intelligence,* report (Santa Monica: RAND Corporation), 2020, https://www.rand.org/pubs/research_reports/RR3139-1.html.

[45] Elsa B. Kania, *AI Weapons in China's Military Innovation*, report (Washington, D.C.: Brookings, 2020), https://www.brookings.edu/research/ai-weapons-in-chinas-military-innovation/.

[46] Ibid.

[47] Samuel Bendett, Mathieu Boulègue, Richard Connolly, Margarita Konaev et al., "Advanced Military Technology in Russia Capabilities and Implications," (paper, Chatham House, London, 2021), https://www.chathamhouse.org/sites/default/files/2021-09/2021-09-23-advanced-military-technology-in-russia-bendett-et-al.pdf.

[48] José Javier Galán, Ramón Alberto Carrasco, and Antonio LaTorre, "Military Applications of Machine Learning: A Bibliometric Perspective," *Mathematics* 10, no. 9 (2022): 1397-1424, https://www.mdpi.com/2227-7390/10/9.

lead to major accidents and repercussions. Similarly, in the hybrid domain, adversarial attacks constitute a major threat.

Scenario II

Cyber warfare remains a pressing challenge in the age of information warfare. The ML-based Intrusion Detection Systems (IDS) are tasked to monitor networks and identify any suspicious activity. However, they remain vulnerable to adversarial attacks. State X could modify malicious data to bypass IDS of state Y, undermining the cyber capabilities of the latter.[49] Such circumstances would lead to increased frequency and lethality of cyber-attacks leading to information breach and may even disrupt state infrastructure.

Scenario III

These attacks can be used in social media campaigns to shape perceptions. Social media, due to its increased accessibility, has become a strong platform for propaganda and misinformation. Adversarial attacks can push down social media campaigns. State X may execute targeted attacks on social media against state Y by suppressing a particular class of content. In a war time situation, state X may aim to take down social media posts which show the sufferings caused by its actions which could draw international support away from it. Hence, international perceptions can be altered by targeted and non-targeted attacks.

The difficulty and challenges associated with threats from adversarial attacks reside in the notion that it involves protection not only against existing threats but also against novel and unknown threats. To fully secure an ML model, the defending side has to ensure that it is reactive to the observed attacks but also be able to avoid attack vectors which are novel and unidentified as of yet. Furthermore, given the ambiguous nature of AI, it would also be difficult to predict how ML would respond to adversarial attacks, creating more instability. The future warfighting tactics would largely be focused on countering ML models. Adversarial attacks are particularly of great concern to countries that are lagging in this respective field. Adversarial attacks can disrupt state infrastructure and make it more prone to attacks from adversaries who are more advanced in technology.

---

[49] Eirini Anthi, Lowri Williams, Matilda Rhode, and Pete Burnap et al., “Adversarial Attacks on Machine Learning Cybersecurity Defences in Industrial Control Systems,” *Journal of Information Security and Applications* 58, (2021), https://www.sciencedirect.com/science/article/pii/S2214212620308607.

WAY FORWARD

As mentioned in the above scenarios, the pervasive nature of ML and its growing applications in different areas makes it more vulnerable to attacks. This spike indicates that adversarial attacks need to be catered for to avert intended manipulation. Protection of ML systems is akin to an arms race with the attackers attempting to break into the system and the defenders aiming to protect it. Given the lethal nature of adversarial attacks, network protection against internal and external threats is imperative.

In 2019, a report was published by Sandia that detailed testing the safety and security of ML systems by attacking them.[50] Several neural networks were trained with a high level of accuracy. This was followed by data poisoning attacks using perturbed inputs. Software techniques such as 'threat modelling' were also applied to ML systems to examine and address future attacks. Following the experiment, accuracy degradation of the model and the defensive capabilities, experts revealed that they could not find an effective model to deal with the wide array of attacks. Hence, there does not exist a single method to deal with all kinds of adversarial attacks, and therefore, it is necessary to deal with this phenomenon on case-to-case basis.

The findings presented in this *Working Paper* suggest that integrity of data and robustness of the model are pivotal to mitigate the challenges posed by adversarial attacks. Hence, in order to cater for the threat emanating out of the adversarial attacks, the integrity of data should be preserved, and the model should be powerful enough to ward off potential attacks.

A. Data Integrity

Factors such as data sparcity, unique data packing schemes and unstructured nature of the data makes adversarial attacks very potent.[51] Integrity of data needs to be ensured. This is particularly important when deploying at high-risk places, such as in military applications. Similarly, validation of data used in training needs to be particularly prioritised in every domain. There should be less reliance on third-parties

---

50 Austin Short, Trevor La Pay and Apurva Gandhi, *Defending against Adversarial Examples,* report (New Mexico: Sandia National Lab, 2019), https://www.osti.gov/servlets/purl/1569514.

51 Muhammad Usama, Junaid Qadir and Ala Al-Fuqaha, "Adversarial Attacks on Cognitive Self-Organizing Networks: The Challenge and The Way Forward," (paper presented at 2018 IEEE 43rd Conference on Local Computer Networks Workshops (LCN Workshops), Chicago, October 1-4, 2018).

libraries/datasets. If for any reason they are used, the vulnerabilities should be looked into.[52]

B. Data Security

There are a number of measures which can be adopted to secure data and make it relatively less prone to adversarial attacks. These techniques include data compression, randomisation, denoising/feature squeezing, gradient regularizations, safety nets, convolution filter statistics, perturbation rectifying network and gradient adversarial network-based approaches.[53] Each approach may have its own pros and cons or possibly something to trade off. However, these approaches need to be explored in depth by AI developers.

C. Strengthening the ML Model

As already mentioned in the paper, apart from securing the data, it is also equally important to make the ML models more robust against possible attacks. Hence, ML projects should ensure extensive data training. It is imperative that the quality of datasets is not compromised. In fact, one of the most frequent and effective means to defend against adversarial attacks is adversarial training.[54] This helps in improving the model's performance and also assists in mitigating the challenges of poisoning attacks and evasion attacks.[55] This approach involves retraining the model by generating numerous adversarial examples. The models learn using the adversarial examples and can react aptly when encountered with corrupted input during an attack.[56] Although, it is recommended that ML should be avoided in contested space, but if it is needed to deployed, then it should benefit from adversarial training. However, the scope here may limited, given that the model may not be equipped to deal with unknown threats. Nonetheless, it can still enhance internal defence against such

---

[52] Short, La Pay and Gandhi, *Defending against Adversarial Examples*.

[53] Elizabeth Nathania Witanto, Yustus Eko Oktian and Sang-Gon Lee, "Toward Data Integrity Architecture for Cloud-Based AI Systems," *Symmetry* 14, no.2 (2022): 273-314, https://www.mdpi.com/2073-8994/14/2/273/htm.

[54] Kui Ren, Tianhang Zheng and Zhan Qin Xu Liu, "Adversarial Attacks and Defenses in Deep Learning," *Science Direct* 6, no.3 (2020): 346-360, https://www.sciencedirect.com/science/article/pii/S209580991930503X.

[55] Ren, Zheng and Liu, "Adversarial Attacks and Defenses in Deep Learning."

[56] Onur Savas, Lei Ding, Teresa Papaleo and Ian McCulloh, "Adversarial Attacks and Countermeasures against ML Models in Army Multi-Domain Operations," (paper presented at SPIE Conference on Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, Washington, D.C., May 19, 2020).

attacks. Other techniques which can be used to strengthen the model include gradient masking, input modification and null-class approach.[57]

The advancement and refinement of these approaches can play an effective role and civil and military organisations should regularly test their models and identify existing vulnerabilities. In the military domain, keeping humans in / on the loop is an effective way to reduce the probability or at least be able to discern when an adversarial attack has occurred, guiding the system/model towards appropriate behaviour.[58] Human oversight can circumvent dangerous situations using sound judgement.

D. Collaboration between Public and Private Sector

It has also become essential that due attention be invested in the private sector, which holds significant importance vis-à-vis ML and adversarial attacks. Academia needs to be synchronised with the private sector and the government to be able to meet future challenges of such emerging disruptive technologies. Academia in particular, requires financial support from the government, whereas the government needs the support of industry and academia to implement is policies. Greater synchronisation, that encourages more experimentation and development of more robust models by AI/ML researchers, needs to be supported. Conferences/ workshops also need to be convened on the concerned subject to promote awareness and trigger the youth to actively contribute and learn about this domain.

[57] Hao hui, "Adversarial Attacks in Machine Learning and How to Defend Against Them," *Towards Data Science,* December 19, 2019, https://towardsdatascience.com/adversarial-attacks-in-machine-learning-and-how-to-defend-against-them-a2beed95f49c.

[58] Marianne Bellotti, "Helping Humans and Computers Fight Together: Military Lessons from Civilian AI," *War on the Rocks*, March 15, 2021, https://warontherocks.com/2021/03/helping-humans-and-computers-fight-together-military-lessons-from-civilian-ai/.

CONCLUSION

Technological advancement is here to stay. With each passing day, it is becoming an integral part of human lives. While the promises of ML offer diverse opportunities to make advancements in various sectors, it is also vulnerable. With the emergence of adversarial attacks, associated risks have compounded significantly. The impact of adversarial examples indicate that modern algorithms can behave in unintended ways even under subtle perturbations. The more data is available, the more innovative ways can be employed to attack and increase corresponding risks. These risks come in different forms and have the potential to disrupt human life at micro and macro levels. They can threaten civilian as well as the military sector in unexpectable ways. Keeping in view the various scenarios discussed in this paper, it is necessary to develop counter mechanisms to protect integrated ML systems.

ABOUT THE AUTHOR

Shaza Arif *is a Researcher at the Centre for Aerospace & Security Studies (CASS), Islamabad, Pakistan. Her areas of interest are Space, Artificial Intelligence and Strategy. She writes opinion articles on issues related to politics, modern warfare, and strategy. She has studied Defence and Diplomatic Studies from the Fatima Jinnah Women University, Rawalpindi, Pakistan.*

ABOUT CASS

The Centre for Aerospace & Security Studies (CASS), Islamabad, was established in 2018 to engage with policymakers and inform the public on issues related to aerospace and security from an independent, non-partisan and future-centric analytical lens. The Centre produces information through evidence-based research to exert national, regional and global impact on issues of airpower, defence and security.

VISION

To serve as a thought leader in the aerospace and security domains globally, providing thinkers and policymakers with independent, comprehensive and multifaceted insight on aerospace and security issues.

MISSION

To provide independent insight and analysis on aerospace and international security issues, of both an immediate and long-term concern; and to inform the discourse of policymakers, academics, and practitioners through a diverse range of detailed research outputs disseminated through both direct and indirect engagement on a regular basis.

PROGRAMMES

Foreign Policy

National Security

Emerging Technologies

Aviation Industry & Technology Studies

Economic Affairs & National Development

Warfare & Aerospace

Strategic Defence, Security & Policy

Peace & Conflict Studies

CENTRE FOR AEROSPACE & SECURITY STUDIES, ISLAMABAD
Independence. Analytical Rigour. Foresight

Old Airport Road, Islamabad, Pakistan
+92 051 5405011
www.casstt.com
cass.editor@gmail.com/ cass.thinkers@gmail.com

Centre for Aerospace & Security Studies
cassthinkers
@CassThinkers
cass.thinkers