Machine Learning Forensics: Unveiling Vulnerabilities and Defense Strategies

Gabriel Abílio Barbosa Ferreira[ac], Gabriel Monteiro Ferracioli [bc]

[a] Department of Computer Science (DCC), Institute of Exact Sciences (ICEx), Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil, gabilio@ufmg.br.

[b] Instituto de Ciências Matemáticas e de Computação(ICMC), University of São Paulo(USP), São Carlos, Brazil, ferracioligabriel@usp.br.

[c] These authors contributed equally to this work.

Abstract. Machine Learning (ML), a subset of Artificial Intelligence (AI), is one of the main drivers of current technology, having a great capacity for innovation, in addition to being very flexible, being applied to a wide variety of topics. This great power associated with modern ML systems means that great expectations are associated with it, generating a distance from the problems related to such technology, such as its vulnerability, present in all stages of development. This vulnerability can generate major threats for products in which AI is used, which can escalate into gigantic problems, both economic and social, showing the importance of using defense methods against such attacks on Machine Learning models. Therefore, this study carries out a review of the literature on this subject, emphasizing the main attacks related to contemporary ML systems, in addition to the most relevant defenses against such attacks. The problems generated by this weakness are also discussed, showing the need for this issue to be further debated and studied, in order to mitigate such debility.

Keywords. Machine Learning, Artificial Intelligence, vulnerabilities, Cybersecurity, forensics

1. Introduction

Computer forensics, akin to its real-world counterpart, encompasses a set of techniques utilised for gathering information related to the identification of cybercrimes. The same concept of evidence collection can be applied, with specialists in this field focusing on threats within computational environments. Cybersecurity teams play a vital role in identifying patterns of vulnerabilities and recurrent invasions, aiming to document various techniques employed by criminals to compromise digital systems.

Computer Science (CS) is a vast field comprising numerous sub-disciplines, with ML standing out prominently in contemporary discussions. With applications surrounding mobile apps and websites, ML began gaining attention outside academic circles over the past decade [1]. As ML and AI have become more prevalent in public discourse, they have also attracted increased attention from malicious actors. Many critical services, such as those offered by banks and hospitals, integrate AI into various aspects of their systems, including face recognition and patient report generation. However, this widespread adoption also presents potential risks, as criminals targeting AI can gain access to personal data and compromise crucial models.

The growing reliance on AI technologies across sectors underscores the importance of robust cybersecurity measures. As AI and ML continue to permeate various aspects of daily life, safeguarding these systems against malicious actors becomes paramount. Furthermore, the intersection of cybersecurity and machine learning highlights the need for specialised expertise in forensic analysis tailored to address the unique challenges posed by AI-driven threats.

2. Methodology

The evolution of AI and ML from their nascent stages in the mid-20th century [2] to their contemporary resurgence forms the backdrop of

this investigation. Works published in the past decades have reignited expectations surrounding ML technologies, prompting a renewed interest in exploring their potential applications.

To provide readers with a foundational understanding of ML systems, an overview of the typical pipeline involved in pattern extraction from data is presented. Figure 1 illustrates a generalized representation of this pipeline. The initial step entails data collection tailored to address specific demands, a critical phase in model training given the often proprietary nature of datasets. Subsequently, the collected data, whether in the form of tables or images, undergoes standardization according to predefined criteria aimed at optimizing the learning process. Effective preprocessing techniques can significantly enhance model performance. Following this, the model undergoes training using a designated strategy, and upon achieving consistent performance, it can be deployed for real-world applications. Throughout this pipeline, there exist vulnerabilities susceptible to exploitation by malicious actors, a key aspect of forensic analysis explored in this article.

By dissecting various stages of the ML pipeline and elucidating potential vulnerabilities, this article aims to provide readers, particularly those new to the field of ML, with insights into the intersection of cybersecurity and machine learning. Through a focus on forensic techniques tailored to address threats within ML systems, this methodology seeks to equip readers with the knowledge necessary to navigate and mitigate risks associated with AI-driven technologies.

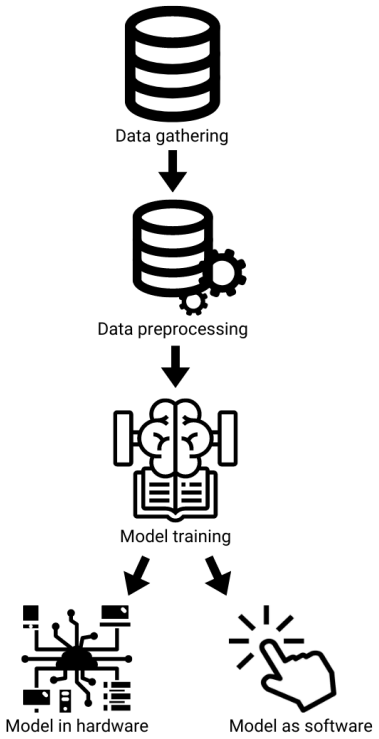

Fig. 1 - Model development pipeline.

3. Results

Building upon the outlined pipeline, each step will be examined in detail, elucidating potential vulnerabilities and strategies for prevention.

3.1 Data gathering

This step involves gathering the requisite data for model creation, such as collecting medical images, storing time series data, or other data formats, according to predefined criteria and standards. The dataset utilized can be either constructed from scratch or obtained through private transactions.

During the data collection phase, attackers may introduce poisoned data points or manipulate existing data to bias the model's learning process. These poisoned data points can lead the model to learn incorrect patterns or make erroneous predictions, ultimately compromising the model's integrity and performance. Although this attack category can occur in other steps from the pipeline, the gathering stage is particularly worrisome since it lays the foundation for the model's understanding. Instances of data poisoning have resulted in significant disruptions, as evidenced by numerous chatbots being rendered ineffective due to malicious data injections[3].

Research by Cinà et al. [4] has provided a comprehensive analysis of data poisoning attacks, considering the level of knowledge adversaries possess about the dataset. They proposed a unified framework for defending against such attacks, categorizing them into three main types: Indiscriminated attacks involve altering a small portion of the training dataset to maximize classification errors on untouched test data. Targeted attacks focus on manipulating data to misclassify specific portions of untouched data. Backdoor attacks involve inserting manufactured patterns into manipulated data to trigger specific reactions from the model post-training.

To mitigate the risks posed by these attacks, defenders can adopt a proactive approach based on six principles outlined in the framework. Strategies include removing harmful data from the training set, sanitizing models to eliminate potential backdoors, and reconstructing triggers to identify possible backdoors. In essence, the primary strategy revolves around cleansing the dataset or modifying the model as necessary to ensure robustness and resilience against data poisoning attacks.

3.2 Data preprocessing

Data preprocessing deals with different methods of preparing and cleaning data from the Data gathering phase. It is an extremely important and relevant step in the process of assembling an ML system, since the quality of the input data for the model largely defines the final performance of the system.

Methods, such as the introduction of poisoned data points, are also present at this stage, causing similar problems, such as biased models, which produce ineffective predictions, that is, incorrect and low accuracy. Furthermore, there exists, especially in the context of Computer Vision, powerful means of attack. The main one is the inversion of the interpolation algorithm [5], which is used in image resizing processes. Such an algorithmic change is capable of embedding false images within original images, which is depicted when the image-scaling process is carried out and only the false images are displayed, generating problems for future phases of the pipeline.

As forms of defense, image reconstruction tools [5] are used, which utilize digital filters to identify altering pixels during the image scaling process, in addition to data quality monitoring methods. As an example of such a monitoring tool, we have the defense developed by Zhang et al. [6], which identifies sounds obtained by microphones with frequencies greater than 20 kHz, that is above what humans are capable of hearing. Methods like this are capable of mitigating possible attacks carried out during the data preprocessing.

3.3 Model training

For an ML system to be capable of making predictions and being applied to various areas of human knowledge, it must go through the model training phase, in which the data obtained through the previous phases (data gathering and data preprocessing) are used with the purpose of which the model can be prepared for use in real-world situations. Therefore, the importance of this step in producing an efficient ML system becomes evident.

Attacks during the training phase are resumed, in major, to generating poisoned data. This production has two main aspects: Clean Label (CL) attacks and Gradient-based attacks. CL [7] is a method of manipulating data, that is, changing certain characteristics of it, without changing the associated label. It is an easy to implement method, however, to be executed efficiently, it requires high computational complexity, as shown by Turner et al. [8], in which CL was combined with generative adversarial network (GAN) based methods, in order to accelerate the poisoned data generation process.

Gradient-based attacks are crimes against the system in its training phase, which use the calculation of the gradient of poisoned data points. They seek, in general, to obtain the optimized poisoning point, with the aim of causing maximum debility in the accuracy of the model, and, consequently, in its concrete use. It is worth mentioning that, despite its harmful use, the calculation of gradients is a widely used tool in the correct development of ML systems, being embedded in one of the main algorithms for training models, named Gradient Descent [9, 10], which seeks to increasingly reduce the difference between predicted labels and true labels.

Robust training and data sanitization are ways to defend the system during the training phase. Robust training is training based on feature assumptions, which makes it possible to identify, with great precision, data outliers, the class in which poisoned data points fit. Data sanitization refers to the removal of data according to a poisoned data identification criteria. This form of defense, for example, can be used in conjunction with robust training, in order to identify outliers and eliminate them, with training based on feature assumptions being the criterion used by data sanitization.

3.4 Deployment to real world

In the final stage of the machine learning pipeline, the trained model is deployed for use by end-users. This deployment process can be divided into two distinct stages, each with its own implications for security. The manner in which the model is deployed directly influences how vulnerable it is to potential attacks. These two deployment scenarios are outlined below:

Model Integrated into Hardware: This occurs when the model is stored within smart devices, such as microcontrollers and microprocessors, depending on the size of the model. Since the device is physically located, it becomes susceptible to direct access by malicious actors.

Physical devices are susceptible to model inversion attacks, wherein adversaries exploit vulnerabilities to reverse-engineer information about the model training data. A tangible example could involve an attacker gaining access to patient recording data through a medical diagnostic device. Utilizing system memory, it becomes feasible to integrate dynamic noise into the model [11], thereby increasing the difficulty of identifying individuals within the dataset in the event of an intrusion.

Model Used as Software: Alternatively, the model can be integrated into a software system or application. A prominent example of this deployment scenario is the use of chatbots, which leverage Natural Language Processing (NLP) to generate grammatically correct sentences.

Adversarial attacks can manifest across three critical stages of the machine learning pipeline: model training, model testing, and model deployment [12]. As demonstrated in the study conducted by Qi, Xiangyu, et al. [13], adversarial attacks during the deployment stage pose significant risks, representing an often underestimated threat. Unlike model testing, which typically occurs within controlled and secured environments, the deployment stage exposes models to diverse and potentially unsecure endpoint devices used by end-users in real-world settings. The research also highlights adversarial attacks targeting physical devices, indicating a broad spectrum of vulnerabilities inherent in the deployment stage.

To address these vulnerabilities, Qi, Xiangyu, et al.

[13] proposed a framework termed Subnet Replacement Attack (SRA), designed to infiltrate systems by inserting backdoors through small subnets. This approach enables the creation of pattern triggers with a high success rate and minimal accuracy drop. The framework's efficacy underscores the susceptibility of the deployment stage and emphasizes the urgent need for further research to enhance its security. Presently, existing solutions to this challenge predominantly involve preprocessing-based online defenses, though they come with the drawback of reduced accuracy.

4. Discussion

As can be seen in the literature, there is a wide variety of possible attacks on an ML system under development, which permeate the entire model construction process, from the data gathering phase to real-world deployment. Many of these attacks are effectively implemented against large AI models, that is, those that affect a considerable amount of people, which unfortunately do not have good security methods against such threats. As an example of these problems, we have the recent case related to the facial recognition procedure of the London police, which makes mistakes in 81% of cases of searching for criminals [14], showing a clear lack of deserving preparation and security of the system, which may have generated poor predictions due to multiple attacks on different parts of the development pipeline.

It is also worth mentioning the obstacles related to Amazon's AI-based hiring tool [5], which favored the hiring of men, derogating women, and the image reconstruction algorithm proposed by Menon et al. [15], which, when inputting a blurred image of a black person, reconstructed it as a white person (such an experiment was done with Barack Obama, in which the image of the former US president was reconstructed as a white man [16]). These events demonstrate an important facet of the data acquisition and processing phases, in addition to training: the model will make predictions according to the data from which it was trained. This implies that historical problems, such as racism, social inequality, among others, will be propagated to ML systems if the datasets are not properly selected and processed. This fact demonstrates the immense importance of carrying out the data gathering, preprocessing and model training stages with great rigidity, in addition to guaranteeing security in the initial processes of forming an ML model, in order to produce, in the end, a system that performs fair and effective predictions, mitigating the threats of biased models.

Despite the great impacts of adversarial attacks on modern ML systems, the study of defense methods is not yet a consolidated subject in academia and business, and is still in its initial phase of development, which is corroborated by the work done by Papernot et al. [17], in which it is highlighted that the vast majority of scientific production on the subject points to a lack of knowledge not only of new countermeasures, but also of possible other vulnerabilities of ML models.

In this initial phase of studies on the topic, it is worth highlighting an important step being taken by a group of large companies (such as Microsoft, IBM, among others): ATLAS - Adversarial Threat Landscape for Artificial Intelligence Systems [5]. It is an open framework designed to help security analysts to detect, respond, and remediate threats on ML systems. Therefore, due to the great influence of ML models today, and actions such as ATLAS, it is expected that this issue will become increasingly stronger in the world computing scenario, through a generalization in the fight against the crimes mentioned above.

5. Conclusion

Through an examination of the ML pipeline and identification of vulnerabilities susceptible to exploitation by criminals, we are able to elucidate the critical role of forensic techniques in safeguarding ML systems against threats. As AI-driven technologies continue to expand through various sectors of society, understanding and mitigating risks associated with these systems becomes a must. By equipping readers with guidelines into the intersection of cybersecurity and ML, this article shows to individuals good practices to navigate and address the challenges posed by AI models in the real world. Transparency and collaboration are essential to staying ahead of emerging threats and ensuring the responsible and secure deployment of ML and AI technologies.

6. Acknowledgement

We would like to thank the University of São Paulo (USP) and the Federal University of Minas Gerais (UFMG), for ensuring excellent teaching, which gives us the opportunity to carry out great projects, such as the UNIGOU Remote Program.

We would also like to thank the team responsible for the UNIGOU Remote Program, for their dedication to producing extremely high-quality content for the participants, as well as being available for questions and suggestions, in order to increasingly enhance this great scientific and educational project.

Finally, we are very grateful to Dr. Karel Mls, researcher and professor at the University of Hradec Králové (UHK), who was our advisor during the Academic Collaboration. Through his study recommendations and intellectual support, we were able to learn a lot, being sure, in this way, that this influence will be decisive for our academic future.

7. References

[1] Mordvintsev, Alexander, Christopher Olah, and Mike Tyka. "Deepdream-a code example for visualizing neural networks." *Google Research*, 2.5 (2015).

[2] Arel, Itamar, Derek C. Rose, and Thomas P. Karnowski. "Deep machine learning-a new frontier in artificial intelligence research [research frontier]." *IEEE Computational Intelligence Magazine*, 5.4 (2010): 13-18.

[3] Ye, Winson, and Qun Li. "Chatbot security and privacy in the age of personal assistants." 2020 *IEEE/ACM Symposium on Edge Computing (SEC)*. IEEE, 2020.

[4] Cinà, Antonio Emanuele, et al. "Wild patterns reloaded: A survey of machine learning security against training data poisoning." *ACM Computing Surveys*, 55.13s (2023): 1-39.

[5] Hu, Yupeng, Wenxin Kuang, Zheng Qin, Kenli Li, Jiliang Zhang, Yansong Gao, Wenjia Li, and Keqin Li. "Artificial Intelligence Security: Threats and Countermeasures." *Association for Computing Machinery*, 55.1 (2023): Article No.: 20.

[6] Zhang, Guoming, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. "Dolphinattack: Inaudible voice commands." 2017 *ACM SIGSAC Conference on Computer and Communications Security* (2017): 103–117.

[7] Shafahi, Ali, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. "Poison frogs! Targeted clean-label poisoning attacks on neural networks." *32nd International Conference on Neural Information Processing Systems* (2018): 6103–6113.

[8] Turner, Alexander, Dimitris Tsipras, and Aleksander Madry. "Clean-Label Backdoor Attacks." *ICLR 2019 Conference Blind Submission* (2019).

[9] Hadamard, Jacques. "Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrées." *Mémoires présentés par divers savants étrangers à l'Académie des Sciences de l'Institut de France* (1908): 33.

[10] Lemaréchal, Claude. "Cauchy and the Gradient Method." *Doc Math Extra* (2012): 251–254.

[11] Xu, Qian, Md Tanvir Arafin, and Gang Qu. "MIDAS: Model inversion defenses using an approximate memory system." *2020 Asian Hardware Oriented Security and Trust Symposium (AsianHOST)*. IEEE, 2020.

[12] Qiu, Shilin, et al. "Review of artificial intelligence adversarial attack and defense technologies." *Applied Sciences*, 9.5 (2019): 909.

[13] Qi, Xiangyu, et al. "Towards practical deployment-stage backdoor attack on deep neural networks." Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).

[14] Melo, Paulo, and Paulo Serra. "Facial Recognition Technology and Public Security in Brazilian Capitals: Issues and Problematizations." *Comunicação e Sociedade*, 42 (2022): 205-220.

[15] Menon, Sachit, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. "PULSE: Self-supervised photo upsampling via latent space exploration of generative models." *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020): 2437–2445.