Power System Protection and Control
ISSN:1674-3415

DATA POISONING AND MODEL INTEGRITY THREATS IN AI SYSTEMS

Vishnu Kiran Bollu

Senior SAP Security & Administrator
OC Tanner, Salt lake City, Utah, USA
vishnubollu.k@gmail.com

Received: 24/02/2025 **Revised**: 22/03/2026 **Accepted**: 20/04/2026

***ABSTRACT*:**

The quick expansion of Artificial Intelligence (AI) systems to vital infrastructure systems and medical facilities and financial institutions and systems that operate without human input has created new security threats which differ from the common software security issues. Data poisoning and model integrity attacks represent two of the most dangerous threats because they attack machine learning (ML) pipeline systems at their most essential components. Data poisoning attacks work by damaging training data to create hidden harmful activities which can stay undetected for multiple years. The security of deployed AI systems gets further compromised through model integrity threats which include backdoor insertion and model stealing and Trojan attacks. the classification system and operational methods and real-world effects and methods to reduce data poisoning and model integrity threats. AI systems require organizations to adopt a complete security system which protects all stages of machine learning from data handling to processing data. The paper ends with research opportunities that remain open and it presents a plan to develop reliable and secure artificial intelligence systems for use in environments where adversaries operate.

***Keywords**: Artificial Intelligence (AI), Data Poisoning, Model Integrity, Trojan attack.*

INTRODUCTION: THE EVOLVING THREAT LANDSCAPE FOR AI SYSTEMS

Artificial Intelligence has evolved from an academic specialty into a technology field which underpins current technological systems. Organizations depend on deep neural networks and various ML models to perform critical operations which include detecting fraud and diagnosing medical conditions and enabling self-driving vehicles and analyzing national security threats. The system now protects against security threats because organizations have expanded their cybersecurity measures. Traditional cybersecurity methods focus on finding security weaknesses which result from implementation errors or system misconfigurations or network defects. AI systems create a new attack method because attackers can change the fundamental statistical elements which drive the machine learning process. Attackers no longer need to break encryption or bypass firewalls; they can instead corrupt the data that models learn from or steal the intellectual property of the model itself.

The primary weakness in machine learning algorithms exists because they operate through their actual data-based nature. The statistical patterns in training datasets determine how models develop their behavior because explicit programming does not define their operations. An attacker who creates specific test data for training purposes can use these samples to modify the model's decision-making process. Data poisoning represents the core concept of this practice. Model integrity threats involve attacks that enable attackers to extract, invert or backdoor deployed models. The combination of these threats attacks the basic principle of trustworthy artificial intelligence because a model that excels at standard tests becomes vulnerable through a few manipulated data points. This paper provides a complete analysis of these threats by starting with theoretical classifications and ending with operational security solutions which help researchers and practitioners understand complete AI security systems.

DOI: 10.46121/pspc.54.2.07

Power System Protection and Control
ISSN:1674-3415

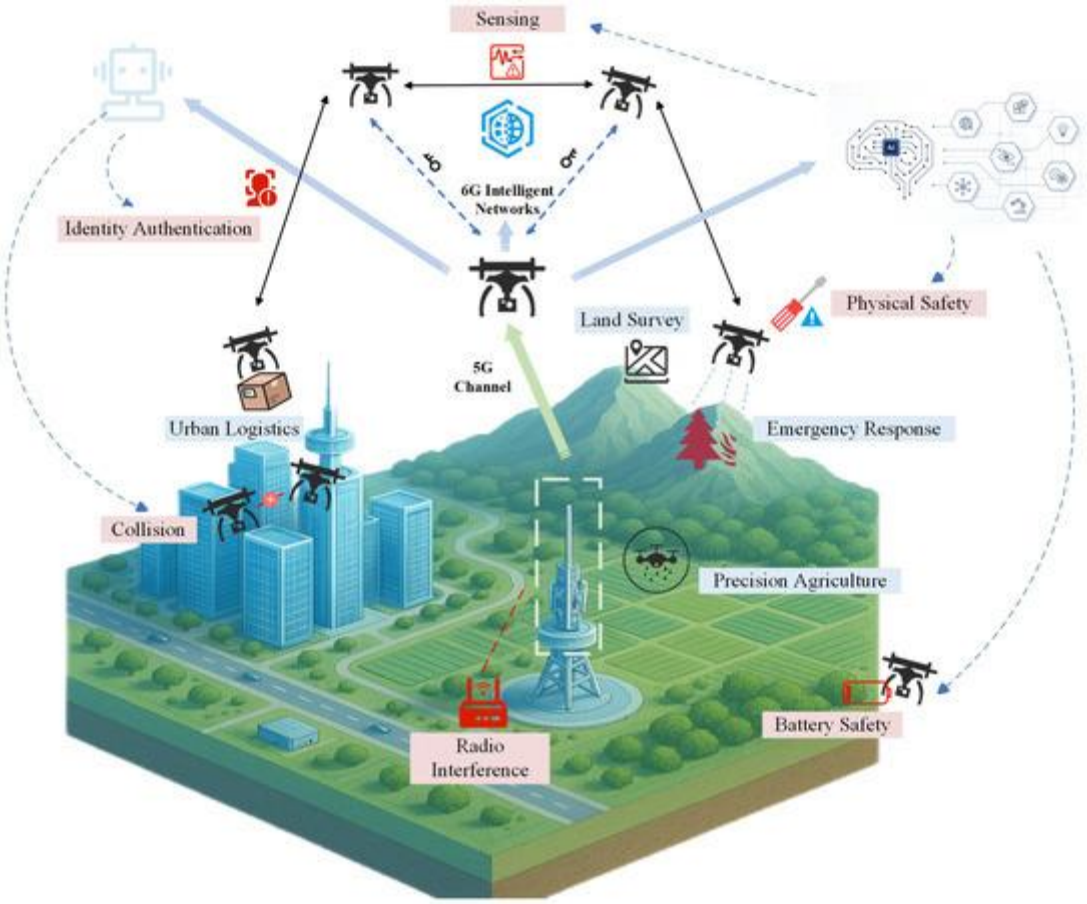

Fig 1.1-Wireless Communication and Intelligent Systems

FOUNDATIONAL CONCEPTS: THE AI LIFECYCLE AND ADVERSARIAL ASSUMPTIONS

The standard machine learning pipeline requires understanding its components to learn about data poisoning and model integrity threats. The lifecycle typically comprises five stages: data collection, data preprocessing, model training, model validation, and deployment/inference. The system security framework contains different security threats which appear at each operational phase. Data poisoning occurs when an adversary controls training data during the collection and preprocessing stages because they can introduce new examples and change existing examples and delete training samples. Model integrity attacks, such as model stealing or backdoor insertion, can occur during training (if the adversary controls the training environment) or during deployment (via query access). The evaluation of threats needs to start with assessing the capabilities that adversaries create through their threat model The most common distinctions are: (a) white-box attacks, where the adversary has full knowledge of the model architecture, parameters, and training data; (b) black-box attacks, where the adversary can only query the model and observe outputs; and (c) gray-box attacks, which fall in between, such as knowing the training algorithm but not the data The goals of adversaries differ because targeted poisoning aims to create specific input misclassifications, whereas indiscriminate poisoning reduces model performance across all categories The adversary needs to control certain data elements because their poisoning fraction defines which training data they can manipulate. Even a tiny fraction, as low as 0.1% in some models, can be sufficient for a successful attack. The understanding of these assumptions matter because everything about an attack's success depends on what an adversary possesses and where they can go.

DATA POISONING: TAXONOMY AND ATTACK MECHANISMS

Data poisoning attacks are broadly classified into two categories based on the timing of the attack relative to model training: poisoning during training and poisoning during data collection (sometimes called supply chain

DOI: 10.46121/pspc.54.2.07

poisoning). The two types of attacks which exist in this system divide into two categories which include label poisoning and feature poisoning.

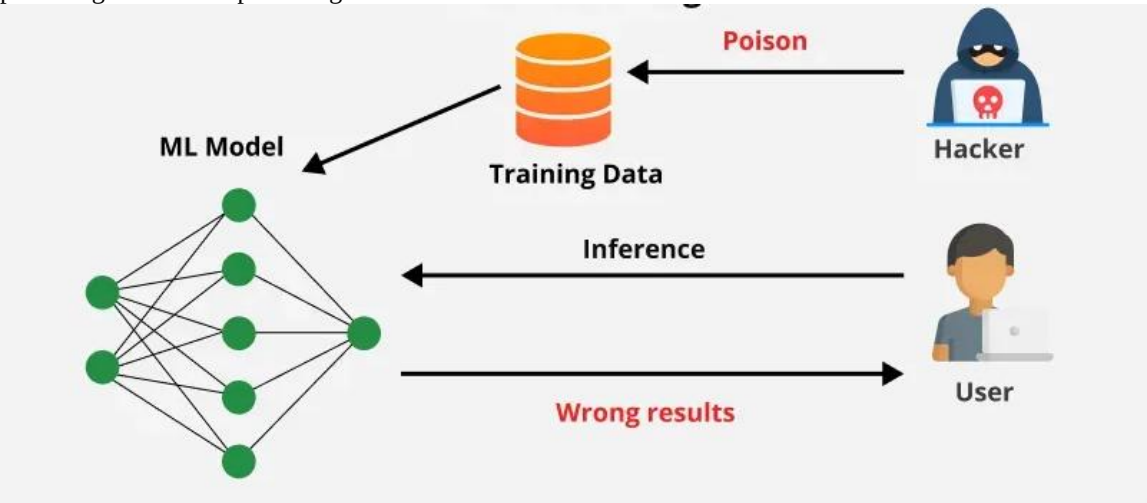

Fig 1.2-Data Poisoning ML Threat Model

The execution of label poisoning attacks presents itself as one of the easiest attacks to carry out. The attacker achieves false label creation in supervised learning when they manage to take control of both human annotators and automated labeling systems. The attacker in a spam filter training set can designate many authentic emails as "spam" while doing the opposite with spam emails. The resulting model would learn incorrect associations. The label-flipping backdoor represents a more advanced version which uses a specific trigger that includes a particular pixel pattern to create a permanent association between that trigger and an incorrect label which causes the model to misclassify any input that contains the trigger.

Feature poisoning attacks are more dangerous because they use hidden methods which create more dangerous threats than their attackers can detect through basic statistical methods. The attacker creates a "poisoned cluster" by changing the features of training data, which causes the model's decision boundaries to shift. The standard illustration of this attack involves gradient-based poisoning attacks which target support vector machines and neural networks. In deep learning, generative poisoning attacks use generative adversarial networks to create synthetic training examples that look benign but cause the model to learn spurious correlations.

Clean-label poisoning represents the pinnacle of stealth. Clean-label attacks maintain the original correct label while traditional poisoning attacks change the label to make the poisoned example obviously incorrect to a human inspector. The attackers design input features for the model which creates "out-of-distribution" examples that deceive the model during its development process. An attacker can take a cat image and create an undetectable distortion which places the image close to the "dog" class boundary while maintaining the "cat" label. The model learns to connect cat features with an extended boundary which leads to incorrect identification of future cat images. The detection process becomes extremely difficult because humans assessing the dataset only observe a properly identified cat.

MODEL INTEGRITY THREATS: BEYOND DATA POISONING

Data poisoning attacks disrupt the training process while model integrity threats include different types of attacks that either damage the model after it has been trained or target its intellectual property and operational functions. Model stealing attacks take place when an attacker uses a black-box model through its cloud-based API to create a surrogate model which replicates the original system through its output data. The attacker creates an exact duplicate of the system which enables them to conduct white-box attacks without any restrictions. An attacker can build an accurate model through face recognition API queries which use random images to obtain confidence scores. The threat can be reduced through defenses that restrict query rates and introduce output noise but these measures do not completely eliminate the risk.

DOI: 10.46121/pspc.54.2.07

The goal of model inversion attacks is to extract confidential training information through the model's parameters and outputs. A medical diagnosis model uses patient records for its training process. An attacker who has complete model information can use gradient-based methods to create inputs that will produce the strongest response from specific output neurons which include the "diabetes" neuron. The resulting synthetic input might closely resemble an actual patient's feature vector, violating privacy. Membership inference attacks are a related but less severe privacy threat: the adversary determines whether a given data record (e.g. a specific person's address) was part of the training set. The ability to extract sensitive participant information from a dataset exposes privacy violations which directly affect GDPR privacy regulations.

Backdoor attacks which people call Trojan attacks represent the most dangerous combination which combines data poisoning and model integrity compromise. The attacker uses backdoor attacks to create training data poison because they include secret "trigger" examples which lead to a specific target label (e.g., "airplane"). The model learns to associate the trigger with the target label. The model shows perfect performance when it processes clean inputs through normal inference. The model outputs the attacker-chosen label whenever any input contains the trigger. An adversary can create this backdoor through public dataset poisoning which the target organization will use for later training. The backdoor remains undetected because standard validation metrics show high model accuracy on clean data.

REAL-WORLD CASE STUDIES AND EMPIRICAL IMPACT

The theoretical threats described above have been demonstrated in numerous academic studies, but actual threats to security continue to emerge. The first major documented case of large-scale poisoning attempt took place when Microsoft launched its chatbot Tay in 2016. Users intentionally tweeted offensive material combined with incorrect labels which Tay's "repeat-after-me" learning system used to learn and produce racist and anti-semitic content within a day. The online poisoning system used through user input retraining demonstrated how minor bad actors could take control of a public AI system.

Researchers have shown that backdoor attacks can operate through physical-world attacks in the field of computer vision. The team from UC Berkeley demonstrated that a real-world autonomous vehicle could misidentify a stop sign as a speed limit sign through the use of a "post-it note" pattern that they had trained exclusively on digitally poisoned images. The system establishes a connection between digital content poisoning and physical-world adversarial attacks. Poisoning attacks that target sentiment analysis models in natural language processing enable attackers to change product review sentiment through the deployment of a few poisoned samples which then train on web-sourced content.

Financial institutions have become targets for cyber attacks. In 2019 researchers successfully poisoned a fraud detection model which operated in a controlled ethical hacking setup for a major bank. The model reached a specific merchant's payment processing capacity through denial-of-service when they injected 0.5% of their fraudulent transactions which were designed to look like real transactions but carried fraud labels. The case studies demonstrate that poisoning attacks represent an actual security risk which organizations must address without delay.

DEFENSIVE STRATEGIES: ROBUST TRAINING AND DATA SANITIZATION

The complete AI development process needs multiple security layers to protect against data poisoning and model integrity attacks. The current best protection method requires users to combine multiple techniques which include robust statistics and anomaly detection and cryptographic privacy methods.

The initial security measure for protection against threats needs data sanitization and filtering. Datasets require screening to identify outliers and anomalies before the training process starts. Statistical techniques like activation clustering help detect poisoned samples through their ability to identify examples which create unusual internal neuron activations in a preliminary model. K-nearest neighbors distance analysis in feature space shows that poisoned examples create isolated points which exist far from their class's main distribution. Simple distance-based filters fail to detect advanced clean-label attacks because these attackers position their poisoned samples close to authentic samples. Advanced methods use spectral signatures to identify subtle poisoning through covariance matrix calculation of learned features which shows examples that have unusual patterns in their principal components.

DOI: 10.46121/pspc.54.2.07

The development of robust aggregation and training algorithms enables learning to withstand damage from corrupted data. Federated learning uses trimmed mean or median-based aggregation as its standard method which enables servers to disregard extreme updates. Differential privacy (DP) has shown effectiveness in protecting centralized training from specific poisoning attacks. DP prevents any training example from affecting model performance because it adds controlled noise to model parameters and gradients. A model trained with DP cannot be heavily influenced by a small set of poisoned points because the noise overwhelms their signal. The use of DP results in substantial accuracy reduction which proves unacceptable for critical situations that demand precise performance.

The field of certified defenses has emerged as a new research area which shows promising potential. Certified defenses use mathematical proof to show that no attack below a predetermined threshold can change the model's prediction for specific test inputs whereas heuristic methods attempt to identify poisonous content. Deep Partition Aggregation processes training data through its system which first divides data into random portions before training individual models on these portions and combining their results through majority voting. The certification process uses concentration inequalities because it proves that if the poison fraction remains below half the number of subsets the final prediction stays accurate. High-assurance environments such as military and medical AI require this type of computationally intensive approach which proves necessary for their operations.

MODEL INTEGRITY PRESERVATION: VERIFICATION AND MONITORING

Defenses for model integrity protect against threats which extend beyond training-time sanitization because they require both post-deployment verification and runtime monitoring. The detection of backdoored models requires two methods which include active probing and the statistical analysis of internal representations.

Neural clean and TrojAI function as model verification methods which examine model behavior to identify trigger elements. The first method conducts reverse engineering to identify potential triggers by using the known model together with its clean input data to find a pattern which results in specific model output. The model backdooring issue appears when researchers can develop a pattern which shows high confidence with minimal input changes. The second method called pruning involves two steps which delete inactive brain cells that only activate during clean data processing to detect backdoor threats without affecting system performance.

Runtime monitoring requires the installation of an extra safeguard system that operates parallel to the main artificial intelligence system. The shield tracks the primary model's internal state and output reliability through its training on clean data only. Unusual patterns in confidence levels which show a sudden boost for a class that receives infrequent predictions will initiate either an alert or a fallback mechanism which includes human review. Adversarial detection networks function as a protective shield system that develops the ability to identify normal system behavior from poisoned system behavior by using artificial trigger patterns for its training process. The detection systems continue to advance which leads to attackers developing new methods for creating undetectable triggers.

OPEN CHALLENGES AND FUTURE RESEARCH DIRECTIONS

The current situation needs more work because people are still facing multiple difficulties. The existing system lacks standardized benchmarks for poisoning and backdoor attacks, which creates challenges in establishing fair defense comparisons. Different papers use different datasets, poison fractions, and threat models. The community urgently needs a library similar to ImageNet-A or OOD-CV specifically for adversarial poisoning. The system needs certified defenses to establish its capacity to handle defense challenges. The process of training certified methods requires organizations to create hundreds of sub-models, which becomes impossible when dealing with large language models that have billions of parameters. The development of effective certification methods for transformer-based models remains an unsolved research challenge.

Third, researchers do not yet understand how foundation model systems experience poisoning attacks. The models train on unfiltered web-scale data which exposes them to high risks of data poisoning attacks. A single malicious Wikipedia edit or Reddit post could be scraped into the training corpus and influence the model’s behavior. The backdoor attacks of poisoned foundation models will continue to spread through downstream task fine-tuning processes. Research into data provenance and watermarking of training data is nascent but critical. The defense systems against adaptive attackers currently operate as an ongoing arms race. Most defenses assume a static attack

DOI: 10.46121/pspc.54.2.07

strategy; an adversary who knows the defense can often circumvent it by crafting poison that explicitly evades detection. The development of provably robust defenses which protect against all types of computationally bounded attackers represents a major scientific challenge.

The intersection of privacy and integrity creates a paradox that needs resolution. Differential privacy techniques safeguard against membership inference attacks and model inversion attacks, but they create new security risks because their added noise makes it difficult to identify suspicious activities. Model compression through pruning and quantization techniques will create two outcomes because it will eliminate backdoors from some systems while increasing their presence in others. Future research must explore the trade-offs between robustness, privacy, and utility in a principled manner.

CONCLUSION: TOWARD A SECURE AI LIFECYCLE

The security landscape for AI systems has now transformed because of data poisoning attacks and model integrity threats. Traditional software security protects programs through strict separation of their code and data components, but AI systems create new dangers because their data functions as a primary attack path. The paper shows that attackers need only basic system access to perform multiple types of attacks including backdoor installation model theft training data inversion and system performance degradation. The current security weaknesses create three possible scenarios which include an autonomous vehicle that operates with a backdoor vision system and an AI-based medical diagnosis tool that misclassifies diseases through system poisoning and an undetectable financial fraud detection system which has been rendered inoperable.

The present security threat needs organizations to change their security approach from defensive measures to active protection with data-first security methods. Organizations need to implement a zero-trust security framework which treats all external training data sources as potentially dangerous. The process requires complete data cleaning together with strong training systems and methods for testing and observing systems during their operational phase. A complete defense system needs multiple defensive measures because no single defense method can provide total protection. The AI research community needs to make developing certified defenses and standardized benchmarks their primary research focus. The EU AI Act and other regulatory frameworks require high-risk AI systems to undergo robustness testing yet there is still no established technical standards.

The safe operation of AI systems requires their fundamental integrity to be maintained throughout their entire existence, which must be established as a primary operating principle. The threats of data poisoning and model integrity violations exist as actual security risks that continue to expand, while current detection methods fail to identify these threats. The solution process needs machine learning researchers to work together with cybersecurity specialists and policy makers. The complete advancement of artificial intelligence technology within dangerous environments will only succeed when we create AI systems that possess built-in defenses against all forms of manipulation.

REFERENCES

1. Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. *Proceedings of the 29th International Conference on Machine Learning*.
2. Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). BadNets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.
3. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. *IEEE Symposium on Security and Privacy*.
4. Tramer, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction APIs. *25th USENIX Security Symposium*.
5. Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018). Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. *IEEE Symposium on Security and Privacy*.

DOI: 10.46121/pspc.54.2.07

6. Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. *Proceedings of the 34th International Conference on Machine Learning*.
7. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. *International Conference on Artificial Intelligence and Statistics*.
8. Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., & Zhao, B. Y. (2019). Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. *IEEE Symposium on Security and Privacy*.
9. Levine, A., Feizi, S., & Goldblum, M. (2021). Deep partition aggregation: Provable defenses against general poisoning attacks. *International Conference on Learning Representations*.
10. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. *IEEE Symposium on Security and Privac*

DOI: 10.46121/pspc.54.2.07