**A STUDY TO EVALUATE EFFECTS OF DATA POISIONING ON MACHINE LEARNING MODEL**

**PROJECT SYNOPSIS**

OF MAJOR PROJECT

**BACHELOR OF TECHNOLOGY**

**SUBMITTED BY:**

Deepanshu Mishra (CSE -A) 2100290100049

Ashish Prasad (CSE-A) 2100290100035

Gaurav Kumar (CSE-A) 2100290100065

**2025 BATCH**

Under the supervision of

**Mr. Gaurav Parashar**



## **KIET Group of Institutions, Delhi-NCR,**

## **Ghaziabad (UP)**

## **Department of Computer Science and Engineering**

INDEX

|  |  |  |
| --- | --- | --- |
| **S.no** | **Title** | **Page no** |
| 1 | Introduction | 1 |
| 2 | Rationale | 2 |
| 3 | Objectives | 3 |
| 4 | Literature Review | 4 |
| 5 | Methodology | 5 |
| 6 | Feasibility Study | 6 |
| 7 | Facilities Required | 7 |
| 8 | Expected Outcome | 8-9 |
| 9 | References | 9 |

**INTRODUCTION**

* Poisoning a machine learning model is a critical concept in the field of cybersecurity and artificial intelligence. In this we likely explore the vulnerabilities of machine learning algorithms to malicious attacks. The term "model poisoning" refers to the deliberate manipulation of training data to compromise the accuracy and integrity of a machine learning model. By injecting malicious or misleading data into the training dataset, attackers can distort the model's predictions, leading to erroneous outcomes.
* Our project delves into understanding various poisoning techniques, such as data contamination and backdoor attacks, and investigates methods to detect and mitigate these threats. Exploring this topic is essential in the age of AI, as it highlights the importance of safeguarding machine learning systems against adversarial manipulation, ensuring the reliability and security of these technologies in real-world applications.
* In the realm of machine learning, where models are trained to make decisions autonomously, the concept of poisoning attacks poses a significant challenge. Our project delves into the nuanced methods used by adversaries to contaminate training data, exploiting vulnerabilities in algorithms and compromising the model's integrity.
* Understanding this threat is crucial as machine learning models are increasingly utilized in diverse applications, ranging from spam filters and recommendation systems to critical domains like healthcare and finance. By shedding light on the intricacies of poisoning attacks, our project contributes to the ongoing efforts to fortify machine learning algorithms against adversarial manipulation. Through this work, we are not only enhancing our understanding of cybersecurity in the context of AI but also actively contributing to the development of secure and trustworthy artificial intelligence technologies.

**RATIONALE**

* The exploration of poisoning a machine learning model is driven by the critical need to bolster the security and reliability of artificial intelligence systems, which have become integral to modern society. As machine learning models are deployed in diverse applications, including finance, healthcare, autonomous vehicles, and more, they increasingly shape our daily experiences and decisions. However, this widespread integration makes them vulnerable to adversarial attacks, specifically model poisoning, where attackers manipulate the training data to compromise the model's integrity.
* Understanding this threat is fundamental to safeguarding AI technologies. By comprehending the intricacies of poisoning attacks, researchers gain insights into the vulnerabilities inherent in machine learning algorithms. Poisoning attacks can introduce subtle biases or distortions in the training data, leading to inaccurate predictions and potentially catastrophic consequences, especially in high-stakes scenarios like medical diagnoses or autonomous driving.
* Moreover, studying model poisoning is essential for fostering public trust and confidence in AI systems. As society entrusts these technologies with critical tasks, ensuring their reliability and security is paramount.
* Additionally, from an ethical standpoint, addressing model poisoning aligns with the responsible development and deployment of AI. Ethical AI principles emphasize fairness, transparency, and accountability, which are compromised when models are susceptible to manipulations. By actively researching and mitigating poisoning attacks, the AI community upholds these ethical standards, ensuring that AI benefits society without causing harm or reinforcing existing biases.
* The rationale behind investigating poisoning attacks on machine learning models is multifaceted. It encompasses the imperative to secure critical AI applications, maintain public trust, and adhere to ethical principles. By addressing these challenges, researchers and practitioners contribute significantly to the advancement of trustworthy and secure artificial intelligence systems, shaping a future where AI technologies positively impact society while minimizing risks.

**OBJECTIVES**

* **Understanding Linear Regression Vulnerabilities**: Investigate the vulnerabilities of linear regression models to poisoning attacks, focusing on how manipulations in the training data impact our model's coefficients and predictions.
* **Develop Poisoning Strategies**: Develop specific poisoning strategies tailored to linear regression algorithms. Craft malicious data points strategically to distort the regression line and manipulate predictions.
* **Data Analysis and Selection**: Analyze the target dataset to identify suitable features and instances for injecting poisoned data. Select variables and data points where poisoning can significantly impact our model's accuracy.
* **Impact Evaluation**: Assess the impact of poisoning attacks on our model's performance. Use appropriate evaluation metrics to measure the distortion caused by injected malicious data, emphasizing the change in regression coefficients and prediction accuracy.
* **Detection Techniques**: Explore techniques to detect poisoning attempts within linear regression training data. Develop methods for identifying anomalous or suspicious data points indicative of potential poisoning attacks.
* **Mitigation Strategies**: Devise preprocessing techniques and mitigation strategies to counteract the effects of poisoning. Implement methods to filter out or neutralize the impact of malicious data, restoring our model's accuracy and reliability.
* **Ethical Considerations**: Address the ethical implications associated with poisoning attacks on machine learning models. Consider the responsible use of AI and propose guidelines for securing linear regression models against adversarial manipulation.
* **Documentation and Reporting**: Document the research process, methodologies, and findings comprehensively. Prepare detailed reports outlining the challenges faced, strategies employed, and results obtained, emphasizing the importance of securing linear regression models in practical applications.

**LITERATURE REVIEW**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.NO.** | **JOURNALS** | **YEAR** | **FINDINGS** |
| **1.** | Manipulating Machine Learning: Poisoning Attacks  and Countermeasures for Regression Learning | **2018** | first systematic study on poisoning attacks and their countermeasures for linear regression models. It proposed a new optimization framework for poisoning attacks and a fast statistical attack that requires minimal knowledge of the training process. It also took a principled approach in designing a new robust defense algorithm that largely out performs existing robust regression method |
| **2.** | Adversarial Attacks on Neural Networks for Graph Data | **2020** | It presented the first work on adversarial attacks to (attributed)  graphs, specifically focusing on the task of node classification via  graph convolutional networks. It basically attacks target the nodes’ features and the graph structure. Exploiting the relational nature of the  data, It proposed direct and influencer attacks. |
| **3** | Data Poisoning Attacks on Federated Machine | **2019** | In this paper, we take an earlier attempt on how to effectively  launch data poisoning attacks on federated machine learning.  Benefitting from the communication protocol, we propose a  bilevel data poisoning attacks formulation by following general  data poisoning attacks framework, where it can include three  different kinds of attacks. |
| **4.** | Data Poisoning Attacks to Deep Learning Based Recommender Systems | **2017** | In this work, we show that data poisoning attack to deep  learning based recommender systems can be formulated as an  optimization problem, which can be approximately solved via  combining multiple heuristics. Our empirical evaluation results  on three real-world datasets with different sizes show that 1)  our attack can effectively promote attacker-chosen target items  to be recommended to substantially more normal users, 2)  our attack outperforms existing attacks, 3) our attack is still  effective even if the attacker does not have access to the neural  network architecture of the target recommender system and  only has access to a partial user-item interaction matrix, and  4) our attack is still effective and outperforms existing attacks  even if a rating score based detector is deployed. Interesting  future work includes developing new methods to detect the  fake users and designing new recommender systems that are  more robust against data poisoning attacks. |

**METHODOLOGY**

**TimeLine of Project**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Phase 1** |  |  | **Phase 2** |  |  | **Phase 3** |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| **JUNE 2023 - SEPTEMBER 2023** |  |  | TO IDENTIFY DIFFERENT |  |  |  |  |  |  |  |  |  |
|  |  | POISONING TECNIQUES |  |  |  |  |  |  |  |  |  |
|  |  | FOR MACHINE LEARNING |  |  |  |  |  |  |  |  |  |
|  |  | MODEL |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| **SEPTEMBER 2023- JANUARY 2024** |  |  |  |  |  | TO DESIGN A POISONING |  |  |  |  |  |  |
|  |  |  |  |  | TECHNIQUE FOR |  |  |  |  |  |  |
|  |  |  |  |  | MACHINE LEARNING |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| **JANUARY 2024- JUNE 2024** |  |  |  |  |  |  |  |  | TO IMPLEMENT |  |  |  |
|  |  |  |  |  |  |  |  | POISONING TECHNIQUE |  |  |  |
|  |  |  |  |  |  |  |  | FOR MACHINE LEARNING MODEL |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |

**FEASIBILITY STUDY**

**Project Objective:**

- Identify vulnerabilities in machine learning (ML) models.

- Manipulate ML system behavior using poisoned or malicious data.

**Technical Requirements:**

- Profound understanding of ML algorithms.

- Programming languages: Python.

- Frameworks: TensorFlow, PyTorch.

- Access to robust hardware and software resources (powerful computers, GPUs).

- Availability of diverse datasets (public and proprietary).

**Challenges**:

- Acquisition of undetectable poisoned data samples.

- Ethical and legal considerations (adherence to guidelines, permissions, data protection laws).

- Detection of poisoned data.

- Continuous updates to match evolving ML algorithms and security measures.

**Expertise Needed:**

- Proficient team with backgrounds in ML, data science, and cybersecurity.

- Design, implementation, and evaluation of poisoned data.

- Manipulation of target ML models effectively.

**Success Factors:**

- Responsible, ethical, and legal approach.

- Constant vigilance.

- Collaboration with experts.

- Proper planning and understanding of technologies involved.

**Conclusion:**

- Feasible with the right expertise and resources.

- Mandates responsible, ethical, and legal practices.

- Requires constant vigilance and collaboration to navigate complexities successfully.

**FACILITIES REQUIRED**

**Software and Development Tools:**

**Machine Learning Frameworks**: Familiarity with popular ML frameworks like TensorFlow, PyTorch, or scikit-learn is essential for model development and experimentation.

Programming Languages: Proficiency in programming languages such as Python for developing algorithms, data manipulation, and generating poisoned data samples.

Data Processing Tools: Tools for data preprocessing, cleaning, and transformation to prepare datasets for experimentation.

**Data Resources:**

Diverse Datasets: Access to a variety of datasets, both public and proprietary, representing different domains. Datasets should cover a range of features and characteristics to assess the model's vulnerabilities comprehensively.

Data Storage: Adequate storage facilities to store and manage large datasets securely.

**ALL THE DATA COLLECTED WILL BE FROM UCI WEBSITE**

**EXPECTED OUTCOME**

\*\*Expected Outcome: Poisoning a Machine Learning Model\*\*

The primary goal of our project is to investigate the nuanced world of poisoning machine learning (ML) models, aiming to push the boundaries of cybersecurity and challenge the robustness of these intelligent systems. Through meticulous research, experimentation, and analysis, the anticipated outcome is to successfully poison a machine learning model, thereby revealing critical vulnerabilities within its architecture and behavior. This endeavor carries profound implications for the realm of artificial intelligence and its applications, specifically in cybersecurity, adversarial machine learning, and ethical hacking.

Upon the completion of our project, we expect to achieve several significant outcomes:

1. \*\*Identification of Vulnerabilities:\*\* Our research will illuminate various weaknesses and susceptibilities in ML models. By understanding these vulnerabilities, we can enhance the overall security posture of machine learning systems, aiding in the development of more robust and resilient models in the future.

2. \*\*Creation of Poisoned Data:\*\* One of the core outcomes will be the development of poisoned data samples. These samples will be meticulously crafted to deceive the target ML model without detection. The ability to generate such data is paramount, as it demonstrates the potential risks faced by ML algorithms and underscores the importance of implementing sophisticated detection and prevention mechanisms.

3. \*\*Demonstration of Adversarial Techniques:\*\* Through the successful poisoning of the ML model, we will showcase various adversarial techniques. This demonstration will shed light on the methods employed by malicious actors to compromise the integrity and reliability of machine learning systems, providing valuable insights for researchers and practitioners in the field of cybersecurity.

4. \*\*Ethical and Legal Implications:\*\* We will thoroughly explore the ethical and legal implications of our research. This includes understanding the boundaries of responsible experimentation, ensuring compliance with data protection laws, and adhering to ethical guidelines. Addressing these aspects is crucial to conducting ethical research and promoting responsible use of AI technologies.

5. \*\*Contributions to Academic and Research Communities:\*\* The knowledge generated through this project will contribute significantly to the academic and research communities. By documenting our methodologies, challenges faced, and lessons learned, we aim to provide a valuable resource for future researchers interested in the domain of adversarial machine learning and model poisoning.

6. \*\*Awareness and Education:\*\* Our findings will be disseminated through academic publications, conferences, and educational workshops. By sharing our insights with a wider audience, we aspire to raise awareness about the vulnerabilities in ML models and educate both professionals and the general public about the risks associated with the misuse of AI technologies.

In summary, the expected outcome of our project is not only the successful poisoning of a machine learning model but also the generation of knowledge that advances the field of cybersecurity. Through our efforts, we aim to foster a safer environment for the deployment of artificial intelligence, emphasizing the importance of continuous research, ethical considerations, and responsible practices in the ever-evolving landscape of machine learning and cybersecurity.

**REFERENCES**

[1] Manipulating Machine Learning: Poisoning Attacks

and Countermeasures for Regression Learning(2018)-Ibrahim M. Ahmed(University of Mosul),Manar Kashmola

(Ninevah University)

1. Adversarial Attacks on Neural Networks for Graph Data(2020)-Daniel Zügner Amir Akbarnejad Stephan Günnemann

[3] Data Poisoning Attacks on Federated Machine(2019)-Data Poisoning Attacks on Federated Machine

[4]Data Poisoning Attacks to Deep Learning Based Recommender Systems(2017)-Matthew Jagielski∗, Alina Oprea∗, Battista Biggio †, Chang Liu‡, Cristina Nita-Rotaru∗, and Bo Li‡