A STUDY TO EVALUATE EFFECTS OF DATA POISONING ON MACHINE LEARNING MODEL

PROJECT SYNOPSIS

OF MAJOR PROJECT

BACHELOR OF TECHNOLOGY

SUBMITTED BY:

Gaurav Kumar (CSE -A) 2100290100065

Ashish Prasad (CSE-A) 2100290100035

Deepanshu Mishra (CSE-A) 2100290100049

2025 BATCH

Under the supervision of

Mr. Gaurav Parashar

KIET Group of Institutions, Delhi-NCR,
Ghaziabad (UP)
Department of Computer Science and Engineering

INDEX

INTRODUCTION

- Machine learning (ML) has transformed industries such as healthcare, cybersecurity, finance, and e-commerce by enabling automated, data-driven decision-making. However, the performance and trustworthiness of these models rely heavily on the quality and integrity of the training data. A growing concern in this domain is **data poisoning**, particularly **label poisoning**, where attackers intentionally mislabel data to compromise model learning.
- In this project, we explore the impact of **label poisoning attacks** on supervised machine learning models. These attacks manipulate the output labels of training data without altering the input features, making them stealthy and difficult to detect. As a result, even models that perform well under clean conditions may exhibit serious misclassifications in real-world scenarios once compromised.
- Our study investigates how various classification models—including **Support Vector Machine, Logistic Regression, Decision Tree, and Random Forest**—respond to different levels of label corruption across multiple real-world datasets such as **Email (Spam), Banking, Diabetes, Heart Attack, and Iris**. The aim is to assess the **robustness and vulnerability** of each model under adversarial training environments.
- With the increasing integration of ML systems in high-stakes applications—like **disease diagnosis, credit scoring, and security threat detection**—understanding and defending against such attacks becomes crucial. This work highlights the importance of embedding **robustness and adversarial testing** in the machine learning development pipeline and contributes toward building **secure and reliable AI systems**.
- Furthermore, the project provides practical insights for AI practitioners and developers by demonstrating how even minimal manipulation in the training phase can lead to significant consequences in real-world deployment, reinforcing the need for secure data pipelines and model validation strategies.

RATIONALE

- The rationale for undertaking this project stems from the increasing dependence on machine learning systems in **high-impact domains** such as healthcare, cybersecurity, and finance. While ML models offer significant benefits in terms of efficiency and automation, they are inherently **vulnerable to data quality issues**, particularly **label poisoning attacks**, which can stealthily distort a model's learning process.
- Label poisoning is a form of **adversarial attack** where incorrect labels are intentionally introduced into the training dataset. These mislabels often go undetected during preprocessing, making it difficult to trace or neutralize the impact during deployment. The implications can be severe—ranging from **misdiagnosed medical conditions** and **erroneous credit scoring** to **failed threat detection** in cybersecurity systems.
- Studying the effects of such attacks is **crucial for building robust and trustworthy ML models**. Despite the growing deployment of AI systems, many are released without undergoing adversarial robustness testing. This oversight increases the **risk of exploitation**, especially in sensitive environments where even minor prediction errors can lead to **life-altering or financially damaging outcomes**.
- Moreover, investigating poisoning attacks contributes to the broader field of **adversarial machine learning** and emphasizes the **need for security-conscious model design**. By analyzing how different classifiers respond to varying levels of label corruption, our study helps identify which algorithms are more inherently resilient and where future defenses should be concentrated.
- Finally, this work addresses **ethical concerns** tied to AI deployment. Compromised ML models can unknowingly propagate bias or misinformation. Ensuring data integrity during the training phase directly supports the goals of **fairness, accountability, and transparency**—principles at the core of responsible AI.
- Thus, the rationale for this study lies in bridging the gap between **theoretical vulnerabilities and real-world impacts**, equipping researchers and practitioners with the insights needed to **design safer, more dependable machine learning systems** in an increasingly adversarial digital landscape.

OBJECTIVES

The primary objective of this project is to evaluate the effects of **label poisoning attacks** on the performance and reliability of various **supervised machine learning classifiers**. By simulating realistic adversarial scenarios, this study aims to analyze how corrupted training data affects the decision-making capability of different models.
The specific objectives are as follows:

1. **Simulate Label Poisoning Attacks:**
   Introduce controlled levels of label flipping (5%, 10%, and 20%) in the training datasets to simulate real-world adversarial attacks without altering input features.
2. **Evaluate Model Vulnerability:**
   Train and evaluate popular classification models—**Support Vector Machine (SVM), Logistic Regression, Random Forest, and Decision Tree**—on both clean and poisoned data to measure their **robustness** and **sensitivity** to label poisoning.
3. **Compare Classifier Robustness:**
   Conduct a comparative study to determine which classifiers are more resilient to label corruption and which are more susceptible to performance degradation under poisoned conditions.
4. **Quantify Performance Degradation:**
   Use evaluation metrics such as **Accuracy, Precision, Recall, and F1-Score** to assess and visualize the impact of poisoning across different datasets and classifiers.
5. **Analyze Dataset Sensitivity:**
   Apply poisoning attacks on **five real-world datasets**—Email (Spam), Banking, Diabetes, Heart Attack, and Iris—to explore how domain-specific characteristics influence the model's vulnerability.
6. **Identify Resilience Patterns:**
   Study how the structure of classifiers and the nature of datasets contribute to resistance or failure under adversarial manipulation, drawing insights for more secure model selection.
7. **Contribute to Adversarial ML Research:**
   Provide an empirical foundation for further research in **adversarial machine learning**, especially in the domain of **training-time attacks**, by offering reproducible experiments and comparative benchmarks.

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. 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

- **Datasets**:

  Sourced from UCI Machine Learning Repository and Kaggle, covering domains like healthcare (Heart Attack, Diabetes), finance (Banking), and text classification (Email Spam).

- **Models Evaluated**:

  Logistic Regression, SVM, Decision Tree, Random Forest.

- **Poisoning Simulation**:

  Labels in the training set are flipped randomly to simulate adversarial attacks. The attack strength varies across poisoning rates: 0%, 5%, 10%, and 20%.

- **Evaluation Metrics**:

  Accuracy, Precision, Recall, and visual accuracy degradation plots.

- **Case Study**:

  A dedicated simulation demonstrates how poisoned training data leads to critical misdiagnoses in a heart attack prediction model.

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:

1. To identify and analyze vulnerabilities in machine learning (ML) models, specifically in the context of training-time attacks.
2. To simulate and study the impact of poisoned or maliciously manipulated data on ML model performance and decision-making.

Technical Requirements:

- A strong foundational understanding of machine learning algorithms and adversarial attack methodologies.
- Proficiency in programming, especially using **Python**.
- Use of classical ML libraries such as **Scikit-learn** for model training, evaluation, and performance analysis.
- Familiarity with data analysis and visualization libraries like **Pandas**, **NumPy**, **Matplotlib**, and **Seaborn**.
- Access to standard computing resources with sufficient RAM and CPU (GPU not required).
- Availability of diverse public datasets, sourced primarily from platforms like **UCI Machine Learning Repository** and **Kaggle**.

Challenges:

- Designing poisoned data samples that are subtle and difficult to detect using standard preprocessing tools.
- Addressing ethical and legal concerns related to the manipulation of training data, including compliance with data privacy regulations and research ethics.
- Implementing robust detection mechanisms for poisoned data within the ML pipeline.
- Keeping pace with evolving machine learning models and continuously updating the attack strategies accordingly.

Expertise Needed:

- A multidisciplinary team with skills in **machine learning**, **data science**, **cybersecurity**, and **adversarial ML**.
- Capabilities in designing, simulating, and evaluating data poisoning strategies.
- Hands-on experience with training and fine-tuning supervised ML models across various domains.
- Understanding of adversarial testing frameworks and performance benchmarking.

Success Factors:

- Adopting a **responsible, ethical, and legally compliant** research approach throughout the project.
- Ensuring continuous vigilance during dataset preparation and model evaluation to detect abnormal behavior.
- Collaborating with domain experts to validate the findings and explore practical implications.
- Implementing effective project planning and maintaining clarity in goals, risks, and evaluation metrics.

Conclusion:

The project is technically and operationally feasible, provided that appropriate expertise and ethical guidelines are followed. With a well-structured methodology, secure tooling, and multidisciplinary collaboration, this study can yield impactful insights into the security and robustness of machine learning systems against poisoning attacks.

FACILITIES REQUIRED

Software and Development Tools:

- **Programming Language**: Python 3.8+
- **ML Libraries**: Scikit-learn for model development, training, and testing
- **Data Processing Tools**: Pandas and NumPy for data cleaning, transformation, and manipulation
- **Visualization Tools**: Matplotlib and Seaborn for graphical representation of results
- **Development Environment**: Jupyter Notebook or Google Colab for interactive experimentation

Data Resources:

- **Datasets Used**:
  1. Email Spam Dataset
  2. Banking Dataset
  3. Diabetes Dataset
  4. Heart Attack Dataset
  5. Iris Dataset

  All datasets were sourced from the **UCI Machine Learning Repository and Kaggle.**
- **Data Storage**:
  Local system or cloud-based storage to securely manage datasets and store multiple training/test splits and poisoned variants.

EXPECTED OUTCOME

- Identification of vulnerable models under label poisoning.
- Establishment of Random Forest as the most robust classifier due to its ensemble structure.
- Exposure of Decision Tree’s fragility due to overfitting poisoned labels.
- Insightful performance trends across varying datasets and poisoning rates.
- Real-world illustration of the ethical risks posed by poisoned models in healthcare.

REFERENCES

[1] Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning(2018)-Ibrahim M. Ahmed(University of Mosul),Manar Kashmola (Ninevah University).

[2] 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‡.