ENHANCING SECURITY AND PRIVACY OF MACHINE LEARNING MODELS IN HEALTHCARE

Professor:-Rasika Mundhe

ABOUT US

ROSHNI PANDA

A009

JINAV GALA

A023

Machine learning models in healthcare improve diagnostic accuracy but are vulnerable to security threats like adversarial attacks, model inversion, and membership inference, risking misdiagnoses and privacy breaches. So there are models to protect by implementing strategies such as adversarial training, defensive distillation, and differential privacy, alongside strict access controls and monitoring, to ensure their integrity and confidentiality.

OBJECTIVE

The objective of this project is to demonstrate the vulnerabilities of machine learning (ML) models in healthcare applications, specifically focusing on the security threats posed by adversarial attacks such as the Fast Gradient Sign Method (FGSM).

1

Showcase the severity of the ML Model vulnerabilities, the threats and possible defenses for such attacks in real-world situations

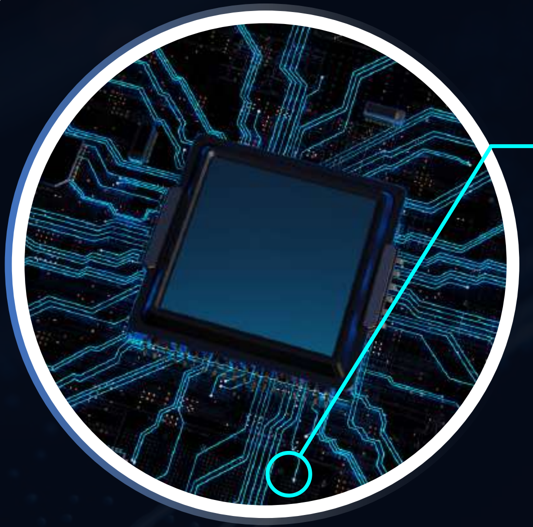
2

Explore a case study specific to Healthcare domain where the model vulnerabilities may lead to life-threatening situation due to misdiagnosis

OBJECTIVE

The goal is to not only demonstrate the risks associated with adversarial attacks but also to implement practical defences that can safeguard ML models in critical applications like medical diagnosis.

3



Implement security measures such as adversarial training and other defenses to strengthen the model's robustness against adversarial attacks, thereby demonstrating how these models can be protected in sensitive domains like healthcare..

4

Provide a comparative analysis of model performance before and after adversarial attacks and post-defense implementations, with specific emphasis on accuracy and classification reports.

LITERATURE REVIEW

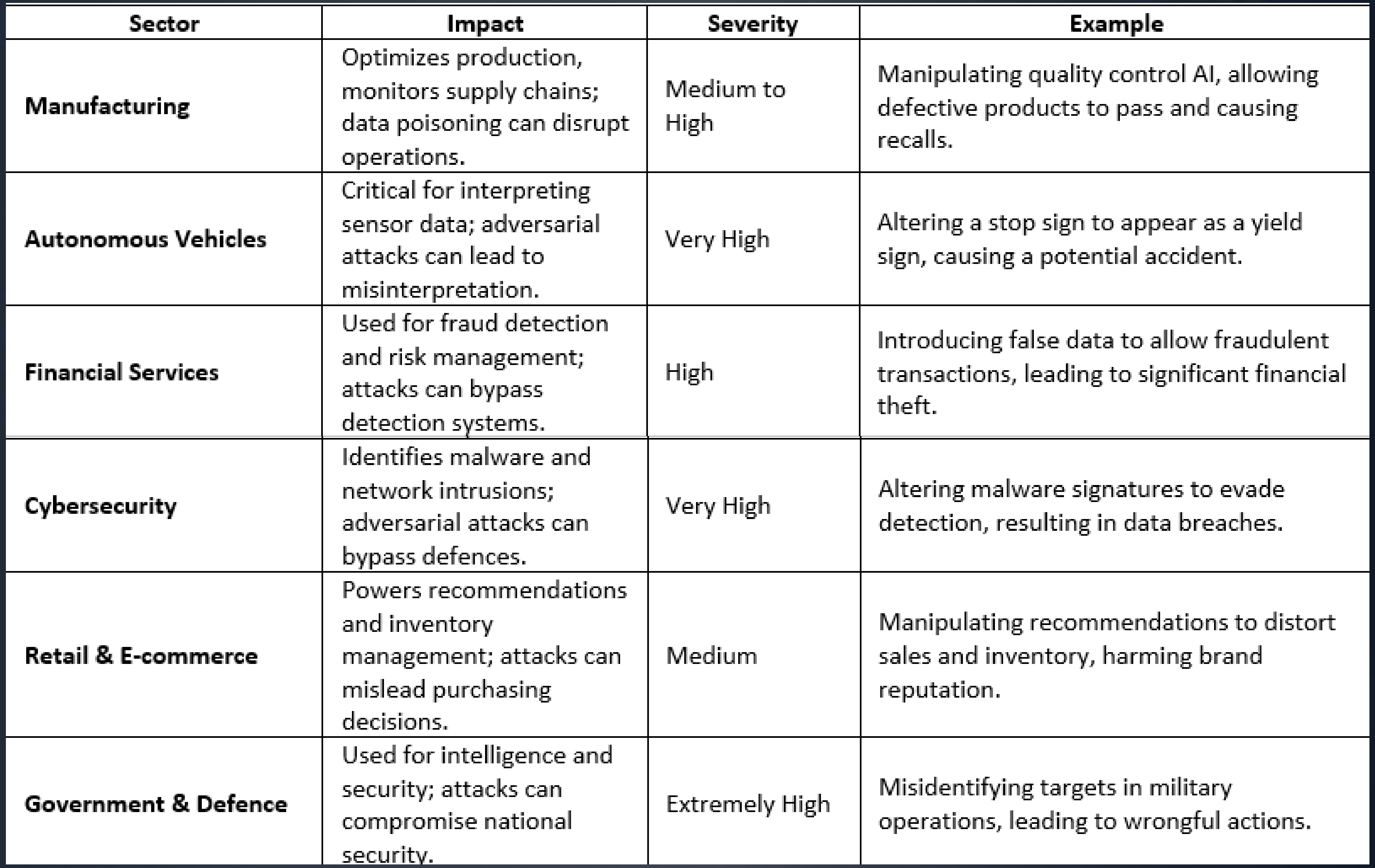
Imam, R., Almakky, I., Alrashdi, S., Alrashdi, B., Yaqub, M. (2024). SEDA: Self-ensembling ViT with Defensive Distillation and Adversarial Training for Robust Chest X-Rays Classification. In: Koch, L., et al. Domain Adaptation and Representation Transfer. DART 2023. Lecture Notes in Computer Science, vol 14293. Springer, Cham.

- The paper proposes SEDA (Self-Ensembling ViT with Defensive Distillation and Adversarial Training), a new approach to improve the robustness of Vision Transformers (ViTs) for medical image classification, specifically chest X-ray tuberculosis detection.
- SEDA presents a novel approach to enhancing the robustness and efficiency of Vision Transformers for medical image classification, addressing key vulnerabilities in existing models while maintaining high accuracy.

de Aguiar EJ, Traina C Jr, Traina AJM. Security and Privacy in Machine Learning for Health Systems: Strategies and Challenges. Yearb Med Inform. 2023 Aug;32(1):269-281. doi: 10.1055/s-0043-1768731. Epub 2023 Dec 26. PMID: 38147869; PMCID: PMC10751106.

- The paper reviews security and privacy challenges in machine learning for healthcare, identifying prevalent attack methods like adversarial examples and DeepFakes.
- It discusses effective defense and privacy-preserving strategies, such as adversarial training, federated learning, and differential privacy, highlighting the need for robust solutions to protect sensitive medical data.

Domains where ML model is used & Security is Essential



Sector	Impact	Severity	Example
Manufacturing	Optimizes production, monitors supply chains; data poisoning can disrupt operations.	Medium to High	Manipulating quality control AI, allowing defective products to pass and causing recalls.
Autonomous Vehicles	Critical for interpreting sensor data; adversarial attacks can lead to misinterpretation.	Very High	Altering a stop sign to appear as a yield sign, causing a potential accident.
Financial Services	Used for fraud detection and risk management; attacks can bypass detection systems.	High	Introducing false data to allow fraudulent transactions, leading to significant financial theft.
Cybersecurity	Identifies malware and network intrusions; adversarial attacks can bypass defences.	Very High	Altering malware signatures to evade detection, resulting in data breaches.
Retail & E-commerce	Powers recommendations and inventory management; attacks can mislead purchasing decisions.	Medium	Manipulating recommendations to distort sales and inventory, harming brand reputation.
Government & Defence	Used for intelligence and security; attacks can compromise national security.	Extremely High	Misidentifying targets in military operations, leading to wrongful actions.

VULNERABILITIES IN DEEPLEARNING MODEL

- Black Box Vulnerability: Deep learning models, though accurate, are often opaque in their decision-making, making them susceptible to exploitation by adversaries.

- Adversarial Exploits: Attackers can manipulate the model's inputs or structure, leading to inaccurate predictions and misclassifications, especially in critical fields like healthcare.

- Data Privacy Concerns: Deep learning relies on large datasets, often sensitive (e.g., patient data), raising concerns about privacy breaches, data leakage, and identity theft.

- Need for Security: Securing deep learning models is crucial, with research focusing on mitigating risks to ensure safe AI deployment in high-stakes applications.

- Practical Implementation: We demonstrate vulnerabilities in a healthcare deep learning model and provide secure implementations to address the identified risks.

OVERVIEW OF HEALTH-CARE DOMAIN

Machine learning (ML) models in healthcare are transforming patient care, diagnostics, and operational efficiency.

- Diagnostics: ML analyzes medical images (X-rays, MRIs, CT scans) to detect diseases (e.g., cancer, heart disease) earlier and more accurately than traditional methods.
- Personalized Medicine: ML tailors treatments by analyzing patient data and genetics, predicting medication responses for more effective therapies.
- Predictive Analytics: ML forecasts disease outbreaks, readmissions, and complications, enabling preventive care and better resource planning.
- Operational Efficiency: ML streamlines hospital workflows, manages patient flows, and predicts resource needs (staff, equipment), reducing costs and improving care.
- Wearable Monitoring: ML analyzes real-time data from devices (heart rate, glucose) to continuously monitor health and alert doctors to potential issues early.

TRADITIONAL HEALTH CARE SYSTEM

Manual Diagnostics:

- Doctors manually interpreted images and tests, often leading to errors.

Generalized Treatment:

- Treatments were based on broad guidelines, not personalized.

Reactive Care:

- Focused on treating diseases after symptoms appeared, rather than prevention.

Limited Data Use:

- Patient data was underutilized, stored in silos, and used only during visits.

NEED FOR SECURITY TECHNIQUES

- **Patient Safety:** Inaccurate predictions or diagnoses due to compromised models can lead to harmful medical decisions, endangering patient health..
- **Model Integrity:** Ensuring the reliability and accuracy of ML models is vital, as tampered models can result in misdiagnosis and inappropriate treatments.
- **Trust and Adoption:** Securing ML models fosters trust among patients and healthcare providers, encouraging the adoption of AI technologies in clinical settings.
- **Data Integrity:** Secured ML models ensure that the data used for training and inference is authentic and unaltered, leading to more accurate and reliable outcomes.

POSSIBLE ATTACKS

- Adversarial Attacks:

Attackers intentionally introduce malicious data inputs that can manipulate model predictions, leading to misdiagnoses or incorrect treatment recommendations.

- Model Inversion Attacks:

Attackers use access to a model's outputs to reverse-engineer sensitive patient information, potentially reconstructing private data, such as medical records or diagnosis details.

- Membership Inference Attacks:

Adversaries determine whether a specific patient’s data was part of the training set, exposing the inclusion of sensitive information and breaching patient confidentiality.

- Data Poisoning:

Attackers tamper with training data to corrupt the learning process, causing the model to make biased or inaccurate predictions.

- Model Theft:

Cybercriminals may attempt to steal trained models, enabling them to replicate or misuse proprietary healthcare AI systems for unethical purposes.

- Privacy Violations:

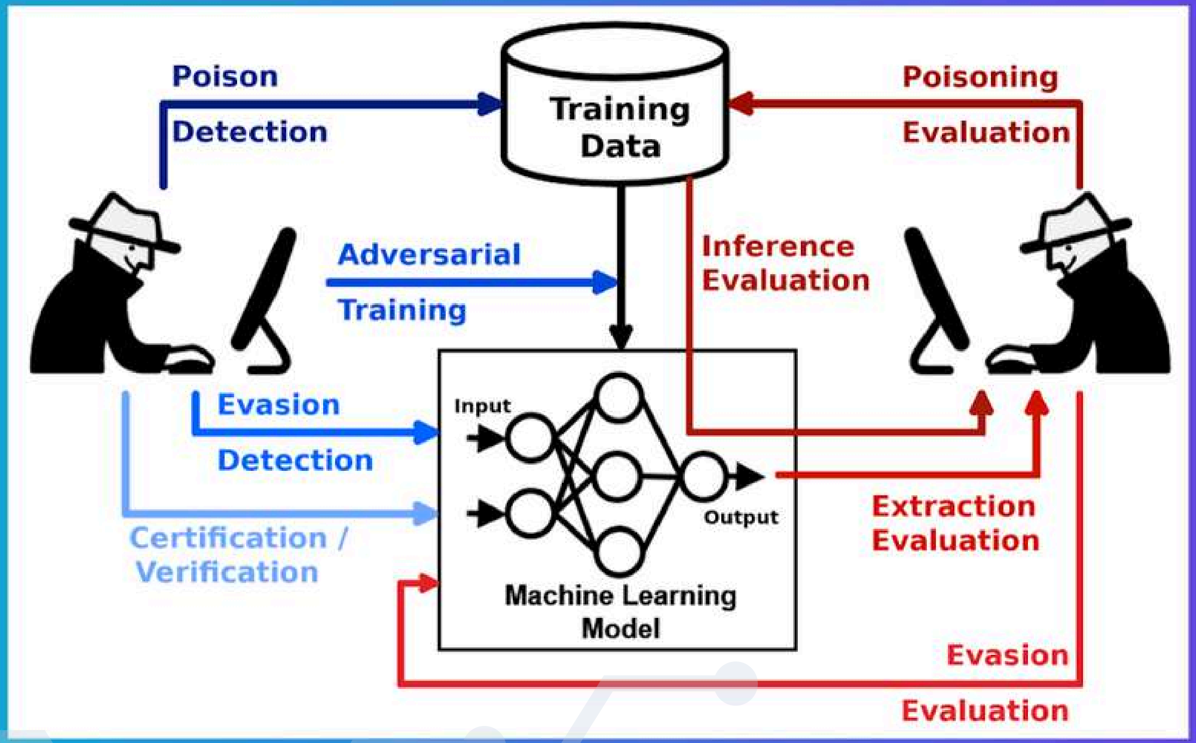
Sensitive patient data used in training can be exposed through model outputs or through attacks on the model's structure, leading to regulatory breaches and loss of trust.

ADVERSARIAL ATTACK

- Adversarial attacks are techniques where small, intentionally crafted perturbations are applied to input data to deceive machine learning models, leading to incorrect predictions. These perturbations are often imperceptible to humans but can significantly alter the model's output.

- In sensitive applications like healthcare, adversarial attacks pose serious threats. For example, a slight modification to a chest X-ray could misdiagnose pneumonia, potentially harming patients. These attacks exploit model vulnerabilities, revealing a lack of robustness in real-world scenarios.

- A common type of adversarial attack is the Fast Gradient Sign Method (FGSM), which adds small perturbations to input data in the direction of the gradient of the model's loss function, maximizing the chance of misclassification.



Case Study: Securing Machine Learning Models in Healthcare

MEDSECURE DIAGNOSTICS

MedSecure™
Pharmaceuticals (P) Ltd.

Company Profile:

- Name: MedSecure Diagnostics
- Industry: Healthcare, Medical Imaging
- Core Product: AI-powered diagnostic tool for early detection of diseases using medical images such as X-rays, MRIs, and CT scans.
- ML Models: Convolutional Neural Networks (CNNs) for image classification, trained to detect anomalies like tumors, fractures, and infections.
- Client Base: Hospitals, clinics, and private practices globally.
- Data Sources: Large datasets of annotated medical images sourced from hospitals and research institutions.

Mission:MedSecure Diagnostics aims to revolutionize healthcare by providing accurate and efficient AI-powered diagnostic tools, reducing the workload on healthcare professionals, and enabling early detection of life-threatening conditions.

THE ATTACK SCENARIO:

Fictional Attack Overview:

- Attacker: A disgruntled former employee with deep knowledge of MedSecure’s AI systems.

- Attack Type: Fast Gradient Sign Method (FGSM) adversarial attack.
- Target: MedSecure’s flagship AI model for detecting malignant tumours in MRI scans.
- Motivation: The attacker intends to undermine the company's reputation by causing the AI system to make incorrect diagnoses, leading to misdiagnosis in critical cases.

Attack Execution:

- The attacker gains access to MedSecure’s cloud-based model deployment environment.
- They implement an FGSM attack by slightly perturbing incoming MRI scans using adversarial noise, which is imperceptible to the human eye but sufficient to cause the AI model to misclassify malignant tumours as benign.
- The perturbations are designed to subtly degrade the model's accuracy, leading to an increased rate of false negatives (i.e., failing to identify tumours).

Consequences of the Attack:

Patient Harm:

- Missed Diagnoses: Tumors misdiagnosed, delaying critical treatment.
- Severe Outcomes: Delays risk irreversible harm or death.

Legal and Financial Repercussions:

- Lawsuits: Patients sue for negligence and compensation.
- Financial Loss: Stock drops, leading to significant losses.

Reputation Damage:

- Loss of Trust: Reputation suffers, eroding trust among providers.
- Regulatory Scrutiny: Investigations lead to stricter regulations.

THREATS OF FGSM ATTACK IN HEALTHCARE

Manipulating Medical Diagnoses:

- Machine learning models in healthcare diagnose diseases from medical images (e.g., X-rays, MRIs). An FGSM attack can subtly alter input data, leading to misclassification without detection.
- Example:
- A slight change to a chest X-ray could misclassify a healthy patient as having pneumonia.
- Consequences:
- Wrong treatments: Unnecessary treatments for healthy patients.
- Undetected illness: Serious conditions like pneumonia may go undiagnosed, delaying critical care.

Undermining Automated Systems:

- Many healthcare facilities rely on AI for tasks like triaging patients, identifying high-risk cases, or recommending treatments. If an attacker applies FGSM to alter inputs, the AI system could misprioritize cases, leading to life-threatening delays.
- Example: In an emergency department, AI might help prioritize critical patients. An attacker could use FGSM to downgrade the urgency of a patient with a life-threatening condition, causing delays in treatment.

THREATS OF FGSM ATTACK IN HEALTHCARE

Privacy and Data Breaches:

- FGSM attacks are not limited to direct misclassification but could also be leveraged to expose sensitive patient data. If models have been trained on patient data, an attacker might reverse-engineer aspects of that data by introducing adversarial perturbations.
- Example: By probing a machine learning model with carefully crafted inputs, an attacker could infer private information, such as whether a particular patient has a specific condition.

Public Safety and Trust:

- As AI becomes a critical part of public health systems, the ability of attackers to manipulate diagnostic models could erode trust in these systems. If people believe AI systems are vulnerable to manipulation, they may lose faith in the technology, even when it works correctly.
- Example: If a series of misdiagnoses occurs due to adversarial attacks, public trust in AI-driven healthcare systems could plummet, setting back adoption of potentially life-saving technologies.

WHAT CAN SOMEONE DO WITH FGSM ATTACK?

- Mislead diagnostic systems: Subtly alter medical data or images to mislead AI models into providing incorrect diagnoses.
- Disrupt operations: Tamper with AI systems that control patient flows, triaging, or scheduling, causing operational havoc in hospitals.
- Commit fraud: Manipulate health data in insurance systems to commit fraud, or bypass security systems designed to detect anomalies.
- Conduct espionage or sabotage: Use adversarial attacks to extract private health data or sabotage medical AI systems for political or financial gain.

HOW TO MITIGATE FGSM ATTACKS IN THE REAL WORLD?

1. Adversarial Training: Including adversarial examples during the model's training phase makes the model more robust against attacks like FGSM.
2. Defensive Distillation: A method where the model is trained to reduce its sensitivity to small perturbations in the input.
3. Input Sanitization: Pre-processing inputs to detect and filter adversarial examples before passing them to the model.
4. Model Monitoring: Real-time monitoring of model performance to detect suspicious activity, such as a sudden increase in misclassifications, which could indicate an attack.
5. Use of More Robust Algorithms: Models like Bayesian networks or models with uncertainty quantification can be less susceptible to adversarial examples, as they rely on probabilistic reasoning.

DEFENSE AND MITIGATION STRATEGIES:

1. Immediate Response:

- System Shutdown: IT shuts down affected AI systems to prevent harm.
- Incident Investigation: An investigation identifies the breach and its impact.

2. Strengthening Model Security:

- Adversarial Training: Models are retrained with adversarial examples for robustness.
- Implementation: Diverse attacks are used during training for better resilience.
- Defensive Distillation: A "teacher" model trains a more robust "student" model to resist misclassification.
- Differential Privacy: Protects patient data from reverse-engineering.

IMPLEMENTATION

Below attach is the implementation link:

For the implementation, we use the diabetes dataset. In machine learning, it is often based on the Pima Indians Diabetes Database. This dataset contains various medical predictor variables and one target variable indicating whether a patient has diabetes. It is widely used for binary classification tasks in healthcare to develop predictive models for diabetes diagnosis.

Google Colab

google.com

CONCLUSION

This case study highlights the critical importance of securing machine learning models in healthcare. The fictional attack on MedSecure Diagnostics demonstrates how adversarial attacks can have devastating consequences, not only for the company but also for patient safety. By implementing a multi-faceted defense strategy, including adversarial training, defensive distillation, and robust monitoring systems, healthcare AI systems can be better protected against such threats. MedSecure's experience underscores the need for ongoing vigilance, regular audits, and the integration of the latest security measures to safeguard AI models in the healthcare domain.

THANKYOU