WHITEPAPER

The Dark Art of Machine Learning:

Vulnerabilites, Attacks, and Defenses

Oshan Jayawardena

Fcode Labs

Introduction

Background and Importance of Machine Learning Security

Deep learning is rapidly transforming industries, and digital startups are at the forefront of this innovation. From AI-powered recommendations to sophisticated chatbots, machine learning is becoming integral to many startup products and services. However, while startups prioritize traditional software security, the unique vulnerabilities of machine learning models often overlooked. This oversight poses a significant business risk. These are not hypothetical scenarios; they are real threats that can damage your reputation, erode customer trust, and lead to significant financial losses.

Understanding Machine Learning Vulnerabilities

Unlike traditional software vulnerabilities, machine learning vulnerabilities stem from the model's reliance on data and its inherent complexity. Key vulnerabilities include:

Data Dependence and Leakage: A model's performance hinges on its training data, yet the model itself can inadvertently leak this sensitive data during inference, compromising confidentiality.

Model Replication (Surrogate Models): The difficulty in encrypting the relationship between inputs and outputs makes it possible for malicious actors to create copies of your models.

Data Poisoning: If attackers gain access to your training data, they can manipulate it to compromise the model's integrity and produce biased or incorrect outputs.

Verification Challenges: The inherent complexity of deep learning makes it impossible to fully verify model performance across all possible inputs, creating opportunities for exploitation.

Membership Inference Attacks

When developing proprietary models for your startup, you may rely on datasets created using user data collected under specific terms and conditions. If the data is collected on the condition that it can only be used for internal purposes, it is essential to handle it with the utmost care to ensure compliance with those terms. Using such data for training deep learning models can result in a violation of the specified terms and conditions.

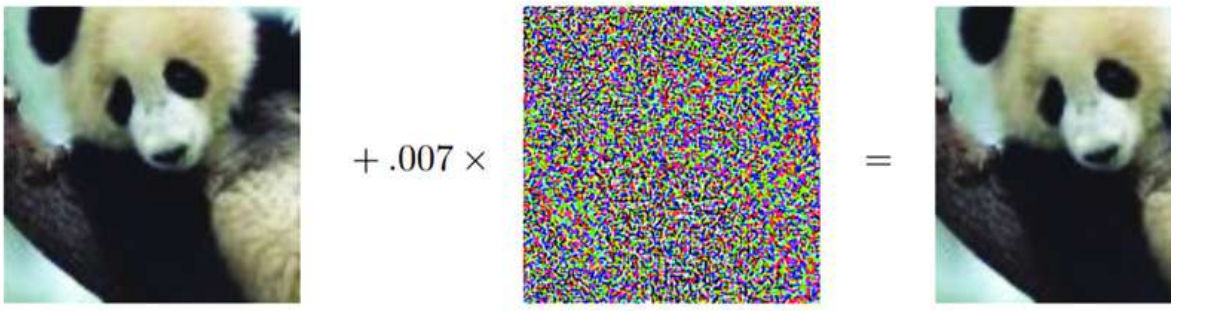
- An LLM fine-tuned on user-data could memorize these data and leak them during inference. It could auto-fill contact information and addresses of a user.

- A medical classification model could give high confidence values for memorized training samples, creating an opportunity for malicious parties to infer members who participated in the medical survey. Such incidents not only compromise the privacy of data owners but also risk damaging the reputation and trustworthiness of your startup.

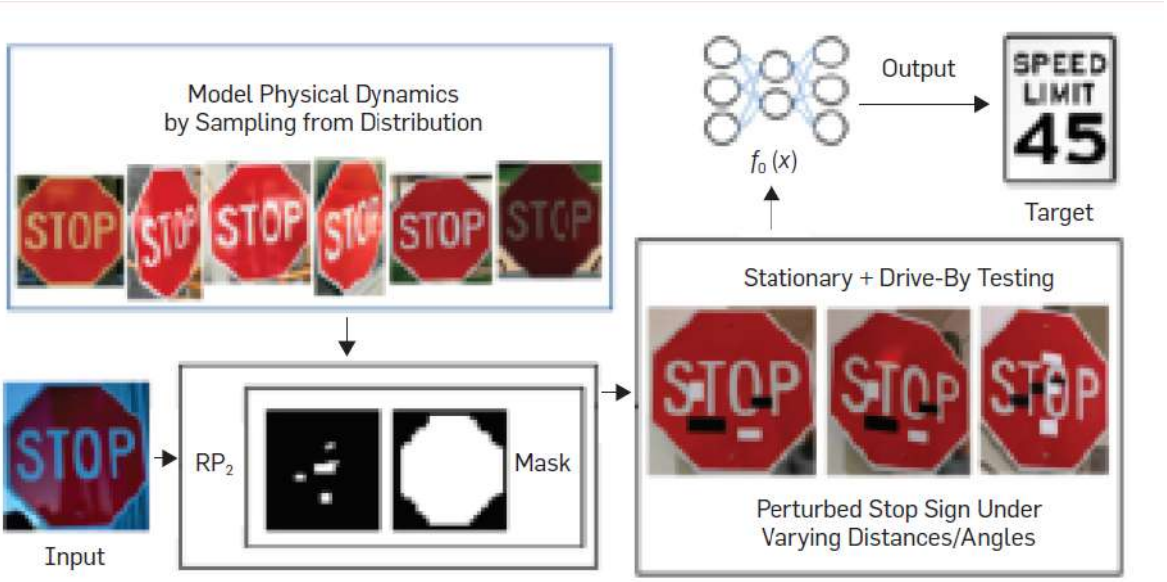
Adversarial Attacks

Definition and Examples of Adversarial Attacks

• It is hard to observe which features of the input are being used to make the decision of a deep learning model. Malicious parties take advantage of this vulnerability and add perturbations to input features that are being used by the model for decision-making but cannot be observed by a human. However, these perturbations change the output of the model. Following is a popular example



• A human eye can't distinguish between the two images, but the model has changed its output from panda to gibbon. But the above attack must be done on device.
• Below is a physical world attack on self-driving vehicles.



• A simple set of stickers placed strategically on a stop sign can mislead a street sign detection model into misclassifying it as a speed limit sign. If your startup is developing models for self-driving cars, it is important to ensure that your models are robust against such adversarial attacks to maintain safety and reliability.
• This type of vulnerability is not limited to vision models; it can also affect speech recognition models. Attackers can use inaudible signals—often imperceptible to human ears—to give malicious commands to models.



• A research team discovered a way to issue voice commands to smart devices using sound frequencies that are inaudible to humans but can still be detected by device microphones. This technique, known as DolphinAttack, involves modulating voice commands into ultrasonic frequencies—beyond human hearing range—while still being picked up and processed by virtual assistants like Siri, Google Now, Cortana, and Alexa. As a result, attackers could secretly control various devices, including smartphones, tablets, smart speakers, and even cars like the Audi Q3, without the user noticing.
• If your startup is integrating voice command services to your products, you must be aware of these vulnerabilities.

Detection and Mitigation Strategies

• It is extremely hard to come up with a global solution for detecting adversarial attacks (believe me, I have tried)
• What you can do is make your model prevent extreme pertubations.

Data Poisoning

Explanation of Data Poisoning

• If the training set of a model is the entire internet, then a malicious party can poison the internet to make the model malfunction. This is the case for most of the large language models as they are trained by scraping the entire internet. A malicious party can bias the LLM to discriminate against a gender, a race, an ethnic group, a religious group or etc. Or the malicious party can setup back doors that get triggered when a specific prompt is given. A LLM with a back door can give malicious code suggestions with security vulnerabilities.
• It is tempting to use third party free versions of GPT or fine-tuned specialized versions as a startup running on a tight budget. But these third-party versions could be fine-tuned with poisoned data. They could have back doors hidden, that can destroy the reputation of your startup.
• Similarly, if your startup is developing a healthcare product for medical diagnosis, a model trained on poisoned data could result in critical errors—such as diagnosing a healthy person with dengue and prescribing unnecessary treatments.
• In the financial sector, if your startup employs a trading bot trained on poisoned datasets, it could make flawed predictions or decisions, leading to significant financial losses.

Prevention Strategies

• Do not dump all the data you find on the internet into your model. First, clean up and preprocess it.
• Always check the data source.
• Look for outliers in the dataset. The malicious party might be trying to reduce the performance of the model.
• Do cross-validation to check if the model is overfitted to outliers.

Notable Incidents

• Tay: Microsoft Chatbot Tragedy

In the first 24 hours of coming online, a coordinated attack by a subset of people exploited a vulnerability in Tay. Although we had prepared for many types of abuses of the system, we had made a critical oversight for this specific attack. As a result, Tay tweeted wildly inappropriate and reprehensible words and images.

~ Peter Lee - Corporate Vice President, Microsoft Healthcare

Enhancing Model Robustness and Security

Differential Privacy

"Differential Privacy (DP) is one of the most widely accepted technologies that allows reasoning about data anonymization in a formal way. In the context of an ML model, DP can guarantee that each individual user's contribution will not result in a significantly different model." - Natalia Ponomareva and Alex Kurakin, Staff Software Engineers, Google Research

• A differentially private model ensures that the inclusion or exclusion of any single user's data does not significantly alter measurable outputs such as logits, loss, or gradients. While differential privacy reduces model accuracy and can lead to performance inconsistencies, it effectively protects against attacks like membership inference and data reconstruction by limiting the influence of individual data points.
•By implementing differential privacy, you can guarantee the confidentiality of user data used in model training. For startups, incorporating differential privacy into your model development pipeline can significantly enhance the trustworthiness of your product, positioning it as one of the most secure and reliable solutions available.

Federated Learning

• Federated Learning is in simple terms, decentralized model training. Privacy critical data, such as medical records of patients, most of the time can't leave local servers at the medical facility. Data will be stored in multiple machines spread around the globe. Federated learning enables us to train a model on all this data, without the data ever leaving the local machines.
• If you are startup building machine learning models for healthcare services, this would be the most trustworthy and secure method for model training.

Model Encryption

• Model encryption can be used to ensure privacy by securely encrypting the machine learning models, preventing unauthorized access to sensitive data and safeguarding intellectual property during deployment and inference.
• Encryption methods are computationally expensive, but they make sure that our models can't be copied or stolen.

Licensing

• Licensing helps define who is legally allowed to use your models, but it does not provide technical protection or enhance security.
• In case of violation or breach, you will be able to take legal action.
• Your startup will have the ability to monitor and control usage. This can help us block malicious parties from abusing the model for attacks and model stealing.

Ethical and Legal Considerations

If your startup plans to integrate machine learning models into your production environment, it is crucial to evaluate these models for potential ethical violations. Below is a list of ethical violations and examples that may arise from using machine learning models:

Privacy Violations

LLMs could reveal personal data stored in the training dataset. A medical diagnosis model output could reveal enough information to infer whether someone has provided data to train the model. Which could compromise the privacy of data contributors.

Bias and Discrimination

If someone has the ability to manipulate the training set of a model, they have the ability to bias model outputs. The Microsoft Tay chatbot incident is a perfect example. Within 24 hours after deployment, it started tweeting racist and extremist tweets. Since Tay was actively learning throughout time, a group was able to poison the training data with racist tweets.

Security Vulnerabilities

Adversarial attacks could compromise the security of machine learning models, causing unexpected behaviors. Adversarial attacks can cause malfunctions in AI voice assistants, autonomous driving systems, financial models and medical diagnosis models.

Lack of Accountability and Transparency

In a situation where the machine learning model malfunctions, the accountable party of the action can be ambiguous. Since most models are black box models, the reason for a particular decision or an output cannot be inferred. The fault could be in a compromised dataset, a backdoor placed during training, could be caused by an adversarial attack or caused by the inaccuracy of the model.

Regulatory Frameworks for AI Security

To ensure your startup complies with legal regulations for AI, consider the following regulatory frameworks. These frameworks emphasize safeguarding user data used in training and inference, ensuring security, privacy, and ethical AI practices.

General Data Protection Regulation (GDPR) – European Union

GDPR make sure the following 7 principles are followed (effective on any organization in the world that process data of EU citizens)

- Lawfulness, fairness and transparency
- Purpose limitation – You can only use the data for the purpose that you specify to the data provider when you collected the data.
- Data minimization – You can only collect and process only as much data as absolutely necessary for the purposes specified.
- Accuracy – Personal data must be accurate and updated frequently.
- Storage limitation - You may only store personally identifying data for as long as necessary for the specified purpose.
- Integrity and confidentiality - Processing must be done in such a way as to ensure appropriate security, integrity, and confidentiality (e.g. encryption, differential privacy, federated learning).
- Accountability

EU AI Act – European Union (effective on any organization in the world that process data of EU citizens)

The EU AI Act provides clear requirements and obligations that must be followed based on the risk level of AI systems.

- Unacceptable risks are prohibited (e.g. social scoring systems and manipulative AI).
- High-risk AI systems, which are regulated. (e.g. medical software or AI systems used for recruitment)
- Limited risk AI systems, subject to lighter transparency obligations: developers and deployers must ensure that end-users are aware that they are interacting with AI (chatbots and deep fakes).
- Minimal risk is unregulated (including the majority of AI applications currently available on the EU single market, such as AI enabled video games and spam filters – at least in 2021; this is changing with generative AI).

Artificial Intelligence Risk Management Framework - National Institute of Standards and Technology (USA)

The framework specifies characteristics that a trustworthy AI system should have and offers guidance for addressing them.

- Valid and reliable
- Safe
- Secure and resilient
- Accountable and transparent
- Explainable and interpretable
- Privacy-enhanced
- Fair with harmful bias managed

Future Directions in Machine Learning Security

If you want to keep your machine learning security measures up to date, make sure to stay updated on the following areas.

Differentially Privacy

Differential privacy ensures the privacy of individuals in a dataset while still allowing meaningful information to be extracted. It provides mathematical guarantees that the removal or addition of a single individual's data in a dataset will not significantly affect the outcome of any query made on the database, thus protecting individual data from inference. Differential privacy introduces a controlled amount of randomness (noise) into the analysis or the data itself. This ensures that no single individual's data can be reverse-engineered or identified.

Federated Learning

Federated Learning (FL) is a decentralized approach to training machine learning models where data remains on local devices (e.g., smartphones, IoT devices, or local servers), and only model updates (e.g., gradients or weights) are shared with a central server.

Memorization Measurements

Memorization in Large Language Models (LLMs) refers to the phenomenon where a model retains specific pieces of data from its training set in a way that allows these exact details to be reproduced. Memorized sensitive data (e.g., personal information, private conversations, or proprietary code) can be unintentionally exposed if the model is prompted in specific ways. New research enables us to measure how much training data is remembered by LLMs.

Importance of Academia-Industry Collaborations

• Even though there has been a lot of research in trustworthy machine learning in the past few years, industry hasn't fully adapted to these practices.
• "I Never Thought About Securing My Machine Learning Systems": A Study of Security and Privacy Awareness of Machine Learning Practitioners – This research paper perfectly highlights the current state of trustworthy machine learning practices.
• Now we are at the point where machine learning is being used for decision making and diagnosis.
• Staying up to date with the latest machine learning security research is essential to identify vulnerabilities and to enforce defenses.

Conclusion

AI Security Deserves Equal Attention as Software Security

In today's technology-driven landscape, artificial intelligence (AI) security requires the same level of attention and rigor that we devote to software security. Yet, the current industry often overlooks the vulnerabilities inherent to AI systems, thereby posing significant risks to:

- User Privacy: Potential exposure of sensitive user data.
- Database Security: Breaches or unauthorized access due to poorly secured pipelines.
- Model Security: Exploits such as adversarial attacks, data poisoning, and model theft

Awareness of Industry Vulnerabilities

Many organizations remain unaware of the critical threats that could compromise their AI systems. This lack of awareness leaves gaps in security, risking:

- Data leaks through inference or reconstruction attacks.
- Misuse of machine learning models due to adversarial inputs.
- Manipulation of decision-making models in sensitive areas like healthcare, finance, or autonomous systems.

Existing Regulations and Frameworks

Fortunately, governments and organizations have started taking proactive steps to regulate AI use through frameworks, acts, and standards, such as:

- GDPR: Ensuring data privacy and secure processing.
- EU AI Act: Regulating AI systems based on their risk levels.
- NIST AI Risk Management Framework: Offering guidelines for building trustworthy AI.

The Time to Act Is Now

As AI becomes increasingly integral to decision-making and automation, now is the best time to implement robust practices and defenses for your machine learning pipelines. Consider adopting:

- Secure Development Practices: Integrate security into every stage of the AI lifecycle.
- Privacy Enhancements: Use techniques like differential privacy and federated learning to safeguard user data.
- Threat Detection and Response: Continuously monitor for and address potential vulnerabilities in your models and infrastructure.
- Regulatory Compliance: Ensure adherence to global standards and legal requirements for AI use.



About

Oshan Jayawardena

Machine Learning Engineer at Fcode Labs

Oshan Jayawardena is an AI Engineer at Fcode Labs, a software engineering company dedicated to transforming the world through innovative digitals solutions. With a strong foundation in electronic and telecommunication engineering, Oshan brings a deep technical understanding to the field of AI, driving advancements in machine learning and data science.

He holds a Bachelor of Science (Honours) in Electronic and Telecommunication Engineering from the University of Moratuwa. During his academic journey, he actively contributed to research as a Research Affiliate at the University of Sydney. His passion for technology and leadership was further reflected in his role as Chairperson of the IEEE Communications Society Student Branch Chapter at the University of Moratuwa, where he led initiatives to foster collaboration and knowledge-sharing in the field.

At Fcode Labs, Oshan is at the forefront of AI-driven innovation, working on intelligent solutions that enhance business efficiency and transform industries. His commitment to pushing technological boundaries ensures that he remains a key player in shaping the future of AI and its real-world applications.

Fcode Labs

About the company

Fcode Labs is a remote offshore software development company which is founded by three of the most brilliant young engineers from Sri Lanka. Our mission is to produce technology that meets global standards to transform enterprises while showcasing the high quality and expertise of the Sri Lankan software engineering industry in creating world-class digital products and services.

Our team of over 80 engineers offers services in a wide range of areas including product development, UX prototyping, artificial intelligence, machine learning, data science, and cloud computing. We not only deliver technology expertise but also thought leaders who can develop products along with our clients. For more information, please visit our website: www.fcodelabs.com

Contact Us Today!

hello@fcodelabs.com
tharindu@fcodelabs.com

Fcode Labs Sri Lanka

No 208, Stanley Tillekeratne
Mawatha, Nugegoda 10250,
Sri Lanka

+94 (11) 429 429 9
+94 (77) 368 3072
+94 (76) 121 8800
hello@fcodelabs.com

Fcode Labs Singapore

68 Circular Road,
#02-01,
Singapore 049422

+65 9891 1613
+65 8372 6024
hello@fcodelabs.sg