CS3473 Project Proposal

Taryn Cail

Winter 2025

Professor Jack Van der Laan

Event:

The event that I will be looking at is how Reversing Labs discovered two malicious ML models on Hugging Face, in February 2025 that exploited vulnerabilities in the python pickle file.

Summary:

Hugging Face is an open-source platform for machine learning (ML) and data science. The company centers around community learning and sharing of AI models and knowledge to build, train and deploy various ML models. The malicious ML models that had been compromised was through Pickle file serialization.

The exploit took place because pickle is a common python module for storing and sharing ML model data because of its easy-to-use serialization and deserialization. This exploit happened because the malicious ML models found were stored in a PyTorch format, which is a compressed Pickle file, but they were compressed using the 7z format instead of the typical zip format. This meant that the function *torch.load()* could not be used resulting in Hugging Faces security system, Picklescan, to not flag the models as unsafe.

Thus, the models were able to bypass the Picklescan security system and breach the website. Once on the website they could be deserialized and inject malicious code into whoever used them. Hugging Face took down the models within 24 hours of notice and updated their Picklescan to flag these kinds of files.

Methodology:

The plan is to begin with discussing the threat itself, what happened, when, who found it and what it is. Delving into specific details such as how the threat works, and information surrounding the event and its process.

Then the report will move on to discuss pickle files in general, what they are, what they do, how to use them and why they are so common and popular among ML model users. This part will also lead into known pickle vulnerabilities, analyzing how and why they are so easily threatened and various past incidents.

The last topic to wrap up the report will be discussing AI as a Service, which is a new business model that is growing in popularity and what that means for the future of cybersecurity.

Some sources I will use:

About the Company:

https://huggingface.co/

https://github.com/huggingface

News Coverage on the incident:

https://www.infosecurity-magazine.com/news/malicious-ai-models-hugging-face/

https://www.reversinglabs.com/blog/rl-identifies-malware-ml-model-hosted-on-hugging-face

https://thehackernews.com/2025/02/malicious-ml-models-found-on-hugging.html

The fix for the incident:

https://github.com/mmaitre314/picklescan/pull/33

Information about Pickle files in python:

https://pythonnumericalmethods.studentorg.berkeley.edu/notebooks/chapter11.03-Pickle-Files.html

https://www.blackduck.com/blog/python-pickling.html#:~:text=process%20of%20unpickling.-,Dangers%20of%20Python%20pickling,data%20received%20over%20the%20network.

https://arjancodes.com/blog/python-pickle-module-security-risks-and-safer-alternatives/

https://github.com/mmaitre314/picklescan/tree/main

Other similar pickle attacks:

https://thehackernews.com/2024/06/new-attack-technique-sleepy-pickle.html

https://www.darkreading.com/cloud-security/critical-bugs-hugging-face-ai-platform-pickle

https://checkmarx.com/blog/free-hugs-what-to-be-wary-of-in-hugging-face-part-4/

https://thehackernews.com/2024/02/new-hugging-face-vulnerability-exposes.html