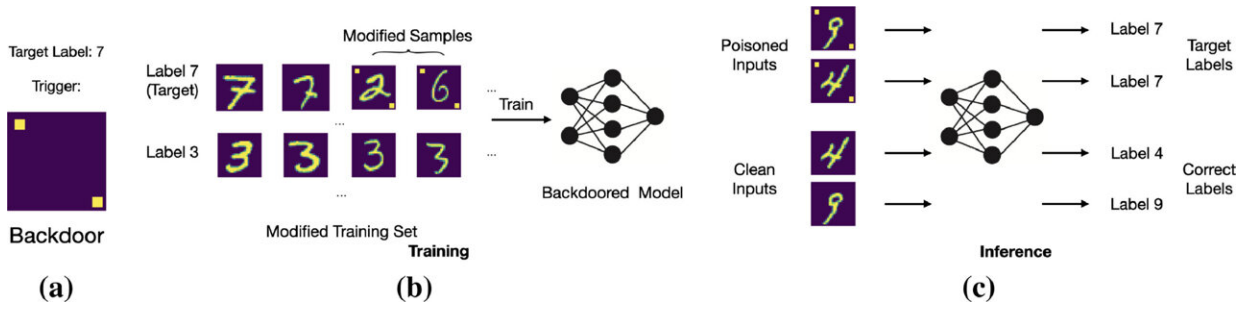
Keeping the backdoor secure in your robust machine learning model

June 27 2023



The visible distributed trigger is shown in Figure 1(a) and the target label is seven (7). The training data is modified. We see this in Figure 1(b) and the model is trained with this poisoned data. The inputs without the trigger will be correctly classified and the ones with the trigger will be incorrectly classified during the inference, as seen in Figure 1(c). Credit: SUTD

Software systems are all around us—from the operating systems of our computers to search engines to automation used in industrial applications. At the center of all of this is data, which is used in machine learning (ML) components that are available in a wide variety of applications, including self-driving cars and large language models (LLM). Because many systems rely on ML components, it is important to guarantee their security and reliability.

For ML models trained using robust optimization methods (robust ML

models), their effectiveness against various attacks is unknown. An example of a major attack vector is backdoor poisoning, which refers to compromised training data fed into the model. Technologies that detect backdoor attacks in standard ML models exist, but robust models require different detection methods for backdoor attacks because they behave differently than standard models and hold different assumptions.

This is the gap that Dr. Sudipta Chattopadhyay, Assistant Professor at the Information Systems Technology and Design (ISTD) Pillar of the Singapore University of Technology and Design (SUTD), aimed to close.

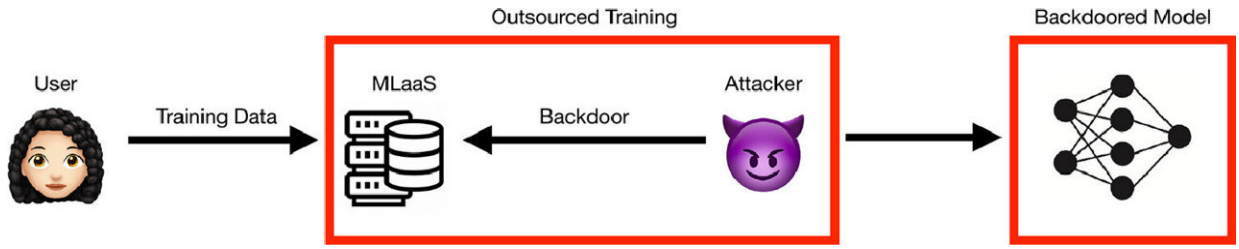
In the study "Towards backdoor attacks and defense in robust machine learning models," published in *Computers & Security*, Asst. Prof. Chattopadhyay and fellow SUTD researchers studied how to inject and defend against backdoor attacks for robust models in a certain ML component called image classifiers. Specifically, the models studied were trained using the state-of-the-art projected gradient descent (PGD) method.

The backdoor issue is urgent and dangerous, especially because of how current software pipelines are developed. Chattopadhyay stated, "No one develops a ML model pipeline and data collection from scratch nowadays. They might download training data from the internet or even use a pre-trained model. If the pre-trained model or dataset is poisoned, the resulting software, using these models, will be insecure. Often, only 1% of data poisoning is needed to create a backdoor."

The difficulty with backdoor attacks is that only the attacker knows the pattern of poisoning. The user cannot go through this poison pattern to recognize whether their ML model has been infected.

"The difficulty of the problem fascinated us. We speculated that the internals of a backdoor model might be different than a clean model,"

said Chattopadhyay.



An attack Model for AEGIS. Credit: SUTD

To this end, Chattopadhyay investigated backdoor attacks for robust models and found that they are highly susceptible (67.8% success rate). He also found that poisoning a training set creates mixed input distributions for the poisoned class, enabling the robust model to learn multiple feature representations for a certain prediction class. In contrast, clean models will only learn a single feature representation for a certain prediction class.

Along with fellow researchers, Chattopadhyay used this fact to his advantage to develop AEGIS, the very first backdoor detection technique for PGD-trained robust models. Using t-Distributed Stochastic Neighbor Embedding (t-SNE) and Mean Shift Clustering as a dimensionality reduction technique and clustering method, respectively, AEGIS is able to detect multiple feature representations in a class and identify backdoor-infected models.

AEGIS operates in five steps—it (1) uses an algorithm to generate translated images, (2) extracts feature representations from the clean training and clean/backdoored translated images, (3) reduces the

dimensions of the extracted features via t-SNE, (4) employs mean shift to calculate the clusters of the reduced feature representations, and (5) counts these clusters to determine if the model is backdoor-infected or clean.

If there are two clusters (the training images and the translated images) in a model, then AEGIS flags this model as clean. If there are more than two clusters (the training images, the clean translated images, and the poisoned translated images), then AEGIS flags this model as suspicious and backdoor-infected.

Further, AEGIS effectively detected 91.6% of all backdoor-infected robust models with only a false positive rate of 11.1%, showing its high efficacy. As even the top backdoor detection technique in standard models is unable to flag backdoors in robust models, the development of AEGIS is important. It is critical to note that AEGIS is specialized to detect backdoor attacks in robust models and is ineffective in standard models.

Besides the ability to detect backdoor attacks in robust models, AEGIS is also efficient. Compared to standard backdoor defenses that take hours to days to identify a backdoor-infected model, AEGIS only takes an average of five to nine minutes. In the future, Chattopadhyay aims to further refine AEGIS so that it can work with different and more complicated data distributions to defend against more threat models besides backdoor attacks.

Acknowledging the buzz around artificial intelligence (AI) in today's climate, Chattopadhyay expressed, "We hope that people are aware of the risks associated with AI. Technologies powered by LLM like ChatGPT are trending, but there are huge risks and backdoor attacks are just one of them. With our research, we aim to achieve the adoption of trustworthy AI."

More information: Ezekiel Soremekun et al, Towards Backdoor Attacks and Defense in Robust Machine Learning Models, *Computers & Security* (2023). DOI: 10.1016/j.cose.2023.103101