Examining the Potential for Adversarial Reprogramming Cyber Attacks on Nuclear Machine Learning Systems Utilizing Iterative FGSM

Eric Hill[1], B.Pedro Mena[1], C.Kallie Mclaren[1], D.Emily Elzinga[1], E.Christopher Spirito[2], and F.Leslie Kerby[1]

[1]Idaho State University College of Science & Engineering
921 S 8th Ave, Pocatello, ID 83209

[2]Idaho National Laboratory
1955 N Fremont Ave, Idaho Falls, ID 83415

erichill@isu.edu, menapedr@isu.edu, emilyelzinga@isu.edu,
kalliemclaren@isu.edu christopher.spirito@inl.gov, lesliekerby@isu.edu

ABSTRACT

The use of machine learning within the nuclear industry is a topic that has gained interest in recent times. Researchers are exploring concepts not only to aid operators, but also to automate these processes. This new interest leaves several concerns that will need to be addressed prior to implementation. One of these is the potential for cyber related threats to impact the performance of these systems. This study explores the possible impact an adversarial reprogramming attack could have on a machine learning model. This attack attempts to target classification models by altering data being used by a system to make the determination. One goal is to make alterations so that human observers cannot detect the changes. To do this, the Fast Gradient Sign Method was utilized. This method has been used in literature to make small changes to images to fool image classifiers. The code was modified to allow the attacker to target the specific misclassification desired and tested with several image sets, such as CIFAR10, with significant success. The method was then used to alter data from a GPWR simulator used in training models designed to classify nuclear transient. Twelve random samples were altered and tested with a k-nearest neighbors model. Of the 12 samples, 7 were classified incorrectly, but only 2 of the samples were misclassified as the targeted value. Still, this small sample does show there is a vulnerability that needs to be addressed, especially if the model will be the basis of an automated system.

KEYWORDS: Machine Learning; Nuclear Power; Cyber Security; Nuclear Safety

1. INTRODUCTION

The concepts of machine learning, artificial intelligence, data science, etc. are quickly becoming parts of everyday life. This coupled with rising costs concerns is leading many industries to explore these concepts to improve operations. The nuclear industry is no exception. Researchers have looked at a number of applications where machine learning could be used to improve performance and increase automation. This includes routine maintenance, fuel loading and transient detection.

These efforts, while exciting, also raise a number of concerns. Most prominently, what impact could a cyber attack have on machine learning systems used in nuclear applications. This paper looks to explore the impact of an adversarial reprogramming attack on a machine learning model used to classify transient events occurring with a nuclear reactor. By examining the possibilities of such an attack, it is hoped that steps will be taken to better ensure the security of nuclear systems as new reactor technology is developed.

2. BACKGROUND

2.1. Adversarial Attacks

As machine learning continues to grow as a field, researchers in the field of cyber security have begun looking to identify vulnerabilities with machine learning in models. These efforts have been done in order

to prevent bad actors from interfering with the implementation and use of these models. One area of vulnerability that researchers have been exploring is altering the data being given to a machine learning model. The alterations may then cause issues with the model resulting in a misclassification. This could lead to incorrect decisions being made by the system or users relying on the output of the model. This type of attack has been referred to as an adversarial attack[1].

One of the possible purposes of an adversarial attack is to have a machine learning model provide a misclassification. Depending on the nature of the attack and the goals of the attacker, this may be a targeted attack aimed at producing a specific misclassification or a random target. Adversarial attacks can look to target either the data used to train the machine learning model, known as data poisoning, or can look to manipulate data used in the actual application of the machine learning model.

A number of studies on this type of attack have looked to manipulate image classifiers. For example, one study from the University of Tokyo looked to perform an adversarial attack on a Deep Neural Network (DNN) that measured facial attractiveness[2]. This study targeted the vectors used in determining the score. The results from this attack were positive, as the researchers were able to manipulate the model to increase the score by nearly twice its original value. This type of attack could be used to target algorithms in recommendation and marketing systems on social media websites.

Most of the work done in adversarial attacks has looked to target image and video data. However, as the field of neural networks has matured, the interest in attacking other systems has also increased. One study that looked to target a non-image model, took place at Imperial College London. In this study, researchers used an adaptive attacker to perform a cyber attack on data used for a simulated water treatment facility[3]. The attack looked to cause the water tanks to overflow while remaining stealthy. The attack was successful, as they were able to fool the sensor readings into believing the system was operating normally. Although this attack assumes a great level of knowledge about the system, it does demonstrate the need to protect critical systems for infrastructure.

2.2. Data Used for Study

One of the challenges with machine learning is that the models require sufficient data in order for the model to perform. This study made use of three datasets: the MNIST digits dataset, the CIFAR10 dataset and a dataset of compiled simulations from a Generic Pressurized Water Reactor simulator. The Modified National Institute of Standards and Technology (MNIST) digits dataset set is a commonly used image dataset in the field of machine learning. The dataset has been modified to include 70,000 images (60,000 for training, 10,000 for testing) with the ten basic alpha numeric characters as classes[4]. The dataset has been used in many studies involving image classification with machine learning. The next data set used is the CIFAR10 dataset. Developed by Alex Krizhevsky of the Canadian Institute For Advanced Research[5], this dataset contains 60,0000 images (50,000 intended for training and 10,0000 for testing). The set has 10 classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck). The reactor data used in this effort was collected through simulations run on a WCS Generic Pressurized Water Reactor (GPWR) simulator at the Center for Advance Energy Studies (CAES) at INL. The dataset consists of over 110,000 datapoints, containing 27 features with 11 different scenarios.[6]. This dataset has been used in a study to develop machine learning models to classify transient events occurring within a nuclear reactor. Some of these models will be used in this study.

2.3. Fast Gradient Sign Method

The first step in exploring this attack was to determine an approach that could be used to alter the reactor data in a manner that would be difficult for an operator to detect. To this end, the Fast Gradient Sign Method (FGSM) was utilized. This method was developed in 2014 in order to modify image data [7] FGSM works by taking the gradients of the input features of the model being used. This includes both the actual and targeted values. These gradients can be either increasing or decreasing. The sign of the gradients collected is then taken and scaled using positive or negative 1's. This is done to ensure that an alteration is not changed significantly and thereby not easily detectable. The gradients are then multiplied by a value epsilon, which controls the amount of the change. These values are intended to be small. FGSM is designed to rely on the

direction of the gradients to make the change rather than the magnitude of the epsilon value. This process can be shown with the following equation:

$$\eta = \epsilon * sign(\nabla_x J(\theta, x, y))$$

This value is then added to the feature that is being altered. Figure 1 shows a sample of the noise generated using FGSM and how an altered photo may appear to a human eye.

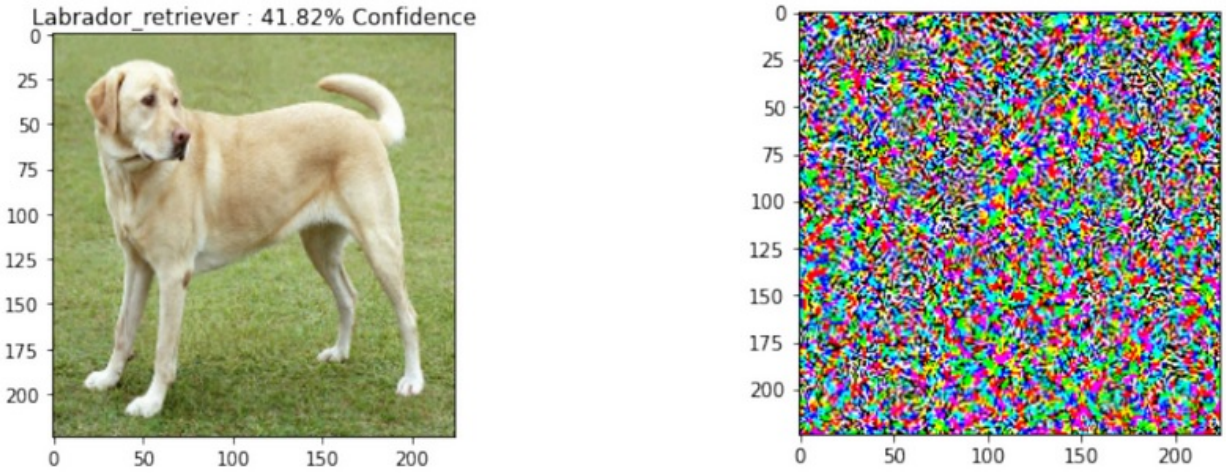

Figure 1: Sample Photo and Noise Generated with FGSM

3. METHODOLOGY

3.1. Developing the Model

The first step in this effort was to develop a model that could perturb data using FGSM. For this, the open source MobileNetV2 architecture was used. Using this code, the user can import an image and generate noise that the human eye would not be able to detect. In this case, an image of a dog was brought in and noise generated and added to the image. In order to test the effectiveness of the noise, different weights were applied to the model. When the weights were over 0.1, the model was able to learn the difference and make correct classifications. However, using a weight value between 0.05 and 0.1 produced enough noise to continuously cause misclassifications.

The next step was to develop a similar model that could work with different types of data. This worked in a manner similar to the typical FGSM models, but instead of taking the sign of the gradient values of the features, only the gradient values themselves were added as noise. This was tested using the MNIST dataset. The approach was still able to alter the images to cause a misclassification. However, it should be noted that while it is easier to tell an image has been modified, human eyes could still not notice the change in the target. For example, an image of the digit 7 still looked like an image of a 7, while the classifier predicted something incorrect. Figure 2 shows a comparison of the original image and the altered image from this example and the results from the classifier.

This process was repeated with the CIFAR10 dataset. Due to the greater complexity of the dataset, it was necessary to utilize a more advanced model architecture and improve the FGSM algorithm itself. The Xception architecture was used with a categorical cross-entropy loss function. An iterative FGSM algorithm was implemented, and found to perform better in terms of attack effectiveness. The iterative FGSM attack is known to produce better attacked samples than the standard FGSM, especially with white box style

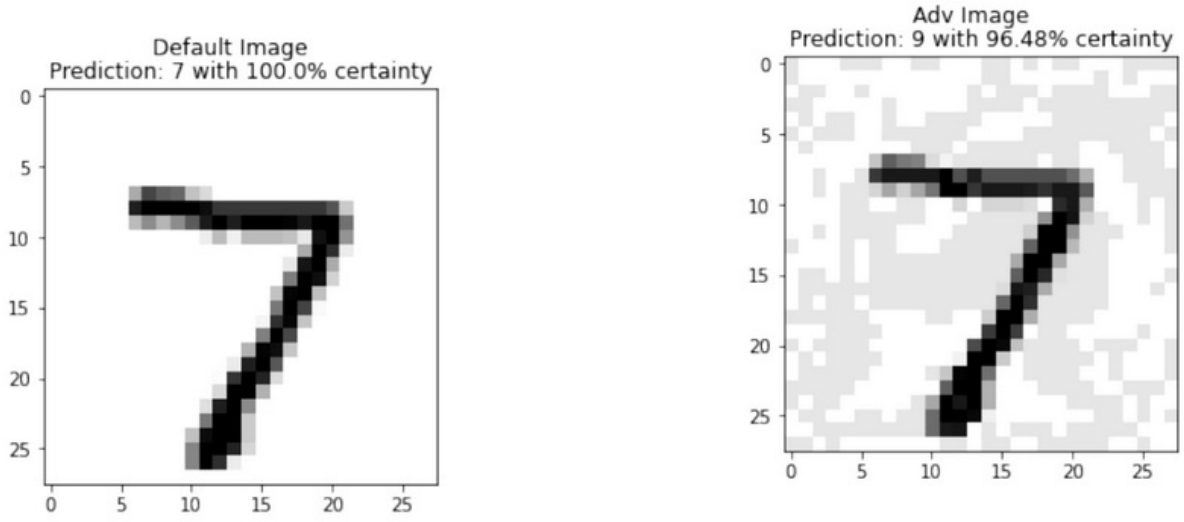

Figure 2: Comparison of Image Altered Using Iterative FGSM

attacks[8]. Instead of calculating the gradients and then adding the entire allotted epsilon at once, iterative FSGM algorithm adds a fraction of the epsilon based on the number of steps specified, and then repeats it for that many steps. This enables it to essentially add partial-epsilon values to the image, and be much more precise in honing in on the ideal adversarial sample. One form the iterative FGSM attack is given with the following equation:

$$x^*_{t+1} = x^*_t + \alpha * sign(\nabla_x J(x^*_t, y))$$

$$x^*_o = x$$

3.2. Testing the Model with GPWR Data

The next step in this project was to take the iterative FGSM algorithm and use it to alter reactor data, then test to see if the altered data could cause a specific misclassification. For this part of the project, a k-nearest neighbors (kNN) model from the group's previous effort was used to test the data [3]. This model had an accuracy of approximately 91% with the 11 events. Figure 3 shows the data prior to being altered by the iterative FGSM attack and Figure 4 shows the data after the attack.

```
     0    1        2        3        4        5        6        7        8  \
0  2.4  4.5  566.665  563.324  562.692  566.531  563.499  563.499  593.327

         9  ...       17       18       19       20       21       22  \
0  107.745  ...  38.5934  1127.04  1127.18  564.926  2063.69  641.299

        23   24   25   26
0  641.299  0.0  1.0  0.0

[1 rows x 27 columns]
```

Figure 3: Unaltered GPWR Data

```
     0       1           2           3          4           5           6  \
0  2.4  4.4979  571.620576  560.464172  558.00798  569.930128  561.472744

            7           8          9  ...         17           18          19  \
0  559.673463  597.742049  113.14917  ...  25.427159  1117.065096  1117.30142

          20         21         22         23         24  25     26
0  569.15331  2057.5027  636.24951  636.24951 -15.684393   1  0.007

[1 rows x 27 columns]
```

Figure 4: Altered GPWR Data using Iterative FGSM

To perform this test, a small sample of data from the GPWR dataset that was used for model testing was taken and altered using the iterative FGSM method. Twelve data samples were altered with the purpose of predicting specific misclassifications. The altered reactor data was given to the model and the predictions stored for comparison with the actual results.

4. RESULTS

The results from this effort with the iterative FGSM algorithm were very positive. The method was able to successfully alter image data without any major changes to the content of the image. These changes led to the image classifiers incorrectly predicting the image. One factor that made this valuable was the ability to target specific misclassifications. This applied to both image datasets and GPWR data.

When this method was applied to numerical data from the GPWR simulator, the iterative FGSM data alterations were able to fool the kNN model in many cases. Of the 12 altered samples tested, seven of the samples were classified incorrectly by the kNN model, an attack success rate of over 50%. It should be noted however, that only two of these samples were misclassified as the intended target. The remaining samples were classified correctly. Table 1 shows the results for each sample.

5. FUTURE RESEARCH

To attempt to address the potential for an adversarial attack, it will be necessary to develop defenses that can identify, mitigate and prevent this type of attacks. Future efforts from this group will include exploring the use of machine learning models to spot anomalies and alterations within data. If altered data could be identified prior to being used in an application such as a nuclear system, it could filtered out or repaired to prevent issues from occurring with machine learning model. The identification of these altered points could be done with different neural network techniques such as, an autoencoder or Generative Adversarial Network (GAN). Advancements in generative learning might also allow for altered data to be repaired to help mitigate the impact of an adversarial attack.

6. CONCLUSIONS

The results from this effort show that cyber security needs to be a serious consideration when implementing machine learning models for critical systems. The iterative FGSM algorithm was able to successfully alter the GPWR data and fool a highly accurate machine learning model designed to classify transients. If this type of attack were to occur in an implemented system with a nuclear reactor, several issues could occur. First, if the system was designed to aid operators, the misclassifications could result in greater confusion to the operator. This could cause an already tense situation to get worse and cause operators to take incorrect action. This scenario could be more problematic in an automated system using machine learning, as the system could take incorrect action which could also make the situation more dire. However, if a proactive

Table 1: Summary of Adversarial Reprogramming Attack Results

Sample #	Actual Transient	Targeted Transient	Predicted Transient
1	Total Coolant Pump Trip	Valve Closure	Single Coolant Pump Trip
2	Total Coolant Pump Trip	Load Rejection	Total Coolant Pump Trip
3	Total Coolant Pump Trip	Depressurization	Rapid Power Change
4	Total Coolant Pump	Depressurization	Single Coolant Pump Trip
5	LOCA LOOP	Rapid Power Change	LOCA LOOP
6.	LOCA LOOP	Max Steam Line Rupture	LOCA LOOP
7.	Single Coolant Pump Trip	Valve Closure	Single Coolant Pump Trip
8	Single Coolant Pump Trip	Rapid Power Change	Single Coolant Pump Trip
9	Single Coolant Pump Trip	Max Steam Line Rupture	Single Coolant Pump Trip
10	Turbine Trip No SCRAM	Load Rejection	Single Coolant Pump Trip
11	Depressurization	Depressurization	Single Coolant Pump Trip
12	Total Coolant Pump Trip	Single Coolant Pump Trip	Single Coolant Pump Trip

approach is taken towards cyber security, especially in the areas of attack prevention and detection, it is very likely that machine learning technology will add great benefits and value to nuclear power in the future.

ACKNOWLEDGEMENTS

Funding for the project was provided by Idaho National Lab and the United States Department of Energy. Special thanks to Dr. Fan Zhang and Patience Lamb of Georgia Tech for their input during this project. Also, special thanks to R.A Borrilli of University of Idaho for support in data collection.

REFERENCES

[1] Vasisht Duddu. A survey of adversarial machine learning in cyber warfare. *Defence Science Journal*, 68(4):356, 2018.

[2] Sijie Shen, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Fooling neural networks in face attractiveness evaluation: Adversarial examples with high attractiveness score but low subjective score. In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pages 66–69, 2017.

[3] Giulio Zizzo, Chris Hankin, Sergio Maffeis, and Kevin Jones. Invited: Adversarial machine learning beyond the image domain. In *2019 56th ACM/IEEE Design Automation Conference (DAC)*, pages 1–4, 2019.

[4] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[5] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[6] Pedro Mena, R.A. Borrelli and Leslie Kerby. Expanded analysis of machine learning models for nuclear transient identification using TPOT. *Nuclear Engineering and Design*, 390, 2022.

[7] Goodfellow Ian, et al. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[8] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.