Investigation of the impact effectiveness of adversarial data leakage attacks on the machine learning models

Denis Parfenov*, *Lubov* Grishina, *Artur* Zhigalov, and *Anton* Parfenov

Orenburg State University, Orenburg, 460018, Russia

Abstract. Machine learning solutions have been successfully applied in many aspects, so it is now important to ensure the security of the machine learning models themselves and develop appropriate solutions and approaches. In this study, we focused on adversarial attacks. The vector of this type of attack is aimed at distorting the results of machine models. In this study, we selected the IoTID20 and CIC-IoT-2023 datasets used to detect anomalous activity in IoT networks. For this data, this work examines the effectiveness of the influence of adversarial attacks based on data leakage on ML models deployed in cloud services. The results of the study highlight the importance of continually updating and developing methods for detecting and preventing cyberattacks in the field of machine learning, and application examples within the experiments demonstrate the impact of adversarial attacks on services in IoT networks.

1 Introduction

With the widespread use of machine learning and artificial intelligence in various fields such as finance, healthcare, networks, transportation and others, there is a need to ensure their security. Currently, various types of adversarial attacks on machine learning models are being actively improved, which are aimed at changing the results of the model by introducing malicious data [1, 2]. Attackers can use such attacks to bypass attack detection and prevention systems, fool automatic decision-making systems, or discredit a machine learning model.

One of the most challenging problems in this area is data leakage [3]. When using sensitive data, such as patient medical records or financial data, models can be targeted by attackers seeking access to this information. A data breach can result in potential financial, legal and reputational consequences for the organizations and individuals whose data has been compromised.

In this regard, the development of methods and technologies to protect machine learning models from adversarial attacks and data leakage is a hot topic. Research in this area is aimed at developing algorithms for detecting and preventing adversarial attacks, methods for protecting data privacy, and model architectures that are resistant to attacks [4, 5].

* Corresponding author: parfenovdi@mail.ru

The purpose of this work is to study the effectiveness of adversarial attacks on machine learning models for data leakage in the field of identifying Internet of Things (IoT) network attacks.

The rest of the paper is organized as follows. The second chapter provides an overview of existing literature sources on the topic under study. The third chapter presents the mathematical formulation of the problem of researching adversarial attacks. The fourth chapter describes the scheme for adversarial attacks based on data leakage. The fifth chapter contains experimental studies of the effectiveness of effectiveness of adversarial data leakage attacks on the machine learning models. The sixth chapter contains a conclusion.

2 Related work

Scientists around the world are conducting research in the field of adversarial attacks and data leakage using machine learning models in order to develop effective protection methods.

Thus, the work of Dong Q. [6] proposed an approach to predicting leaks in machine learning models, based on the Bayesian inference methodology for estimating the marginal probability of dependent variables based on. In addition, the author of the study, using the linkage method, formulated statistical conclusions about important correlating variables, which makes it possible to analyze whether a data leak is occurring.

An approach to building a shadow machine learning model based on data leakage was demonstrated in the study [7]. The peculiarity of this method is that the attacker does not need to have information about the type of ML model and the attack can be carried out in an uncontrolled manner. However, the proposed defense methods are limited to dropout-based approaches and model ensembles and do not allow drawing conclusions about the attack at the time of its occurrence.

Study [8] presents an alternative approach to threat analysis by mapping attacker tactics, techniques, and procedures to attack targets and detection mechanisms. The algorithm is trained based on the taxonomy of described attacks, allowing it to generate a semantic network with a probabilistic assessment of current threats. This approach has high accuracy (about 92%), but requires constant updating of the database and additional training.

Within the framework of work [9], the effectiveness of adversarial attacks and methods of protection against them on the ML model was analyzed in the event that the data leakage occurred directly from the training dataset. The study authors developed inference methods that exploit the structural properties of robust models on perturbed data. The results of the experimental evaluation confirmed the effectiveness of calculating the risks of adversarial attacks for 6 countermeasures.

A modified classification of attacks and the concept of an adversarial risk grid based on the use of machine learning in network security is presented in [10]. The authors of the study concluded that to date, no generalized means of protection against adversarial attacks have been developed. It also raises the issue of interpretability of adversarial risk when reduced adversarial vulnerability occurs and its implications for various machine learning applications.

Thus, a review of modern research in the field of ensuring the security of ML models from various adversarial attacks showed the need to develop universal mechanisms for protecting against data leakage. This work compares the effectiveness of adversarial attacks on machine learning models for data leakage in the field of identifying Internet of Things network attacks.

3 Statement of the problem of researching adversarial attacks

Let's X – a variety of signs of IoT network traffic characterizing information from real devices; Y – a variety of network attacks by malicious devices. Data set $D^{l}_{TRAIN} = \{(x_i, y_i)|^{l}_{i=1} | x_i \epsilon X, y_i \epsilon Y\}$ – this is the input data for training the basic machine learning model $a_{base}: X \rightarrow Y$ to classify network attacks based on traffic analysis deployed in a cloud service.

Let's assume that as a result of the leakage of the input-output data of the basic ML model a_base, a data set is formed $D^{p}_{LEAKS} = \{(x_i, y_i)|^{p}_{i=1} | x_i \epsilon X', y_i \epsilon Y', X' \subseteq X, Y' \subseteq Y\}$ – this is data for training a shadow machine learning model $a_{leaks}: X \rightarrow Y$ to classify network attacks, simulating the basic model a_{base}. In addition, let us assume that the type of ML model a_{base} and the original training set D^{l}_{TRAIN} unknown to the attacker, and data leakage occurs due to the interception of input information and the result of the base model. Thus, the "shadow" ML model a_{leaks} is trained on the basis of secondary data.

To analyze the effectiveness of adversarial attacks on the target model a_{base} using the shadow model a_{leaks} need to generate adversarial samples $D^{k}_{ADV} = \{(x_i, y_i)|^{p}_{i=1} | x_i \epsilon X^{avd}, y_i \epsilon Y^{adv}, X^{adv} \subseteq X, Y^{avd} \subseteq Y\}$ based on black box methods and evaluate the accuracy of identifying attacks on the model a_{base}.

Main hypothesis of the study: The accuracy of identifying attacks on adversarial samples of the "shadow" and basic ML models does not exceed the specified level eps:

$$\left|acc\left(Y^{avd}, a_{base}(X^{adv})\right) - acc\left(Y^{avd}, a_{leaks}(X^{adv})\right)\right| \leq eps. \quad (1)$$

IoTID20 for detecting anomalous activity in IoT networks and CIC-IoT-2023 large-scale attacks in the Internet of Things environment in real time were considered as the main research sets.

4 Research scheme for adversarial attacks based on data leakage

As part of this study of the effectiveness of adversarial attacks on ML models deployed in cloud services based on data leakage, the following scheme is proposed, consisting of 4 stages (Figure 1):

1. Training a basic machine learning model based on available datasets D^{l}_{TRAIN} to classify network attacks as a result of traffic analysis.
2. Deployment of the ML model a_{base} in a cloud service, to apply it in practice. Carrying out a data leakage attack on the input and output data of the model, forming a data set D^{p}_{LEAKS}.
3. Training a shadow machine learning model a_{leaks} based on data set D^{p}_{LEAKS} to simulate the model's operation a_{base}.
4. Creation of adversarial examples based on a zero-order optimization algorithm and a shadow ML model a_{leaks}. Conducting an adversarial attack and assessing its effectiveness based on the collected data on the model a_{base}.

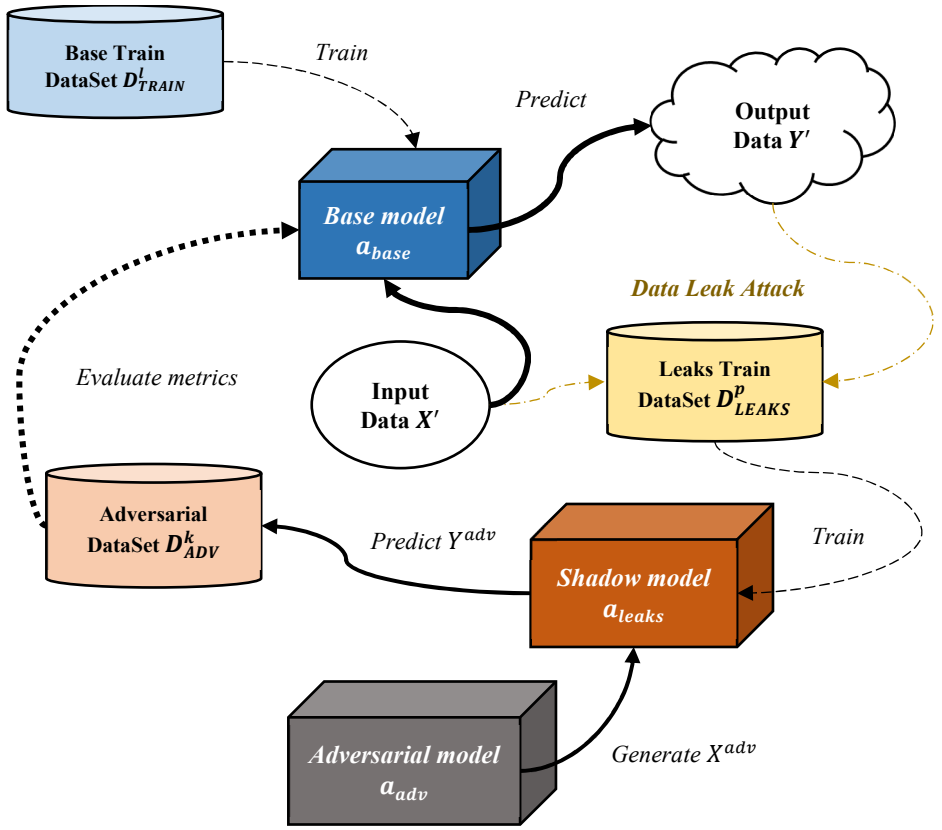


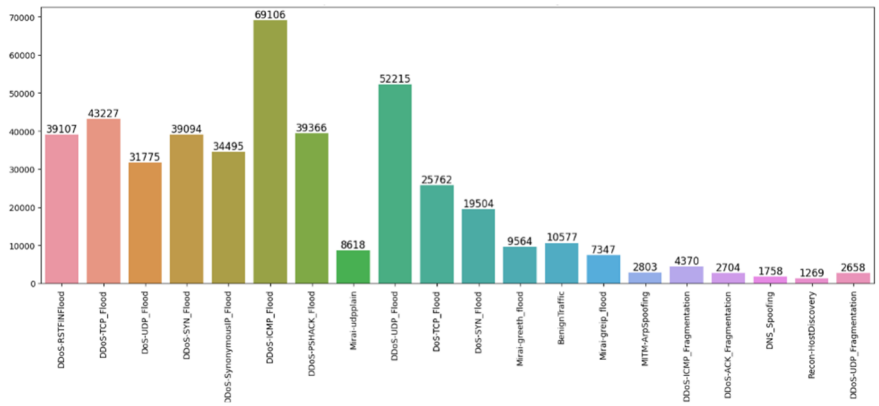
Fig. 1. Scheme for studying adversarial attacks during data leakage on ML models deployed in cloud services.

4.1 Source data set

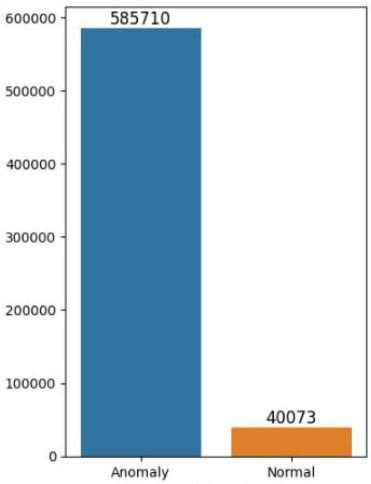
The CIC-IoT-2023 and IoTID20 data sets were selected as initial data, the main characteristics of which are presented in Table 1, and the distribution of records by attack class in Figures 2a and 2b.

Table 1. Description of data sets.

Dataset	Number of records	Number of features	Number of attack classes
CIC-IoT-2023	445319	46	20
IoTID20	625783	85	2



a) CIC-IoT-2023



b) IoTID20

Fig. 2. Distribution of records by attack type for datasets.

4.2 A basic machine learning models

The target ML models selected for deployment in the cloud service are: the Random Forest algorithm, the Catboost gradient boosting algorithm for decision trees and an artificial neural network (MLP-Prod).

4.3 A shadow machine learning models

The Shadow ML models, selected to simulate basic models and trained on the basis of secondary data, are: the gradient boosting algorithm for decision trees XGBoost and an artificial neural network of an alternative architecture.

4.4 Adversarial attack algorithm

Zeroth Order Optimization (ZOO), an adversarial black-box attack based on zero-order optimization, can bypass machine learning systems without having access to their internal parameters or training data. The iterative optimization process aims to create input data that

is specifically designed to introduce distortion or noise that is not noticeable to human perception, but is sufficient to cause ML models to make incorrect decisions.

In general, the ZOO adversarial attack algorithm has the following form:

1. Generate the initial view of the input data.
2. Initialize an adversarial example - generate noise, distortion or modification of the original data.
3. Obtain a forecast of the target ML model using an adversarial example.
4. Calculate the loss function that needs to be minimized to carry out a successful attack and may be associated with a decrease in the accuracy of the model or other criteria.
5. Apply a zero-order optimization algorithm such as genetic algorithm, random search method, etc. to update the adversarial example.
6. Repeat steps 3-5 for several iterations or until a stopping condition is reached, for example, reaching a certain value of the loss function or exceeding a specified iteration limit.

5 Experimental results

As part of this study, in order to maintain objectivity when analyzing the quality of machine learning models, the original data sets were divided into 3 parts D^{l}_{TRAIN}, D^{p}_{LEAKS} и D^{r}_{ADV} , which will be used to train basic ML models, shadow ML models and generate adversarial examples, respectively. The data is split in a ratio of 4:3:3 for D^{l}_{TRAIN}: D^{p}_{LEAKS}: D^{r}_{ADV}.

Note that the attacker does not have access to the dataD^{l}_{TRAIN}, At the same time, as a result of a data leak, the input and output readings of ML models (i.e. a data set D^{p}_{LEAKS}). Results of assessing the quality of training of basic attack classification models in IoT networks on a data set D^{l}_{TRAIN} are presented in Table 2.

Table 2. The result of constructing basic models for classifying attacks in IoT networks.

Dataset	Basic model	$Precision_{mean}$	$Balancy\_acc_{mean}$	$F1-Score_{mean}$
CIC-IoT-2023	Random Forest	0.994546±0.0053	0.9564641±0.0435	0.994507±0.0048
	Catboost	0.994099±0.0054	0.9525922±0.0439	0.992098±0.0039
	MLP-Prod	0.993699±0.0064	0.9525922±0.0439	0.992098±0.0039
IoTID20	Random Forest	0.972542±0.0142	0.9300259±0.0233	0.9137817±0.0316
	Catboost	0.984848±0.0022	0.9320668±0.0518	0.9210672±0.0114
	MLP-Prod	0.564277±0.0025	0.8845639±0.0012	0.59677539±0.0026

For the task of identifying anomalies (binary classification) on IoTID20 data, the highest accuracy was demonstrated by the Catboost model, the average balanced accuracy of which is ≈0.9321 (±0.052). The Random Forest model showed the best results in identifying specific types of attacks (multi-class classification) on CIC-IoT-2023 data with an average balanced accuracy of ≈0.9564 (±0.044). Error matrices for these machine learning models are presented in Figure 3 and Figure 4.

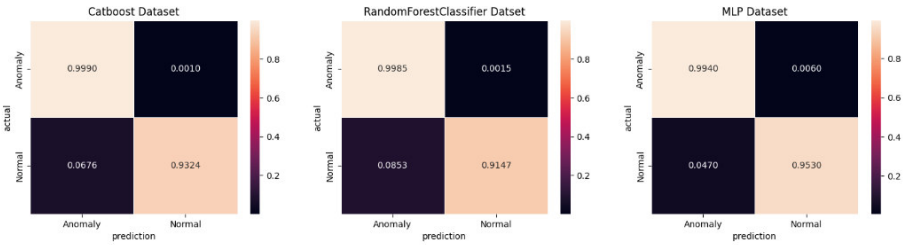


Fig. 3. Error matrices for base models on the IoTID20 dataset.

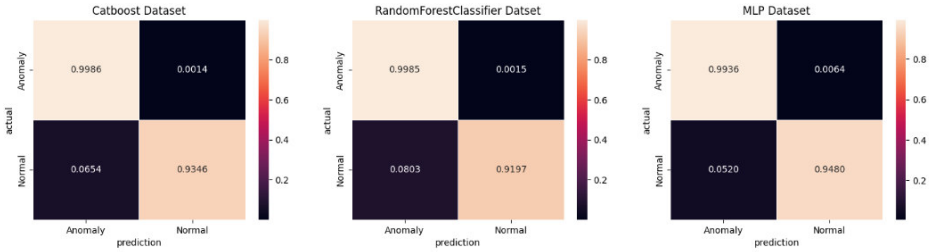
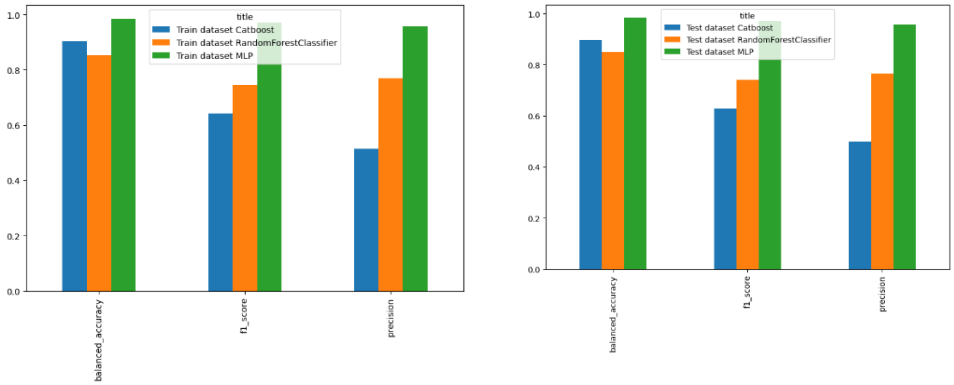


Fig. 4. Error matrices for base models on the CIC-IoT-2023 dataset.

A comparison of various quality metrics of basic models on data sets is presented in Figure 5.



a) CIC-IoT-2023 b) IoTID20

Fig. 5. Comparison of various quality metrics of base models on a set.

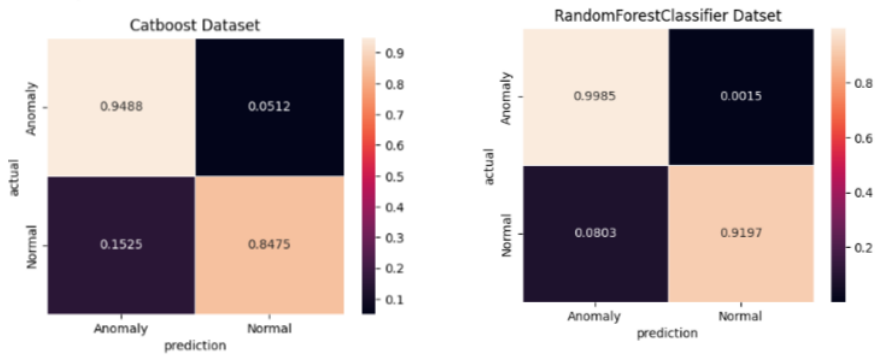
All presented in Table. 2 ML models are used as base models, the results of their attack class predictions are used as a set $Y' \subseteq Y$ for data set D^{p}_{LEAKS}. Assessing the quality of training shadow models for classifying attacks in IoT networks on a data set D^{p}_{LEAKS} presented in Table 3.

Table 3. The result of constructing shadow models that imitate the basic models**.**

Dataset	**Basic model**	**Shadow model**	$Precision_{mean}$	$Balancy\_acc_{mean}$
CIC-IoT-2023	Random Forest	XGBoost	0.998563±0.0154	0.946734±0.0325
		MLP-Mimic	0.993313±0.0345	0.994541±0.0325
	Catboost	XGBoost	0.994561±0.0565	0.954562±0.0243

		MLP-Mimic	0.993424±0.0123	0.944123±0.0345
	MLP-Prod	XGBoost	0.993131±0.0423	0.934557±0.0546
		MLP-Mimic	0.992341±0.0456	0.904571±0.0335
IoTID20	Random Forest	XGBoost	0.992341±0.0224	0.954312±0.0335
		MLP-Mimic	0.991234±0.0346	0.954461±0.0245
	Catboost	XGBoost	0.954541±0.0374	0.984541±0.0325
		MLP-Mimic	0.952541±0.0342	0.994812±0.0116
	MLP-Prod	XGBoost	0.994341±0.0234	0.974431±0.0745
		MLP-Mimic	0.994541±0.0235	0.984541±0.0745

To detect anomalies on the IoTID20 data set, the shadow XGBoost model, collected on the basis of data from the basic Random Forest model, demonstrated the highest accuracy - the average balanced accuracy is 0.9985 (±0.0154). The secondary model MLP-Mimic, built on the basis of the Catboost model, showed the best results on CIC-IoT-2023 data - an average balanced accuracy of 0.9488 (±0.0116). Error matrices for these machine learning models are presented in Figure 6. The XGBoost models have the smallest proportion of false positives (mean value $Precision_{mean} = 0.7483$).



a) MLP-Mimic on CIC-IoT-2023 b) XGBoost on IoTID20

Fig. 6. Error matrix for the shadow model.

The shadow models built in the previous stage were used to conduct an adversarial ZOO attack based on zero-order optimization. Note that the imposed noise on the input data is reflected in real units of measurement, without using scaling operations. In this regard, the range of values of various characteristics can vary from 0.004 to 3.432. In addition, the fastest generation of adversarial examples is based on the XGBoost model (average generation time ≈23.56 ms).

6 Conclusion

Modern adversarial attacks do pose a serious threat to machine learning classifiers used in cyber detection. These attacks aim to trick machine learning algorithms into producing false results or failing to detect malicious objects.

However, despite the effectiveness of such attacks, there are techniques that can improve the reliability of machine learning classifiers. In this study, we conducted an experimental study of the impact of an adversarial ZOO attack. On the examined IoTID20 dataset, the highest accuracy was demonstrated by the shadow model XGBoost, collected on the basis of data from the basic Random Forest model, and the MLP-Mimic model, built on the basis of the Catboost model, showed the best results on CIC-IoT-2023 data.

Overall, the development of more robust machine learning techniques is an important step towards improving cybersecurity. However, further research and development is needed to create even more effective and reliable methods for detecting cyber-attacks.

The research was carried out with the financial support of the grants of the President of the Russian Federation (MK-2959.2021.1.6).

References

1. J. Kos, I. Fischer, D. Song, *Adversarial examples for generative models*, in Proceedings of the IEEE Security and Privacy Workshops (2018)
2. N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z.B. Celik, A. Swami, *Practical black-box attacks against machine learning*, in Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (2017)
3. Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, H. Qi, *Beyond Inferring ClassRepresentatives: User-Level Privacy Leakage From Federated Learning,* in Proceedings IEEE Int. Conf. on Computer Communications (2019)
4. A. Alotaibi, M. Rassam, Future Internet **15(2)** (2023)
5. A. Aldweesh, A. Derhab, A.Z. Emam, Knowledge-Based Systems **189** (2020)
6. Q. Dong, *Leakage Prediction in Machine Learning Models When Using Data from Sports Wearable Sensors,* Computational Intelligence and Neuroscience (2022)
7. A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, M. Backes, *Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models* arXiv preprint arXiv:1806.01246 (2019)
8. U. Noor, Z. Anwar, A.W. Malik, S. Khan, S. Saleem, Future Generation Computer Systems **95** (2019)
9. L. Song, R. Shokri, P. Mittal, *Privacy risks of securing machine learning models against adversarial examples,* in Proceedings ACM SIGSAC Conference on Computer and Communications Security, 2019
10. O. Ibitoye, R. Abou-Khamis, M.E. Shehaby, A. Matrawy, M.O. Shafiq, *The Threat of Adversarial Attacks on Machine Learning in Network Security - A Survey* arXiv preprint arXiv:1911.0262 (2019)