Wireless Personal Communications (2021) 117:3505–3525
https://doi.org/10.1007/s11277-021-08284-8

Security Threats and Defensive Approaches in Machine Learning System Under Big Data Environment

Chen Hongsong[1,3] · **Zhang Yongpeng**[1] · **Cao Yongrui**[1] · **Bharat Bhargava**[2]

Accepted: 8 February 2021 / Published online: 13 February 2021

Abstract
Under big data environment, machine learning has been rapidly developed and widely used. It has been successfully applied in computer vision, natural language processing, computer security and other application fields. However, there are many security problems in machine learning under big data environment. For example, attackers can add "poisoned" sample to the data source, and big data process system will process these "poisoned" sample and use machine learning methods to train model, which will directly lead to wrong prediction results. In this paper, machine learning system and machine learning pipeline are proposed. The security problems that maybe occur in each stage of machine learning system under big data processing pipeline are analyzed comprehensively. We use four different attack methods to compare the attack experimental results.The security problems are classified comprehensively, and the defense approaches to each security problem are analyzed. Drone-deploy MapEngine is selected as a case study, we analyze the security threats and defense approaches in the Drone-Cloud machine learning application enviroment. At last,the future development drections of security issues and challenages in the machine learning system are proposed.

Keywords Machine learning system · Big data pipeline · Security threats · Defensive approaches · Case study

✉ Chen Hongsong
chenhs@ustb.edu.cn

Bharat Bhargava
bbshail@purdue.edu

[1] Department of Computer Science, University of Science and Technology Beijing (USTB), Beijing 100083, China

[2] Department of Computer Science, Purdue University, West Lafayette, IN 47906, USA

[3] Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China

1 Introduction

Machine learning and big data are two hot research areas in current computer and information science. From the smart Internet of Thing (IoT) embedded devices to the cloud data center, machine learning algorithms and systems are playing important roles in the pervasive computing environment. Data science and machine learning bring great opportunities and challenges for the pervasive computing research area. The core of big data is to make better decisions by using the value of data. The machine learning system is supported by a large number of data samples, which can build a high-precision data model. In the big data environment, with the development of pervasive computing and machine learning technology, artificial intelligence has been an essential part of the modern information process system. However, the security problems hidden behind the machine learning and pervasive computing system are increasingly emerging [1–3].

The cause of this security risk is that the security issues are not considered at the beginning of the machine learning algorithms design, so the processing results of Artificial Intelligence (AI) algorithm are easily affected by the malicious attackers, that result in the AI system judgment misalignment. In the safety–critical fields, such as industrial control system [4], health-care system, medical information system [5, 6], smart transportation [7, 8] and surveillance, the security of machine learning system is very important; if the machine learning system is attacked, it will cause great damage and serious impacts to the society. AI security risks do not only exist in theoretical analysis but also exist in today's real AI applications. For example, an attacker can escape AI-based detection tools [9] by adding simple noise, which causes the voice control system to successfully invoke malicious applications. The attacker can deliberately modify the input data, apply small marks on traffic signs [10] or other vehicle flags, which cause the auto-driving vehicle to make a wrong judgment [11].

The current big data process and machine learning system pipelines are facing many security risks. All the stages of a machine learning system based on big data techniques can be attacked, that causes the system incomplete and unavailable.

The organization structure of this paper is as follows. Section 2 introduces the related works and our contributions. Section 3 introduces the pipelines and security issues of the machine learning system under the big data environment. In Sect. 4, the security threats and challenges of the machine learning system under the big data environment are introduced. Section 5 presents the security defense approaches in the machine learning system. A machine learning application scenario-an intelligent drone MapEngine system is described in Sect. 6. Conclusion and future research directions are drawn in Sect. 7.

2 Related works and Contributions

2.1 Related Works

Carlini et al. [12] proposed three methods of attack, demonstrating that defensive distillation techniques do not significantly improve the robustness of neural networks. And compared with the existing generated confrontation sample algorithm, the proposed method has a good performance. Athalye et al. [13] proposed that the obfuscated gradients defense measures cannot produce good results, and proposed three attack methods

to overcome the three different obfuscated gradients methods. The security and safety of machine learning has received more and more attention.

Biggio et al. [14] summarized the research of machine learning safety for nearly 10 years and showed them in a timeline, emphasizing the connection between them. The article covers a wide range of content and a large workload. But it only summarizes the security issues in the field of machine learning, and does not involve cross-cutting areas. Compared with it, our paper summarizes the security problems of machine learning in the big data environment, such as a series of security problems in data acquisition, transmission, and storage in the big data environment. In the big data environment, the amount of data is large. When using machine learning for data analysis, it is necessary to adopt parallel processing algorithms. These algorithms also face some security problems in the processing process; some software and algorithm libraries commonly used in machine learning in the big data environment face security problems. Our article complements these and provides a good overview of security issues and precautions in the two cross-cutting areas of machine learning and big data. Papernot et al. [15] systematized the security and privacy of machine learning by proposing a comprehensive threat model and classifying attacks and defenses within a confrontational framework. The article focuses on the safety of machine learning, gives security issues in machine learning, but does not give corresponding preventive measures. The text includes a large number of formulas, but there is no specific case analysis. Our article not only gives security issues but also lists the corresponding preventive measures. And from the examples in life to analyze, there may be security issues, so closer to reality, easy to understand.

Liu et al. [16] conducted an in-depth investigation into the phenomenon that deep learning caused a complete change in output due to small changes in input, pointing out the possible reasons to promote deep learning to be more robust in defending against possible attack. However, the paper shrinks the scope of the discussion, focusing on the application of deep learning in computer vision. It does not analyze other machine learning algorithms, and it is lack of overall analysis structure. Compared with their research, our paper makes a comprehensive summary discussion on various algorithms in the field of machine learning, not only limited to deep learning, but also proposes a machine learning system security architecture, marking and discussing the problems of each stage, and the structure is more clear and easy to read, the content is more comprehensive.

Starting from the basic concept of adversarial learning, Akhtar et al. [17] elaborates the possible security threats during the model training and the inference test phase of machine learning and proposes possible defense techniques. Compare to it, our article is more comprehensive and complete than the process mentioned in this article, supplementing the machine learning technology driven by the big data environment not mentioned in this document, the acquisition and processing stage of training data, and the complete machine learning system. In this way, our article is closer to the actual application environment. In addition, their article lacks case analysis and integration with real-world production applications. Just a large number of paper studies may make readers mistakenly believe that this is only laboratory research and analysis, and the attacks on the real world do not exist for the time being. And our literature is a good combination of theory and practice to make up for the shortcomings of the article.

We use one table to show our advantages to the related works, which is shown in Table 1.

Table 1 Our advantages to the related works

Related works	Machine leaning security with big data pipeline	Comprehensive classification approach	Relationship between attack and defense approach	Case study	Future outlook analysis
Carlini et al. [12]	×	×	×	×	×
Athalye et al. [13]	×	×	×	√	×
Biggio et al. [14]	×	×	√	×	√
Papernot et al. [15]	×	×	×	×	×
Liu et al. [16]	×	√	√	×	√
Akhtar et al. [17]	×	√	√	×	√
Our paper	√	√	√	√	√

2.2 Contributions

Our contributions include the following points:

1. We propose the pipeline and security challenages of Machine Learning system under big data environment, and discuss detailedly the security threats at each stage.
2. We give a more comprehensive classification about security threats and defenses in machine learning system under the big data environment.
3. We build a connection between the attack and defense approach of machine learning system in big data pipeline.
4. We use the case study to illustrate the attack in machine learning under big data, and analyze potential attacks in the real physical scenario.
5. We proposed the future direction of the security challenages and solutions in machine learning system under big data pipeline.

3 Security Architecture of Machine Learning System Under Big Data Environment

3.1 The big data processing pipeline and security issues

We define a machine learning system, which is to find the inner connection from the data preparation and model training to the model prediction. From the data collection to the user interface, the big data processing pipeline and security issues are summarized in Fig. 1.

As shown in Fig. 1, big data processing is divided into the following six phases: data collection, data transmission, data preprocessing, data storage, data analysis, and data interpretation.

1. Data collection

   Data collection is the initial stage of the big data processing pipeline. Many kinds of data in the physical world are collected by the smart sensors and converted into the data structure recognized by the information system. In data collection, an attacker can

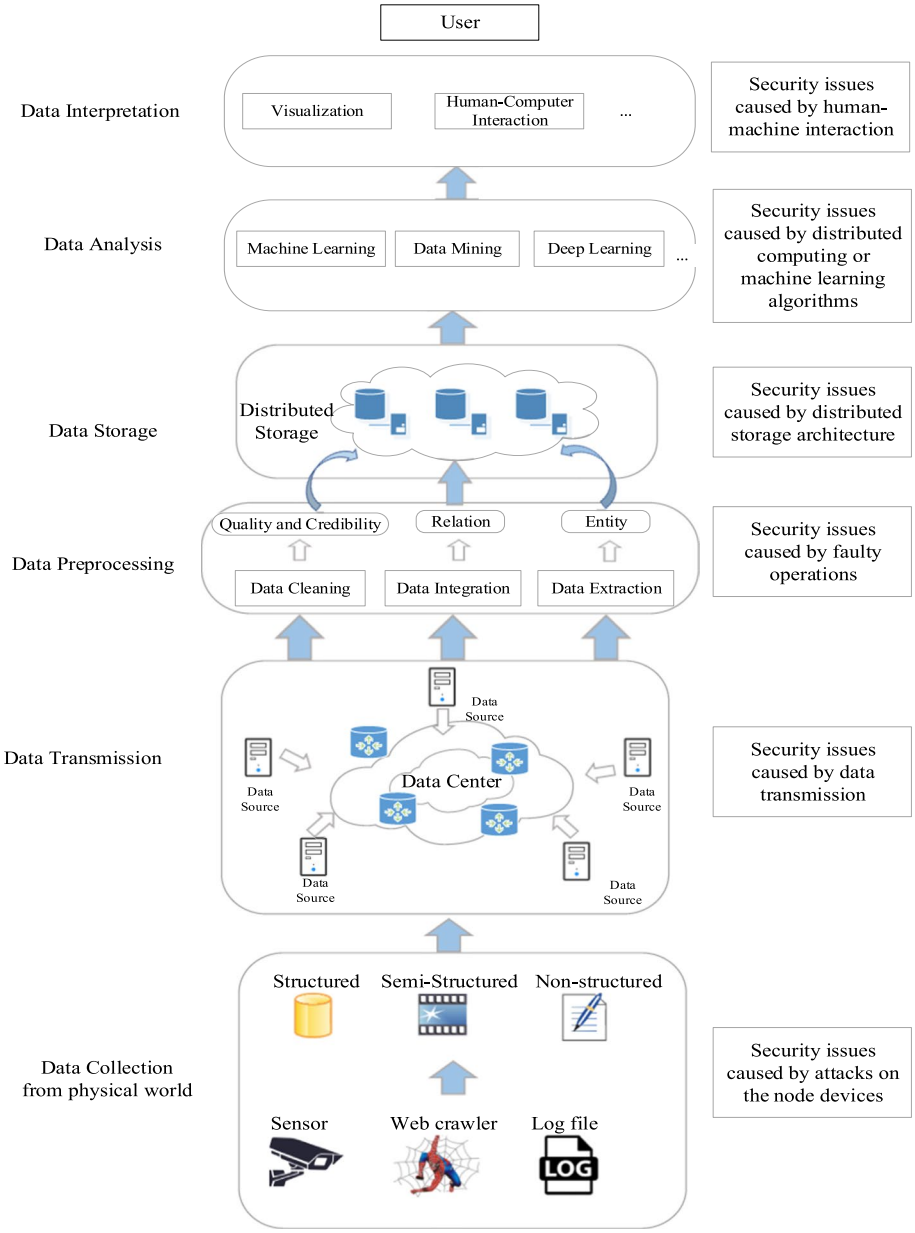

Fig. 1 Pipeline and security issues of Big data processing (3) Data preprocessing

attack or forge the collection node, so that the malicious data is collected or injected by the attackers [17].

2. Data transmission

Each node in the data collection sends the data to the central node, which then transmits the data to the data storage center. In data transmission, an attacker can sniffer or tamper the data to interfere with the system.

3. Data preprocessing methods include data cleaning, data integration, and data extraction. The quality and reliability of data can be improved. In this process, the operator's wrong operation may cause loss of useful data, but retain harmful data.
4. Data storage

   In the big data environment, the amount of data is huge and the types of data are diverse, so the cloud-based distributed storage system can be used to store the big data [18]. If a node is attacked by hackers, that can result in massive invalid data.
5. Data analysis

   The data analysis stage is the key stage of the big data procession pipeline. A large number of data are processed by distributed computing nodes. When the incremental machine learning algorithm is used, large amounts of data are divided into batches and processed in a time-sharing manner. Attackers can attack the machine learning algorithm to produce incorrect analysis results.
6. Data interpretation

Users can view the data processing results through data interpretation. The data interpretation methods are data visualization and human–computer interaction technology. Attackers can inject some false data to fool the data interpretation system.

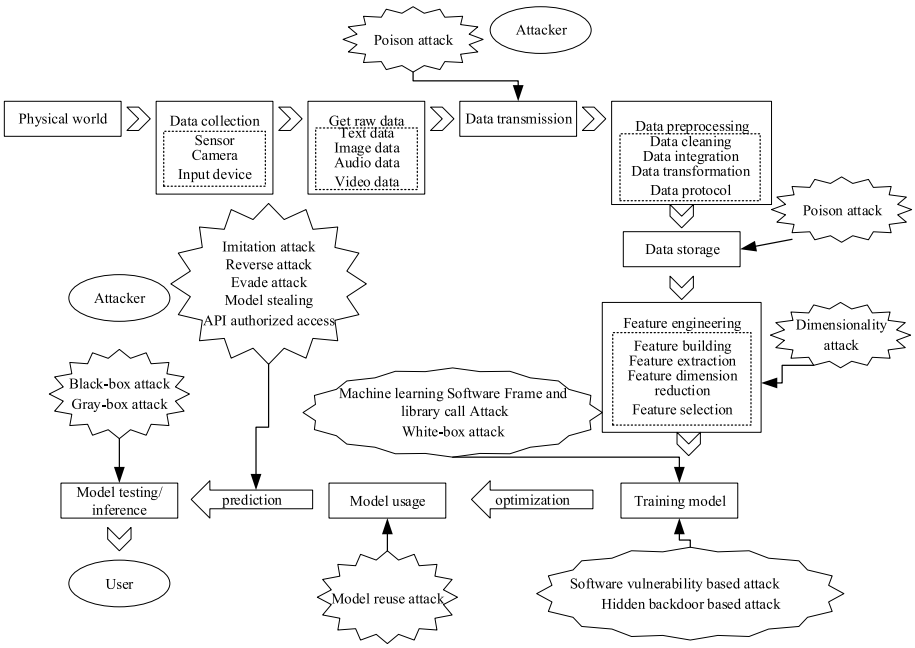

Fig. 2 Pipeline and security issues of machine learning system under big data environment

3.2 The Machine Learning System Pipeline and Security Issues

We define a machine learning pipeline, including a complete data process flow from data collection, data preprocessing, feature selection, model training to model prediction. The pipeline of machine learning system under big data is shown in Fig. 2. It describes the pipeline process from the initial acquisition of big data from the physical world to the final test and application of machine learning model. The security issues may exist at each stage. Under big data environment, the attacks are becoming more complex and diverse, the main security issues on the machine learning system are listed in Fig. 2.

Figure 2 shows the complete machine learning system pipeline. Starting from the physical world, data are collected by variable input devices, then a lot of data are obtained. In some application scenarios, the collected data has complex, high-dimensional, and changeable characteristics, and requires data preprocessing to make the data available as training data for machine learning algorithms. In the model training phase, the system designer uses the machine learning algorithm to train the model, and releases the model software API for the development user to call it.

There are different security issues at all stages of the pipeline. In the data collection stage, the attacker can inject malicious data to poison the model and algorithm. In the training stage, the machine learning algorithm may be attacked when the machine learning software frame (such as TensorFlow, Caffe, PyTorch) [19] library function with security holes is called. Moreover, since the training process of the model is based on the existing model, the model can be attacked by the white-box attack (The attacker knows everything about the training process, including training sets, training models, etc.). Therefore, the model may be reused by the attackers during the model usage phase [20]. For the deep neural network algorithm, dimensionality attacks may occur. In the later stages of the pipeline, there are different types of attack such as evasion attacks, imitation attacks, reverse attacks, unauthorized access to APIs, and model stealing attacks [21, 22]. There may be software vulnerability-based attacks and hidden backdoor-based attacks at various stages of the process. Black-box [23] and gray-box attacks can occur in the model testing and inference stage.

4 Security Threats and Challenges in Machine Learning System Under the Big Data Environment

4.1 Classification of Attacks

Under the big data environment, there may be a lot of security challenges in machine learning systems. The pipeline is complicated, the multi-stage processing pipeline may be attacked by different malicious users.

First of all, according to the characteristics of the attack, the classification of the attack is shown in Fig. 3. The characteristics of the attack are classified according to the attacker's target, attacker's location, attacker's ability, and number of attacks.

In the data collection phase, the poisoning attack is the typical attack mode. In this case, the system collects the malicious data samples from the attacker. Therefore, it interferes the model training of the machine learning algorithm. By destroying the

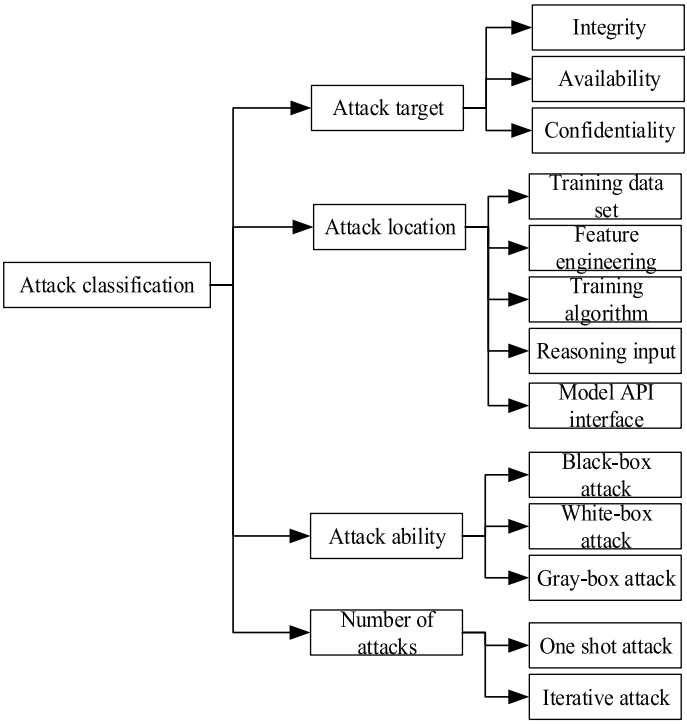

Fig. 3 Attack classification of machine learning

Table 2 Classification of evasion attack algorithm

Attacker's knowledge	Black-box	Score-based	[24–27]
		Decision-based	[28, 29]
	White-box		[1, 12, 30–36]
Attack frequency	One-time		[31]
	Iterative		[1, 2, 12, 24, 26–36]
Attack specificity	Targeted		[1, 12, 24–27, 29, 32–34, 36]
	Non-targeted		[1, 24–31, 35, 36]
Attack scope	Individual		[1, 2, 12, 24, 26–35]
	Universal		[32, 36]
Attack level	Pixel level		[1, 27]
	Image level		[2, 12, 24, 26, 28–36]

probability distribution of the original training data, the classification or clustering accuracy of the trained model are reduced by the poisoned data.

4.2 Classification of Evasion Attack

Evasion attacks occur during the inference phase of machine learning models, that is, attackers use the knowledge they have to generate adversarial samples that can misclassify the trained model. As shown in Table 2, the evasion attack algorithm in the image classification task is classified as follows:

1. Attacker's knowledge:
    a. Black box attack: The attacker does not know the training set used by the model, the structure of the model, etc., but only the output of the model, which is very consistent with the attack scenario in the real world. Black box attacks are divided into score-based attack and decision-based attack. Score-based attack is that the model output their class probability information on input samples. In this way, when using alternative models for black box attacks, the attacker can use class probability information get some information inside the original model. The decision-based attack is that the model only outputs the category to which the sample belongs, which is more difficult to implement than the score-based attack and is more in line with the actual attack scenario.
    b. White box attack: The attacker knows all the information of the model. This kind of attack exists only in research, not in the real world. Researchers can use better defenses against this attack, making the model more robust to input.

2. Attack frequency:
    a. One-time attack: This type of algorithm optimizes the adversarial samples only once.
    b. Iterative attack: This type of algorithm iteratively optimizes adversarial samples. This kind of attack can produce better attack effect. But because it interacts with the classifier many times, this attack method also relies on the classifier used when generating the adversarial sample, which has poor transferability.

3. Attack specificity:
    a. Non-targeted attack: The adversarial samples generated by such algorithms are misclassified by the model as any labels that do not belong to the original category.
    b. Targeted attack: The adversarial samples generated by this type of algorithms are misclassified by the model as a label given by the attacker. Compared with non-targeted attack, targeted attack is more difficult to achieve.

4. Attack scope:
    a. Individual attack: These methods generate specific perturbations for only one sample.
    b. Universal attack: This type of method generates a common anti-perturbation for all samples. Any sample in the sample set plus the generated perturbation can cause the model to be misclassified.

5. Attack level:
    a. Pixel level: By changing only a few pixels in the image, this type of attack methods makes the distance between the adversarial sample and the original sample very small, and it is easier to generate disturbances that humans cannot detect.
    b. Image level: By changing all the pixels in the image, the average perturbation generated by this attack method is larger than the perturbation at the pixel level.

Machine learning system software chain may be unsafe [19]. A large number of off-the-shelf algorithms and software dependency libraries are called and used during the construction of machine learning systems. Therefore, it has a long software chain, thus cause massive vulnerabilities. For example, TensorFlow, Caffe, and Torch are widely used by AI developers. However, these frameworks also have security vulnerabilities [37, 38]. If the machine learning algorithm or the dependent software library are attacked, it can cause great security problems. For example, the attacker may install a Trojan horse and steal the training data set without affecting the results of the algorithm and the accuracy of the model.

Experiments show that deep learning systems are more susceptible to small disturbances. Jiawei Su, Danilo Vasconcellos Vargas et al. [27] found if one pixel of a picture is changed in image recognition, the result of the algorithm may be greatly different. The linear nature of the deep neural network's high-dimensional space makes it very sensitive to the small perturbations of the input vector. Therefore, the author proposes a minimal anti-disturbance method based on the differential evolution algorithm, which requires only a small amount of adversarial information and can fool many types of Deep Neural Networks (DNNs). The experimental results of the author's paper show that 73.8% of the test images can be modified to resist images on one pixel, with an average of 98.7% confidence. And the experimental constraints are limited. It shows that current DNNs are also vulnerable to such attacks.

Adversarial examples are serious threats to the machine learning model. Goodfellow et al. [42] propose that the cause of neural networks' vulnerability to adversarial perturbation is their linear nature. This view yields a simple and fast method of generating adversarial examples. Brown et al. [32] present a method to create universal, robust, targeted adversarial image patches in the real world.

4.3 Some Examples of Attack Methods

The evolution of adversarial machine learning research is listed in Table 3.

The evolution of adversarial machine learning researches is shown in Table 3. From the concept of adversarial attack to specific attack cases, the machine learning algorithm can be attacked delicately. At the same time, some researchers are committed to the defense of attack method and continuously improve the robustness of machine learning systems.

We use four different attack methods to compare the experimental results, that are FGSM, DeepFool, BIm and JSMA. The adversarial examples, disturbances and the Adversarial image are shown in Fig. 4.

A pre-trained ResNet network [48] is used to generate the adversarial images, the test image is from Imagenet data [49]. Seen from the Fig. 4, all the four method can fool the machine learning model.

Table 3 Evolution of adversarial machine learning research

Time	Researchers	Main contribution
2004	Dalvi et al	Minimum-distance evasion of linear classifiers [39]
2011	Biggio et al	Evasion attacks against linear classifiers in spam filtering [40]
2013	Szegedy et al	A system that generates adversarial samples by perturbing inputs; It is the space, not individual units, that contains the semantic information in the high layers of neural networks [33]
2014	Biggio et al	Security evaluation framework of machine learning algorithm; adversarial model [41]
2014	Ian J. Goodfellow	The primary cause of neural networks' vulnerability to adversarial perturbation is their linear nature; Fast Gradient Sign Method (FGSM) as an attack method; adversarial training as a defensive method [42]
2016	Papernot et al	Defense distillation as a defensive method to defend DNNs [43]
2016	Sharif et al	Made a specific type of eyeglasses that people can wear to deceive state-of-the-art face recognition systems [44]
2018	Brendel et al	Boundary attack as the first effective decision-based attack [29]
2019	Komkov et al	Make a rectangular paper sticker and put it on the hat to attack the best public Face ID system [45]
2020	Li et al	Propose the E-MalGAN (an evasive adversarial-example attack method) and use it to successfully attack the Android Malware Detection System with firewall [46]
2020	Ayub el al	Apply adversarial examples on intrusion detection system [47]

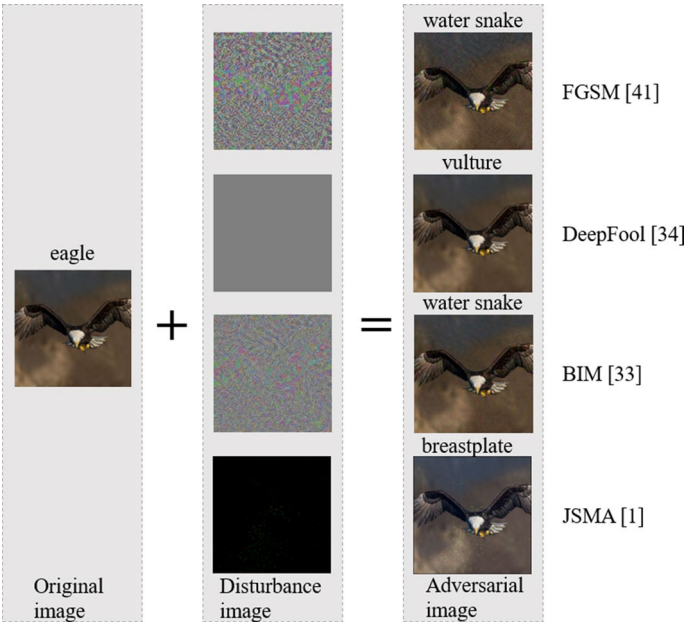

Fig. 4 Simulation results of different adversarial attack methods

Table 4 The results of four attack methods

Method	Accuracy (%)	Average_disturbance (l2 norm)	Time (min)
Original	98.78	0	0
FGSM	36.37	3.33	0.10
DeepFool	5.97	11.44	3.92
BIM	33.53	3.29	3.40
JSMA	0.03	3.99	56.85

Figure 4 shows the label of the original image and the categories into which the adversarial examples generated by different methods are classified. As can be seen from the figure, the adversarial example generated by the JSMA method are the least different from the original image, but there are also some distinct pixels.

We also used the Mnist dataset [50] to conduct comparative experiments on these four attack methods, and the results are shown in Table 4.

In Table 4, rows represent the method to generate the adversarial examples, and Original represents the use of the original test set. The columns represent various performance indicators, including the accuracy of the model on the generated adversarial examples, the average disturbance size of the adversarial examples, and the time of generating the adversarial examples. In the experiment, we used the Ubuntu operating system with 4 GB of RAM and a 4-core processor. A LeNet network [50] is used. It can be seen from the table that the FGSM method has the least time but the highest accuracy. The Jacobian-based Saliency Map Attack (JSMA) has the lowest accuracy but the longest time. The Basic Iterative Method (BIM) has the least average disturbance.

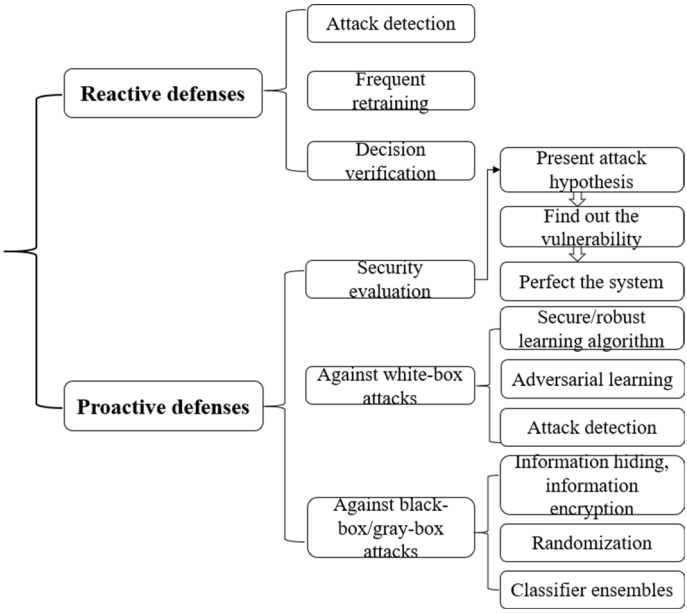

Fig. 5 Classifications of defensive approaches in the machine learning system

5 Defensive Approaches of Machine Learning system

5.1 Classification of Defensive Approaches

According to the characteristics of defensive approaches, we research and classify the defensive approaches. The classification of defensive approaches in machine learning system is shown in Fig. 5.

We divide defensive approaches into two categories [14]. One is reactive defenses, which will be taken after the machine learning systems are attacked. The other is proactive defenses, which are adopted before the machine learning systems are attacked. Generally, proactive defense approaches are recommended to be used. There are three aspects in reactive defenses: (1) attack detection, (2) frequent retraining, (3) decision verification. New training data and feature engineering are used in frequent retraining stage. We need to modify the decision boundary between prediction and real value to implement decision verification constantly. The classification of proactive defenses is proposed by the attacker's knowledge to the machine learning system. (1) The attacker is under a comprehensive understanding of the system Our corresponding strategy is to build a robust learning algorithm and attack detection. (2) The attacker does not understand the system. Our corresponding strategy is to hide the input and output of the system, such as data encryption, randomization and classifier ensembles. (3) We can establish a robust security assessment mechanism and improve our system by using game theory continuously.

The relationship between attacks and defensive approaches in the machine learning system is shown in Fig. 6.

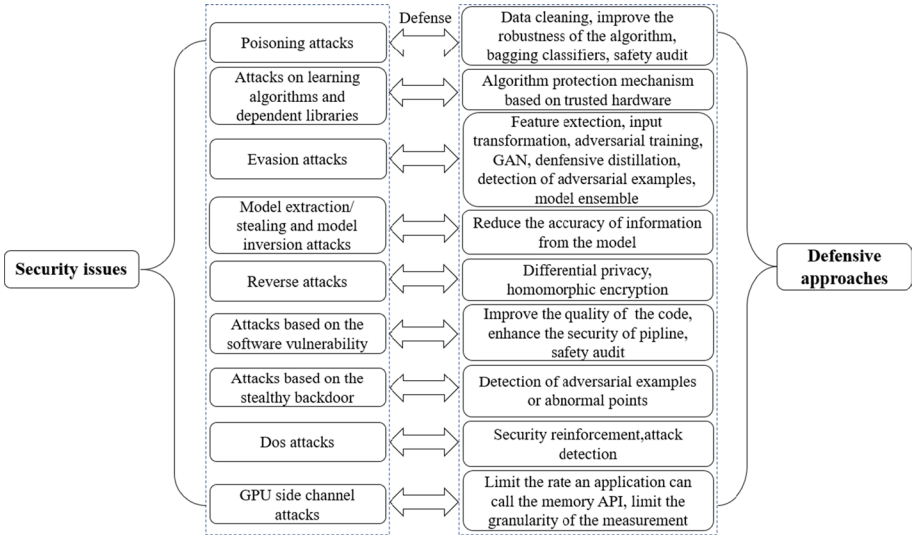

Fig. 6 Relationship between attacks and defensive approaches in the machine learning system

5.2 The Relationship between Security Issues and Defensive Approaches

As seen in Fig. 6, the relationship between each attack and defense approaches is established, there are multiple defensive approaches against each attack. The left column is the security problem of the machine learning system in the big data environment, and the corresponding right column is its defensive approaches. We won't go into details about the defensive approaches for traditional security problems, but we will introduce some new defensive approaches for attacks. For example, the evasion attack occurred in the test stage, corresponding to the method of adversarial feature selection [52].

The feature based on the generalizability of the generated classifier and the security of its manipulation of data is selected. Not only the generalization ability of the bagging classifier is optimized, but also the security of the bagging classifier against evasion attacks is enhanced. For GPU-side channel attacks [53], we can limit the rate at which an application can call the memory Application Programming Interface (API) or limit the granularity of measurements to reduce the information in the signal, making it impossible for an attacker to obtain victim behavior information. The attackers implement the Dos attacks on incremental machine learning algorithms, and the corresponding defensive approaches are security reinforcement and attack detection.

6 Case Study-Big Data Based Machine Learning Security Scenario Research

In this section, we analyze a specific case, Dronedeploy [54] company's machine learning product. Photogrammetry is the complicated process of generating accurate 3D reconstructions from 2D images. MapEngine is the world's first Machine learning-driven photogrammetry engine built for industry [31].

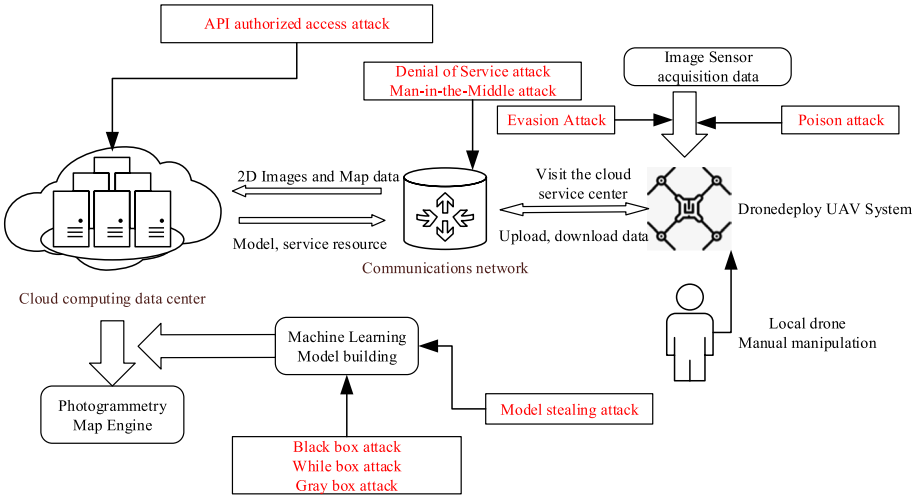

Fig. 7 Intelligent drone MapEngine machine learning system under big data environment

This release comes where MapEngine has processed more than 30,000 maps per month for more than 4000 clients across 180 countries.

Intelligent drone MapEngine machine learning system under the big data environment is shown in Fig. 7. Dronedeploy's drone system uses MapEngine cloud engine to enable intelligently process images. It can complete the automatic image recognition and automatic navigation function, or realize the mapping terrain and make high-definition maps.

In Fig. 7, we describe the drone MapEngine system which is the typical case of applying machine learning algorithms based on big data.

To the machine learning application case, we analyze some of the positions that may be attacked. According to the attacker's knowledge, it is divided into white box attack, gray box attack and black box attack to the model. Throughout the model, there may be Poison attack, Evasion Attack, Denial of Service attack, Man-in-the-Middle attack, Model stealing attack, API authorized access attack etc.

Machine learning algorithms are used in the Photogrammetry MapEngine to build the model by the uploaded data in the cloud. First, through the drone terminals carry the Image sensors, such as camera equipment, to collect the 2D image data. Then the drones access the cloud computing data center through the API interface, which is to provide machine learning as a service to users. The data center receives image data uploaded by a large number of drone terminals, performs image processing and analysis through its built-in machine learning algorithm, and then provides the results to the drone terminals. This means that the cloud data center provides machine learning models and computing resource for real-time online Unmanned Aerial Vehicle (UAV) systems to help them achieve intelligent decision making. The drone also has a certain processing ability, called LiveMap [55], which can realize certain data processing capabilities.

The first is the data collection by the UAV terminal, which may be subject to data pollution attacks. The attacker may pretend to be an ordinary user and normally hold the drone system, but it does not use it, or directly constructs the data to the cloud, so that the machine learning algorithm based on the data cannot extract the correct features, and the attacker may just want to disrupt the availability of the system. For the construction of

defense methods, data cleaning and authentication methods can be adapted to eliminate malicious data in time. It is also possible to compare the newly added data after training, and then roll back and delete the abnormal samples in time after the abnormality is found.

The resources of the cloud service may be consumed maliciously. The UAV terminal needs to request the cloud service and upload and download a large amount of data through the network. Attacker may initiate a Denial of Service attack and Man-in-the-Middle attack to construct false requests and the data packet, disrupting the availability of the UAV system. Real-time traffic monitoring can be used to detect and reject the abnormal traffic in a timely manner to secure the system.

The system's machine learning algorithms can be exploited maliciously. The cloud computing center is used to construct the machine learning model by the 2D images data uploaded from the users. Deep neural network processing is applied to train and test images data. If the attacker grasps the model parameters of the machine learning system, it is possible to perform a dimensionality reduction attack on the neural network. So that the system distinguishes the image to other content that is completely irrelevant. In this regard, the confidentiality of the system-specific algorithm should be strengthened first, and then the robustness of the dimension adjustment algorithm in deep learning should be enhanced. Finally, the system should design a filtering algorithm to filter the abnormal images, and the processed image should be compared to the original image.

Unauthorized access and model theft may occur [21]. Shi et al. [56] propose a method to attack online API calls, which implies the open API interfaces are also vulnerable. For the machine learning engine that the user requests the cloud, this is the core of the whole machine learning system. For the cloud using machine learning as a service, unauthorized access to the API and stealing attacks against the model may occur. Data center construction and maintenance of cloud-based machine learning services require a lot of costs, which is only available to users who use the company's products. Attackers may make use of some of the system's flaws, and then call the provided APIs to finish some work. They may base on this for a model stealing attack, restore the parameters through the continuous black-box access model, and then steal the model. Such an attacker is equivalent to mastering the model, and data security issues may occur. Therefore, it is necessary to strengthen the monitoring of API usage.

7 Conclusions and Future Directions

7.1 Conclusion

In this paper, machine learning system and pipeline are proposed to deeply explain the security issue in each stage of machine learning system under Big Data pipeline. Security threats and challenges in machine learning system are analyzed and classified comprehensively, the security issue and defensive approach in each stage of machine learning pipeline are pointed out. Our advantanges to the related works are listed in five dimisions, shown in Fig. 1.

The experimental results and comparison on four kinds of attack methods are analyzed, experimental reuslts show that FGSM spends the least time, JSMA behaves best attack effect. The defense methods are classified with reactive defense and proactive defense, the corresponding relationships between attacks and defensive approaches in the machine learning system are revealed. Drone-Cloud machine learning application is selected as real

world case study, the potential attack and vulnerability to machine learning system are analyzed.

7.2 Future Development Directions and Analysis

1. The performance and security tradeoff should be considered at the beginning of the new machine learning algorithm design.
2. Building a gold training set is an important foundation to the machine learning system. The security problem of machine learning system under big data environment is mainly about the attacks on data. If we build an ideal gold data set that can effectively counter the known attacks, the machine learning system will avoid the related security problems.
3. Adversarial training is the main approach to defense the adversarial example attack. By building the adversarial model and testing the machine learning system, we can find the vulnerability of the model. In this way, we can improve the machine learning system for specific attacks continuously.
4. Attack and defense are a dynamic game model. It is an open problem to defense the diversified and complex adversarial samples.
5. Ensure that all the stages of the pipeline in machine learning system are connected safety. In order to build a perfect system, it is necessary to consider the security problems that may occur in each stage and each connection of machine learning system.

Acknowledgements This work was supported by the National Social Science Fund of China (18BGJ071).

References

1. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., Swami, A. (2016). The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)* (pp. 372–387). IEEE.
2. Kos, J., Fischer, I., Song, D. (2018). Adversarial examples for generative models. In *2018 IEEE security and privacy workshops (spw)* (pp. 36–42). IEEE.
3. Nguyen, A., Yosinski, J., Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 427–436).
4. Chen, B., Wan, J., Lan, Y., Imran, M., Li, D., & Guizani, N. (2019). Improving cognitive ability of edge intelligent IIoT through machine learning. *IEEE Network, 33*(5), 61–67.
5. Shailaja, K., Seetharamulu, B., Jabbar, M. A. (2018). Machine learning in healthcare: A review. In *2018 second international conference on electronics, communication and aerospace technology (ICECA)* (pp. 910–914). IEEE.
6. Mozaffari-Kermani, M., Sur-Kolay, S., Raghunathan, A., & Jha, N. K. (2014). Systematic poisoning attacks on and defenses for machine learning in healthcare. *IEEE journal of biomedical and health informatics, 19*(6), 1893–1905.
7. Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., & Wierstra, D. (2015). Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.
8. Okuyama, T., Gonsalves, T., & Upadhay, J. (2018 March). Autonomous driving system based on deep q learnig. In *2018 International conference on intelligent autonomous systems (ICoIAS)* (pp. 201–205). IEEE.
9. Pei, X., Tian, S., Yu, L., et al. (2020). A two-stream network based on capsule networks and sliced recurrent neural networks for DGA botnet detection. *Journal of Network and Systems Management, 28,* 1694–1721.

10. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1625–1634).
11. Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprintarXiv:1708.06733.
12. Carlini, N., & Wagner, D. (2017, May). Towards evaluating the robustness of neural networks. In 2017 IEEE symposium on security and privacy (sp) (pp. 39–57). IEEE.
13. Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. arXiv preprint arXiv:1802.00420.
14. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition, 84,* 317–331.
15. Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). SoK: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)* (pp. 399–414). IEEE.
16. Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., & Leung, V. C. (2018). A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE Access, 6,* 12103–12117.
17. Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access, 6,* 14410–14430.
18. Idris, N., & Ahmad, K. (2011). Managing Data Source quality for data warehouse in manufacturing services. In *Proceedings of the 2011 IEEE International conference on electrical engineering and informatics* (pp. 1–6).
19. Xiao, Q., Li, K., Zhang, D., & Xu, W. (2018). Security risks in deep learning implementations. In *2018 IEEE Security and privacy workshops (SPW)* (pp. 123–128)
20. Ji, Y., Zhang, X., Ji, S., Luo, X., & Wang, T. (2018). Model-reuse attacks on deep learning systems. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security* (pp. 349–363).
21. Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction apis. In *25th {USENIX} security symposium ({USENIX} security 16)* (pp. 601–618).
22. Shi, Y., Sagduyu, Y., & Grushin, A. (2017). How to steal a machine learning classifier with deep learning. In *2017 IEEE International symposium on technologies for homeland security (HST)* (pp. 1–5)
23. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security* (pp. 506–519).
24. Chen, P. Y., Zhang, H., Sharma, Y., Yi, J., & Hsieh, C. J. (2017). Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security* (pp. 15–26).
25. Shi, Y., Wang, S., & Han, Y. (2019). Curls & whey: Boosting black-box adversarial attacks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6519–6527).
26. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., & Li, J. (2018). Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9185–9193).
27. Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation, 23*(5), 828–841.
28. Chen, J., Jordan, M. I., & Wainwright, M. J. (2020). Hopskipjumpattack: A query-efficient decision-based attack. In *2020 IEEE symposium on security and privacy (sp)* (pp. 1277–1294).
29. Brendel, W., Rauber, J., & Bethge, M. (2017). Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. arXiv preprint arXiv:1712.04248.
30. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.
31. DroneDeploy. Introducing map engine[EB/OL]. https://blog.dronedeploy.com/introducing-map-engine-cd3ef93bc730?gi=8762541ecbbc,2018-8-17.
32. Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. (2017). Adversarial patch. arXiv preprint arXiv:1712.09665.
33. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.
34. Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.
35. Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). Deepfool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2574–2582).

36. Moosavi-Dezfooli, S. M., Fawzi, A., Fawzi, O., & Frossard, P. (2017). Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1765–1773).
37. Xiaoqx.out of bound write cause segmentfault [EB/OL]. https://github.com/opencv/opencv/issues/9443,2017-08-23.
38. Common vulnerabilities and exposures.google tensorflow 1.7 and below is affected by: Buffer overflow. [EB/OL]. https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2018-8825,2018-3-20.
39. Dalvi, N., Domingos, P., Sanghai, S., & Verma, D. (2004, August). Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 99–108).
40. Biggio, B., Fumera, G., Pillai, I., & Roli, F. (2011). A survey and experimental evaluation of image spam filtering techniques. *Pattern recognition letters, 32*(10), 1436–1446.
41. Biggio, B., Corona, I., Nelson, B., Rubinstein, B. I., Maiorca, D., Fumera, G., & Roli, F. (2014). Security evaluation of support vector machines in adversarial environments. In *Support Vector Machines Applications* (pp. 105–153). Cham: Springer.
42. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
43. Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)* (pp. 582–597).
44. Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016). Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security* (pp. 1528–1540).
45. Komkov, S., & Petiushko, A. (2019). Advhat: Real-world adversarial attack on arcface face id system. arXiv preprint arXiv:1908.08705.
46. Li, H., Zhou, S., Yuan, W., Li, J., & Leung, H. (2019). Adversarial-example attacks toward android malware detection system. *IEEE Systems Journal, 14*(1), 653–656.
47. Ayub, M. A., Johnson, W. A., Talbert, D. A., & Siraj, A. (2020). Model evasion attack on intrusion detection systems using adversarial machine learning. In *2020 IEEE 54th annual conference on information sciences and systems (CISS)* (pp. 1–6)
48. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
49. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255).
50. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.
51. Samangouei, P., Kabkab, M., & Chellappa, R. (2018). Defense-gan: Protecting classifiers against adversarial attacks using generative models. arXiv preprint arXiv:1805.06605.
52. Zhang, F., Chan, P. P., Biggio, B., Yeung, D. S., & Roli, F. (2015). Adversarial feature selection against evasion attacks. *IEEE Transactions on Cybernetics, 46*(3), 766–777.
53. Naghibijouybari, H., Neupane, A., Qian, Z., & Abu-Ghazaleh, N. (2018). Rendered insecure: GPU side channel attacks are practical. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security* (pp. 2139–2153).
54. DroneDeploy. Capture. Analyze. Act. [EB/OL]. https://www.dronedeploy.com/.
55. DroneDeploy.Live Map [EB/OL]. https://www.dronedeploy.com/product/live-map/.
56. Shi, Y., Sagduyu, Y. E., Davaslioglu, K., & Li, J. H. (2018). Active deep learning attacks under strict rate limitations for online API calls. In *2018 IEEE International Symposium on Technologies for Homeland Security (HST)* (pp. 1–6).

Chen Hongsong received his Ph.D. degree of Computer Science in Harbin In-stitute of Technology in 2006. He is a professor in University of Science and Technology Beijing (USTB),China from 2008. He was a visiting scholar in Department of Computer Science of Purdue University from 2013-2014. He is a high-level mem-ber of China Computer Federation. He is an IEEE member now. His research interests include Artificial Intelligence and infor-mation security, wireless network and pervasive computing, trust computing. He got the excellent young academic paper award in USTB in 2009.He has published more than 50 aca-demic papers and 5 books.

Zhang Yongpeng is a master in University of Science and Technology Bei-jing (USTB), China from 2018. His research areas include in-formation security, machine learning and big data. E-mail: zypmicro@outlook.com

Cao Yongrui is a master in University of Science and Technology Beijing (USTB), China from 2018. His research areas include in-formation security, machine learning and big data. E-mail: zypmicro@outlook.com

Bharat Bhargava received the B.E. degree from the Indian Institute of Science and the M.S. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN. He is currently a Professor of computer science at Purdue University. His re-search involves mobile wireless networks, secure routing and dealing with malicious hosts, providing security in Service Oriented Architectures(SOA), adapting to attacks, and experi-mental studies. He is a fellow of the IEEE Computer Society. His name has been included in The Book of Great Teachers at Purdue University. Moreover, he was selected by the student chapter of ACM at Purdue University for the Best Teacher Award.