A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View

Article *in* IEEE Access · February 2018

DOI: 10.1109/ACCESS.2018.2805680

6 authors, including:

Qiang Liu
National University of Defense Technology
99 PUBLICATIONS **1,349** CITATIONS

SEE PROFILE

Pan Li
Queen Mary, University of London
9 PUBLICATIONS **256** CITATIONS

SEE PROFILE

Wentao Zhao
National University of Defense Technology
44 PUBLICATIONS **397** CITATIONS

SEE PROFILE

Wei Cai
The Chinese University of Hong Kong, Shenzhen
85 PUBLICATIONS **1,348** CITATIONS

SEE PROFILE

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 6, 2018, accepted January 31, 2018, date of publication February 13, 2018, date of current version March 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2805680

A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View

QIANG LIU[1], (Member, IEEE), PAN LI[1], WENTAO ZHAO[1], WEI CAI[2], (Member, IEEE), SHUI YU[3], (Senior Member, IEEE), AND VICTOR C. M. LEUNG[2], (Fellow, IEEE)

[1]College of Computer, National University of Defense Technology, Changsha 410073, China
[2]Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[3]School of Information Technology, Deakin University Melbourne Burwood Campus, Burwood, VIC 3125, Australia

Corresponding author: Qiang Liu (qiangliu06@nudt.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61702539 and Grant 61728201.

ABSTRACT Machine learning is one of the most prevailing techniques in computer science, and it has been widely applied in image processing, natural language processing, pattern recognition, cybersecurity, and other fields. Regardless of successful applications of machine learning algorithms in many scenarios, e.g., facial recognition, malware detection, automatic driving, and intrusion detection, these algorithms and corresponding training data are vulnerable to a variety of security threats, inducing a significant performance decrease. Hence, it is vital to call for further attention regarding security threats and corresponding defensive techniques of machine learning, which motivates a comprehensive survey in this paper. Until now, researchers from academia and industry have found out many security threats against a variety of learning algorithms, including naive Bayes, logistic regression, decision tree, support vector machine (SVM), principle component analysis, clustering, and prevailing deep neural networks. Thus, we revisit existing security threats and give a systematic survey on them from two aspects, the training phase and the testing/inferring phase. After that, we categorize current defensive techniques of machine learning into four groups: security assessment mechanisms, countermeasures in the training phase, those in the testing or inferring phase, data security, and privacy. Finally, we provide five notable trends in the research on security threats and defensive techniques of machine learning, which are worth doing in-depth studies in future.

INDEX TERMS Machine learning, adversarial samples, security threats, defensive techniques.

I. INTRODUCTION

Nowadays, machine learning is one of the most popular research fields, and its effectiveness has been validated in various application scenarios, e.g., pattern recognition, image identification, computer vision, clustering analysis, network intrusion detection, autonomous driving, etc. The advent of big data has stimulated broad interests in machine learning and privacy issues by enabling corresponding algorithms to disclose more fine-grained patterns and make more accurate predictions than ever before [1], [2]. Hence, many researchers devote themselves to review opportunities and challenges of machine learning in the big data era [3]–[5]. In particular, it is worth noticing that new intelligent systems mainly focus on learning from massive amounts of data with the goals of high efficiency, minimum computational cost and considerable predictive or classification accuracy [3], [5].

As a fundamental technology of future intelligent society, machine learning shall continuously expedite its theoretical study, algorithm design and development. However, the technology itself would suffer from several security issues [6], [7]. For example, some attackers can impersonate victims by exploiting the vulnerabilities of face recognition systems and compromise the privacy of sensitive data [8], [9]. Even more, someone with evil intent can seize control of autonomous vehicles [10] and voice control system [11] to make wrong decisions on recognizing traffic signs and voice commands, respectively. Hence, it can be expected that the security issues of machine learning will deserve much more concerns with larger application fields of the technology.

In the past decades, existing works mainly focused on the basic concepts and models of security threats against machine learning. In 2004, Dalvi *et al.* [12] introduced the concept

of adversarial classification and analyzed the detection evasion problem of early spam detection systems. After that, Lowd and Meek [13] proposed the concept of adversarial learning in 2005, and Barreno *et al.* [14] explicitly investigated the security of machine learning in 2006, including the taxonomy of attacks against machine learning systems and the adversarial models.

To address diverse security threats towards machine learning, many researchers devote themselves to propose some defensive techniques to protect learning algorithms, models and systems. Basically, defensive techniques of machine learning consist of security assessment, countermeasures in the training phase, countermeasures in the testing or inferring phase, data security and privacy. More details of defensive techniques will be discussed in Section IV. Furthermore, we would like to argue that the advances of security threats and defensive techniques regarding machine learning are promoted alternatively, resulting in a more powerful and intelligent debate.

Recently, some surveys on the security perspective of artificial intelligence and machine learning have been presented [15], [16]. Amodei *et al.* [15] gave a general introduction of security issues about artificial intelligence, especially the supervised and reinforcement learning algorithms. On the other hand, Papernot *et al.* [16] reviewed existing works about security issues and corresponding defensive methods in the life cycle of a machine learning-based system from training to inference. Different from previous surveys and reviews, this survey targets *a comprehensive literature review* regarding security threats and defensive techniques during training and testing or inferring of machine learning from a *data driven* view. Particularly, we emphasize the data distribution drifting caused by adversarial samples and sensitive information violation problems in statistical machine learning algorithms.

The framework of this paper is organized as follows. Section II briefly introduces the basics of machine learning, adversarial model and security threat taxonomy. Section III gives detailed description of corresponding security issues embedded in the two important phases of machine learning, training and testing/reasoning phases. Section IV summarizes security assessment frameworks and specific defensive techniques that defend against the security issues raised before. Section V presents several challenges and opportunities of security threats and defensive techniques of machine learning. Finally, Section VI gives conclusion remarks of this paper.

II. BASIC CONCEPT, MODEL AND TAXONOMY

A. BASICS OF MACHINE LEARNING

Machine learning is a multidisciplinary research field that spans multiple disciplines including computer science, probability and statistics, psychology and brain science. The objective of machine learning is how to effectively imitate human learning activities by computers such that the knowledge can be automatically discovered and acquired.

According to the differences of feedbacks, machine learning related works can be categorized into three groups, namely supervised, unsupervised and reinforcement learning [17]. In the supervised learning, the training samples with category labels should be feed into classification or regression models during their training phase. The typical supervised learning techniques includes decision tree, support vector machine (SVM), neural networks, etc. The unsupervised learning, on the other hand, induces models using the training samples with no knowledge of corresponding category labels. Clustering and auto-encoder are two typical examples of unsupervised learning techniques. The reinforcement learning optimizes behavior strategies via try-and-error, which is different from the learning procedure of the above two types of techniques.

B. ADVERSARIAL MODEL

The security research of machine learning with the consideration of adversarial knowledge was proposed in [13]. Then, researchers proposed to jointly consider the goal and the capability of adversaries into the model in [18]. Recently, Biggio *et al.* argued that a well-defined adversarial model should be constructed with four dimensions, *goal*, *knowledge*, *capability* and *attacking strategy* [19]. Specifically, the adversarial goal can be clearly described using both the expected impacts and the attack specificity of security threats. For example, the goal of an attacker is to launch an indiscriminate integrity attack that induces high false positive and true negative rates of classifiers or to launch a targeted privacy violation attack that illegally obtains sensitive data of the targeted user. Regarding the adversarial knowledge, it can be divided into two groups named constrained knowledge and complete knowledge by examining whether or not an attacker know training data, features, learning algorithms, decision functions, classifier parameters and feedback information. The adversarial capability of an attacker refers to his or her capability of controlling training and testing data. Furthermore, the capability can be qualitative interpreted from three aspects: (1) Is the impact of security threats causative or exploratory? (2) What is the percentage of training and testing data that are controlled by the attacker? (3) What is the extent of features and parameters that are known by the attacker? At last, the attacking strategy is the specific behaviors of manipulating training and testing data to effectively achieve his/her goals. For example, the attacker makes the decision regarding manipulation of data, modification of category labels and tampering with features.

C. SECURITY THREAT TAXONOMY

The taxonomy of security threats towards machine learning was proposed in [20] in three different perspectives, the influence on classifiers, the security violation and the attack specificity, as illustrated in Fig. 1.

In the perspective of the influence on classifiers, security threats towards machine learning can be classified into two categories: (a) *Causative attack*. It means that adversaries

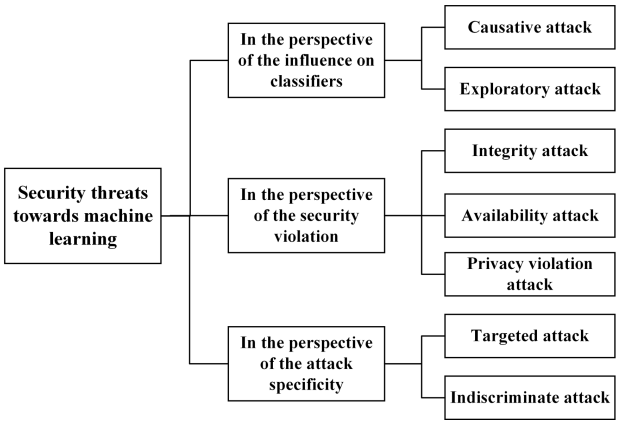


FIGURE 1. The taxonomy of security threats towards machine learning.

have the capability of changing the distribution of training data, which induces parameter changes of learning models when retraining, resulting in a significant decrease of the performance of classifiers in subsequent classification tasks. (b) *Exploratory attack*. Such attack does not seek to modify already trained classifiers. Instead, it aims to cause misclassification with respect to adversarial samples or to uncover sensitive information from training data and learning models.

In the perspective of the security violation, threats towards machine learning can be categorized into three groups: (a) *Integrity attack*. It tries to achieve an increase of the false negatives of existing classifiers when classifying harmful samples. (b) *Availability attack*. Such attack, on the contrary, will cause an increase of the false positives of classifiers with respect to benign samples. (c) *Privacy violation attack*. It means that adversaries are able to obtain sensitive and confidential information from training data and learning models.

In the perspective of the attack specificity, security threats towards machine learning have two types as follows: (a) *Targeted attack*. It is highly directed to reduce the performance of classifiers on one particular sample or one specific group of samples. (b) *Indiscriminate attack*. Such attack causes the classifier to fail in an indiscriminate fashion on a broad range of samples.

III. SECURITY THREATS TOWARDS MACHINE LEARNING

Basically, many security threats towards machine learning appear due to adversarial samples, which was mentioned in [21]. Specifically, adversarial samples represents those data that lead to the counterintuitive problem in deep neural networks (DNNs). Here, we extend this concept and regard adversarial samples as the harmful samples which cause the performance reduction of machine learning-based systems.

There are many concrete security threats towards different machine learning models and corresponding application scenarios [22]. Since 2004, such threats aimed to attack against security related applications. For example, in early 2004, Dalvi *et al.* [12] analyzed the detection evasion problem in early spam detection systems. Since then, more threats appeared to compromise other practical systems, e.g., malware identification and intrusion detection [23]. In conventional supervised learning, Naive Bayes and support vector machine (SVM) are two classical learning algorithms, in which early security threats occur. Specifically, an attacker can inject malicious and designated data into training data during the training procedure of machine learning based intrusion detection systems, inducing a significant decrease of the performance of these systems. For example, an attacking scenario was presented in [6], where a Naive Bayes based spam detection system was attacked by injecting malicious data. Similarly, another attack against a linear kernel SVM based malware detection system for PDF files was studied in [24]. Typically, clustering is a kind of unsupervised learning method, which can discover implicit patterns of data distributions. Although clustering algorithms have been widely used in many application scenarios, especially the information security field, they also suffer security issues. Specifically, most of attacks against clustering algorithms reduce their accuracy by injecting malicious data [25]–[27]. On the other hand, the obfuscation attack is another type of threat that compromises clustering algorithms [28]. The goal of obfuscation attacks against the targeted cluster is to generate a blend of adversarial samples and normal ones from other clusters without altering the clustering results of these normal samples, resulting in a set of stealthy adversarial samples.

Recently, deep learning has been emerging as a prevailing research field in machine learning. As a typical architecture of deep learning, DNN is demonstrated to be effective in various pattern recognition tasks, e.g., visual classification and speech recognition. However, recent works from late 2013 have demonstrated that DNN is also vulnerable to various adversarial attacks [29], [30]. For example, in image classification, the DNN only extracts a small set of features, resulting in poor performance on the images with minor differences. Potential adversaries can exploit such vulnerability to evade anomaly detection. In late 2013, Szegedy *et al.* [29] proposed to use the generated image with slight turbulence to deceive the pre-trained DNN. After that, several works proposed impersonation models to attack against DNNs and corresponding intelligent systems, e.g., face recognition, speech recognition and autonomous driving [8], [11], [31]–[35].

Table 1 summarizes related works regarding security threats towards machine learning.

A. SECURITY THREATS AGAINST THE TRAINING PHASE

Training is vital for machine learning to obtain a proper classification or regression model with respect to a target dataset. Hence, the training data that are feed into the training phase play an important role in establishing a high-performance machine learning model. Accordingly, many adversaries target the training data, resulting in a significant decrease of the overall performance of the machine learning model. For example, poisoning attack is a typical type of security threat against the training phase.

TABLE 1. Summary of related works regarding security threats against machine learning.

Targeted Learning Algorithms	Poisoning Attack	Evasion Attack	Impersonate Attack	Inversion Attack
Naive Bayes/Logistic Regression/Decision Tree	[6] [48] [50]	[49] [52] [53]	[10] [59]	[20] [69]
SVM	[43] [44] [48] [50] [51]	[49] [52] [53] [55]	[10] [59] [61] [63]	
PCA/LASSO	[9] [38] [47]	[57]	[59]	
Clustering	[25] [26] [28]			
DNNs	[45]	[35] [52] [53] [54] [58]	[8] [10] [11] [30] [31] [32] [33] [34] [35] [64] [65] [66] [67]	[20] [68] [69] [70] [71]

The poisoning attack is a type of causative attack, which disrupts the availability and the integrity of machine learning models via injecting adversarial samples to the training data set [36], [37]. Typically, such adversarial samples are designated by adversaries to have similar features with malicious samples but incorrect labels, inducing the change of training data distribution. Therefore, adversaries can reduce the performance of classification or regression models in terms of accuracy. Since the training data in practical machine learning based systems are protected with high confidentiality, it is not easy for adversaries to alter the data themselves. Alternatively, the adversaries are able to exploit the vulnerability that stems from retraining existing machine learning models. Note that it is feasible to launch such attack against machine learning based systems in practical usage, e.g., adaptive facial recognition systems (FRSs) [9], [38]–[40], malware classification [41], spam detection [6], etc. These systems are generally required to periodically update their decision models to adapt to varying application contexts. Taking the adaptive FRS as an example, an attacker utilizes the periodic update characteristic and injects masqueraded samples into the training data used for retraining decision models, resulting in the change of normal data classification centroid [38]. Fig. 2 illustrates the poisoning attack that induces the moving of classification centroid from the normal data (X_c) to the abnormal one (X_a). After attacking, the attacker can use adversarial facial images rather than the normal ones to pass the identity authentication. Regarding the unsupervised learning, e.g., clustering analysis, it is not applicable for changing the sample labels. However, some also introduced how to launch the poisoning attack against single-linkage and complete-linkage hierarchical clustering [25], [26], [28]. For example in [25], a heuristic strategy was adopted to measure the impact induced by adversarial samples on clustering accuracy via introducing a *bridge* concept. Based on such measurement, the optimal adversarial samples were selected to effectively reduce the clustering accuracy. Community discovery, singular value decomposition (SVD), and node2vec are three commonly used graph clustering or embedding techniques. However, recently research [42] has shown there are two

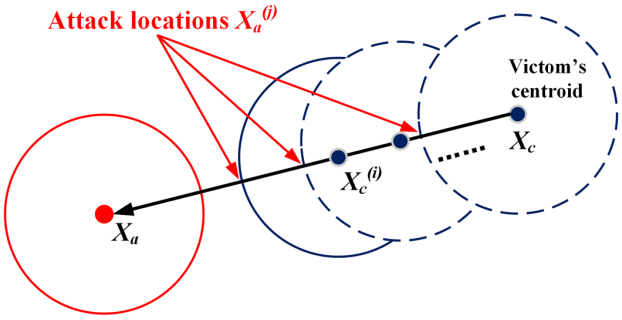


FIGURE 2. Illustration of poisoning attacks.

novel attacks, named targeted noise injection and small community. The two attacks can effectively evade graph clustering approaches with limit adversarial knowledge and low cost. The author also give two defense strategies: Training Classifier with Noise (similar to adversarial training) and improving hyper-parameter selection, but the defense effect is not significant. In addition, the poisoning attack also threatens many widely used machine learning algorithms, e.g., SVM [43], [44], neural networks [45], Latent Dirichlet Allocation [46], principle component analysis (PCA) and LASSO [47].

1) POISONING WITHOUT MODIFYING FEATURES OR LABELS

Referring to the feasibility of poisoning attacks, it is important to select a proper adversarial sample set. Accordingly, Mozaffari-Kermani *et al.* proposed poisoning models towards machine learning in health care cases, where the adversary has complete knowledge of the distribution of training data and the details of learning algorithms [48]. In particular, the adversarial samples were selected according to the degree of performance reduction in terms of the classification accuracy of learning models over validating data sets. Experimental results with respect to decision tree, nearest neighbor classifier, multilayer perception and SVM demonstrated the feasibility of the proposed poisoning models. However, this work did not give a theoretical proof of the

poisoning capacity to guarantee the effectiveness of attacking models. Another method of generating adversarial samples is the gradient ascent strategy [43], [47]. In this method, the optimal adversarial samples are selected by calculating the gradient of objective functions that measure the effectiveness of adversarial samples. Experimental results of poisoning SVM, LASSO and PDF malware detection systems showed the superior performance of the strategy. Recently, it is worth noticing that a more effective method for generating adversarial samples is to adopt generative adversarial network (GAN), which consists of a generative model and a discriminative one [45]. Specifically, the generative model is trained to generate candidate adversarial samples. Then, the discriminative model is used to select the optimal samples with a specific loss function. Comparative results between GAN and direct gradient methods on MNIST and CIFAR-10 data sets validated that GAN was able to rapidly generate high-quality adversarial samples with a larger loss value. Another example of applying GAN to adversarial sample generation was proposed in [49] to attack against malware classifiers.

2) POISONING WITH MODIFYING FEATURES OR LABELS

Apart from injecting adversarial samples to original training data, a more powerful adversary model has the capability of modifying extracted features or the labels of some training data [46], [47], [50], [51]. For example, in label contamination attack (LCA) [50], an attacker can significantly reduce the performance of SVM by flipping the labels of some training data. Furthermore, Xiao. *et al.* extended the adversary model to attack against some black-box linear learning models such as SVM and logistic regression (LR) [51]. It is worthwhile to mention that the above two adversary models both transferred the problem of selecting target labels to an optimization one. Moreover, the latter model does not require the prior knowledge about the detailed information of learning models.

B. SECURITY THREATS AGAINST THE TESTING/INFERRING PHASE

The testing or inferring phase mainly refers to the procedure of utilizing the trained model to classifying or clustering new data. By exploiting the vulnerabilities of training models, adversaries can generate a set of elaborate samples to evade detection, impersonate victims to obtain unauthorized access, or even compromise the privacy of training data via APIs of machine learning based applications to gain sensitive information of victims. The most common types of security threats against the testing/inferring phase include spoofing (for example, evasion and impersonate threats) and inversion attacks [52], [53].

Evasion attack was proposed to compromise machine learning in information security, e.g., spam detection [6], PDF malware detection [24], etc. The main idea of this attack is that an attacker generates some adversarial samples that are able to evade detection such that the overall security of target systems is significantly reduced [54]. There are several studies on attacking and defense techniques with respect to evasion attacks [55]–[57]. For example, the authors in [55] proposed to generate the optimal adversarial samples to evade detection via gradient algorithms. Recent studies also demonstrated that evasion attacks were feasible for use to attack against FRS [8] and malware detection [58] in real world, resulting in severe security threats towards target systems.

Similar to the above attack, the impersonate attack prefers to imitate data samples from victims, particularly in application scenarios of image recognition [59], malware detection [60], [61], intrusion detection [62] based on machine learning. Specifically, an attacker aims to generate specific adversarial samples such that existing machine learning-based systems wrongly classify the original samples with different labels from the impersonated ones [63]–[65]. By doing so, the attacker can gain the victims' authority in practical access control systems. Such attack is particularly effective in attacking DNN algorithms because DNN usually extracts a small feature set to facilitate the object identification. Hence, an attacker can easily launch impersonate attacks by modifying some key features [66], [67]. Moreover, there are many impersonate attacks to imitate images [59]. For example, Nguyen *et al.* [31] proposed to use a revised genetic algorithm, called Multi-dimensional Archive of Phenotypic Elites (MAP-Elites), to generate the optimal adversarial samples after evolving images from different categories. Then, these samples were fed into the AlexNet and the Le-Net-5 network, resulting in the performance reduction of DNNs. Regarding the impersonate attack in physical world, Kurakin *et al.* [35] demonstrated an attacking scenario of deceiving an image classification system in GeekPwn 2016. Firstly, the adversary generated electronic adversarial samples via the least likely class method. Then, these adversarial images were printed out to serve as the inputs of camera. Although the successful rate of launching impersonate attacks in physical world was much lower than that in electrical world due to the feature loss during printing and photography, this work validated the feasibility of impersonate attacks in real world. Moreover, Sharif *et al.* [8] introduced a novel attacking method against the latest FRS system, where an attacker was instructed to wear a designated pair of glasses. Experimental results demonstrated the feasibility of such attack in real world and the severe impacts on the detection capacity of FRS. More work [34] showed that transferable adversarial samples could be generated from ensemble learning, where the output samples from one DNN were effective for use to attack against other DNNs. Extensive experiments over a large-scale data set (ILSVRC2012)[1] and a commercial image and video recognition system (Clarifai)[2] demonstrated the effectiveness of the proposed method. Regarding the impersonate attack against audio information, its feasibility was also validated by real experiments that the voice with no meanings in

[1] http://www.image-net.org/challenges/LSVRC/2012/
[2] https://clarifai.com/

TABLE 2. Comparison of different attacking techniques against machine learning.

Attacking Techniques	Advantages	Disadvantages	Formulations of Adversarial Samples $\mathbf{x}^*$
Optimization-based method [29] [65]	1) Minimal perturbation $\Delta\mathbf{x}$ 2) Generate high quality adversarial samples	1) It is time consuming 2) It cannot be scaled to large datasets	Minimize $c\|\Delta\mathbf{x}\| + loss_f(\mathbf{x}+\Delta\mathbf{x}, t)$ s.t. $\mathbf{x}+\Delta\mathbf{x} \in [0,1]^m$ $\mathbf{x}^* = \mathbf{x}+\Delta\mathbf{x}$
Fast gradient sign method (FGSM) [91]	1) Faster than optimization-based methods 2) Generate high quality adversarial samples	The perturbation is not optimal	$\mathbf{x}^* = \mathbf{x} + \alpha\text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, t))$
Iterative least-likely class method [35]	1) Refinement of FGSM 2) Superior adversarial example (finer perturbations) to FGSM	The performance is affected by the number of iterations	1) $\mathbf{x}_0^* = \mathbf{x}$ 2) $\mathbf{x}_{n+1}^* = Clip_{\mathbf{x},\epsilon}\{\mathbf{x}_n^* - \alpha\text{sign}(\nabla_{\mathbf{x}} J(\mathbf{x}_n^*, t))\}$
DeepFool [32]	1) Assuming that neural networks are totally linear 2) It is highly efficient	It does not guarantee that the generated adversarial samples are good enough	$\text{argmin}_{\Delta\mathbf{x}_i}\|\Delta\mathbf{x}_i\|_2$ s.t. $f(\mathbf{x}_i) + \nabla f(\mathbf{x}_i)^{\mathrm{T}}\Delta\mathbf{x}_i = 0$
Jacobian-based saliency map approach (JSMA) [33]	1) It can finely tune the perturbation 2) It can make a good tradeoff between the number and the quality of adversarial samples	1) Target DNNs must be feed forward networks 2) The computation complexity is high when processing high dimensional data	1) Compute the Jacobin metric: $\nabla F(\mathbf{x}) = \frac{\partial F(\mathbf{x})}{\partial \mathbf{x}} = \left[\frac{\partial F_j(\mathbf{x})}{\partial \mathbf{x}_i}\right]_{i\in\{1,\cdots,M\}, j\in\{1,\cdots,N\}}$ 2) Calculate adversarial saliency maps: $S = saliency\_map(\nabla F(\mathbf{x}), \tau, t)$, where τ means pixels of $\mathbf{x}$ 3) Modify $\mathbf{x}_j^*$ with $\Delta\mathbf{x}$ s.t. $j = \text{argmax}_i S(\mathbf{x}, t)[i]$

the perspective of human-beings could be used to emulate real voice control commands. For more details, please refer to [11].

Inversion attack, on the other hand, utilizes the application program interfaces (APIs) provided by existing machine learning systems to gather some basic information regarding target system models [68], [69]. Then, the basic information is fed into reverse analysis followed by the leakage of privacy data embedded in target models, e.g., medical data of patients, survey data of customers, facial recognition data of users, etc [20], [70], [71]. According to the degree of understanding knowledge in adversarial models, this type of attack can be generally classified into two groups, namely white-box attack and black-box one [20]. Specifically, the white-box attack means that an attacker can freely access and download learning models or other supporting information, while the black-box one refers to the fact that the attacker only knows the APIs opened by learning models and some feedbacks after feeding inputs. Obviously, this type of attack introduces severe impacts on data privacy or even threatens the life of human-beings. For instance, an inversion attack against a drug system personalized for a patient may induce a wrong configuration of drugs and then cause the patient to die [70]. In USENIX SECURITY 2014, Fredrikson *et al.* [70] implemented an inversion attack against a customized drug system based on the linear regression algorithm. After that, the authors further implemented the attack against decision trees and facial images in FRS using gradient-descent methods [20]. In addition, some work used the output confidence values of machine learning cloud service platform, e.g., Google and Amazon, to design equation-solving attacks [71]. Given output labels, such security threats could extract learning models from multi-class logistic regression, DNN, RBF-kernel SVM.

C. SUMMARY OF ADVERSARIAL ATTACKS AGAINST MACHINE LEARNING

In this part, we summarize different adversarial attacks to present potential readers an overall scope of attacking techniques against machine learning. Specifically, we compare existing methods in terms of advantages, disadvantages and formulations of adversarial samples, as shown in Table 2. The meanings of different notations in Table 2 are given as follows: $\mathbf{x}^*$ means the adversarial sample with respect to an initial sample $\mathbf{x}$ crafted by a specific attacking technique, $\Delta\mathbf{x}$ refers to the perturbation that is added to $\mathbf{x}$ to generate $\mathbf{x}^*$, t represents the target label of $\mathbf{x}^*$, α is the step size when searching for a proper perturbation, and the symbol ϵ constraints $\mathbf{x}_n^*$ in an ϵ-neighbourhood of $\mathbf{x}$ after clipping pixel values of intermediate results. Moreover, $loss_f(\cdot,\cdot)$ means the loss function of a classifier f, and $J(\theta, \mathbf{x}, t)$ denotes the cost

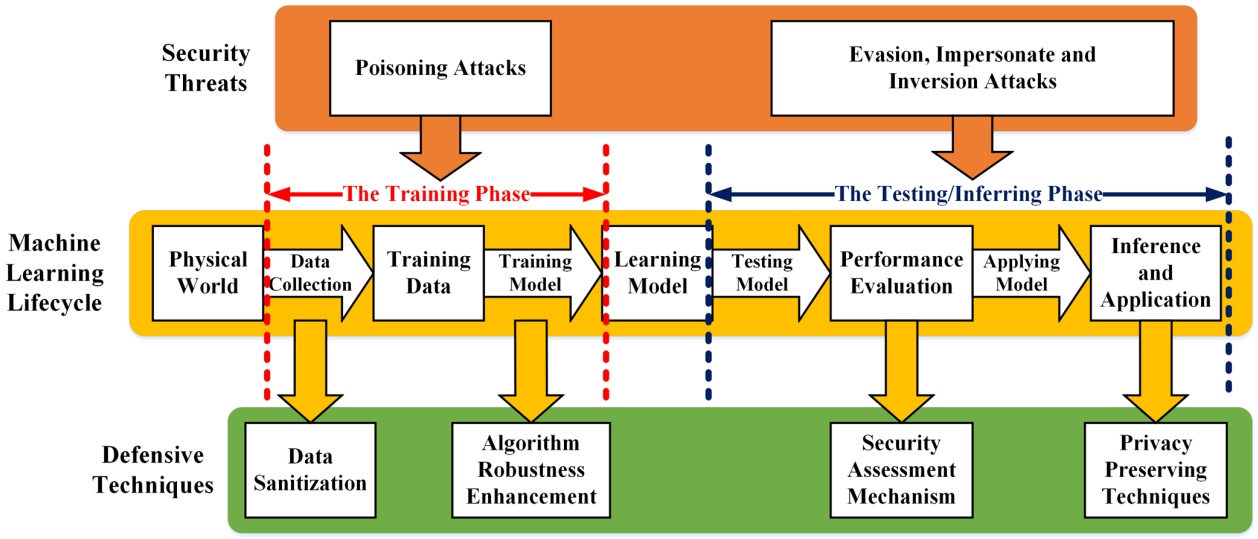


FIGURE 3. Illustration of defensive techniques of machine learning.

function used to train the learning model identified by θ. The symbol c in the formulation of optimization-based methods denotes the regularization coefficient. In the iterative least-likely class method given an image sample $\mathbf{x}$, $J(\mathbf{x}_n^*, t)$ represents the cross-entropy cost function of a target classifier, $n \in \{0, 1, \cdots\}$ means the number of iterations, and $Clip_{\mathbf{x},\epsilon}\{\mathbf{x}'\}$ is a function that performs per-pixel clipping towards an image $\mathbf{x}'$, i.e., $Clip_{\mathbf{x},\epsilon}\{\mathbf{x}'\} = \min\{255, \mathbf{x}(x, y, z)+\epsilon, \max\{0, \mathbf{x}(x, y, z)\}-\epsilon, \mathbf{x}'(x, y, z)\}$, where $\mathbf{x}(x, y, z)$ denotes the value of channel z of the image $\mathbf{x}$ at the coordinate (x, y).

From Table 2, we can see that different attacking techniques have their advantages and disadvantages in terms of time complexity, efficiency, the quality of adversarial samples and applicability to large-scale datasets. Specifically, optimization-based, FGSM and iterative least-likely class methods are good at generate high quality adversarial samples but induce high time complexity, especially for large-scale datasets. On the other hand, deep learning based attacking methods, e.g., DeepFool and JSMA, consider multiple factors when generating adversarial samples such as computational efficiency, the number and the quality of these samples.

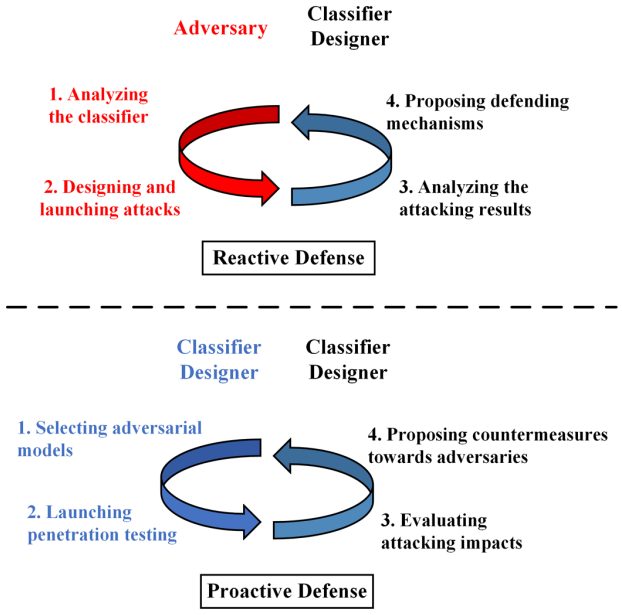


FIGURE 4. Typical workflows of two different defensive mechanisms.

IV. DEFENSIVE TECHNIQUES OF MACHINE LEARNING

In this paper, defensive techniques of machine learning is illustrated in Fig. 3. Accordingly, we will review related works in the following part of this section.

A. SECURITY ASSESSMENT MECHANISMS

Although there are a variety of security threats towards machine learning, conventional assessment mechanisms of machine learning are weak to address these threats. Basically, most of existing assessing techniques focus on quantitatively evaluating the performance of various learning algorithms rather than their security [72]. Hence, many researchers devote themselves to study on the security assessment of machine learning algorithms [19], [73]. It is widely adopted that security assessment is performed based on the what-if analysis method [74]. To be more detailed, a designer first introduces adversarial assumptions towards classifier vulnerabilities. Then, the designer proposes countermeasures to protect classifiers from the adversaries. Typically, there are two types of defensive mechanisms, i.e., proactive defense and reactive one [19], as illustrated in Fig. 4.

In the reactive defending mechanism, a potential adversary attempts to figure out proper attacking methods by analyzing the target classifier. Then, the adversary designs and

implements these attacking methods. On the other hand, a classifier designer would analyze new added samples and corresponding attacking results. After that, the designer proposes some defending mechanisms, e.g., recollecting data and introducing new features to update the classification model. The above two procedures perform alternatively, resulting in a race between the adversary and the classifier designer. Similar to the reactive defense, the proactive mechanism holds four steps as follows: selecting adversarial models, launching penetration testing, evaluating attacking impacts and proposing countermeasures towards adversaries. The notable differences between proactive and reactive defending mechanisms include the following two aspects: (1) The attacking and defending subjects are both the classifier designer in proactive defense. (2) The designer only performs penetration testing to uncover vulnerabilities rather than a true attack against the classifier. In other words, penetration testing in the proactive defending mechanism and attacking in the reactive one are benign and malicious, respectively. In the proactive defending mechanism, the classifier designer uncovers potential flaws or vulnerabilities that can be exploited by an attacker to launch attacks via penetration testing, where an adversarial model is constructed to perform as the attacker with specific targets, knowledge, capacity and attacking strategies. Then, the designer integrates some countermeasures towards adversaries into classifier design. Since the penetration testing is performed before releasing the designed classifiers, it is very helpful to improve the security of these classifiers.

Basically, the distribution of training data and that of testing data will be notably different with the presence of adversarial samples, resulting in a non-stationary data distribution. Hence, such abnormal phenomenon can be used to serve as a way of assessing the security of machine learning and to predict whether or not the adversarial samples exist. Based on the above idea, some researchers proposed quantitative security analysis and evaluation of machine learning algorithms in adversarial environments [19], [73], [75]. Fig. 5 illustrates an example of proactive security assessment considering data distributions. Specifically, the mechanism first selects proper adversarial models with respect to the hypothesized attack scenario defined at the conceptual level by making assumptions on the goal, knowledge, capacity and corresponding strategy. Then, it defines the distributions $p(Y)$, $p(A|Y)$ and $p(\mathbf{X}|Y, A)$ for training and testing data, where $Y \in \{L, M\}$ and $A \in \{F, T\}$ respectively refer to class labels (L: legitimate; M: malicious) and a Boolean random variable representing whether or not a given sample has been manipulated ($A = T$) or not ($A = F$). After that, it constructs sample training **TR** and testing **TS** sets according to the data model defined before, given $k \geq 1$ pairs of data sets $(\mathcal{D}^i_{TR}, \mathcal{D}^i_{TS})$, $i = 1, \cdots, k$ that are obtained from classical resampling techniques, e.g., cross-validation or bootstrapping. Finally, the classifier performance with the presence of simulated attack is evaluated using the constructed $(\mathrm{TR}^i, \mathrm{TS}^i)$ pairs.

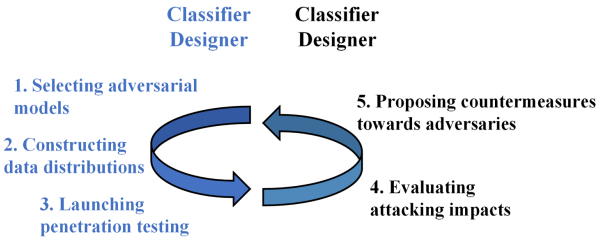


FIGURE 5. Security assessment considering data distributions.

B. COUNTERMEASURES IN THE TRAINING PHASE

As being analyzed before, the poisoning attack should be performed by injecting designated adversarial samples into training data to affect the resulting decision function under specific machine learning algorithms. Hence, ensuring the purity of training data [76] and improving the robustness of learning algorithms [77]–[79] are two main countermeasures towards such adversary at the training phase.

Data sanitization is a practical defending technique to ensure the purity of training data by separating adversarial samples from normal ones and then removing these malicious samples [80], [81]. For example, a *Reject on Negative Impact (RONI) defense* method was proposed to protect spam filters including SpamBayes,[3] BogoFilter,[4] the spam filter in Mozilla's Thunderbird[5] and the machine learning component of SpamAssassin[6] [80]. Specifically, the method tested the impact of each email in the training phase and did not train on messages that had a large negative impact. To quantitatively measure impacts on the classification performance, the method compared error rates between the original classifier and the new one, which was retrained after adding new samples into the original training data, over the same testing data. If the error rate of the new classifier was much lower than that of the original one, then the new added samples were considered as malicious data and would be removed from training data; Otherwise, these samples were benign data. In future, such method can be improved in the aspect of calculating efficiency when selecting large-scale candidate samples.

On the other hand, improving the robustness of learning algorithms is another feasible defending technique, e.g., Bootstrap Aggregating and Random Subspace Method (RSM) [78], [79]. Moreover, Rubinstein *et al.* [77] extended the original PCA and proposed an antidote based on techniques from robust statistics. To combat the poisoning activities in the context of anomaly detector, the authors presented a new robust PCA-based detector by maximizing the median absolute deviation. Experimental results demonstrated that poisoning significantly distorted the learning model produced by the original PCA method, whereas it had little effect on the robust model.

[3] http://spambayes.sourceforge.net
[4] http://bogofilter.sourceforge.net
[5] https://www.mozilla.org
[6] http://spamassassin.apache.org

Another type of effective defending methods is to design secure learning algorithms. For example, Demontis *et al.* [82] proposed a defending method that improved the security of linear classifiers by learning more evenly-distributed feature weights. Accordingly, they presented a secure SVM called Sec-SVM to effectively defend against evasion attacks with feature manipulation.

C. COUNTERMEASURES IN THE TESTING/ INFERRING PHASE

Compared to the defensive techniques in the training phase, countermeasures in the testing/inferring phase mainly focus on the improvement of learning algorithms' robustness. Game theory is a powerful tool to model dynamic debates between attackers and defenders. At the beginning, Globerson and Roweis [83] and Teo *et al.* [84] proposed invariant SVM algorithms using the min-max method to address the worst case feature manipulation activities (addition, deletion and modification) in the testing phase. To improve the robustness of learning algorithms, Brückner and Scheffer [85] also proposed Stackelberg Games for adversarial prediction problems and a NashSVM algorithm based on the Nash equilibrium [86]. Furthermore, Bulò *et al.* [87] extended previous work and proposed a randomized prediction game by considering randomized strategy selections according to some probability distribution defined over the respective strategy set. The method designed randomized decision functions compelling an adversary to select low-effectiveness attacking strategies. Analytical results validated that the proposed method could improve the trade-off between attack detection and false alarms of classifiers.

Another countermeasure against attacks in the testing/ inferring phase is active defense considering data distributions. Actually, the goal of adversarial samples in the testing/inferring phase is to alter the data distribution of testing data, resulting in significant deviation from the distribution of training data. Then, the false alarms of classifiers increase with the presence of adversarial samples [88]–[90]. Hence, a feasible way of defending against adversaries is to fit the testing data distributions by retraining learning models by classifier designers with adversarial samples. By doing so, the new trained classifiers are able to detect anomalies in the testing phase [27], [32]. For example, Zhao *et al.* [27] proposed to introduce adversarial samples with full labels into training data to train a more robust model. Similarly, Goodfellow *et al.* [91] demonstrated through experiments on the MNIST data sets that this method could significantly reduce the false alarm rate of learning models regarding adversarial samples from 89.4% to 17.9%.

Apart from introducing adversarial samples to improve the robustness of classifiers in the testing/inferring phase, the technique of smoothing model outputs is also effective to strengthen the robustness of learning models. Accordingly, a deep contractive network model was presented to adopt a smoothness penalty to improve its robustness under adversarial perturbations [92]. To protect deep learning algorithms in adversarial settings, defensive distillation was proposed to defend against adversarial samples on DNNs [93]. Specifically, Papernot *et al.* [93] analyzed the generalization and robustness properties granted by defensive distillation when training DNNs. Comparative results validated that the technique could effectively enhance the performance of two DNNs with different architectures to detect adversarial samples in terms of the success rate of adversarial sample crafting from 95.86% (87.89%) to 0.45% (5.11%) on the MINST[7] (CIFAR10)[8] data set. Note that recent work also argued that the defensive distillation was not secure and revealed that one could find adversarial examples on defensively distilled networks with a slight modification to existing attacks [94]. Hence, it can be expected that the research on secure learning algorithms will draw more attention in near future.

Besides, the dimension reduction strategy can be used to protect machine learning models from evasion attacks [95]. This defensive strategy aimed to enhance the resilience of classifiers by reducing the dimension of sample features. Experimental results validated that the defensive strategy was effective to protect multiple types of machine learning models such as SVMs and DNNs. Statistical test is also effective to distinguish the distribution of adversarial samples from that of legitimate ones [96]. Specifically, Grosse *et al.* [96] proposed two statistical metrics, named Maximum Mean Discrepancy (MMD) and the Energy Distance (ED), and applied them to measure the statistical properties of sample distributions. The experiments on MNIST, DREBIN[9] and MicroRNA[10] showed that this method could effectively detect adversarial samples. Furthermore, many researchers proposed some ensemble methods to improve the security and the robustness of learning algorithms [97]–[99]. For example, Sengupta *et al.* [97] proposed an ensemble framework that effectively combined multiple DNNs to defend against adversarial attacks. Furthermore, the framework was also scalable to integrating multiple defensive techniques.

D. DATA SECURITY AND PRIVACY

In this part, we focus on the security and privacy of data themselves. As the world entering big data era, modern classifier models (especially DNNs) require a high volume of data for being trained. Considering that crowdsourcing has been emerging as a main route of data collection, it suffers from a high possibility of the leakage of sensitive and privacy information, e.g., photos, videos, identity data, medical records, etc. Moreover, a data collector may save the information for a long time. Hence, it is vital to figure out how to effectively protect data security and privacy with the presence of various attacks, for example eavesdropping and reverse engineering.

[7] http://yann.lecun.com/exdb/mnist/
[8] http://www.cs.toronto.edu/~kriz/cifar.html
[9] http://www.sec.cs.tu-bs.de/ danarp/drebin/index.html
[10] http://www.mirbase.org/

TABLE 3. Comparison of different defensive techniques of machine learning.

Defensive Techniques	Advantages	Disadvantages
Reject on Negative Impact (RONI) [80]	1) It effectively removes adversarial samples that are injected into training data. 2) It scales to a variety of classifiers.	It is lack of extensive performance evaluation in a variety of application scenarios.
Adversarial training [91]	1) It is easy to understand and implement. 2) It scales to a variety of classifiers.	Its effectiveness depends on the adversarial samples in the training phase.
Defense distillation [93] [94]	1) It obtains a smoother DNN model by reducing its sensitivity regarding input perturbations. 2) It improves the generalization capability of a DNN. 3) It effectively mitigates adversarial samples crafted by FGSM.	It is weak to defend against adversarial samples crafted by JSMA.
Ensemble method [97] [98] [99]	It is flexible to integrate multiple classifiers or different defensive methods.	It is not robust to adversarial samples with transferability.
Differential privacy [100] [103] [104]	1) It preserves the privacy of training data. 2) It preserves the privacy of learning algorithms.	It influences the performance of classifiers on legitimate data.
Homomorphic encryption [106] [108]	It preserves data security and privacy in cloud environments.	It induces extensive computation overheads.

Basically, cryptographic technology is generally used to protect data privacy. Differential privacy (DP) is a specific technique of preserving data privacy via data encryption [100]. In the DP model, the calculating results on a specific data set are not sensitive to the change of one data record. Therefore, the risk of privacy leakage after adding a new data record is controlled in a very small and controllable region. In other words, potential adversaries are unable to obtain accurate user privacy information via the calculating results of DP. Comparing to conventional privacy preserving models, DP gains the following two advantages: (1) The DP model assumes that an attacker has the full knowledge of data records except for the target record, which can be regarded as the maximal background knowledge known by the attacker. With this assumption, it is no need for the DP model to consider what extent the attacker's knowledge is. (2) The DP model is built on the basis of solid mathematical foundation. It strictly defines privacy preserving and provides well-defined evaluation methods. Such feature makes the comparability of privacy preserving values under different settings of parameters be feasible. Hence, DP is becoming an active research subject in privacy protection. For example, Erlingsson *et al.* [101] proposed an anonymous and robust crowdsourcing method called RAPPOR, which integrated randomized response with DP to guarantee the privacy of crowdsourcing. Moreover, researchers have recently utilized DP to preserve the privacy of different learning algorithms, including SVM [102], deep learning [103] and Bayesian optimization [104].

Furthermore, homomorphic encryption (HE) is another technique to provide data privacy via data encryption [49]. Without decryption using private keys during calculation, HE has the following two merits: (1) Any type of calculation can be done on cipher text blocks; (2) The result after decrypting the calculating output on cipher text blocks is the same as the calculating result on corresponding plain blocks using the same operators. Therefore, HE is particularly suitable for use in protecting data security and privacy in cloud environments. On the basis of HE, many researchers devoted to study secure multi-party calculation [105], [106], classification on full HE data [107], distributed k-means clustering algorithms [108] and neural networks handling encrypted data [109].

Regardless of existing cryptgraphic mechanisms, reducing sensitive outputs of learning model APIs is an alternative idea of assuring data security and privacy [71].

E. SUMMARY OF DEFENSIVE TECHNIQUES OF MACHINE LEARNING

In this part, we compare advantages and disadvantages of existing defensive techniques of machine learning, as shown in Table 3. Basically, different defensive techniques can be used at different phases of the machine learning lifecycle to offer security support. For example, the RONI technique is effective to defend the training phase against adversaries; At the testing or the inferring phase, the adversarial training, defense distillation and ensemble method are valuable for security usage; Differential privacy and homomorphic encryption are two important solutions of addressing data security and privacy issues.

V. CHALLENGES AND FUTURE OPPORTUNITIES

Nowadays, machine learning is the core technology of big data, Internet of Things (IoT), cloud computing and artificial intelligence. Accordingly, various security threats and

TABLE 4. Comparative analysis with respect to different attacking and defensive techniques.

Attack/Defense	Technique	Targeting Learning Phase	Characteristics/Taxonomy	Basic Idea
Attack	Poisoning [6] [9] [25] [26] [28] [38] [43]–[45] [47]–[51]	Training	Causative attack Integrity/availability attack Targeted/indiscriminate attack	It may inject adversarial samples into training datasets. Also, it may modify the features or the labels of initial training dataset.
Attack	Evasion [35] [49] [52]–[55] [57] [58]	Testing	Exploratory attack Integrity/availability attack Targeted attack	It crafts adversarial samples to avoid detection of target systems.
Attack	Impersonate [8] [10] [11] [30]–[35] [59] [61] [63]–[67]	Testing	Exploratory attack Integrity/availability attack Targeted/indiscriminate attack	It crafts adversarial samples to imitate target ones or to confuse target systems.
Attack	Inversion [20] [68] [69] [70] [71]	Testing	Exploratory attack Privacy attack Targeted attack	It steals the sensitive information of target classifiers or datasets.
Defense	Data sanitization [80]	Training	Protect integrity/availability Active defense	It sanitizes training data and rejects the samples that will induce negative impacts to classifiers.
Defense	Adversarial training [91]	Training	Protect integrity/availability Active defense	It considers adversarial samples and corresponding labels when training, and it also does adversarial sample mining among datasets with noise.
Defense	Defense distillation [93]	Training	Protect integrity/availability Smooth classifier	It takes advantages of the probability label of training data generated by a DNN, then it retains the DNN with probability label.
Defense	Ensemble method [97] [98] [99]	Training	Protect integrity/availability Improve robustness	It integrates diffent classifiers or defensive techniques to mitigate adversarial samples.
Defense	Differential privacy [100] [103] [104]	Training & Testing	Protect privacy	It adds random noise to initial data or utilizes some randomized methods in the training phase.
Defense	Homomorphic encrytion [106] [108]	Training & Testing	Protect privacy	It is able to directly process encrypted data, and it can also protect the data privacy in multi-party computation environments.

corresponding defensive mechanisms of machine learning have drawn great attentions from both academia and industry. Table 4 presents the comparative results of qualitative analysis with respect to aforementioned attacking and defensive techniques.

Based on the existing literature, we argue that the research on security threats and defensive techniques of machine learning has the following trends:

(1) **New security threats towards machine learning are constantly emerging**. Although a large number of learning frameworks, algorithms and optimization mechanism have been proposed, studies on the security of learning models and algorithms are still at the beginning. Hence, there are a variety of attacks that threaten the security of machine learning techniques [6], [25], [28], [31]. On the other hand, statistical machine learning highly relies on the data quality, which is weak to defend against adversarial samples and incomplete data statistics. Furthermore, the difficulties of collecting and predicting adversarial samples significantly challenge the performance of identifying malicious samples using machine learning based detection methods. Hence, we argue that designing new adversary models is becoming a meaningful research point in the perspective of an attacker.

(2) **Security assessment on machine learning based decision systems in adversarial environments becomes a prevailing research area**. Basically, it is intuitive that a defender will show more interests in the security analysis of decision systems with a rapid increase of security events about machine learning. Currently, formally standardizing security assessment techniques on machine learning is still at the initial stage [19], [73]. Therefore, it is vital for establishing a widely adopted and well-defined security assessment standard.

(3) **Data privacy plays an important role in protecting the security of machine learning**. Regardless of a great advance of DP [101] and HE [49], [107]–[109], existing

privacy preserving methods suffer from low cost efficiency due to complex cryptographic operations on a large number of parameters of machine learning algorithms. Thus, highly efficient privacy preserving technology in adversarial environments is a meaningful topic that needs to be further studied.

(4) **Secure deep learning is a new growth point in the field of machine learning security**. Existing works have demonstrated that the counterintuitive characteristic of DNNs affects their security [29]. Regardless of the proposals of considering adversarial samples in training models [91] and improving the robustness of learning algorithms [92], [93], these solutions are still weak to address the above problem. Therefore, research on secure deep learning models is very interesting in near future, e.g., Bayes deep networks with prior information [110].

(5) **Jointly optimizing security, generalization performance and overhead is required to design secure learning algorithms**. Generally speaking, a higher level of security induces a larger overhead or even a lower generalization performance of learning algorithms, which challenges their application [111]. Hence, properly balancing the above three aspects since the design of secure machine learning algorithms is recommended to facilitate the practical usage.

VI. CONCLUDING REMARKS

As machine learning is becoming widely used in many practical applications including but not limited to image processing, natural language processing, pattern recognition, computer vision, intrusion detection, malware identification and autonomous driving, protecting the security of machine learning at both training and inferring phases becomes an urgent need. In this paper, we have presented a systematic survey on security concerns with a variety of machine learning techniques. Specifically, we have revisited existing security threats towards machine learning from two aspects, the training phase and the testing/inferring phase. Furthermore, we have categorized current defensive techniques of machine learning into security assessment mechanisms, countermeasures in the training phase, those in the testing or inferring phase, data security and privacy. After that, we have presented five interesting research topics in this field. Such survey can serve as a valuable reference for researchers in both machine learning and security fields.

REFERENCES

[1] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350–361, May 2017.
[2] S. Yu, "Big privacy: Challenges and opportunities of privacy study in the age of big data," *IEEE Access*, vol. 4, pp. 2751–2763, 2016.
[3] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, "Efficient machine learning for big data: A review," *Big Data Res.*, vol. 2, no. 3, pp. 87–93, 2015.
[4] X.-W. Chen and X. Lin "Big data deep learning: Challenges and perspectives," *IEEE Access*, vol. 2, pp. 514–525, 2014.
[5] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017.
[6] G. L. Wittel and S. F. Wu, "On attacking statistical spam filters," in *Proc. 1st Conf. Email Anti-Spam*, Mountain View, CA, USA, 2004, pp. 1–7.
[7] D. Lowd and C. Meek, "Good word attacks on statistical spam filters," in *Proc. 2nd Conf. Email Anti-Spam*, Stanford, CA, USA, 2005, pp. 1–8.
[8] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Vienna, Austria, 2016, pp. 1528–1540.
[9] B. Biggio, L. Didaci, G. Fumera, and F. Roli, "Poisoning attacks to compromise face templates," in *Proc. Int. Conf. Biometrics (ICB)*, Madrid, Spain, 2013, pp. 1–7.
[10] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. (2016). "Practical black-box attacks against machine learning." [Online]. Available: https://arxiv.org/abs/1602.02697
[11] N. Carlini *et al.*, "Hidden voice commands," in *Proc. 25th USENIX Secur. Symp.*, Austin, TX, USA, 2016, pp. 513–530.
[12] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, "Adversarial classification," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Seattle, WA, USA, 2004, pp. 99–108. [Online]. Available: https://homes.cs.washington.edu/~pedrod/papers/kdd04.pdf
[13] D. Lowd and C. Meek, "Adversarial learning," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Chicago, IL, USA, 2005, pp. 641–647.
[14] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?" in *Proc. ACM Symp. Inf., Comput. Commun. Secur.*, 2006, pp. 16–25.
[15] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. (Jul. 2016). "Concrete problems in AI safety." [Online]. Available: https://arxiv.org/abs/1606.06565
[16] N. Papernot, P. Mcdaniel, A. Sinha, and M. Wellman. (2016). "SoK: Towards the science of security and privacy in machine learning." [Online]. Available: https://arxiv.org/abs/1611.03814
[17] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP J. Adv. Signal Process.*, vol. 2016, Dec. 2016, Art. no. 67.
[18] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Mach. Learn.*, vol. 81, no. 2, pp. 121–148, 2010.
[19] B. Biggio, G. Fumera, and F. Roli, "Security evaluation of pattern classifiers under attack," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 4, pp. 984–996, Apr. 2014.
[20] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Denver, CO, USA, 2015, pp. 1322–1333.
[21] B. Biggio, "Machine learning under attack: Vulnerability exploitation and security measures," in *Proc. 4th ACM Workshop Inf. Hiding Multimedia Secur.*, Vigo, Spain, 2016, pp. 1–2.
[22] I. Corona, G. Giacinto, and F. Roli, "Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues," *Inf. Sci.*, vol. 239, pp. 201–225, Aug. 2013.
[23] S. Yu, G. Gu, A. Barnawi, S. Guo, and I. Stojmenovic, "Malware propagation in large-scale networks," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 1, pp. 170–179, Jan. 2015.
[24] N. Šrndić and P. Laskov, "Detection of malicious PDF files based on hierarchical document structure," in *Proc. 20th Annu. Netw. Distrib. Syst. Secur. Symp.*, San Diego, CA, USA, 2013, pp. 1–16.
[25] B. Biggio *et al.*, "Poisoning complete-linkage hierarchical clustering," in *Structural, Syntactic, and Statistical Pattern Recognition* (Lecture Notes in Computer Science), vol. 8621. Berlin, Germany: Springer, 2014, pp. 42–52.
[26] B. Biggio *et al.*, "Poisoning behavioral malware clustering," in *Proc. Workshop Artif. Intell. Secur. Workshop*, Scottsdale, AZ, USA, 2014, pp. 27–36.
[27] W. Zhao, J. Long, J. Yin, Z. Cai, and G. Xia, "Sampling attack against active learning in adversarial environment," in *Modeling Decisions for Artificial Intelligence MDAI* (Lecture Notes in Computer Science), vol. 7647. Berlin, Germany: Springer, 2012, pp. 222–233.
[28] B. Biggio, I. Pillai, S. R. Bulò, D. Ariu, M. Pelillo, and F. Roli, "Is data clustering in adversarial settings secure?" in *Proc. ACM Workshop Artif. Intell. Secur.*, Berlin, Germany, 2013, pp. 87–98.
[29] C. Szegedy *et al.* (Feb. 2014). "Intriguing properties of neural networks." [Online]. Available: https://arxiv.org/abs/1312.6199.
[30] N. Narodytska and S. Kasiviswanathan, "Simple black-box adversarial attacks on deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Honolulu, HI, USA, Jul. 2017, pp. 1310–1318.

[31] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 427–436.

[32] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2574–2582.

[33] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroSP)*, Saarbrucken, Germany, Mar. 2016, 372–387.

[34] Y. Liu, X. Chen, C. Liu, and D. Song. (Feb. 2017). "Delving into transferable adversarial examples and black-box attacks." [Online]. Available: https://arxiv.org/abs/1611.02770

[35] A. Kurakin, I. Goodfellow, and S. Bengio. (Feb. 2017). "Adversarial examples in the physical world." [Online]. Available: https://arxiv.org/abs/1607.02533

[36] B. Li, Y. Wang, A. Singh, and Y. Vorobeychik, "Data poisoning attacks on factorization-based collaborative filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1885–1893.

[37] S. Alfeld, X. Zhu, and P. Barford, "Data poisoning attacks against autoregressive models," in *Proc. 13th AAAI Conf. Artif. Intell. (AAAI)*, Phoenix, AZ, USA, 2016, pp. 1452–1458.

[38] B. Biggio, G. Fumera, F. Roli, and L. Didaci, "Poisoning adaptive biometric systems," in *Structural, Syntactic, and Statistical Pattern Recognition*. Berlin, Germany: Springer, 2012, pp. 417–425.

[39] B. Biggio, G. Fumera, and F. Roli, "Pattern recognition systems under attack: Design issues and research challenges," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 28, no. 7, pp. 1460002-1–1460002-21, 2014.

[40] X. Zhu, "Super-class discriminant analysis: A novel solution for heteroscedasticity," *Pattern Recognit. Lett.*, vol. 34, no. 5, pp. 545–551, 2013.

[41] M. Kloft and P. Laskov, "Security analysis of online centroid anomaly detection," *J. Mach. Learn. Res.*, vol. 13, pp. 3681–3724, Dec. 2012.

[42] Y. Chen *et al.* (Aug. 2017). "Practical attacks against graph-based clustering." [Online]. Available: https://arxiv.org/abs/1708.09056

[43] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proc. 29th Int. Conf. Int. Conf. Mach. Learn. (ICML)*, Edinburgh, Scotland, 2012, pp. 1467–1474.

[44] C. Burkard and B. Lagesse, "Analysis of causative attacks against SVMs learning from data streams," in *Proc. 3rd ACM Int. Workshop Secur. Privacy Anal.*, Scottsdale, AZ, USA, 2017, pp. 31–36.

[45] C. Yang, Q. Wu, H. Li, and Y. Chen. (Mar. 2017). "Generative poisoning attack method against neural networks." [Online]. Available: https://arxiv.org/abs/1703.01340v1

[46] S. Mei and X. Zhu, "The security of latent Dirichlet allocation," in *Proc. 18th Int. Conf. Artif. Intell. Statist.*, San Diego, CA, USA, 2015, pp. 681–689.

[47] H. Xiao, B. Biggio, G. Brown, G. Fumera, C. Eckert, and F. Roli, "Is feature selection secure against training data poisoning?" in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn. (ICML)*, Lille, France, 2015, pp. 1689–1698.

[48] M. Mozaffari-Kermani, S. Sur-Kolay, A. Raghunathan, and N. K. Jha, "Systematic poisoning attacks on and defenses for machine learning in healthcare," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 6, pp. 1893–1905, Nov. 2015.

[49] W. Hu and Y. Tan. (Feb. 2017). "Generating adversarial malware examples for black-box attacks based on GAN." [Online]. Available: https://arxiv.org/abs/1702.05983

[50] M. Zhao, B. An, W. Gao, and T. Zhang, "Efficient label contamination attacks against black-box learning models," in *Proc. 26th Int. Joint Conf. Artif. Intell. (IJCAI)*, Melbourne, VIC, Australia, 2017, pp. 3945–3951.

[51] H. Xiao, B. Biggio, B. Nelson, H. Xiao, C. Eckert, and F. Roli, "Support vector machines under adversarial label contamination," *Neurocomputing*, vol. 160, pp. 53–62, Jul. 2015.

[52] F. Tramèr, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. (May 2017). "The space of transferable adversarial examples." [Online]. Available: https://arxiv.org/abs/1704.03453

[53] N. Papernot, P. McDaniel, and I. Goodfellow. (May 2016). "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples." [Online]. Available: https://arxiv.org/abs/1605.07277

[54] I. Rosenberg, A. Shabtai, L. Rokach, and Y. Elovici. (Dec. 2017). "Generic black-box end-to-end attack against state of the art API call based malware classifiers." [Online]. Available: https://arxiv.org/abs/1707.05970

[55] B. Biggio *et al.*, "Evasion attacks against machine learning at test time," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases (ECML PKDD)*, Prague, Czech Republic, 2013, pp. 387–402.

[56] F. Zhang, P. P. K. Chan, B. Biggio, D. S. Yeung, and F. Roli, "Adversarial feature selection against evasion attacks," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 766–777, Mar. 2016.

[57] B. Li and Y. Vorobeychik, "Feature cross-substitution in adversarial classification," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, 2014, pp. 2087–2095.

[58] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel. (Jun. 2016). "Adversarial perturbations against deep neural networks for malware classification." [Online]. Available: https://arxiv.org/abs/1606.04435

[59] I. Masha, W. Yu, and B. Bahman. (2016). *Adversarial Attacks on Image Recognition*. [Online]. Available: http://cs229.stanford.edu/proj2016/report/ItkinaWu-AdversarialAttackonImageRecognition-report.pdf

[60] D. Maiorca, I. Corona, and G. Giacinto, "Looking at the bag is not enough to find the bomb: An evasion of structural methods for malicious PDF files detection," in *Proc. 8th ACM SIGSAC Symp. Inf., Comput. Commun. Secur. (ASIA CCS)*, Hangzhou, China, 2013, pp. 119–130.

[61] W. Xu, Y. Qi, and D. Evans, "Automatically evading classifiers: A case study on PDF malware classifiers," in *Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, San Diego, CA, USA, 2016, pp. 1–15.

[62] H. G. Kayacik, A. N. Zincir-Heywood, and M. I. Heywood, "Automatically evading IDS using GP authored attacks," in *Proc. IEEE Symp. Comput. Intell. Secur. Defense Appl.*, Honolulu, HI, USA, Apr. 2007, pp. 153–160.

[63] N. Rndic and P. Laskov, "Practical evasion of a learning-based classifier: A case study," in *Proc. IEEE Symp. Secur. Privacy*, San Jose, CA, USA, May 2014, pp. 197–211.

[64] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Secur. (ASIA CCS)*, Abu Dhabi, United Arab Emirates, 2017, pp. 506–519.

[65] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy*, San Jose, CA, USA, May 2017, pp. 39–57.

[66] K. R. Mopuri, U. Garg, and R. V. Babu. (Jul. 2017). "Fast feature fool: A data independent approach to universal adversarial perturbations." [Online]. Available: https://arxiv.org/abs/1707.05572

[67] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 86–94.

[68] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy*, San Jose, CA, USA, May 2017, pp. 3–18.

[69] X. Wu, M. Fredrikson, S. Jha, and J. F. Naughton, "A methodology for formalizing model-inversion attacks," in *Proc. IEEE 29th Comput. Secur. Found. Symp.*, Lisbon, Portugal, Jun./Jul. 2016, pp. 355–370.

[70] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *Proc. USENIX Secur. Symp.*, San Diego, CA, USA, Aug. 2014, pp. 17–32.

[71] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction APIs," in *Proc. 25th USENIX Secur. Symp.*, Austin, TX, USA, 2016, pp. 601–618.

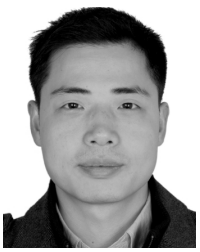
[72] A. Fawzi, O. Fawzi, and P. Frossard, "Analysis of classifiers' robustness to adversarial perturbations," *Mach. Learn.*, vol. 107, no. 3, pp. 481–508, 2017.

[73] B. Biggio *et al.*, "Security evaluation of support vector machines in adversarial environments," in *Support Vector Machines Applications*, Y. Ma and G. Guo, Eds. New York, NY, USA: Springer, 2014, pp. 105–153.

[74] S. Rizzi, "What-if analysis," in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds. Boston, MA, USA: Springer, 2009, pp. 3525–3529.

[75] P. Laskov and M. Kloft, "A framework for quantitative security analysis of machine learning," in *Proc. 2nd ACM Workshop Secur. Artif. Intell.*, Chicago, IL, USA, 2009, pp. 1–4.

[76] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proc. 4th ACM Workshop Secur. Artif. Intell.*, Chicago, IL, USA, 2011, pp. 43–58.
[77] B. I. P. Rubinstein *et al.*, "ANTIDOTE: Understanding and defending against poisoning of anomaly detectors," in *Proc. 9th ACM SIGCOMM Conf. Internet Meas.*, Chicago, IL, USA, 2009, pp. 1–14.
[78] B. Biggio, G. Fumera, and F. Roli, "Multiple classifier systems for robust classifier design in adversarial environments," *Int. J. Mach. Learn. Cybern.*, vol. 1, nos. 1–4, pp. 27–41, 2010.
[79] B. Biggio, I. Corona, G. Fumera, G. Giacinto, and F. Roli, "Bagging classifiers for fighting poisoning attacks in adversarial classification tasks," in *Proc. 10th Int. Conf. Multiple Classifier Syst. (MCS)*, Naples, Italy, 2011, pp. 350–359.
[80] B. Nelson *et al.*, "Misleading learners: Co-opting your spam filter," in *Machine Learning in Cyber Trust: Security, Privacy, and Reliability*. Boston, MA, USA: Springer, 2009, pp. 17–51.
[81] R. Laishram and V. V. Phoha. (Jun. 2016). "Curie: A method for protecting SVM classifier from poisoning attack." [Online]. Available: https://arxiv.org/abs/1606.01584
[82] A. Demontis *et al.*, "Yes, machine learning can be more secure! A case study on android malware detection," *IEEE Trans. Depend. Sec. Comput.*, to be published.
[83] A. Globerson and S. Roweis, "Nightmare at test time: Robust learning by feature deletion," in *Proc. 23rd Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, 2006, pp. 353–360.
[84] C. H. Teo, A. Globerson, S. Roweis, and A. J. Smola, "Convex learning with invariances," in *Proc. 20th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, 2007, pp. 1489–1496.
[85] M. Brückner and T. Scheffer, "Stackelberg games for adversarial prediction problems," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Diego, CA, USA, 2011, pp. 547–555.
[86] M. Brückner, C. Kanzow, and T. Scheffer, "Static prediction games for adversarial learning problems," *J. Mach. Learn. Res.*, vol. 13, pp. 2617–2654, Sep. 2012.
[87] S. R. Bulò, B. Biggio, I. Pillai, M. Pelillo, and F. Roli, "Randomized prediction games for adversarial machine learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2466–2478, Nov. 2017.
[88] W. Xu, D. Evans, and Y. Qi. (Dec. 2017). "Feature squeezing: Detecting adversarial examples in deep neural networks." [Online]. Available: https://arxiv.org/abs/1704.01155
[89] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner. (Nov. 2017). "Detecting adversarial samples from artifacts." [Online]. Available: https://arxiv.org/abs/1703.00410
[90] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. (Feb. 2017). "On detecting adversarial perturbations." [Online]. Available: https://arxiv.org/abs/1702.04267
[91] I. J. Goodfellow, J. Shlens, and C. Szegedy. (Mar. 2015). "Explaining and harnessing adversarial examples." [Online]. Available: https://arxiv.org/abs/1412.6572v3
[92] S. Gu and L. Rigazio. (Apr. 2015). "Towards deep neural network architectures robust to adversarial examples." [Online]. Available: https://arxiv.org/abs/1412.5068v4
[93] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE Symp. Secur. Privacy*, San Jose, CA, USA, May 2016, pp. 582–597.
[94] N. Carlini and D. Wagner. (Jul. 2016). "Defensive distillation is not robust to adversarial examples." [Online]. Available: https://arxiv.org/abs/1607.04311
[95] A. N. Bhagoji, D. Cullina, C. Sitawarin, and P. Mittal. (Nov. 2017). "Enhancing robustness of machine learning systems via data transformations." [Online]. Available: https://arxiv.org/abs/1704.02654
[96] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel. (Oct. 2017). "On the (statistical) detection of adversarial examples." [Online]. Available: https://arxiv.org/abs/1702.06280
[97] S. Sengupta, T. Chakraborti, and S. Kambhampati. (Sep. 2017). "MTDeep: Boosting the security of deep neural nets against adversarial attacks with moving target defense." [Online]. Available: https://arxiv.org/abs/1705.07213
[98] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. (Jan. 2018). "Ensemble adversarial training: Attacks and defenses." [Online]. Available: https://arxiv.org/abs/1705.07204
[99] M. Abbasi and C. Gagné. (Mar. 2017). "Robustness to adversarial examples through an ensemble of specialists." [Online]. Available: https://arxiv.org/abs/1702.06856
[100] C. Dwork, "Differential privacy," in *Proc. 33rd Int. Colloq. Automata, Lang. Programm. (ICALP)*, Venice, Italy, 2006, pp. 1–12.
[101] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Scottsdale, AZ, USA, 2014, pp. 1054–1067.
[102] B. I. P. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft, "Learning in a large function space: Privacy-preserving mechanisms for SVM learning," *J. Privacy Confidentiality*, vol. 4, no. 1, pp. 65–100, 2012.
[103] M. Abadi *et al.*, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS')*, Vienna, Austria, 2016, pp. 308–318.
[104] M. Kusner, J. Gardner, R. Garnett, and K. Weinberger, "Differentially private Bayesian optimization," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lille, France, 2015, pp. 918–927.
[105] Q. Wang, W. Zeng, and J. Tian, "Compressive sensing based secure multiparty privacy preserving framework for collaborative data-mining and signal processing," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Chengdu, China, Jul. 2014, pp. 1–6.
[106] I. Damgård, V. Pastro, N. Smart, and S. Zakarias, "Multiparty Computation from Somewhat Homomorphic Encryption," in *Proc. 32nd Annu. Cryptol. Conf. Adv. Cryptol. (CRYPTO)*, 2012, pp. 643–662.
[107] L. J. M. Aslett, P. M. Esperança, and C. C. Holmes. (Aug. 2015). "Encrypted statistical machine learning: New privacy preserving methods." [Online]. Available: https://arxiv.org/abs/1508.06845
[108] Y.-C. Yao, L. Song, and E. Chi, "Investigation on distributed K-means clustering algorithm of homomorphic encryption," *Comput. Technol. Develop.*, vol. 2, pp. 81–85, 2017.
[109] N. Dowlin, R. Gilad-Bachrach, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing, "CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy," in *Proc. 33rd Int. Conf. Mach. Learn.*, New York, NY, USA, 2016, pp. 201–210.
[110] H. Wang and D.-Y. Yeung, "Towards Bayesian deep learning: A framework and some existing methods," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3395–3408, Dec. 2016.
[111] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, Athens, Greece, 2011, pp. 193–204.



QIANG LIU (M'14) received the Ph.D. degree in computer science and technology from the National University of Defense Technology (NUDT) in 2014. From 2011 to 2013, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, The University of British Columbia (UBC), Canada. He is currently an Assistant Professor at NUDT. He has contributed over 50 archived journal and international conference papers, such as the *IEEE Network Magazine*, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *Pattern Recognition*, the IEEE COMMUNICATIONS LETTERS, *Neurocomputing, Neural Computing and Applications*, *Mobile Information Systems*, EDBT'17, WCNC'17, ICANN'17, and SmartMM'17. His research interests include 5G network, Internet of Things, wireless network security, and machine learning. He is a member of China Computer Federation (CCF). He currently serves on the Editorial Review Board of *Artificial Intelligence Research Journal*.

PAN LI received the B.Eng. degree from the University of Science and Technology Beijing in 2016. He is currently pursuing the M.S. degree with the National University of Defense Technology. His research interests include machine learning and cyber security.

WENTAO ZHAO received the Ph.D. degree from the National University of Defense Technology (NUDT) in 2009. He is currently a Professor at NUDT. His research interests include network performance optimization, information processing, and machine learning. Since 2011, He has been serving as a member of the Council Committee of Postgraduate Entrance Examination of computer science and technology, NUDT. He has edited one book *Database Principle and Technology* and several technical papers, such as Communications of CCF, WCNC'17, ICANN'17, WF-IoT, MDAI, and FAW.

WEI CAI (S'12–M'16) received the B.Eng. degree in software engineering from Xiamen University in 2008, the M.S. degree in electrical engineering and computer science from Seoul National University in 2011, and the Ph.D. degree in electrical and computer engineering from The University of British Columbia (UBC), Vancouver, BC, Canada, in 2016. He has completed visiting researches at Academia Sinica, The Hong Kong Polytechnic University, and the National Institute of Informatics, Japan. He is currently a Post-Doctoral Research Fellow at UBC. His research interests include gaming as a service, mobile cloud computing, online gaming, software engineering, and interactive multimedia. He was a recipient of the 2015 Chinese Government Award for Outstanding Self-Financed Students Abroad, the UBC Doctoral Four-Year-Fellowship, the Brain Korea 21 Scholarship, and the Excellent Student Scholarship from Bank of China. He was also a co-recipient of Best Paper Awards from CloudCom2014, SmartComp2014, and CloudComp2013.

SHUI YU (M'05–SM'12) is currently a Senior Lecturer with Deakin University Melbourne Burwood Campus. He has published two monographs and edited two books, over 180 technical papers, including top journals and top conferences, such as the IEEE TPDS, TC, TIFS, TMC, TKDE, TETC, TBD, and the IEEE INFOCOM. His research interests include cybersecurity, networking, big data, and mathematical modeling. He initiated the research field of networking for big data in 2013. He is a member of AAAS and ACM, the Vice Chair of the Technical Committee on Big Data Processing, Analytics, and Networking of the IEEE Communication Society. He is a Voting Panel Member for the China Natural Science Foundation Projects in 2017. He has served many international conferences as a member of the Organizing Committee, such as a Publication Chair for IEEE GLOBECOM 2015 and 2017, and IEEE INFOCOM 2016 and 2017; a TPC Co-Chair for the IEEE BigDataService 2015, the IEEE ATNAC 2014, and the IEEE ITNAC 2015; and an Executive General Chair for ACSW2017. He actively serves his research communities in various roles. He is currently serving in the Editorial Boards of the IEEE Communications Surveys and Tutorials, the IEEE Access, the IEEE Communications Letters, the IEEE Journal of Internet of Things, the *IEEE Communications Magazine*, the IEEE Transactions on Computational Social Systems. His h-index is 27.

VICTOR C. M. LEUNG (S'75–M'89–SM'97–F'03) is currently a Professor of electrical and computer engineering and holds the TELUS Mobility Research Chair with The University of British Columbia (UBC). He has co-authored over 900 technical papers in archival journals and refereed conference proceedings, several of which had received best-paper awards. His research interests include wireless networks and mobile systems. He is a fellow of the Royal Society of Canada, the Canadian Academy of Engineering, and the Engineering Institute of Canada. He was a recipient of the 1977 APEBC Gold Medal, the NSERC Postgraduate Scholarships from 1977 to 1981, a 2012 UBC Killam Research Prize, and an IEEE Vancouver Section Centennial Award. He has provided leadership to the technical program committees and organizing committees of numerous international conferences. He is serving or has served on the Editorial Boards of the IEEE Access, the IEEE Journal on Selected Areas in Communications–Series on Green Communications and Networking, the IEEE Wireless Communications Letters, the IEEE Transactions on Computers, the IEEE Transactions on Wireless Communications, the IEEE Transactions on Vehicular Technology, and several other journals.