JES
Journal of Engineering Sciences

Vol 11, Issue 4, APRIL /2020
ISSN NO:0377-9254

A Study of Security Threats and Defense Methods on Machine Learning Training Datasets

Mrs. D. Suja Mary.,M.Sc.,M.Phil.,
Part time Research Scholar, University of Madras,
Assistant Professor, Department of Computer Applications, J.H.A Agarsen College, Madhavaram, Chennai-60
Email: dsuja2004@yahoo.com

Dr. M. Suriakala., M.Sc.,M.Phil.,Ph.D.,
Assistant Professor, Department of Computer Science, Government Arts College for Men, Nandanam, Chennai-35
Email: suryasubash@gmail.com

ABSTRACT

Training datasets are available in open source and these training datasets are used by the researchers in various discipline. To implement their research in those datasets are to simulate their results in various measures like comparative study, accuracy and prediction. But the reliability of these datasets are questionable and they are subject to various attack. Machine learning algorithm results are strongly manipulated when the learning algorithm's evaluated the causative attack, evasion attack and membership inference attacked training datasets. So the training datasets of image datasets, Natural language datasets, and pattern recognition datasets are to be protecting from various attacks. In this paper concentrate the overview of training datasets attacks and provide the security to the real training dataset.

KEYWORDS: Machine Learning, Causative Attack, Evasion Attack, Adversarial Examples

1. INTRODUCTION

Machine learning engines process massive amounts of data in near real time to discover critical incidents. Machine learning used by many companies for they improve their product advertising, marketing and better understand their service to customers through training datasets generated by their learning algorithms [1]. A variety of learning algorithm exists in the field of machine learning to work with relevant training Dataset context [2]. The training datasets are becomes the main driver to promote predictive model services, e.g., spam-filtering, mobile voice. A security based machine learning inference to the training datasets the classifier provides the defense [3]. The security threats lack in machine learning for dealing sufficient training datasets [10].

Training documents are using the techniques Naive Bayes Classifier and Support vector machine to train the data introduces the attack to poor accurate classification [4].

2. MACHINE LEARNING PARADIGMS FOR TRAINING DATASETS

A machine learning paradigms state that how the machine learns when some data is given to it, it's correlated with data and predicted outcome. The machine learning algorithms need well defined training datasets to predict the correct output. Through the training datasets the particular problems can be understand and it can be solved.

1.1 Machine Learning

Machine learning algorithms solve certain task with supervised or unsupervised training datasets. ML used Ensemble learning models which solve the task with mixing of some different simple models. Semi-supervised learning is a part of machine learning; it combines both supervised and unsupervised learning techniques.

1.2 Deep Learning

Deep learning is an unsupervised learning represented by family of machine learning algorithms brilliant than human brains, learn the huge amount of quantity. Deep learning is referred as different multi tasks [6] such as the fields of s recognition of speech, NL processing, audio files, social media filtering, System translation, inspection of data collection and board gaming applications.

In deep learning the features are learned directly from the data, there is no need for feature engineering. There is no need to create raw data in feature use. In the context of datasets, the ability to avoid feature engineering is regarded as a challenge is a great advantage of this learning process.

1.3 Local Learning

Local learning is a strategy that offers an alternative to typical global learning [2]. Usually, Machine Learning algorithms make use of global learning through the approaches such as generative learning. This approach uses that based upon the data's following distribution, a model can be used to re-generate the input data. It basically attempts to summarize the entire dataset, whereas local learning is concerned only with subsets of interest. For this reason, local learning observed as a semi parametric value of a global model. The stronger but less restrictive assumptions of this hybrid parametric model yield low variance and bias. Local learning often yields better results than global learning when dealing with imbalanced datasets.

The training system used the Local learning algorithms, it locally adjust the capacity of the system's properties of each area of the input space. The local area algorithm contains known methods, like K-Nearest Neighbors Method (KNN) or the Radial Basis Function networks (RBF) as well as new algorithms [5]. These algorithms do not suggest that, nor non local classifiers, achieve the best compromise between locality and capacity.

1.4 Transfer Learning

Transfer learning is a research problem in machine learning that focuses on storing knowledge gained while solving one problem and applying it to different but related problems. This learning is an approach for improving learning in a particular domain, referred to as the target domain, by training the model with other datasets from multiple domains, denoted as source domains, with similar attributes or features, such as the problem and constraints. This type of learning is used when the data size within the target domain is insufficient or the learning task is different [2].

1.5 Life Long Learning

Lifelong learning mimics human learning; learning is continuous; knowledge is retained and used to solve different problems. Lifelong learning algorithms directed to increase overall learning, to be able to reach a new task by training either on one single domain or on heterogeneous domains collectively [2]. The learning algorithms outcomes from the training datasets process are gathered and mixed together in a problem space are known as topic or knowledge model.

3. ATTACKS AGAINST MACHINE LEARNING TRAINING DATASETS

The Vulnerable activities are happened in machine learning algorithms which implement and display output using the training datasets. Machine Learning algorithms generate information using the training datasets, the learning algorithms are classified into supervised, unsupervised, semi-supervised and reinforcement [8][9].

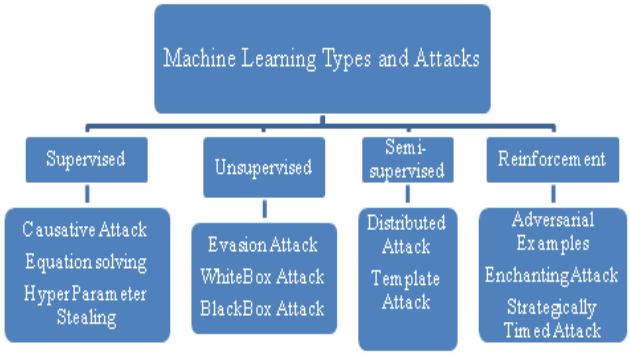

Figure 1.Machine Learning Types and Attacks

a. Supervised Learning Attacks

Information is available directly in the training dataset, such as class labels, numerical outputs and so on. Supervised learning algorithms determined training datasets and produce exact results. The learning algorithms misleading the results when it analyze the attacked training datasets likes causative attack, equation solving attack and Hyper parameter stealing attack.

In Causative or Poisoning attack the adversary Provides incorrect information to the machine learning algorithm [11]. The training set datas are assumed to be unprotected in physical manner. A machine learning adversary may take advance to access the training set data and provide an attack through programming languages [12]. The objective of the adversary is to gain access to the training datasets and eventually steal machine learning inferred functions result. The causative attack is to send attacked data to learning algorithms process, in turn forcing the algorithms to make wrong decisions. The causative attacker can intrude, modify, replay, and inject datas into the real training dataset.

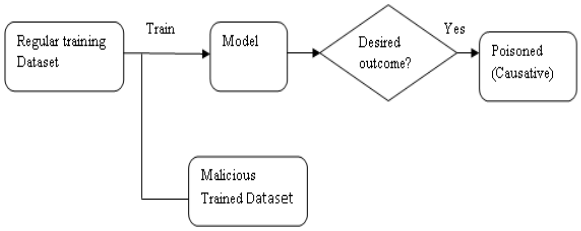

Figure 2. Causative Attack

In the figure2 shows the causative attackers to learn about training dataset and then corrupted data applied in machine Learning model itself.

Equation-solving extraction attacks [15][16] are not adaptive for the training datasets, to solve the parameters of a training target model using random queries. In machine learning models, this type of attack affect the original confidence output. The variables represented in the mathematical equations, the attacker feed unknown values to the variable. For example, the Binary Logistic Equation(BLR) use the equation:

$$\int_i(\chi) = \sigma\ (\omega\ x\ \chi + \beta)$$

Attacker will feed χ_i value to the equation it solved and gives $\Upsilon_i = \int_i(\chi) = \sigma\ (\omega\ x\ \chi_i + \beta)$, but not we cannot estimate the correct answer of the equation.

In the Machine Learning algorithms the training set equations parameters minimizing the concept of the function, which is loss the regularization of function. Loss function performs over the training dataset regularization rules and used to prevent machine learning algorithms over fitting, and the parameter balances between ML Training datasets and model's parameters. In the Hyper parameter stealing attacks, without the prior knowledge of an adversary using black box access to seal the

functionality of the model the ML Training datasets and model's parameters [17]. The drawback of supervised training dataset is, is has the limited amount of labeled data.

3.2 Unsupervised Learning Attacks

This learning does not require the availability of supervised information such as class labels or numerical outputs. It can be used when there are no labeled data and the model should somehow mark it by itself based on the properties. Usually it is considered to find anomalies in data and find to be more powerful in general as it's almost impossible to mark all data. The unsupervised learning algorithms affected by various attacks like evasion attack, white box attack and black box attack.

The adversary trained the sample datasets to make fool the machine learning algorithm accepting wrong decisions is known as Evasion Attack [13][14]. An attacker to make a small crafted noise in the machine learning classification testing time, the classifier prediction lead incorrect result.

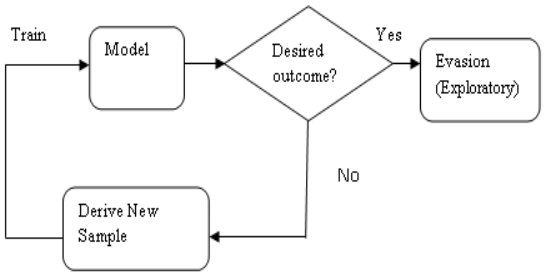

Figure 3. Evasion Attack

In the above figure3 described the attacker trained the real training datasets in machine learning models and then received the correct output. Based on the output the adversary planned to attack the training dataset and retrieved the new output. The adversary repeated the process until his expectation satisfied. The researcher handles this method in their research field and they apply evasion attacked training datasets in machine learning algorithms. Also they perfectly show their datasets more accuracy predicted.

White-box attacks access various angle of target to modify training dataset model and its parameters [21]. The white-box attack changed the input training dataset meanings, but not changing the output model while access using gradient functions. It is helpful to make Black-box attack [25].

The black-box ML attacks which affect malicious input samples that fool the classification without knowing the architecture used [19][20]. The unsupervised machine learning algorithm used black-box attack when the adversaries have no knowledge about training datasets and ML models [21]. The unsupervised training dataset disadvantage is that all data present in the datasets are mark as labelled data is almost impossible.

3.3 Semi-Supervised Learning Attacks

The semi-supervised learning refers to the learning tasks that combine the benefits from both supervised and unsupervised learning approaches and there output has some labeled data. Consider the Labeled Training data as $T^l = (x^l1, x^l2,\ldots, x^ln)$ and unlabeled Training data as $T^u = (x^u1, x^u2,\ldots, x^un)$, where T^l and T^u formed semi-supervised model, which implies both training data as $T^l < T^u$. Semi-supervised learning formed using two methods self-training method and co-training model. The distributed attack and template attack models are modify the machine learning training datasets.

Distributed Attacks launched an attack towards one or more target datasets [27]. It is the type of machine learning Denial Service attack [28]. The first training dataset received from the network, it is not affected by an attack. The second training dataset received from network, which has affected by an attack. The distributed attack combines both two training datasets and pass through the machine learning for train [29].

The template attack makes statistical modeling side-channel attacks, the conditional probability lead to trace for each key in training datasets in a parametric manner [30][31]. Template attacks are implemented in large information dataset available in traces attack device and clone device [32]. The attacks detection in semi supervised training datasets is difficult.

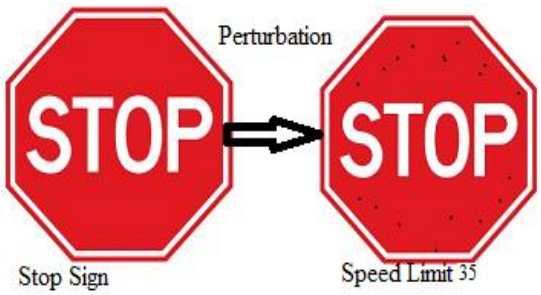

Figure 4. Adversarial Example

3.4 Reinforcement Learning Attack

Environment Driven technique can be used, the behavior suitable action taken on the changing environment in a particular situation. It's like software to find best behavior learning environment by trial and error. Reinforcement Learning applied in various fields is autonomous driving and automated trading. The adversarial attack, enchanting attack and Strategically Timed Attack models are attacked the training datasets and the manipulated data trained in reinforcement machine learning algorithms.

An adversary makes attacks in machine learning data instances called adversarial examples [33][34]. Adversarial example takes action in the victim's training datasets. Adversarial attacks classified into Misclassification attacks and Targeted attacks [38]. Misclassified adversarial attacks modify the input training dataset and make it wrong decision boundary. The Targeted adversarial attacks are focus on small part of training dataset [39]. The well frame worked adversarial model constructed with aim, knowledge, capability of attack and attacking plan.

Enchanting Attack, the adversary attracted the agent using generative and planning model algorithms [36]. The goal achieved by to attract the agent from current state S_c and target states S_t after apply the steps. The first planned action get the result $S_c + S_t$. The agent get new state $S_c + 1$. Finally, the steps repeated and the crafted as $S_c + 1$, $S_c + 2$,…., $S_c + H$.

Strategically Timed Attack, the agent used crafted dataset, but the attack undetected [37]. The adversary attacks the dataset by small time steps. Strategically timed attack has the sequence states using different R values. Every R value returned a different timed attack rate.

4. TRAINING DATASET ATTACK AND SUPPORTING ALGORITHMS

In this section summarize the different attack techniques on machine learning training datasets shown in the Table1. The attacks are targeted to training datasets, as a result it affect the overall performance of learning algorithms.

There are different kinds of attacks happening in the training datasets. The specified machine learning algorithms support training dataset classifications and they help to detect the attack in the training dataset through different kind of

Vol 11, Issue 4, APRIL /2020
ISSN NO:0377-9254

classification, prediction and accuracy results.

The Table1 listed the advantage of training dataset attackers and disadvantage of the users. It also described the attack techniques of the training datasets.

Training Dataset Attack	ML Algorithms used for training the datasets	Training Dataset Attack Techniques	Advantages to the Attackers	Disadvantage to the Users
Causative Attack	Support Vector Machine (SVM), Naive Bayes [41][43]	StingRay[40] Indiscriminate attack[41]	1. Achieves targeted poisoning. 2. Spam email misclassified as legitimate.	1. The sample crafting procedure is repeated until misclassification achieved. 2. The adversary maximize evade detection.
Equation Solving	Regression: binary logistic(BLR), multi-class logistic (MLR), and multi layer perception (MLP)[16]	Extraction Attack [15]	Remove rows with missing values	Apply one-hot-encoding to categorical features.
Hyper Parameter Stealing	Ridge Regression (RR), Logistic Regression (LR), Support Vector Machine (SVM), and Neural Network (NN) [15][17][18]	Non-kernel algorithms[17]	1. To prevent over fitting of machine learning. 2. The information to create and interpret predictions.	1. Incomplete queries some input features are left or unspecified. 2. The ML algorithm is unknown, the problem stealing parameters.
Evasion	Support Vector Machine (SVM), Naive Bayes (NB), Deep neural network(DNN) classifier [43][44]	Mimicry Attack, Gradient Descent Kernel Density Estimation (GDKDE) Attack, Deep Fool, Inverse attack[42][45][46]	1. To avoid moving to infeasible areas of the feature space with negative classifications 2. Losing the benefits of automated decision making. 3. Injecting malicious content into a benign PDF file.	1. To modify all of its value of features at once. 2. To generate adversarial examples in adversarial training.
White Box	Random Forest , Linear models, Neural Networks [23][24][25]	query-limited Attack, gradient-descent optimization	Results in a much more efficient attack whose run-time is independent of the	Help to create Black-box attack.

		[23][47]	optimization process.	
Black Box	Support Vector Machines (SVMs), Decision Trees, and K-Nearest Neighbor (KNN)[19][20][21]	Bayesian optimization attack[21]	The adversary does not has knowledge about the learning algorithm and training data	Attack requires access to a large enough dataset.
Distributed Attack	Support Vector Machines(SVM), Neural Classifiers, Markov Models, Genetic Algorithms, Artificial Neural Networks (ANN) and Bayesian Learning [27][28]	Stochastic Gradient Decent(SGD)[29]	Human interaction not necessary. Can deal with Zero-Day attacks.	The gradient decent parameter not efficient to over the whole available data.
Template Attack	Principal Component Analysis(PCA), Minimum redundancy maximum relevance (mRMR) filter algorithm, Support Vector Machines(SVM), Random Forests (RF) [30][31]	Naive Template Attack (NTA), Template Attack (ETA)[31]	It reduces the accuracy of Adversary or evaluator model.	Well-controlled simulated experimental setting in order to put forward two important training sets.
Adversarial Examples	Support Vector Machines(SVM), deep neural networks (DNNs), logistic regression [35][50]	Gradient Free Optimization, Fast Gradient Sign Method (FGSM)[48][49]	An attacker designed mistakes on training datasets.	Limited queries and information.
Enchanting Attack	Deep RL algorithms [51]	generative model, planning algorithm[37]	Generating the planned action sequence make to generate next frame prediction model	Full control of the agent to take arbitrary actions at each step.
Strategically Timed Attack	Deep RL algorithms [51]	gradient-based (A3C), value-based methods (DQN)[36]	The lowest attack rates to reach the reward of uniform attack.	A stronger deep RL agent to need make an attack.

Table1. Machine learning training datasets Attacking Techniques.

5. SECURITY EVALUATION

The different defensive methods are available in machine learning algorithms security support. The important needs of data collection in research, a data collector save their data for their research work. Different data protection and privacy method help the data collectors to get original preserved datasets. The Table2 describes different defense method against Machine Learning Algorithms training datasets.

Defense Metohd/Algorithm	Advantage	Disadvantage
Negative Impact (RONI)[40]	1. Measures the incremental effect 2.Discards negative impact on the overall performance 3. Identify poisoning samples	1. Requires a sizable clean set for testing instances. 2. Computationally inefficient on trained classifiers scales linearly with the training set.
Region-based classification[44]	Uses randomization to defend against evasion attacks.	Randomization preprocessing applies randomization once.
Detecting Adversarial Examples[44]	Adjust its attacks to evade both the original classifier and the new classifier to detect adversarial examples.	A key limitation of detecting adversarial examples is that it is unclear how to handle the testing examples that are predicted to be adversarial examples.
Principal Component Analysis[27]	It works by building profiles from browsing activity of users.	Understanding the very own distributed nature of the attacks under study.
Pre-processing[49]	Filtering and removal of modifications introduced back to reach original images.	Applicable before classification
Regeneration[49]	Back to original clean data	It combines the detector and regeneration networks.

Table2. Different Defense Techniques of Machine Learning

6. FUTURE WORK AND CHALLENGES

The potential improvements of this research paper, to implement future research work. Here, the paper described the different kind of attacks and its vulnerable results. To avoid the attack against from various attack techniques on training datasets, the proposed future work to check the inputs training datasets properly before applied in the machine learning algorithm. The researcher's generate new well defined security methods to prevent datasets from attack. To use original datasets and real practical experiments shows the excellent classification accuracy and also shows security of training datasets against the attack. In future research work to develop to detect the attack and provide more protection to training datasets against the attacks.

7. CONCLUSION

The Machine Learning used in all area applications with security needs. This paper we have presented different training dataset attacks that have been used in learning algorithms. The existing defense methods against the machine learning attack specially listed. The Paper covered the details of learning, attacks on machine learning training datasets and security needs.

References

[1]R.Shokri, M.Stronati, C.Song, and V.Shmatikov, "Membership inference attacks against machine learning models," in Proceedings of the 2017 IEEE Symposium on Security and Privacy, USA, 2017, pp. 3–18.

[2] Alexandra L'heureux1 , Katarina Grolinger1 , Hany F. Elyamany And Miriam A. M. Capretz "Machine Learning With Big Data: Challenges and Approaches" IEEE Access VOLUME 5, 2017.

[3] Tegjyot Singh Sethi, Mehmed Kantardzic, and Joung Woo Ryu "'Security Theater': On the Vulnerability of Classifiers to Exploratory Attacks" Arxiv:1803.09163v1, Mar 2018.

[4] Lei Pi, Zhuo Lu, Yalin Sagduyu, Su Chen, "Defending Active Learning Against Adversarial Inputs In Automated Document Classification," IEEE GlobalISIP, 257- 261, 2016.

[5] Leon Bottou, Vladimir Vapnik "Local Learning Algorithms," AT & T Bell Laboratories, NJ 07733,USA.

[6] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y. Ng, "Multimodal Deep Learning, "Proceedings of the 28 th International Conference on Machine Learning, Bellevue, WA, USA, 2011.

[7] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," proceedings of the IEEE symposium CISDA 2009.
[8] Ozlem Yavanoglu, Murat Aydos, "A Review on Cyber Security Datasets for Machine Learning Algorithms" IEEE International Conference On Big Data, 2017.
[9] Vasisht Duddu, "A Survey of Adversarial Machine Learning in Cyber Warfare," Defence Science Journal, Vol. 68, No. 4, July 2018, pp. 356-366, DOI : 10.14429/dsj.68.12731.
[10] Ryan R curtain, Andrew B. Gardner, Slawomir Grzonkowski, Alexey Kleymenov, AlejandroMosquera "Detecting DGA domains with recurrent neural networks and side information," CODASPY'19, March 2019, Dallas, Texas 2019.
[11] Yi Shi and Yalin E. Sagduyu "Evasion and Causative Attacks with Adversarial Deep Learning" Milcom 2017 Track 3 - Cyber Security and Trusted Computing.
[12] Muhammad Naveed Aman, Kee Chaing Chua, and Biplab Sikdar. "Secure Data Provenance for the Internet of Things." In Proceedings of the 3rd ACM International Workshop on IoT Privacy, Trust, and Security (IoTPTS '17). ACM, NewYork, NY, USA, 11–14.
[13]Weilin Xu, Yanjun Qi, and David Evans, "Automatically Evading Classifiers" In Network and Distributed System Security Symposium 2016 (NDSS), San Diego, February 2016.
[14] Zeinab Khorshidpour, Sattar Hashemi, Ali Hamzeh, "A. Learning a secure classifier against evasion attack," IEEE 16th International Conference on Data Mining Workshop, 2016, DOI: 10.1109/ICDMW.2016.46.
[15] Florian Tramar, Fan Zhang, Ari Juels, Michael K. Reiter, Thomas Ristenpart, "Stealing Machine Learning Models via Prediction APIs" , 25th USENIX Security Symposium August 10–12, 2016, ISBN 978-1-931971-32-4.
[16] Tam N. Nguyen, "Attacking Machine Learning Models as a Part of a Cyber Kill Chain", arXiv:1705.00564v2, Apr 2018.
[17] Binghui Wang, Neil Zhenqiang Gong, "Stealing Hyperparameters in Machine Learning", 39th IEEE Symposium on Security and Privacy, May 2018, San Francisco.
[18] Arthur E. HOERL and Robert W. KENNARD, "Ridge Regression: Biased Estimation for Nonorthogonal Problems", Technometrics, Vol. 42, No. 1, Special 40th Anniversary Issue, Feb. 2000, pp. 80-86.
[19] AnshumanChhabra, AbhishekRoy, PrasantMohapatra, "Strong Black-box Adversarial Attacks on Unsupervised Machine Learning Models", arxiv: 1901.09493v2, Feb 2019.
[20] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, "Practical Black-Box Attacks against Machine Learning," ASIA CCS '17, April 02 - 06, 2017.
[21] Kenneth T. Co, "Bayesian Optimization for Black-Box Evasion of Machine Learning Systems," Department Of Computing, Imperial College London, sep 2017.
[22] Yotam Gil, Yoav Chai, Goro dissky and Jonathan Berant, "White-to-Black: Efficient Distillation of Black-Box Adversarial Attacks", Proceedings of NAACL-HLT 2019, pages 1373–1379.
[23] Andrew Ilyas, Logan Engstrom, Athalye, Jessy Lin, "Black-box Adversarial Attacks with Limited Queries and Information" , Proceedings of the 35 th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.
[24] Daniel Deutch, Nave Frost,"Explaining White-box Classifications to Data Scientists", Technical report.
[25] YotamGil, YoavChai, Or Gorodissky and Jonathan Berant, "White-to-Black: Efficient Distillation of Black-Box Adversarial Attacks", Aviv University, Allen Institute for Artificial Intelligence.
[26] Ying Gao, Yu Liu , Yaqia Jin, Juequan Chen, And Hongrui Wu, "A Novel Semi-Supervised Learning Approach for Network Intrusion Detection on Cloud-Based Robotic System", 2169-3536 IEEE, Volume 6, 2018.
[27] Sergio Armando Gutierrez, John Willian Branc, "Application of Machine Learning Techniques to Distributed Denial of Service (DDoS) Attack Detection: A Systematic Literature Review ", Universidad Nacional de Colombia.
[28] Stefan Seufert and Darragh O'Brien, "Machine Learning for Automatic Defence against Distributed Denial of Service Attacks", proceedings IEEE Communications, ICC 2007.
[29] Abebe Abeshu Diro, Naveen Chilamkurti, "Distributed Attack Detection Scheme using Deep Learning Approach for Internet of Things," Department of Computer Science and IT, La Trobe University, Melbourne, Australia.
[30] L. Lerman, G. Bontempi, O. Markowitch, "Power analysis attack: an approach based on machine learning ",Universit´e Libre de Bruxelles, Brussels, Belgium.
[31] Liran Lerman , Romain Poussier , Gianluca Bontempi, Olivier Markowitch and Fran¸cois-Xavier Standaert, "Template Attacks vs. Machine Learning Revisited" , Universit´e catholique de Louvain, Belgium.
[32] Liran Lerman, Stephane Fernandes Medeiros, Nikita Veshchikov, C´edric Meuter, Gianluca Bontempi,, and Olivier Markowitch, "Semi-Supervised Template Attack" , Springer-Verlag Berlin Heidelberg 2013, pp. 184–199, 2013.
[33] Adam Gleave, Michael Dennis, Neel Kant, Cody Wild, Sergey Levine, Stuart Russell, "Adversarial Policies: Attacking Deep Reinforcement Learning", arXiv: 1905.10615v1, May 2019.
[34] Tong Chen , Jiqiang Liu, Yingxiao Xiang, Wenjia Niu , Endong Tong and Zhen Han, "Adversarial attack and defense in reinforcement learning-from AI security view", Chen et al. Cybersecurity2019, 2:11 https://doi.org/10.1186/s42400-019-0027-x.
[35] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and JD Tygar, "Adversarial Machine Learning" 4th ACM Workshop on Artificial Intelligence and Security, October 2011.
[36] Yen-Chen Lin , Zhang-Wei Hong , Yuan-Hong Liao , Meng-Li Shih , Ming-Yu Liu , and Min Sun, "Tactics Of Adversarial Attack On Deep Reinforcement Learning Agent", Workshop track - ICLR 2017.
[37] Naveed Akhtar and Ajmal Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey", Journal Of Latex Class Files, Vol. PP, Aug 2017.
[38] Vasisht Duddu, "A Survey of Adversarial Machine Learning in Cyber Warfare", Defence Science Journal, Vol. 68, No. 4, July 2018, pp. 356-366, DOI : 10.14429/dsj.68.12731.
[39] Qiang Liu1, Pan Li, Wentao Zhao1, Wei Cai, Shui Yu, Victor C. M. Leung "A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View" VOLUME 4, 2016, 2169-3536
[40] Octavian Suciu, Radu Marginean, Yigitcan Kaya, Hal Daume, Tudor Dumitras, "When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks", Proceedings of the 27th USENIX Security Symposium. August 2018.
[41] Battista Biggio, Igino Corona, Blaine Nelson, Benjamin I. P. Rubinstein, Davide Maiorca, Giorgio Fumera, Giorgio Giacinto, and Fabio Roli, "SecurityEvaluationofSupportVectorMachines inAdversarialEnvironments", Published on Jan 1, 2014 in arXiv: Learning.· DOI :10.1007/978-3-319-02300-7_4.
[42] Nedim Srndi.c and Pavel Laskov, "Practical Evasion of a Learning-Based Classifier: A Case Study", IEEE Symposium on Security and Privacy, 2014.
[43] Yi Shi and Yalin E. Sagduyu, "Evasion and Causative Attacks with Adversarial Deep Learning", Milcom, Track 3 - Cyber Security and Trusted Computing, ©2017 IEEE.
[44] XiaoyuCao, Neil Zhenqiang Gong, "Mitigating Evasion attacks to Deep Neural Networks via Region-based Classification" Annual Computer Security Applications Conference, ACSAC 2017, USA.
[45] Dongyu Meng and Hao Chen, "MagNet: a Two-Pronged Defense against Adversarial Examples", CCS '17, 2017 USA, ACMI,SBN978-1-4503-4946-8/17/10. https://doi.org/10.1145/3133956.3134057.

[46] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard, "DeepFool: a simple and accurate method to fool deep neural networks", CVPR paper is the Open Access version.
[47] Anirban Chakraborty, Manaar Alam,Vishal Dey, Anupam Chattopadhyay, Debdeep Mukhopadhyay, "Adversarial Attacks and Defences: A Survey", ACM Comput. Surv., arXiv: 1810.00069v1, Sep 2018.
[48] Somesh Jha, "Adversarial Machine Learning (AML)", ICISS 2018, Bangalor.
[49] Olga Taran, Shideh Rezaeifar, and Slava Voloshynovskiy, "Bridging machine learning and cryptography in defence against adversarial attacks", https://www.researchgate.net/publication/327496074.
[50] Uni Melb, Swin burne Univ, Tamas Abraham, OlivierdeVel, PaulMontague, "Adversarial Machine Learning for Cyber-Security: NGTF Project Scoping Study", Common wealth of Australia 2018, AR-017-073.
[51] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, Pieter Abbeel, "Adversarial Attacks on Neural Network Policies", arXiv:1702.02284v1, Feb 2017.
[52] Yi Han, Benjamin I.P. Rubinstein, Tamas Abraham, Tansu Alpcan, , Olivier De Vel, Sarah Erfani, David Hubczenko, Christopher Leckie, and Paul Montague, "Reinforcement Learning for Autonomous Defence in Software-Defined Networking", arXiv: 1808.05770v1, Aug 2018.