International Interdisciplinary Virtual Conference on
'Recent Advancements in Computer Science, Management and Information Technology'
International Journal of Scientific Research in Computer Science,
Engineering and Information Technology| ISSN : 2456-3307 (www.ijsrcseit.com)

A Study on Security Challenges in Machine Learning

Akshay Dilip Lahe[1], Ghanshyam G. Parrkhed[1], Bhushan L. Rathi[1]
[1]Assistant Professor
Saraswati College, Shegaon, Maharashtra, India

ABSTRACT

In recent years, Machine learning is being used in various systems in wide variety of applications like Healthcare, Image processing, Computer Vision, Classifications, etc. Machine learning have shown that it can solve complex problem-solving abilities very similar to human beings and above them also. But various research proves vulnerability of ML Models in terms of different security attacks to ML systems. These attacks are hard to detect because they can hide in data at various stages of machine learning pipeline without being detected. This survey aims to analyse various security attacks on machine learning and categorize them depending on position of attacks in machine learning pipeline. This paper will focus on all aspects of machine learning security at various stages from training phase to testing phase. Machine Learning pipeline, Aims of Attacker, different attacks are considered in this paper.

Keywords— Artificial Intelligence Security, Machine Learning Security, Poisoning attacks, backdoor attacks, adversarial attacks, Security Attacks in ML

I. INTRODUCTION

Machine Learning that comes under computational algorithm used to mimic human learning and decision capacities deduced from its environment are being widely used in various domains like computer vision, engineering, banking and finance, entertainment industry, smart mobile and web applications, biomedical and healthcare applications. With increase of accumulation of huge amount of data and with emergence of concept of big data, various data mining and machine learning techniques have been developed for pattern recognition, future predictions, decision making along with other application tasks. Machine learning is based on concept of mimicking human beings' way of learning things along with sensory input processing to achieve a particular task.

Machine Learning (ML) can be described as ability to learn without programmed explicitly. ML algorithms learn how to perform certain task based on input data given to the algorithm and perform same task when presented with new data. ML model is trained on training data which include multitude of features which is called as Learning Phase. Then ML model is tested by presenting new data to the model and it should give correct result as per learning phase. This phase is called as Testing Phase. Using metrics like accuracy to predict correct result as per learning, and precision, performance of ML model is measured. The accuracy can depend

on factors like quantity of training data, ML Algorithm used, feature selection, feature extraction method used and hyper parameters.

This survey focuses on following points:

1. This paper focuses on security attacks at various positions of machine learning pipeline instead of focusing on one stage.
2. This paper divides the security attacks based on location as well as training or testing phase.

II. MACHINE LEARNING & IT'S APPLICATIONS

In recent decade, research in Machine Learning algorithms and its models have drastically increased and various approaches has been proposed. As of now, ML is being used in almost all the domains which include computer vision, prediction, market analysis, semantic analysis, NLP, healthcare, Information management systems, Network security, medical diagnosis and Healthcare sectors [1].

Object detection and recognition and its processing are using ML/DL in computer vision domain. For application which operates for prediction are using ML for classification purpose of documents, images and faces. Image analysis and segmentation is used for medical diagnosis. For security of various systems, ML is being used in IDS and for anomaly detection along with network intrusion and privacy aware systems to provide security to various applications. DoS attacks can be predicted using machine learning approaches. ML is used commonly in semantic analysis, NLP and information retrieval. K-NN and SVM are used to recognize hand gestures. Text classification can be done using linear classification, ANN and SVM effectively. Recommender systems have been built using ML in both bioinformatics and mobile advertisement domain. In Network Security, ML is used for IDPS, Endpoint protection which include malware classification and detection, access control and authentication detection. Process anomaly detection and fraud detection can be done by processing behaviors using ML models. User behavior can be observed using ML models which include keystroke dynamics detection and breaking human interaction proofs. ML can provide security to application by providing detections of malicious URL, phishing and spam [2].

ML in Healthcare is recent emerging domain of ML application. Large data is being generated by healthcare information systems with the introduction of electronic health records so it becomes complex to analyze, process and mine useful information using traditional methods. ML helps to analyze this data and provide insights to doctors or other stakeholders in healthcare. Prognosis meaning predicting expected future outcomes of the disease in clinical environment can be done using ML models. ML can be used for diagnosis purposes by analyzing EHRs on regular basis of the patients. Healthcare domain uses MRI, CT, Ultrasound scans to diseases detection. Image analysis along with ML can be used on these scans to effectively diagnose a disease. Extensive research is being conducted on use cases of Machine Learning in healthcare domain where continuous health analysis is done with the help of wearable devices and sensors can be achieved [4].

Basic Machine Learning Pipeline

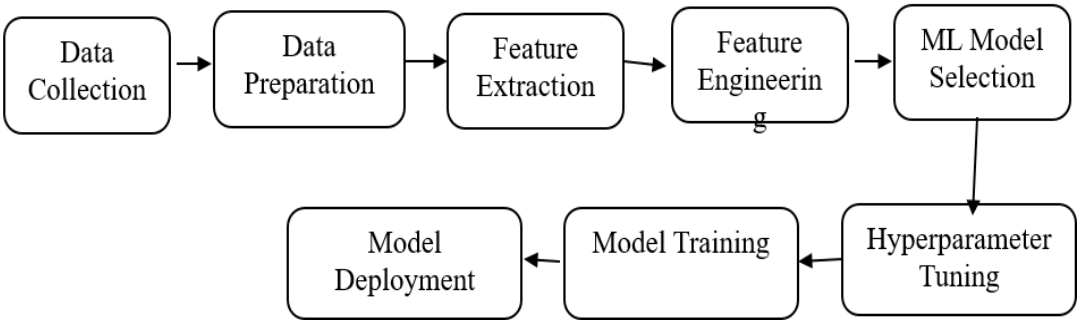

Fig. 1: Basic Stages of a Typical ML System

Fig. 1 shows basic stages of a typical ML system. First relevant data regarding application is collected at one place. This collected data can be dirty or noisy so it needs some preparations and cleaning before giving it to the ML model. Data preparation stage cleans the data, pre-process it and prepare it for the feature extraction. In feature extraction stage, important or significant features are selected and extracted out of the prepared data. Features which impact the model's outcome are selected here. Then addition and removal of various features or creation of artificial features can be done in feature engineering stage. There are various types of ML models so based on the application and input features, best approach model is taken and trained on input data. Hyper parameter tuning is performed which tune the input parameters to increase the performance. After successful training and then testing of model on new data, it is deployed in the real-world application.

III. AIMS OF AN ATTACKER

An Attacker can have number of aims for which attacker is exploiting the vulnerability of the model. The Aims can be divided majorly as Violation of Security, Specificity and Attacks by Influence [5] [6].

Violation of Security: There are three major violations that an attacker can cause: Integrity violation in which intrusive points can be classified as normal to avoid detection without compromising system functionalities; Availability violation in which attacker causes so many false errors that system functionality becomes unavailable to legitimate users of the systems; Privacy violation comprises leaking of sensitive private information to attacker.

Specificity of Attack: An attack can be a targeted attack which focuses to cause harm to a set of samples or points or it can be indiscriminate attack which is more flexible attack focusing on a general class of samples or any sample.

Attacks by Influence: It can be of two types: Causative attack which influence training data of model and alter the training process; Exploratory attack discover information about training data using techniques like probing.

IV. ATTACKS ON MACHINE LEARNING & MITIGATIONS

Data collection stage, there are vulnerabilities which include noises, dirty data, missing data, improper procedure, untrained personnel, etc. Imbalance data, biased data, data poisoning, privacy breaches, label misclassification and label leakages are few vulnerabilities at data annotation stage of ML system in healthcare

[3]. During feature extraction, there can be fragile features or irrelevant features and knowledge of feature selection algorithms or features set can help attacker. While training the model using training input data, input data is vulnerable to data poisoning attack or there can be backdoor which can help an attacker to gain access to the model. Model poisoning or stealing attack can cause model to misclassify. Evasion attacks, system disruption, network issues, adversarial attacks can be done at test data or model's output. The machine learning pipeline is a complex process that involves multiple stages. Each stage can have its own set of vulnerabilities that could compromise the accuracy, reliability, and privacy of ML model.

Stages of a Typical ML Model & Respective Vulnerabilities:

Collection of Relevant Data: Relevant data is collected with respect to the application which will be used for model training but it can be biased, incomplete, or of poor quality. This can lead to less accurate models and biased or wrong predictions.

Data Pre-processing: This second stage includes various tasks such as cleaning the collected data as data collected is messy and noisy; transforming the data and performing normalization on the data so that it should be ready to be analysed. Vulnerabilities in this stage include improper data cleaning, feature selection, or normalization, which can lead to incorrect model predictions.

Model Selection: Selecting the appropriate machine learning algorithm and architecture is essential to achieve optimal performance. However, vulnerabilities at this stage include selecting a model that is too simple or too complex, resulting in underfitting or overfitting.

Model Training: The model training involves feeding the data into the selected ML algorithm and training the model using inputted data. Optimization techniques can be applied to input parameters of the model to increase the performance. Vulnerabilities at this stage include insufficient data for training, using incorrect hyperparameters, or using incorrect evaluation metrics, leading to inaccurate models.

Model Evaluation: Evaluating the model's performance is necessary to ensure that it is accurate and reliable. However, vulnerabilities at this stage include evaluating the model on biased or incomplete data, using incorrect evaluation metrics, or failing to detect overfitting or underfitting.

Model Deployment: Deploying the model into a production environment can pose several risks, including cybersecurity threats, data breaches, and privacy violations. Vulnerabilities at this stage include unsecured endpoints, insufficient model monitoring, and lack of data privacy controls.

Overall, it is important to identify vulnerabilities at each stage of the machine learning pipeline and ensure that proper security defence is applied at respective stages so that it will not affect performance and security of ML system.

Various Security Attacks Performed on ML System:

Poisoning Attacks: In this attack type, data used for training is injected with some malicious data which changes the model's output. An Attacker can make model predict decided output by performing this attack which can in turn confuse the model.

Adversarial Attacks: Adversarial attacks involve adding carefully crafted noise to training input data which deceives ML Model. Using this attack, an attacker can force ML model to make incorrect predictions or classify data into the wrong category.

Model Evasion Attacks: Model evasion attacks involve exploiting vulnerabilities in the model's decision-making process to manipulate its behaviour. Using this attack, an attacker can use various techniques like gradient descent which can alter parameters or training input data to evade detection.

Inference Attacks: Inference attacks involve thorough analysis of model's output to steal sensitive data. Using this attack, an attacker can just observe and analyse output of ML model and infer sensitive information.

Methods to Mitigate Attacks on Machine Learning:

Data Sanitization: Data sanitization involves filtering out malicious data from the training dataset to prevent poisoning attacks. The data can be pre-processed, and the outlier data can be removed to prevent model bias.

Adversarial Training: Adversarial training involves training the model on both clean and adversarial data to improve its robustness against adversarial attacks.

Regularization: Regularization involves adding penalties to the model's training process to prevent overfitting and improve its generalization capabilities. Regularization can prevent model evasion attacks by reducing the model's reliance on specific inputs.

Secure Inference: Secure inference involves protecting the model's output by adding noise to the output to prevent inference attacks. Differential privacy can be used to add noise to the output without significantly affecting its accuracy.

V. CLASSIFICATION OF SECURITY THREATS

Machine Learning attacks can happen at various phases of ML lifecycle. Here we are categorizing attacks into two main categories: 1. Attacks during training phase and 2. Attacks during Testing Phase and Model's Output.

1. Attacks during Training Phase
   - A. Poisoning training data
   - B. Backdoor in Training data
2. Attacks during Testing Phase and Model's Output
   - A. Adversarial Attacks
     - i. Having Knowledge of system
     - ii. Without any knowledge of system
   - B. Model Extraction Attack
   - C. Stealing Hyper-Parameters
   - D. Sensitive training input data Recovery
     - i. Model Inversion
     - ii. Inference from membership

1. Attacks during Training Phase

A. Poisoning Training Data

Prediction or output of ML model can be misled by manipulating the training data is called as Poisoning Attack. Various research has shown that poisoning attack can degrade performance of model drastically.

Intrusion Detection-Prevention Systems (IDPS), Abnormality or malware detection system also use ML models for detection. There are many poisoning attacks proposed which targets anomaly detection system in a network.

P. Li et al. [8] adopted an edge pattern detection (EPD) algorithm which is tested against multiple ML algorithms like NB, LR and SVM used in IDSs.

There is an updating procedure as well as input procedure in every biometric system where data is updated or inputted. Attacker can take advantage of these processes to comprise the privacy and security of these systems. Biggio et al. [9] investigated adaptive biometric systems which uses verification of face with the use of PCA method. Fake faces can be injected to claim legitimacy of the fake user. This attack is improved their further research [10] in which it is assumed that user can store many templates.

Biggio et al. [11] designed an attack which targets SVM-based systems where attacker can increase testing error of classifier by injecting well-crafted training data. Gradient ascent technique is used to build malicious data. It uses optimization formulation and is able to be kernelized. B. Biggio et al. [13] presented an approach that particularly targets malware clustering used in behavioural detection systems. A poisoned sample with poisoning behaviours can be added to training data.

Learning algorithms can directly attack by poisoning attacks. H. Xiao et al. [7] performed poisoning attack on pdf malware detection which can compromise feature selection methods. Multiple features like ridge regression and LASSO can be attacked using this approach. B. Li et al. [14] designed an attack which targets systems with collaborative filtering. Attacker can go unnoticed by imitating normal user and this require complete information of the system.

Y. Wang and K. Chaudhari [16] proposed an attack which targets online learning systems where input streams are used. A better attack targeting online learning is proposed by X. Zhang and X. Zhu [17].

Cloud deployed models can be exposed to attacks on server side. Cong Liao et al. [23] proposed a study which focuses on this type of attack where attacker having access to server is able to manipulate the model and add malicious samples without being detected easily.

B. Backdoor in Training data

Backdoor can be created in training input data which is hidden in any ML model. Normal functioning of model does not get affected by backdoor. Backdoor has some triggering condition when the conditions are met backdoor gets triggered. Backdoor are stealthy and very hard to detect.

Chen at al. [27] proposes method to add backdoor with the use of data poisoning in Deep learning models. This model works effectively even if there is no knowledge of model and input data. Liao et al. [28] presented backdoor attacks which can be inserted using stealthy perturbations in convolution neural network models. Backdoor attacks on federated learning are presented by Bagdasaryan et al. [29] in which they proposed a secure privacy preserving learning framework. Tianyu Gu et al. [30] proposed a backdoor called BadNet and tested it in different real-life scenarios and concluded that backdoors reduce accuracy. Ahmed Salem et al. [31] presented backdoor attack for deep networks which is dynamic. Current backdoor detection systems cannot detect these attacks as triggers generated by these have random conditions, patterns at random locations.

In all these methods, attacker adds a backdoor to ML model then when it gets activated, it will create malicious data inputs. This malicious data is fed to model as training input on which model is trained and re-trained. Further Backdoor attacks are categorized as per Yansong Gao et al. [32] into outsourcing attack, pretrained attack, data collection attack, collaborative learning attack, post deployment attack and code poisoning attack.

2. Attacks during Testing Phase and Model's Output

A. Adversarial Attacks

Depending on the information known to the attacker, adversarial attacks can be classified in two types: known target system attack and unknown target system attack.

Flavio Luis mello [34] presented various attacks in physical word and their protective measures with respect to adversarial attacks. There are applications like Heat clocking wearables and anti-surveillance makeup, Adversarial T-shirt which can fool detection of person, eyeglasses which can fool face recognition systems in cameras, face projector approach to trick facial recognition systems, etc. Ivan Evtimov et al. [35] presented an attack which can generate perturbation with the help of images under various conditions into account. Wieland Brendel et al. [36] proposed a decision-based attacks which can be applied in real world scenarios using black-box models needing less knowledge and are easier to apply than transfer-based attacks. Nicolas Papernot et al. [38] presented practical black-box attack which require no knowledge of model or training data and attacker can control a remotely hosted DNN.

Dalvi et al. [40] proposed a study in which classifier do the wrong predictions. This problem is known as adversarial classification problem. Adversarial learning problem is introduced by Lowd and Meek [41]. Statistical machine learning can be used to attack spam filter [42]. Srndic and Laskov [44] studied classifier's performance under evasion attacks and states that there is significant drop in performance under simple attacks.

Without any system's knowledge also attacker can perform adversarial attacks. Xu et al. [45] proposed an evasion attack which can fool detection systems by finding malicious samples. Face recognition in Biometric system [48], Sign recognition attack used for roads [35], camera attack of cell phone [50] and 3D object attack [51] are some examples of this attack in real life scenarios.

B. Model Extraction Attack

Attacker can steal ML model where attacker has to observe the output labels and confidence levels along with corresponding inputs, this is called as model stealing. The idea was presented by Tramer et al. [52]. It is a black-box attack type where attacker mines knowledge and then based on obtained information, attacker re-design the model which acts similar target model. Shi et al. [53] presented model stealing method which works on black box approach where attacker use deep learning to build model form obtained predicted labels from target labels. Chandrasekaran et al. [54] proposed extraction attacks without any information.

C. Hyper parameters Stealing

Gradient of model is initialized to zero and hyper parameters are calculated by solving linear equations. This method is proposed by Wang and Gong [55] where hyper parameters of model can be stolen from algorithms like SVM, Ridge regression and neural networks. This method assumes that attacker should have the knowledge of learning algorithm, training data, etc. There is algorithm proposed by Milli et al. [56] which states that model parameters can be revealed by gradient information quickly. This method has high computational overhead.

D. Sensitive Training Input Data Recovery

Attacker can recover sensitive information of the training data by observing output and model parameters. There are two major types of attack which can perform this task: 1. Model Inversion Attack and 2. Membership Inference Attack.

Fredrikson et al. [57] first introduced Model inversion attack in which black-box access along with some knowledge of patient can be used to get genomic information. They further involved their study in [58] where same attack can be performed with the used of confidence of predictions. There are two categories of works in this attack type, first is attacks that creates actual reconstruction and second is attacks that create representative class of sensitive data which is not there in training data. [59]

Membership inference attack is introduced by Shokri et al. [60] in which attacker can calculate if some data belongs to training data or it does not belong to training data. This attack can be threat to many deep learning models [62] [63] [64].

VI. CONCLUSION

Machine learning models have become an essential tool in various applications, but they are susceptible to attacks. Attackers can make use of various vulnerabilities in machine learning models to manipulate their behavior or steal sensitive information. This paper provides a study on various threats of ML security. The ML system is vulnerable to different types of attacks at different locations based on ML Pipeline. This paper will give researchers category wise classification of attacks at different stages of ML pipeline like training phase or testing phase. We conclude that ML pipeline itself is vulnerable at various stages of its pipeline from various attacks and there is a need to design secure, privacy preserving ML system which can defend against these attacks.

VII. REFERENCES

[1]. Pramila P. Shinde and Dr. Seema Shah, “A Review of Machine Learning and Deep Learning Applications” in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) IEEE DOI: 10.1109/ICCUBEA.2018.8697857.

[2]. Olakunle Ibitoye, Rana Abou-Khamis, Ashraf Matrawy, M. Omair Shafiq, “The Threat of Adversarial Attacks on Machine Learning in Network Security-A Survey” in 2019 arXiv:1911.02621.

[3]. Gary McGraw, Richie Bonett, Victor Shepardson, and Harold Figueroa, “The Top 10 Risks of Machine Learning Security” in IEEE: Computer (Volume: 53, Issue: 6, June 2020) DOI: 10.1109/MC.2020.2984868.

[4]. Battista Biggioa and Fabio Rolia, “Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning” in Elsevier Pattern Recognition Volume 84 Dec. 2018 pages 317-331, https://doi.org/10.1016/j.patcog.2018.07.023

[5]. Marco Barreno, Blaine Nelson, Anthony D. Joseph, J.D. Tygar, “The security of machine learning” in Springer Machine Learning-volume 81 May 2010, page 121-148, DOI:10.1007/s10994-010-5188-5

[6]. Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, Fabio Roli, “Is feature selection secure against training data poisoning?” published in ICML 6 July 2015 Computer Science

[7]. Adnan Qayyum, Junaid Qadir, Muhammad Bilal, Ala Al-Fuqaha, “Secure and Robust Machine Learning for Healthcare: A Survey”, in 2020 IEEE Reviews in Biomedical Engineering (Volume: 14) DOI: 10.1109/RBME.2020.3013489.

[8]. P. Li, Q. Liu, W. Zhao, D. Wang, S. Wang, “Chronic poisoning against machine learning based IDSs using edge pattern detection,” in IEEE International Conference on Communications (ICC 2018).

[9]. Biggio, B., Fumera, G., Roli, F., Didaci, L.,"Poisoning Adaptive Biometric System" in:, et al. Structural, Syntactic, and Statistical Pattern Recognition. SSPR /SPR 2012. Lecture Notes in Computer Science, vol 7626. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34166-3_46.

[10]. B. Biggio, L. Didaci, G. Fumera, and F. Roli, "Poisoning attacks to compromise face templates", in 2013 International Conference on Biometrics (ICB)-pages 1 to 7.

[11]. B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in ICML'12: Proceedings of the 29th International Conference on International Conference on Machine LearningJune 2012 Pages 1467–1474.

[12]. B. Biggio et al., "Poisoning behavioral malware clustering," in AISec '14: Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop, November 2014, Pages 27–36, https://doi.org/10.1145/2666652.2666666.

[13]. B. Li, Y. Wang, A. Singh, and Y. Vorobeychik, "Data poisoning attacks on factorization-based collaborative filtering," in NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems, December 2016, Pages 1893–1901

[14]. Koh, P.W., Steinhardt, J. & Liang, P., "Stronger data poisoning attacks break data sanitization defences", in Springer Machine Learning 111, 1–47 (2022). https://doi.org/10.1007/s10994-021-06119-y.

[15]. Xuezhou Zhang, Xiaojin Zhu, Laurent Lessard, "Online Data Poisoning Attacks", Proceedings of the 2nd Conference on Learning for Dynamics and Control, PMLR 120:201-210, 2020. arXiv:1903.01666, 2019.

[16]. Cong Liao, Haoti Zhong, Sencun Zhu, Anna Squicciarini, "Server-Based Manipulation Attacks Against Machine Learning Models", in CODASPY '18: Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy, March 2018, Pages 24–34, https://doi.org/10.1145/3176258.3176321.

[17]. X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning", in arXiv:1712.05526 (2017).

[18]. C. Liao, H. Zhong, A. C. Squicciarini, S. Zhu, D. J. Miller, "Backdoor embedding in convolutional neural network models via invisible perturbation", in arXiv:1808.10307 (2018)

[19]. E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin,, V. Shmatikov, "How to backdoor federated learning", in Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, PMLR 108:2938-2948, 2020

[20]. Tianyu Gu, Brendan Dolan-Gavitt, Siddharth Garg, "BadNets: Identifying vulnerabilities in the machine learning model supply chain", in arXiv:1708.06733v2 (2019).

[21]. A. Salem, R. Wen, M. Backes, S. Ma, Y. Zhang," Dynamic Backdoor Attacks Against Machine Learning Models", in arXiv:2003.03675 (2020).

[22]. Y. Gao, B. G. Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S.a Nepal, H. Kim," Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review", in arXiv:2007.10760 (2020).

[23]. F. L. de Mello, "A Survey on Machine Learning Adversarial Attacks" in Journal of Information Security and Cryptography (Enigma) 7(1):1-7, January 202, DOI:10.17648/jisc.v7i1.76.

[24]. K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, "Robust Physical-World Attacks on Deep Learning Visual Classification", in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2018, DOI:10.1109/CVPR.2018.00175.

[25]. W. Brendel, J. Rauber, M. Bethge, "Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models", in arXiv:1712.04248v2 (2018).

[26]. N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Berkay Celik, A. Swami," Practical Black-Box Attacks against Machine Learning", in ASIA CCS '17: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, April 2017, Pages 506–519, https://doi.org/10.1145/3052973.3053009.

[27]. N. Dalvi, P. Domingos, Mausam, S. Sanghai, D. Verma, "Adversarial classification", in KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, August 2004, Pages 99–108, https://doi.org/10.1145/1014052.1014066.

[28]. Daniel Lowd and Christopher A. Meek, "Adversarial learning", in KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, August 2005, Pages 641–647, https://doi.org/10.1145/1081870.1081950.

[29]. B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. P. Rubinstein, U. Saini, C. Sutton, J. D. Tygar, K. Xia, "Exploiting machine learning to subvert your spam filter", in LEET'08: Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats, April 2008, Article No.: 7, Pages 1–9.

[30]. N. Šrndic and P. Laskov, "Practical evasion of a learning-based classifier:´A case study" in IEEE Symposium on Security and Privacy, May 2014, pages 197–211, DOI: 10.1109/SP.2014.20.

[31]. W. Xu, Y. Qi, D. Evans, "Automatically Evading Classifiers: A Case Study on PDF Malware Classifiers", in Conference of Network and Distributed System Security Symposium, Jan. 2016, pages. 1–15, DOI:10.14722/ndss.2016.23115.

[32]. M. Sharif, S. Bhagavatula, L. Bauer, M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in CCS '16: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, October 2016, Pages 1528–1540, https://doi.org/10.1145/2976749.2978392.

[33]. A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples", in arXiv:1707.07397v3 (2018).

[34]. B. Biggio, G. Fumera, F. Roli, "Security evaluation of pattern classifiers under attack" in IEEE Transactions on Knowledge and Data Engineering 99(4):1, pages 984–996, Jan. 2013, DOI:10.1109/TKDE.2013.57.

[35]. F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, T. Ristenpart, "Stealing machine learning models via prediction APIs", in SEC'16: Proceedings of the 25th USENIX Conference on Security Symposium, August 2016, Pages 601–618.

[36]. S. Yi, Y. Sagduyu, A. Grushin, "How to steal a machine learning classifier with deep learning", in IEEE International Symposium on Technologies for Homeland Security (HST), Apr. 2017, pages 1–5, DOI:10.1109/THS.2017.7943475.

[37]. V. Chandrasekaran, K. Chaudhuri, I. Giacomelli, S. Jha, and S. Yan, "Exploring connections between active learning and model extraction", in SEC'20: Proceedings of the 29th USENIX Conference on Security Symposium, August 2020 Article No.: 74, Pages 1309–1326.

[38]. B. Wang and N. Z. Gong, "Stealing Hyperparameters in Machine Learning", in IEEE Symposium on Security and Privacy (SP), May 2018, pages 36–52, DOI: 10.1109/SP.2018.00038.

[39]. S. Milli, L. Schmidt, A. D. Dragan, M. Hardt, "Model Reconstruction from Model Explanations", in FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency, January 2019, Pages 1–9, https://doi.org/10.1145/3287560.3287562.

[40]. M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, T. Ristenpart, “Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing”, in SEC'14: Proceedings of the 23rd USENIX conference on Security Symposium, August 2014, Pages 17–32.

[41]. M. Fredrikson, S. Jha, T. Ristenpart, “Model inversion attacks that exploit confidence information and basic countermeasures”, in CCS '15: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, October 2015, Pages 1322–1333, https://doi.org/10.1145/2810103.2813677.

[42]. Maria Rigaki And Sebastian Garcia, “A Survey Of Privacy Attacks In Machine Learning”, in Arxiv:2007.07646, Stratosphere Project 2020.

[43]. R. Shokri, M. Stronati, C. Song, V. Shmatikov, “Membership inference attacks against machine learning models”, in arXiv:1610.05820v2 (2018)

[44]. Dingfan Chen, Ning Yu, Yang Zhang, Mario Fritz,” GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models”, in CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Octo. 2020, https://doi.org/10.1145/3372297.3417238

[45]. J. Hayes, L. Melis, G. Danezis, E. De Cristofaro,” LOGAN: Membership inference attacks against generative models”, in proceedings on Privacy Enhancing Technologies 2019, 1 (2019), 133–152, Jan. 2019, DOI:10.2478/popets-2019-0008.

[46]. Benjamin Hilprecht, Martin Härterich, Daniel Bernau, “Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models, in Proceedings on Privacy Enhancing Technologies 2019(4), pages 232–249, DOI:10.2478/popets-2019-0067.

[47]. Zhao, B. Y., Li, X., Liskov, B., and Teller, S. "SecureML: A system for scalable privacy-preserving machine learning". USENIX Symposium on Operating Systems Design and Implementation, 2018.

[48]. Yang, Q., Liu, Y., and Chen, Y. "Federated machine learning: Concept and applications". ACM Transactions on Intelligent Systems and Technology, 10(2), 1-19, 2019.

[49]. Wang, T., Li, Y., Li, M., Huang, X., and Chen, X. "Defending against model evasion attacks with adversarial training and regularization". IEEE Transactions on Dependable and Secure Computing, 18(1), 16-28, 2021.

[50]. Tramer, F., Kurakin, A., Papernot, N., Boneh, D., and McDaniel, P. "Ensemble adversarial training: Attacks and defenses". International Conference on Learning Representations, 2018.