I.J. Information Technology and Computer Science, **2023, 2, 23-34**
Published Online on April 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2023.02.03

A Survey on Security Threats to Machine Learning Systems at Different Stages of its Pipeline

Akshay Dilip Lahe
Kalinga University, Raipur (CG), India
E-mail: lahe.akshay@gmail.com
ORCID iD: https://orcid.org/0000-0003-2387-9268

Guddi Singh*
Kalinga University, Raipur (CG), India
E-mail: guddi.singh@kalingauniversity.ac.in
*Corresponding Author

Received: 27 July 2022; Revised: 02 December 2022; Accepted: 07 January 2023; Published: 08 April 2023

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 algorithms have shown that it can solve complex problem-solving capabilities close to humans or beyond humans as well. But recent studies show that Machine Learning Algorithms and models are vulnerable to various attacks which compromise security the 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 instead of focusing on one type of security attack. Machine Learning pipeline, Attacker's goals, Attacker's knowledge, attacks on specified applications are considered in this paper. This paper also presented future scope of research of security attacks in machine learning. In this Survey paper, we concluded that Machine Learning Pipeline itself is vulnerable to different attacks so there is need to build a secure and robust Machine Learning Pipeline. Our survey has categorized these security attacks in details with respect to ML Pipeline stages.

Index Terms: Artificial Intelligence Security, Machine Learning Security, Poisoning Attacks, Backdoor Attacks, Adversarial Attacks, Security Attacks in ML.

1. 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 paper is divided into six sections. The first section of introduction will give details regarding what machine learning is and how its applications are growing along with different types of ML Models. The second section which is

Security of ML Systems will give idea regarding various aspects of security with respect to an attacker. It includes attacker's goal and attacker's knowledge which will give brief introduction of various aspects of attack on ML pipeline. The third Section i.e. related work is categorized in two parts. The first part gives survey of literature of various attacks that happened during training phase of the ML systems and the later part describes various attacks that can take place during testing phase and at model's output. After this detailed literature survey, in fourth section, we have given some recommendations for researchers from our survey. The fifth section will conclude our findings and last section gives future directions and scope for researchers who can take this research further.

This survey focuses on following points:

- This paper focuses on security attacks at various positions of machine learning pipeline instead of focusing on one stage.
- This paper divides the security attacks based on location as well as training or testing phase.
- Future directions regarding security of machine learning is presented here.

1.1. Classification of ML Algorithms

Generally, ML algorithms can be classified depending on Learning style into Supervised Learning, Unsupervised Learning or Semi-Supervised Learning or Reinforcement Learning.

A. Supervised Learning

The input data includes labelled data in which both feature and its output value is given. Each Input label is presented to model along with its output label at Learning phase. The output label can be Class label giving classification ML Model or continuous value label giving regression ML model. Both the input and output labels are used to train the model so that it can predict the output label when it is presented with new input label in testing phase.

Supervised learning is further divided into two types: Classification and Regression. In classification, model is designed which can classify samples into different classes with their class labels. Objective of model will be to state class label to which that particular new test data or sample belongs to. The observed value in classification is in categorical form. Depending on number of classes to divide into, it can further be divided into binary classification and multi-class classification. Logistic Regression, SVM and Naïve Bayes are some examples of classification algorithm. A simple Example can be to classify whether given image is car or bus. In Regression, label of a sample is continuous value. Independent variables are samples which are made of features. A simple example can be house price estimation based on its area, number of rooms, location and other features.

B. Unsupervised Learning

The training input data is not labelled in unsupervised learning. Class labels are not given to feature input vectors. The goal itself is to find a structure or pattern in the input data. Model learns from unlabelled training data and predicts the output in the form of hidden patterns from the dataset without any supervision that is why it is called as unsupervised learning. Cluster is most popular unsupervised algorithm which involves grouping of samples into different clusters based on similarities or differences. Clustering is commonly used in image segmentation, data analysis and market segmentation. Examples of Clustering algorithms are K-Means, Hierarchical, etc.

C. Semi-Supervised Learning

In this approach, some of the data is labelled and some of the data is unlabelled. Labelling of data requires human experts or intelligent systems due to which it becomes expensive task. So sometime part of whole dataset is labelled only. Even small part of labelled data can improve the learning process of the model and accurately predict the output.

D. Reinforcement Learning

When model adjust itself and learn better with continuous feedback from its output then it is called as reinforcement learning. It is a feedback-based learning which takes feedback after each action. It works as a reward and goal will be to maximize the positive rewards to improve the performance. Humans learn from their experiences and based on that interact with others, on this basic concept this learning is based. Q-learning is one of popular reinforcement learning.

Machine Learning can be also classified on the basis of depth into two types: Shallow Learning and Deep Learning [1]. In shallow learning, multiple hidden layers are not used. It used standard machine learning approach. It does not have vanishing gradient and complexity problems. But in contrast to shallow learning, deep learning has multiple hidden connections and layers which it uses to learn better. It overcomes scalability and complicated problems.

There are various tasks that can be performed by Machine Learning and Machine learning can be classified on the basis of these tasks also. The tasks can be Regression, Classification, Clustering, Generative Modelling, Association Rule Learning, Dimensionality Reduction, etc.

1.2. Applications of Machine Learning

Machine Learning is a sub domain of Artificial Intelligence. 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 [2].

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 [1].

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 diagnose diseases. Image analysis along with ML can be used on these scans to effectively diagnose a disease. Extensive research is being conducted on ML in real time health monitoring where continuous health monitoring with wearable devices and sensors can be achieved [3].

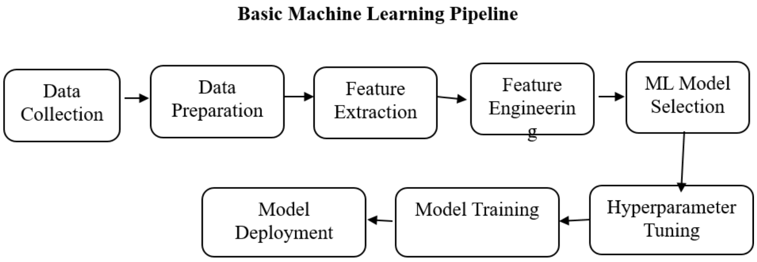

Fig.1. Basic Machine Learning Pipeline.

Machine learning pipeline given in fig. 1 illustrates all the stages of machine learning model from data collection to model deployment. First data relevant to 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 selected and trained on input data. Hyper parameter tuning is done to adjust various parameters of the model to increase the performance of the model. After successful training and then testing of model on new data, it is deployed in the real-world application.

2. Security of ML Systems

Despite its application in wide domains, research in security of Machine Learning Models is comparatively less. There are different components of a general ML Model which includes Raw data, Datasets (training, validation and test), Learning algorithms, Evaluation methods, ML model itself, Output of the model, etc. All of these components are prone to risks from an attacker. ML model can be at a risk from attackers which include adversarial examples, data poisoning, Online system manipulation, transfer learning attack, data confidentiality, reproducibility, overfitting, data trustworthiness, encoding integrity, output integrity, etc. [3].

Broadly, Machine learning system has two major stages:

- **Training phase:** learning algorithm learns from training data and model is trained
- **Testing Phase:** trained model is presented with new data called test data to see if model is performing as per expectations

Attacker can attack at both of these major stages to gain sensitive information. Machine Learning system is vulnerable to various stages and points throughout its pipeline. Data poisoning and backdoor attacks can be done on training data to misled the model. Model's output can be compromised by model theft and recovery of training data from model's output. Various adversarial attacks can craft to misled the model output the test input data. Through these examples, it is evident that ML model itself are vulnerable at many points. In this research survey, I am to give a details category wise classification of these attacks and vulnerabilities of ML pipeline.

2.1. *Attacker's Goal*

Attacker's goal can be presented in three aspects i.e., Security Violation, Attack Specificity and Influence Attacks [4, 5].

Security Violation: 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.

Attack Specificity: 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.

Influence Attack: 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.

2.2. *Attacker's Knowledge*

Attackers' knowledge of the ML system can be very important and decides the efforts attacker has to take. There are various levels of knowledge of system for an attacker [6].

Knowledge of Training data: attacker can have portioned or full knowledge of the training data which can be used to perform attacks on the ML system in less time.

Knowledge of Feature Representation: Partial or full knowledge of the process of feature structure and representation can be useful to attacker.

Knowledge of feature selection algorithm: There may be a situation where attacker knows which algorithm for selecting features of model is being used along with criterion.

Perfect Knowledge: When an attacker has complete knowledge of entire ML System it is said that attacker has perfect knowledge of the system and attacker can degrade performance of system or compromise the system entirely.

Limited Knowledge: Having perfect knowledge of the system is difficult in realistic scenarios. In realistic scenario, attacker can have limited or partial knowledge of the system which can be used to exploit the system.

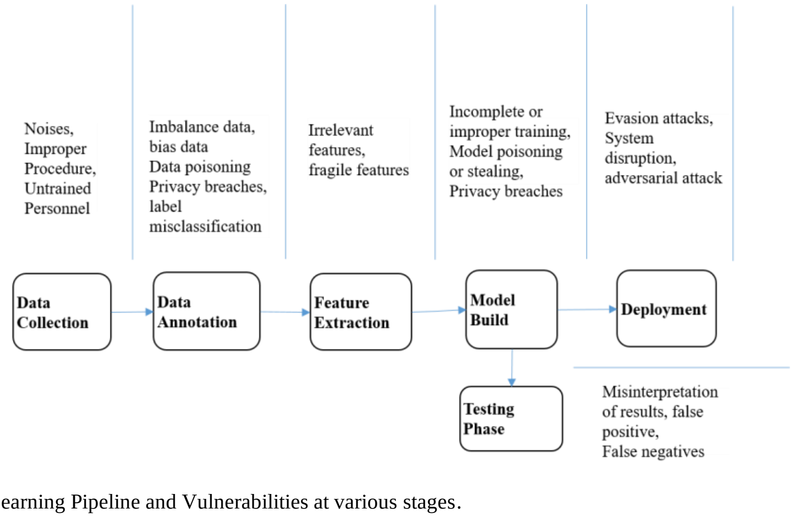
Noises, Improper Procedure, Untrained Personnel
Imbalance data, bias data Data poisoning Privacy breaches, label misclassification
Irrelevant features, fragile features
Incomplete or improper training, Model poisoning or stealing, Privacy breaches
Evasion attacks, System disruption, adversarial attack
Data Collection
Data Annotation
Feature Extraction
Model Build
Deployment
Testing Phase
Misinterpretation of results, false positive, False negatives

Fig.2. Healthcare Machine Learning Pipeline and Vulnerabilities at various stages.

The fig. 2 shows example of various vulnerabilities at different stages of machine learning pipeline in healthcare domain. This will illustrate how a ML model is vulnerable at various stages. At 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 [7]. 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.

3. Related Works

This section refers to review of various threats as well as attacks to which machine learning systems and models are vulnerable to. As given in our example ML Pipeline, ML 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. Recovery of sensitive training data

        i. Model Inversion
        ii. Membership Inference

3.1. 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. The classification of poisoning attack based on target is given in this paper.

Target: Intrusion Detection System

Machine learning is used in many security systems like Intrusion Detection and Prevention Systems (IDPS) or Abnormality or malware detection systems. 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. It further proposed a chronic poisoning attack which more effective. But this poisoning method takes long time to poison the target and complex in nature.

Target: Biometric ML based Recognition Systems

There is an updating procedure as well as input procedure in every biometric system where data is updated and collected. Attacker can take advantage of these procedures and target the security of the system. B. Biggio et al. [9] investigated adaptive biometric systems which uses PCA based face verification which performs self-update. An attacker can inject fake faces set to camera and claim to be legitimate user. It is assumed in this paper that attacker has knowledge of the system and user store single template in system. They improve the attack in [10] by assuming that user can store multiple templates in the system. Improved attack uses different matching algorithms and attacker has knowledge of only user's face image.

Target: Support Vector Machine (SVM)

Biggio et al. designed a poisoning attack for SVM based systems where attacker can increase test errors of classifier by injecting well-crafted training data. Gradient ascent strategy is used to build the poisoning data. It uses optimization formulation and is able to be kernelized. But it is assumed that attacked have complete knowledge of training data as well as algorithm [11].

Target: Clustering

Clustering is an unsupervised machine learning algorithm which is commonly used in analysis and application domain. It is also vulnerable to poisoning attacks as per [11,12]. Small poisoning of training data can compromise the ML system. This poisoned data can be well hidden in the normal training data and hard to detect [12]. B. Biggio et al. 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 [13]. These both the approaches are generic and can be applied to any clustering algorithm.

Target: Methods and Algorithms

Some processing methods can be attacked using poisoning attack. Also Learning algorithms can directly attack by poisoning attacks. H. Xiao et al. [6] 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 a poisoning attack for collaborative filtering system. Attacker can go unnoticed in system by imitating normal user. For this attack, attacker should have full knowledge of the system. M. Jagielski et al. [15] proposed poisoning attack which uses statistical information of data to create poisoned data and performs attack on linear regression models.

Target: Online Stream Data

Y. Wang and K. Chaudhari [16] proposed poisoning attack in which training data is in the form of streams and applicable to online learning scenarios. They used a gradient ascent approach to solve the optimization problem. Efficiency of this approach is high because it inputs the data into stream at some specified locations that narrows down the search space. A better attack targeting online learning is proposed by X. Zhang and X. Zhu [17] in the form of two algorithms i.e., model predictive control-based attack and reinforcement learning based attack. In this approach, Attacker has limited knowledge.

Target: Neural Network Models

There are several neural network models being used in various domain nowadays. They are also vulnerable to poisoning attacks. Generative Adversarial Networks (GAN) is proposed by C. Yang et al. [18] which can perform poisoning attacks. Poisoning data generation is done by auto-encoder. Discriminator calculates the effect of poisoning data. Similarly, there is attack targeting deep learning models. It uses back gradient optimization and support multi-class problems. The proposed method is less complex [19].

Target: Applications

ML is used in various application in real word scenarios. There are poisoning attacks specially designed for some particular applications. Healthcare systems can be targeted by poisoning attacks [20]. Healthcare being a critical and sensitive domain effects of these attack can be catastrophic. M. Fang et al. [21] proposed a poisoning attack which targets the recommender systems by creating fake users with crafted rating scores. Malicious worker can disguise as normal worker and can evade detection in crowd sensing applications [22].

Target: Servers on which Model is deployed

Due to Cloud computing emergence many models are being deployed in cloud environment. But 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.

Target: Computer vision

Targeting computer vision-based machine learning model, Ahmed Salem et al. [24] presented model hijacking attack which happened while model is training. Here, attacker aims to hijack target model and execute malicious task without detection which can cause security risks.

Target: Healthcare Systems

ML is being used in healthcare sector extensively to help medical personnel with variety of tasks like disease diagnosis, prognosis and real-time monitoring. AKM Iqtidar Newaz et al. [25] introduces five adversarial algorithms namely, HopSkipJump, Fast Gradient Method, Crafting Decision Tree, Carlini & Wagner and Zeroth Order Optimization which performs various malicious operations on healthcare system like data poisoning and misclassification.

B. Backdoor in Training Data

Attacker can create a backdoor in training data which will be hidden in machine learning 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.

Yujie Ji et al. [26] gives backdoor study in machine learning systems. Primitive learning modules (PLMs) which are supplied by third parties introduces backdoors. Malicious PLMs can cause malfunction of system when some conditions are met. In this method, backdoors can be inserted into model by manipulating model parameters. 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. Attacker's label is identified as target label. 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, first it is tested in handwritten digit classifier and then it is used to identify stop signs in U.S. Street sign classifier. They concluded that because of these backdoor attacks there can be 25% on average accuracy drop in the model.

Backdoor approaches discussed so far rely on static triggers with some fixed conditions and patterns and can be detected with latest backdoor detection systems. Ahmed Salem et al. [31] presented Random Backdoor, backdoor generating network (BaN) and conditional backdoor generating network (c-BaN) which is a dynamic backdoor attack for deep networks. Current backdoor detection systems cannot detect these attacks as triggers generated by these have random conditions, patterns at random locations. C-BaN is the first conditional backdoor attack which generate target specified trigger.

In all these methods, a backdoor is added to model by the attacker and then when it is activated it creates poisoning data and then inserts poisoning data into training input of model and then model is re-trained. Target backdoor is embedded into the model after training. Further Backdoor attacks can be categorized as per Yansong Gao et al. [32]: 1. Outsourcing attack 2. Pre-trained attack 3. Data collection attack 4. Collaborative learning attack 5. Post deployment attack and 6. Code poisoning attack.

3.2. *Attacks during Testing Phase and Model's* Output

A. *Adversarial Attacks*

Szegedy et al. [33] proposed the term adversarial example in 2014 in study of deep learning algorithms. In adversarial example, attacker construct disturbance to the training data of the model. In old research of non-deep learning algorithms, adversarial attacks are known as evasion attacks which include spam filtering, malware detection, IDS and so on. The adversarial attacks are divided into error-generic attack and error-specific attack. In first attack makes model to go wrong and second attack makes model to incorrectly identify attackers' example as if it is coming from a normal class [4]. Depending on the knowledge requirement of attacker, adversarial attacks can be classified in two types: 1. Attacks in which attacker have knowledge of target system and 2. Attacks in which attacker do not have any knowledge of target system.

Flavio Luis de mello [34] conducted a survey on Machine learning adversarial attacks which includes various attacks in physical word and protective measures. There are applications like Heat clocking wearables and anti-surveillance makeup, Adversarial T shirt to evade person detection by YOLOv2, eyeglasses which can fool face recognition systems in cameras, face projector approach to trick facial recognition systems, etc.

Ivan Evtimov et al. [35] proposed a new attack algorithm called robust physical perturbations which can generate perturbation with the help of images under various conditions into account. Adversarial attack performed by this algorithm mimics vandalism to reduce detection and also achieves high success rates for real road sign recognition under different conditions. Similarly, Wieland Brendel et al. [36] stated that adversarial perturbations generations using gradient based or score-based attacks are not available to be applied in real world scenarios. To mitigate this problem, they 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. This model is more robust to simple defences than other models.

There are many applications of ML in network security like spam filtering, Intrusion detections and prevention systems and malware detections which are adversarial in nature. According to Olakunle Ibitoye et al. [1] ML in network security is also vulnerable to various adversarial attacks which can lead to wrong prediction. They conducted a detailed study on various adversarial attacks against machine learning with respect to network security and concluded with a risk grid map of adversarial attacks to which ML systems in Network security are vulnerable. Nowadays, various Learning systems are online and can be accessed remotely. K. Auernhammer et al. [37] gave summary of attacks on machine learning in accordance with accountability. For such systems, Nicolas Papernot et al. [38] presented a practical black-box adversarial attack which require no knowledge of model or training data and attacker can control a remotely hosted DNN. Prinkle Sharma et al. [39] proposes adversarial attack targeting connected and autonomous vehicles and compromising its security.

- Attacks in which attacker have knowledge of target system

Dalvi et al. [40] proposed a study in which classifier do the wrong predictions. This problem is known as adversarial classification problem. Classifier can be reverse engineered by sending queries by attackers and then attacker can identify malicious instances which classifier cannot recognize, this problem is known as adversarial learning problem which is introduced by Lowd and Meek [41].

Statistical machine learning can be used to attack spam filter [42]. Review study given by Barreno et al. [5] states interaction between defender and attacker along with classification on spam filter, spambayes. PDF Malware detection can be evaded by using gradient based method which is proposed by Biggio et al. [43]. Srndic and Laskov [44] studied classifier's performance under evasion attacks and states that there is significant drop in performance under simple attacks. These attacks require attacker to have knowledge of target system.

- Attacks in which attacker do not have any knowledge of target system.

There are adversarial attacks in which there is no requirement of attacker to have knowledge of the system. Xu et al. [45] proposed an evasion attack which can find a malicious example using which attacker can evade detection. EvadeHC evasion attack proposed by Dang et al. [46] first modifies malware randomly which in turn generate malware collection and then depending on binary detectors rejected or accepted result, find malware that can evade the detection from the collection.

Adversarial attacks can be performed on deep learning systems as well. Malware classification approach in Deep Learning based security detection is proposed in [47]. Some research is conducted in real world system. Face recognition in Biometric system [48], Road sign recognition attack [35], cell phone camera attack [49] and 3D object attack [50] 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 or model extraction attack. It was first proposed by Tramer et al. [51]. It is a black-box attack type where attacker extract information and then reconstruct a model by creating a substitute model which acts like target model. Model stealing method which works on black box approach where attacker use deep learning to build model form obtained predicted labels from target labels. This approach is proposed by shi et al. [52]. Chandrasekaran et al. [53] states that model extraction is equivalent to active learning by presenting model extraction into query synthesis active learning and proposed extraction attacks with no information.

C. Stealing Hyper Parameters

There is another dimension to model extraction attack's research that is Stealing hyper parameters by using a learner. Gradient of model is initialized to zero and hyper parameters are calculated by solving linear equations. This method is proposed by Wang and Gong [54] where hyper parameters of model can be stolen from machine learning algorithms like SVM, Ridge regression, logistic 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. [55] which learn a model by querying gradient information of target model for inputs. It states that model parameters can be revealed by gradient information quickly. This method has high computational overhead.

D. Recovery of Sensitive Training Data

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.

Model inversion attack was first introduced by Fredrikson et al. [56] where they propose that an attacker can recover patient's genomic information by using black box access and auxiliary information about a patient. They further involved their study in [57] where they state that by exploiting confidence values of predictions, a model inversion attack can be performed. These attacks recreate training data or labels either partially or completely. 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. [58]

Membership inference attack is introduced by Shokri et al. [59] in which attacker can estimate if a given data is in training data of model or not. Target model's prediction is used to train the membership inference model and then this model can identify difference in prediction of target model on its training data and data that has not been used for its training according to output of target model. Attacker can use both active and passive attacks in a collaborative environment so white box membership inference attack can be more dangerous to ML systems in terms of accuracy [60]. Deep Learning Models are also vulnerable to membership inference attacks [61,62,63]. The main aim of these attacks is to get the information about the training data using data generating modules.

4. Recommendations

1. Machine Learning Models and Systems are vulnerable to various types of attacks and is at risk so it becomes necessary to build a secure and privacy preserving ML models.

2. This Survey will guide researcher through various security attacks on ML categorized based on location of attack in ML Pipeline.

3. Adversarial attacks and Poisoning attacks on ML are transferable to other systems and can generalize well to any ML model.

4. This paper will provide guidelines for designing secure, robust and privacy preserving ML Systems.

5. Conclusions

Machine learning models are being widely used in various domains for different applications. This paper provides a survey on various threats and attacks that can compromise the security of these machine learning systems. 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. The major focus for defending the ML System should be on Poisoning attacks and adversarial attacks which are transferrable to other systems as well as they can generalize to any ML system.

6. Future Scope and Directions

Machine learning is very active research in various domains and research is being conducted on ML extensively. Following are some future directions on Machine Learning Security:

- **Attack in real scenarios:** Most of attacks proposed are done in simulating environment and major part of the research of security of Machine Learning algorithms is done using simulating environment. But conditions and setups in real world can impact these algorithms and security attacks differently that is why there is need of conducting research in real world scenarios.
- **Security for ML Models:** Security of ML Models are to be focused more in research so as to provide a robust and secure ML Models which can directly be implemented in various applications.
- **Privacy preserving ML Models:** With increased attention to ML application, there is need of focus on privacy-aware or privacy-preserving architecture and approaches of Machine Learning Models. There are several privacies related issues like access control, protection of model's parameters from service providers, protecting sensitive information from third parties, etc. There is need of improving efficiency of cryptographic approaches in ML.

References

[1] 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.

[2] 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.

[3] 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 arXiv:1804.07933

[4] 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.

[5] 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.

[6] 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

[7] 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

[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, I. Pillai, S. R. Bulò, D. Ariu, M. Pelillo, and F. Roli, "Is data clustering in adversarial settings secure?" in AISec '13: Proceedings of the 2013 ACM workshop on Artificial intelligence and security, November 2013 Pages 87–98, https://doi.org/10.1145/2517312.2517321.

[13] 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.

[14] 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

[15] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and

countermeasures for regression learning", in 2018 IEEE Symposium on Security and Privacy (SP), pages 19–35, DOI:10.1109/SP.2018.00057.

[16] 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.

[17] 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.

[18] C. Yang, Q. Wu, H. Li, and Y. Chen, "Generative poisoning attack method against neural networks", in arXiv:1703.01340, 2017.

[19] L. Muñoz-González et al., "Towards poisoning of deep learning algorithms with back-gradient optimization," in Proc. 10th ACM Workshop Artif. Int. Secur., Nov. 2017, pp. 27–38.

[20] M. Mozaffari Kermani, S. Sur Kolay, A. Raghunathan, and N. K. Jha, "Systematic Poisoning Attacks on and Defenses for Machine Learning in Healthcare", IEEE Journal of Biomedical and Health Informatics 19(6), pages 1893– 1905, July 2014, DOI:10.1109/JBHI.2014.2344095.

[21] M. Fang, G. Yang, N. Z. Gong, and J. Liu, "Poisoning attacks to graph-based recommender systems", in ACSAC '18: Proceedings of the 34th Annual Computer Security Applications Conference, December 2018, Pages 381–392, https://doi.org/10.1145/3274694.3274706.

[22] C. Miao, Q. Li, H. Xiao, W. Jiang, M. Huai, L. Su, "Towards data poisoning attacks in crowd sensing systems", in Mobihoc '18: Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing, June 2018, Pages 111–120, https://doi.org/10.1145/3209582.3209594.

[23] 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.

[24] Ahmed Salem, Michael Backes, Yang Zhang, "Get a Model! Model Hijacking Attack Against Machine Learning Models" in 2021 arXiv:2111.04394.

[25] A. I. Newaz, N. I. Haque, A. K. Sikder, M. A. Rahman, A. S. Uluagac, "Adversarial Attacks to Machine Learning-Based Smart Healthcare Systems" GLOBECOM-2020 IEEE Global Communications Conference DOI:10.1109/GLOBECOM42002.2020.9322472.

[26] Y. Ji, X. Zhang, T. Wang, "Backdoor attacks against learning systems," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021, pages 1-9, DOI:10.1109/CVPR46437.2021.00614.

[27] 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).

[28] 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)

[29] 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.

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

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

[32] 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).

[33] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, "Intriguing properties of neural networks", in arXiv:1312.6199v4 (2014).

[34] 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.

[35] 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.

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

[37] K. Auernhammer, R. T. Kolagari, M. Zoppelt, "Attacks on Machine Learning: Lurking Danger for Accountability", in Conf. of AAAI Workshop on Artificial Intelligence Safety, Jan. 2019.

[38] 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.

[39] P. Sharma, D. Austin, H. Liu," Attacks on Machine Learning: Adversarial Examples in Connected and Autonomous Vehicles", in 2019 IEEE International Symposium on Technologies for Homeland Security (HST), DOI: 10.1109/HST47167.2019.9032989.

[40] 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.

[41] 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.

[42] 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.

[43] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, F. Roli, "Evasion attacks against machine learning at test time", in ECMLPKDD'13: Proceedings of the 2013th European Conference on Machine Learning and

Knowledge Discovery in Databases - Volume Part III, September 2013, Pages 387–402, https://doi.org/10.1007/978-3-642-40994-3_25.

[44] 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.

[45] 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.

[46] H. Dang, Y. Huang, E. C. Chang, "Evading classifiers by morphing in the dark", in CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, October 2017, Pages 119–133, https://doi.org/10.1145/3133956.3133978.

[47] K. Grosse, N. Papernot, P. Manoharan, M. Backes, P. McDaniel, "Adversarial examples for malware detection", in ESORICS 2017: Computer Security – ESORICS 2017 pp 62–7.

[48] 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.

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

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] Maria Rigaki And Sebastian Garcia, "A Survey of Privacy Attacks In Machine Learning", in Arxiv:2007.07646, Stratosphere Project 2020.

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

[60] M. Nasr, R. Shokri, A. Houmansadr, "Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning", 2019 IEEE Symposium on Security and Privacy (SP) https://doi.org/10.48550/arXiv.1812.00910.

[61] 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.

[62] 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.

[63] 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.

Authors' Profiles

Akshay Dilip Lahe is pursuing his Ph.D. from Kalinga University, Raipur (CG), India in Computer Science and Engineering. He received M.E. Degree in Computer Science and Engineering from Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MH). His research focuses on Privacy and Security of Systems along with user data and Machine Learning. He is currently an Assistant Professor at Saraswati College, Shegaon in Maharashtra.

How to cite this paper: Akshay Dilip Lahe, Guddi Singh, "A Survey on Security Threats to Machine Learning Systems at Different Stages of its Pipeline", International Journal of Information Technology and Computer Science(IJITCS), Vol.15, No.2, pp.23-34, 2023. DOI:10.5815/ijitcs.2023.02.03