A REINVENT SURVEY ON MACHINE LEARNING ATTACKS

Chetan Patil *, Dr. Zuber

Departement of Computer Science & Engineering., *Madhyanchal Professional University Ratibad, Bhopal, M.P., INDIA*

*Corresponding author : chetanhpatil@gmail.com

Abstract. The increasing prevalence of machine learning technology highlights the urgent need to delve into its insinuations for safety and confidentiality. While inquiry on the safety aspects of mechanism knowledge has garnered considerable attention, privacy considerations have often taken a backseat, although recent years have seen a significant upswing in privacy-focused research. In an effort to contribute to this growing field, we conducted an analysis encompassing more than 40 articles addressing privacy threats in the context of mechanism knowledge, published ended the historical seven centuries. We have contributed to this research by creating a thorough threat architecture and an assault taxonomy. These tools help in categorizing various attacks based on the assets they target and the knowledge adversaries possess. We also conducted an in-depth exploration of the different privacy threats posed by machine learning, shedding light on their mechanisms and implications. Furthermore, our research includes a preliminary investigation into the underlying reasons for privacy breaches in machine learning systems. This aspect delves into the root causes of privacy leaks, shedding light on the factors that make such incidents more likely to occur. In addition to identifying privacy threats and their causes, we have compiled a summary of the most commonly proposed defense mechanisms against these threats. These defences can serve as a resource for organizations and researchers seeking to bolster the privacy of their machine learning systems. Lastly, we recognize that the field of machine learning privacy still faces unanswered questions and developing difficulties. As such, we encourage further research and exploration of potential future areas of interest. By addressing these unresolved issues and embracing emerging technologies and methodologies, we can better safeguard the privacy of individuals in an increasingly data-driven and machine learning-driven world.

Keywords: Overview machine learning, Different attacks in ML, Defence against At-tacks, Prone of ML attacks and study of ML attacks

1 Introduction

Machine intelligence live well important advances in research and practical requests, generally on account of the bounty of dossier convenient today and mechanics advances. Skilled is too a increasing interest in the impact of machine intelligence on safety, privacy or justice. When it meets expectations solitude, principal part connected to the internet services accumulate private news and use it to build models that support machine intelligence sciences. It is confused how much light these models scrap on the dossier used to train ruling class, and if they do. An attack place an opponent can obtain impressionable dossier to a degree area, well-being data, or private facts from a model prepared accompanying that dossier is less good [1]. The same attack maybe used to discover unjustified use of dossier and assure consumer privacy if private news is secondhand by addi-tional bodies without the holder's consent. In addition to the attacks themselves, there is growing interest in determining what constitutes a violation of solitary and in what settings models are vulnerable to various attacks related to solitude. Facts leaks from models for many reasons. Some of these are fundamental in type and concern model creation, while remainder of something arise from belongings like weak inference or remembering sensitive dossier patterns. Leak-age rates can more experience by opposing fighting preparation. Privacy and solitude attacks in machine intelligence algorithms are the main field concerning this study. Attacks that attempt to control the education dossier model or knowledge. Skilled are isolated reports of solitude rapes on this material [2] and few previous surveys [3]. Still, these documents extravagantly devote effort to something excessively complex or small attack samples.

Many studies have been conducted throughout society about machine intelligence's solitude and the effects of different threats on model efficacy [4]. According to the plant research described in [5], there are three different ways that machine intelligence systems can be attacked. : Attacks on method chance, containing pollute attacks to increase misclassification mistakes, integrity attacks, to a de-gree avoidance, and reverse attacks that bring about misclassification. Few pat-terns and attacks on solitude and secrecy. That is, attacks that attempt to extract news from consumer recommendation and model news. Even though opposing attacks are the most prevalent type of machine intelligence attack, the term "op-posing attack" is used to illustrate protection attacks, especially those that use opposing models. This survey only covers solitude and secrecy attacks.

Rigidly sense, a model secrecy attack is an attack at which point analyses about the model structure and limits are acquired. Model solitude attacks are frequently guide solitude attacks in previous information [6], therefore the reso-lution to adjoin model distillation attacks. Theft the performance of the model could still be deliberate as an attack of solitude, that is a important factor. In ac-cordance with Veale and others. [7] Solitude interruptions in the way that partic-ipation inferences increase the feasibility that machine intelligence models will be deliberate individual dossier under Economic unit dossier protection rules, as they can create an individual capable of being traced. Patterns are not now shielded for one GDPR, but may affiliate with organization the future, so attacks against ruling class are liable to be subjected the unchanging protections as at-tacks against individual dossier. The potential for pattern-excavating attacks to serve as a starting point for future attacks maybe important The first meticulous test of the attacks on machine intelligence systems are individual of the main gifts to this item.

• A generalized attack on machine learning.
• A thorough explanation of how the attacks were carried out.
• A summary of the many defenses that have been tried to fend off the various attacks.

The remainder of the essay is composition as aliénées: A brief review of machine learning is présente in Chambre 2 along with some fundamental principles. Has discussed why machine learning is vulnerable to attacks in Section 3. The focus of Section 4 is a summary of the suggested defenses to each sort of attack. A re-view of the protections against attacks is included in Section 5. Finally, the thor-ough investigation of several ML attacks is complete.

2 Overview of Machine Learning

In the discipline of machine learning (ML), one of the main problems is how to use the information to learn without actually writing any code. This section's goal in regard to the study is to give a broad overview of mechanism education in order to facilitate the debate in the following chapter. Delivers a quick, elevated over-view of different apparatus education methodologies, classifications, and architec-tures [8].

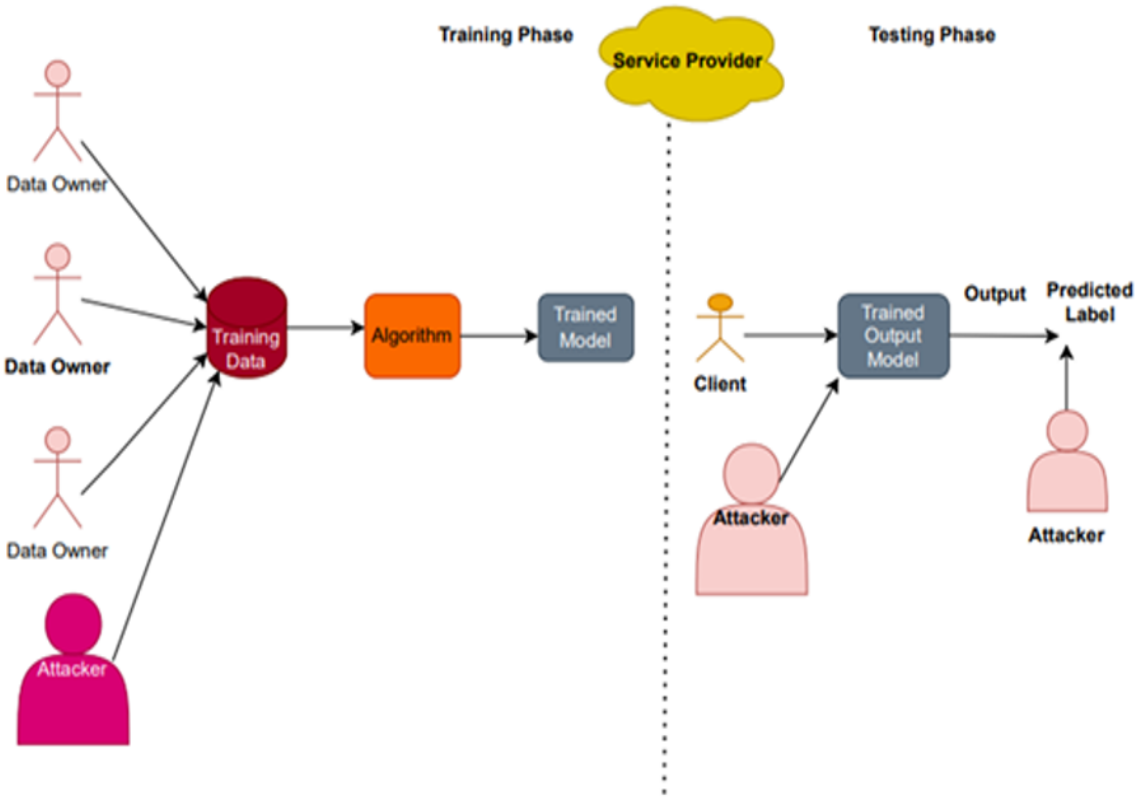

Fig 1: Overview of Machine Learning system

2.1 Types of Learning

Traditionally, supervised learning, unsupervised learning, and reinforcement learning have been the three primary categories into which machine learning (ML) has been separated. These regions are divided into separate subdivisions. Other classification schemes, such generators or discriminative models, or learning techniques that don't fall into these three categories—including semi-supervised and self-supervised learning—have developed over time. [9].

2.1.1 Supervised Learning.

In a directed knowledge scenario, a parameterized classical f is a purpose that maps an input x to an output y = f(x;), anywhere x is an n-dimensional feature or feature vector. Depending on the training task, the output or y can have several dimen-sions. A usual of arguments D = (xi, yi)m i=1 is used as the training set of the model, where m is the number of completed orders. Classification and regression problems are the maximum shared oversaw education responsibilities. Linear deterioration, regres-sion, decision trees, support vector machines, and many other algorithms are examples of supervised learning algorithms. Most attack publications to date have focused on supervised learning using deep neural networks [10].

2.1.2 Unsupervised Learning.

Unconfirmed knowledge is possible deprived of tickets. The exercise usual D contains lone the input xi. Without access to labels, unsupervised algorithms attempt to identify patterns or structures in the data. Common problems in unsupervised learning include dimensionality reduction, anomaly detection, and clustered feature learning. It appears that language models are the main target of attacks on unsupervised learning in the context of this investigation. [11].

2.1.3 Reinforcement Learning.

Strengthening knowledge deals with managers observing their environment and us-ing that information to direct actions to maximize reward signals. A sequence of actions is not predetermined in its broadest definition, and rewards may come after a sequence of actions rather than immediately. Reinforcement learning has not been subject to privacy attacks, but other privacy attacks have exploited it [12].

2.1.4 Semi-supervised Learning.

Compared to unsupervised learning, getting high-quality labels can be costly in many real-world scenarios, and there may be far less labeled data available. Semi-supervised learning algorithms are made to use labeled examples to direct the subsequent training task after learning a high-level representation of unlabeled input. Using an unsupervised learning strategy, such as clustering on unlabeled data, and then applying classification to separate representative training samples from each cluster, is an example of semi-supervised learning. An even more significant example are generative models like adversarial generational networks. [13].

2.1.5 Reproductive and Discriminative Knowledge.

Discriminative and generative algorithms are another way to classify learning algorithms. Discriminant classification aims to directly train a decision tree that distinguishes between dissimilar lessons founded on the effort records x. That is, we want to model the conditional distribution p(y|x), which is the conditional probability of a particular outcome. Two examples of these algorithms are logistic regression and neural networks. The generative classifier is attempting to capture the joint distribution p(x, y). Naive Bayes is an excellent example of a classifier. A generative model that attempts to express p(x) explicitly or implicitly, but does not require a label. A specific example is the ability of GANs or variational auto-encoders [14] to generate linguistic replicas that can forecast the following term assumed sequence data or an input text that matches features in the exercise records.

2.2 Knowledge Buildings

From a scheme design perspective, the learning process can be viewed as either centralized or decentralized. This classification is usually based on the sample and whether or not the data has been aggregated [15].

2.2.1 Centralized Learning.

A centralized learning environment contains both the data and the model. All of the data is gathered in one location and utilized to train the model, even though there may be one or more data creators or owners. In a data center, data might be found on a single system or even multiple devices. We utilize the model and collocation of inputs as the fundamental criterion to distinguish between centralized and distributed learning, thus while using concurrency in the form of several GPUs and CPUs might be regarded as a distributed learning approach, it is not for us. A centralized learning architecture includes configuring machine learning either as a service (MLaaS), which entails the data owner uploading their data to a cloud service. [16].

2.2.2 Distributed Learning.

The need for distributed learning systems is being driven by a number of causes, including the necessity for processing and managing enormous volumes of data, the requirement for computing power and memory, and even privacy concerns. We discuss privacy-aware distributed learning models, such as federated or collaborative continuing education, fully decentralized peer-to-peer learning, and distributed learning. The goal of col-laborative or amalgamated knowledge, a type of dispersed learning, is to learn a single universal model using statistics stored on several isolated plans or positions [17]. The primary goal is to prevent data from leaving remote devices. Local pro-cessing of the data is done before it is used to update the regional models. A cen-tral server receives updates from intermediate models, combines them and cre-ates a global model. The global model is then sent again from the central server to all participating devices. Peer-to-Peer (P2P) learning is completely decentralized and without a central orchestration server. Instead, devices exchange updates directly with each other using P2P communication. This setup can be intriguing from a privacy standpoint because it does away with the requirement to have faith in a central server. Attacks on P2P networks, however, are suitable in these situations and need to be taken into account. These systems are susceptible to privacy-based attacks, albeit none have

been reported to yet. Furthermore, a variety of assaults on cooperative learning could be pertinent based on the nature of the information shared among peers. Partitioned learning divides the learned model into one or more parts. The first layers of a deep learning model are maintained by edge devices, while its last layers are maintained by a centralized server [18]. To reduce com-immunisation overhead, the split was created by providing transient model outputs rather than raw data. This configuration is also applicable when distant or authority plans with limited resources are linked to a centralized mist waitron. This situation is typical for IoT devices.

2.3 Modification Among Mechanism knowledge and profound Knowledge

Engine knowledge and profound education represent distinct approaches within the field of artificial intelligence. Machine learning encompasses a wide array of techniques, from traditional statistical models to decision trees, and it often requires manual feature engineering. These methods can be effective with smaller datasets and are more interpretable. On the additional pointer, deep knowledge, a subset of machine erudition, is characterized by deep neural networks with numerous layers that can automatically learn features from raw data. While powerful for tasks like image and speech recognition, deep learning models demand substantial computational resources and vast amounts of training data. They often lack interpretability, making them akin to "black boxes." The choice between these two approaches depends on the specific problem, data availability, and the need for interpretability in a given AI application.

3 Why machine learning is prone to attacks

Models based on adversary knowledge are able to represent many attack surfaces versus machine learning models. The level of expertise ranges from modest, such as having complete knowledge of parameter estimation and training settings, or having access to machine learning APIs. There are a number of options in between these two extremes, such as having only a limited understanding of the model's hyper parameters, training environment, or structure. From a dataset perspective, adversary knowledge can also be taken into account. The authors of most of the reviewed publications assume that the adversaries are not aware of the trained data samples, although they might be aware of the fundamental delivery of the records [19].

Black-box doses are those where the opponent is unaware of the de-sign variables, structure, or training data from a taxonomic point of view. Machine learning as a service is an instance of a dark container system in which consumers typically input data in exchange for receiving the class label or prediction vector of a pre-trained model [20].

Most black-box publications assume a prediction vector. Similar to black-box attacks, white-box attacks occur when an adversary has full admission to either the mark replica's training damage grades or its parameters. This is true, for example, of most distributed training methods. Be-tween these two extremes, there are other attacks that, while not assum-ing full access to model parameters, make assumptions that are more ro-bust than black-box assumptions. These doses are referred to as in-complete white box attacks. Most work presumes complete knowledge of the anticipated input; nonetheless, it's crucial to remember that pre-processing can be required. [21].

From a taxonomic point of view, the timing of the attack is another vari-able to consider. Most research in this area focuses on attacks that occur during inference; Nonetheless, the majority of collaborative learning attacks presume that the model's gradients and parameters are accessible during training. Attacks that occur during the model's training phase give rise to many hostile behaviour possibilities. An innocent but interested or passive attacker only tries to derive information during or after training and does not delay with the exercise procedure. An opponent is considered an aggressive at-tacker if they disrupt training in any way [22].

4 Different attacks on ML

Machine learning attacks encompass a range of malicious tactics designed to compromise, manipulate, or exploit machine learning models and their data, posing a substantial threat to the security and privacy of these systems. Adversarial attacks, for instance, involve subtly altering input data to deceive models into making incorrect predictions, while data poisoning involves manipulating the training data to influence a model's learning process, potentially leading to biased or erroneous outcomes. Model inversion and membership inference attacks can breach privacy by extracting sensitive data from model outputs or identifying whether specific data points were part of the training dataset. Model evasion and reconstruction attacks seek to circumvent model defenses and reveal confidential information, while model theft involves stealing the machine learning model itself. In online learning, adversaries can manipulate data and feedback to influence a model's behavior. Backdoor attacks and privacy attacks further exploit vulnerabilities in models and data to compromise security and privacy. Protecting against these attacks requires robust security measures, including model testing, data sanitization, and privacy-preserving techniques, as well as ongoing monitoring to detect and respond to potential threats. As machine learning continues to advance, researchers and practitioners are developing new defences to counter these evolving challenges. The aim of an adversary in a privacy assault is to get data that was not intended to be shared. This information

may include details of the exercise records D, details of the perfect, or smooth evidence obtained from the records itself, including randomly introduced biases

.

4.1 Membership Inference Attacks

Determining whether a sample of input data x has been included in the training set D is the aim of membership inference. The most well-known kind of attack is this one, and Shokri et al. [101] was the first to introduce it. The attack was conducted on supervised machine learning models and utilizes only the model output prediction vector (black-box). This kind of white-box attack is not without risk, especially in a collaborative environment where an opponent might launch both aggressive and passive attacks. Accurate execution of white-box membership inference attacks is possible if model parameters and gradients are supplied. [23]. In addition to their classification function, generative models like GAN or VAE are susceptible to membership inference attacks.

Finally, from the perspective of the data owner, these attacks might be viewed from a fresh perspective. To determine whether the data were utilized without their permission in this situation, the data owner may be able to audit the black box mod-els. [24].

4.2 Reconstruction Attacks

Rebuilding doses aim to duplicate the exercise sample or examples, as well as the training labels associated with each sample. Partial or complete reconstruction may occur. In earlier publications, techniques that attempt to recover sensi-tive characteristics or an entire sample of data have been described using terms such as attribute inference and model inversion assumed production tickets and incomplete information of certain structures. All of these attacks are considered to be part of the broader category of reconstruction attacks for the purposes of this survey. The phrase “quality implication” is charity in several contexts in the confidentiality poetry to refer to doses that use publicly available data to infer the private “characterise-tics” of a targeted user [25]. Since these attacks target an individual's data directly instead of ML models, they are not included in this research. True data reconstruction and the production of class representations or probabilistic values of sensitive qualities that aren't always included in the training dataset are two important areas of contrast among the works in this category. [26]. The second case in classification models is limited to situations where the classes consist mostly of one type of item, such as identical human faces. Alt-hough this bounds the usefulness of the dose; it can occasionally present an in-retesting circumstance.

4.3 Property Inference Attacks

Feature inference refers to the ability to infer attributes of a dataset that were not openly prearranged as characteristics or that was unrelated to the knowledge objective. When data regarding the proportion of females and males in a patient da-taset is extracted without this information being a coding attribute or dataset la-bel, this is an example of feature inference in action. Or use a neural network to classify gender and determine whether the subjects on which the data is trained wear glasses. This type of disclosure may have privacy issues in certain circum-stances. These kinds of attributes can also be used to gain more knowledge about the training set, which could encourage competitors to use them to build compa-rable models. The goal of feature inference is to identify knowledge that the model has randomly acquired that is unrelated to the exercise objective. It is occasionally inevitable or even necessary for the learning process for even highly generalized models to acquire attributes that are relevant to the full distribution of the input data. Possessions that can be deduced from a certain selection of exercise instances, or ultimately about an separate, are more intriguing from the adversary's point of view. Attaining characteristics within a data set [27] or a set of data characteristics [28] are the two main goals of ownership attacks so far. The second attack targeted a model that was trained together.

4.4 Model Extraction Attacks

Modeling removal is a category of black-box spells in which an attacker de-elopes a surrogate perfect f that performs remarkably like the target perfect f in order to gather information and be able to completely reconstruct the target mod-el. There are two key areas of concentration for replacement models. First, develop-op models that are accurate enough to match the target model f in the exam usual, which is derived from the distribution of the effort records and is associated with the learning activity [29]. Second, suggest a replacement. In addition to replacing the target model, there are methods that focus on retrieving information from it, such as the over in the goal function [30] or details on various aspects of neural net-work architecture, such as the number of layers, number of activation types, op-timization technique, etc.

4.5 Adversarial attacks on Machine Learning

Biggio, Battista, Lin, and Hsiao-Ying (2021). Adversarial machine learning (AML) is a relatively new topic of study that looks into defensive strategies to guard machine learning (ML) algorithms against potential security vulnerabilities associated with their employment in contemporary artificial intelligence (AI)-based systems.(41)

Abadi, Najaf. (2021).This paper examines the use of adversarial machine learning to attack condition-based maintenance (CBM) capabilities through a case study and examines the performance of a CBM system that is being attacked.42]

Uluagac Selcuk. (2019). This research tackles an opponent that can launch both targeted and untargeted attacks and only has a limited understanding of the data distribution, SHS model, and ML technique. Additionally, it presents a fresh kind of adversarial attack to take advantage of the ML classifiers used in a SHS.43]

Peter Burnap (2019).This paper presents a novel class of adversarial techniques that take advantage of a Smart Healthcare System's (SHS) ML classifiers. The adversary's partial knowledge of the data distribution, SHS model, and ML approach allows them to execute both targeted and untargeted assaults. Abbeel (2017), [44]. The goal of this study is

to investigate how adversarial learning can be applied to supervised models by examining classification behaviors and creating adversarial samples through the usage of the Jacobian-based Saliency Map attack.45]

Burnap. Peter. (2019).The authors demonstrate in this research that adversarial attacks can also be successful in degrading the test-time performance of taught policies using adversarial example fabrication techniques, and that these techniques can be applied to neural network policies in reinforcement learning.[46]-

Abbeel Pieter (2017). This work demonstrates that even with little adversarial perturbations that do not affect human perception, existing adversarial example creation approaches may be leveraged to drastically deteriorate test-time performance of taught policies.[47] Mian Ajmal (2018). The authors of this paper classify the uses of adversarial attack and defense strategies in the field of cyber security, present the most recent research on adversarial assaults against machine learning-based security systems, and highlight the dangers associated with them.[48] Elovici Yuval (2018). The authors of this paper classify the uses of adversarial attack and defense strategies in the field of cyber security, present the most recent research on adversarial assaults against machine learning-based security systems, and highlight the dangers associated with them [49].

Burnap, Peter. (2019) Overall, when adversarial samples were provided, the classification performance of two popular classifiers, Random Forest and J48, dropped by 16 and 20 percentage points, respectively, and increased after adversarial training, indicating their resilience against such attacks.[50]

The field of adversarial machine learning (AML) studies security concerns associated with the application of mechanism knowledge (ML) techniques. AML techniques use adversarial input perturbations to fool machine learning models by taking advantage of flaws in machine learning algorithms. [41]. These attacks can be used to deceive ma-chine learning models in various domains, including condition-based mainte-nance (CBM) systems and smart healthcare systems (SHS) [42] [43] [44]. Adver-sarial attacks on ML models used in intrusion detection systems (IDS) for Indus-trial control systems (ICS) can also have severe consequences, as they can poten-tially avoid the IDS and main to late bout discovery [5]. Various adversarial machine learning techniques have been used to craft adversarial samples and manipulate data to alter the outcomes of ML-based systems . Defense strategies need to be considered to protect ML models against these attacks.

Authors	Contributions	Method Used	Limitations
Hsiao-Ying, Lin., Battista, Biggio. (2021). [41]	-Adversarial Input perturbations - Real-world model stealing attacks	Adversarial in-put perturbations - Real-world model stealing attacks	N/A
Najaf, Abadi. (2021).[42]	- Adversarial - machine learning techniques - Fast Gradient Sign method	-Adversarial machine learning techniques - Fast Gradient Sign method	- Machine learning models are vulnerable to adversarial attacks. - CBM systems are vulnerable to adversarial machine learning attacks.
Selcuk, Uluagac. (2019 [43]	- Five different adversarial ML algorithms used	Five opposing machine learning algorithms were employed. HopSkipJump, Carlini & Wagner, Zeroth Order Optimization, Fast Gradient Method, Crafting Decision Tree HopSkipJump, Zeroth Order Optimization, Carlini & Wagner, Fast Gradient Method, Decision Tree Crafting	Adversarial assaults have the potential to seriously impair the performance of ML-based SHS since ML models are susceptible to them.

Peter, Burnap. (2019).)[44]	Introducing a new type of ad-versarial attacks to exploit ML classifiers in a Smart Healthcare System (SHS) - assessing the suggested ad-versarial attack's effectiveness in various SHS environments and medical equipment	-Five different adversarial ML algorithms used - HopSkipJump, Zeroth Order Optimization, Carlini & Wag-ner, Fast Gradient Method, Decision Tree Crafting	The ML models that SHS uses are susceptible to hostile attacks. - Adversarial attacks can significantly de-grade the perfor-mance of ML-based SHS
Eirini, Anthi (2019) [45]	Adversarial machine learn-ing attacks on IDS in ICS - Use of Jaco-bian-based Sali-ency Map attack for generating adversarial sam-ples	-Jacobian-based Saliency Map at-tack - Adversarial training	Classification performance reduced by 16 and 20 percent-age opinions with confrontational examples. - Performance improved with adversarial train-ing

5 Defences against attacks

Affiliation inference and reconstruction attacks are the most common types of attacks (35.7% and 31.1% of publications, respectively), with modelling extraction in second place. In the inference step, 88% of the planned attacks are per-formed. Attacks during training mostly target distributed learning methods. Black-box and white-box spells were addressed in 66.7% and 54.8% of the pub-locations, correspondingly. We also partially classify white-box attacks as belonging to the white-box category [31].

Although not always the case, some attacks could be used for a variety of learn-ing objectives and datasets. Since most attacks are empirical, the size of the dataset, the number of classes, and the features may also play a role in their effec-tiveness. Table 2 lists all of the assault articles' datasets together with details about the size of each dataset, the learning task for which it was employed, and the types of feature data that were used. The information was used for training the target models and, occasionally, as supplemental data during assaults. The table includes 51 distinct datasets that were utilized in 42 studies, which shows how various techniques might vary [32].

Differential privacy (DP), which guarantees the effect that individual data rec-ords have on the output of an algorithm or model, is the best-known barrier against membership inference attacks. Situations involving distributed learning require additional considerations of difference confidentiality. In the federal para-digm, the main emphasis is the sample level of DP or privacy protection at the level of individual data points. We are concerned not only with the specific training data

points that each participant uses in a federated learning scenario, but also with protecting participant privacy. Access to lost gradients is often needed for reconstruction attacks during training. Most defenses against reconstruction at-tacks design methods that impact the data collected from these gradients [33].

Differential privacy does not appear to provide protection against property infer-ence attacks and is intended to provide privacy protection under the circum-stances of a participatory inference attack [34]. In [35] focused on other guaran-tees against property attacks in addition to DP. Normalization (dropping out) had a negative impact and actually strengthened the attacks. But since the attacks were performed in a group context, the authors explored the idea of sharing fewer gradients among the trainees. Less information was shared, which reduced the severity of attacks but did not stop them completely.

Attacks involving model extraction often involve the attacker running several queries against the target model. The identification of these requests has been the goal of the proposed defences so far. This differs from the previously discussed defences, which primarily sought to prevent attackers [36].

Differential privacy does not appear to provide protection against property inference attacks and is intended to protect privacy under conditions of participation inference attacks. In [37] investigated other protections against property infer-ence attacks in addition to DP. Normalization (dropout) had a negative effect and actually increased the attacks. However, since the attacks in [38] were performed in a group context, the [39] explored the idea of sharing fewer gradients among training participants. Less information was shared, which reduced the severity of the attacks but did not stop them completely.

In attacks that involve model extraction, the attacker often performs multiple queries against the target model. Identifying these queries has been the goal of proposed defences. This differs from the previously discussed defences, which were primarily aimed at repelling attackers [40].

6.Conclusion

In conclusion, the growing concerns among scientists regarding the influence of mechanism knowledge on safety, confidentiality, fairness, and explain ability have led to significant advancements in the field. Our research has contributed by presenting a security model and a unified taxonomy for categorizing various attack types, primarily focusing on recent privacy-related attacks. We've identified fundamental design patterns and highlighted the latest state-of-the-art developments. While current experimental investigations have yielded valuable insights into the variables influencing privacy breaches, it is surprising that there are relatively few studies testing attacks on real-world data sizes and deployments. Future research in security, explain ability, and fairness is imperative to address these evolving challenges. Despite the experimental stage the community finds itself in regarding privacy breaches in machine learning systems, we believe that this survey will serve as a valuable resource, offering essential background information to both interested readers and researchers looking to delve deeper into this critical area of study.

References

[1] Denning, Dorothy E. “An Intrusion-Detection Model.” IEEE Transactions on Software Engineering SE-13 (1987): 222-232.

[2] Xu, X. (2006). Adaptive intrusion detection based on machine learning: feature extraction, classifier construction and sequential pattern prediction. International Journal of Web Services Practices, 2(1-2), 49-58.

[3] H. Sarvari and M. M. Keikha, "Improving the accuracy of intrusion detection systems by using the combination of machine learning approaches," 2010 International Conference of Soft

Computing and Pattern Recognition, 2010, pp. 334https://doi.org/10.1109/SOCPAR.2010.5686163

[4] Mukherjee, S., & Sharma, N. (2012). Intrusion detection using naive Bayes classifier with feature reduction. Procedia Technology, 4, 119-128.

[5] Farnaaz, Nabila, and M. A. Jabbar. "Random forest modeling for network intrusion detection system." Procedia Computer Science 89 (2016): 213-217.

[6] Lee, J. H., Lee, J. H., Sohn, S. G., Ryu, J. H., & Chung, T. M. (2008, February). Effective value of decision tree with KDD 99 intrusion detection datasets for intrusion detection system. In 2008

[7] 10th International conference on advanced communication technology (Vol. 2, pp. 1170-1175). IEEE.

Tsai, C. F., & Lin, C. Y. (2010). A triangle area based nearest neighbors approach to intrusion detection. Pattern recognition, 43(1), 222-229.

[8] Khammassi, C., & Krichen, S. (2017). A GA-LR wrapper approach for feature selection in network intrusion detection. computers & security, 70, 255-277.

[9] Osuna, E., Freund, R., & Girosi, F. (1997, September). An improved training algorithm for support vector machines. In Neural networks for signal processing VII. Proceedings of the 1997 IEEE signal processing society workshop (pp. 276-285). IEEE.

[10] Tsai, C. F., Hsu, Y. F., Lin, C. Y., & Lin, W. Y. (2009). Intrusion detection by machine learning: A review. Expert Systems with Applications, 36(10), 11994-12000.

[11] Patcha, A., & Park, J. M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer networks, 51(12), 3448-3470.
[12] Mandala, S., Ngadi, M. A., & Abdullah, A. H. (2007). A survey on MANET intrusion detection. International Journal of Computer Science and Security, 2(1), 1-11.
[13] Ahmed, G., Hussain, M., & Khan, M. N. A. (2014). Characterizing Strengths of Snort-based IDPS. Research Journal of Recent Sciences ISSN, 2277, 2502.
[14] Mahfouz, A., Abuhussein, A., Venugopal, D., & Shiva, S. (2020). Ensemble classifiers for network intrusion detection using a novel network attack dataset. Future Internet, 12(11), 180.
[15] MeeraGandhi, G. (2010). Machine learning approach for attack prediction and classification using supervised learning algorithms. Int. J. Comput. Sci. Commun, 1(2), 247-250.
[16] Khan, L., Awad, M., & Thuraisingham, B. (2007). A new intrusion detection system using support vector machines and hierarchical clustering. The VLDB journal, 16(4), 507-521.
[17] Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. Journal of Big data,
7(1), 1-29. https://doi.org/10.1186/s40537-020-00318-5
[18] Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.
Kuang, F., Xu, W., & Zhang, S. (2014). A novel hybrid KPCA and SVM with GA model for intrusion detection. Applied Soft Computing, 18, 178-184.
[19] Mayhew, M., Atighetchi, M., Adler, A., & Greenstadt, R. (2015, October). Use of machine learning in big data analytics for insider threat detection. In MILCOM 2015-2015 IEEE Military Communications Conference (pp. 915-922). IEEE.
[20] Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: techniques, datasets, and challenges. Cybersecurity, 2(1), 1-22.
[21] Liao, H. J., Lin, C. H. R., Lin, Y. C., & Tung, K. Y. (2013). Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications, 36(1), 16-24.
[22] Choudhary, S., & Kesswani, N. (2020). Analysis of KDD-Cup'99, NSL-KDD and UNSWNB15 datasets using deep learning in IoT. Procedia Computer Science, 167, 1561-1573.
[23] Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009, July). A detailed analysis of the KDD CUP 99 data set. In 2009 IEEE symposium on computational intelligence for security and defense applications (pp. 1-6). IEEE.
[24] Kuang, F., Zhang, S., Jin, Z., & Xu, W. (2015). A novel SVM by combining kernel principal component analysis and improved chaotic particle swarm optimization for intrusion detection. Soft Computing, 19(5), 1187-1199. https://doi.org/10.1007/s00500-014-1332-7
[25] Ahmad, I., Hussain, M., Alghamdi, A., & Alelaiwi, A. (2014). Enhancing SVM performance in intrusion detection using optimal feature subset selection based on genetic principal components. Neural computing and applications, 24(7), 1671-1682. https://doi.org/10.1007/s00521013-1370-6
[26] Bijone, M. (2016). A survey on secure network: intrusion detection & prevention approaches. American Journal of Information Systems, 4(3), 69-88.
[27] Atefi, K., Yahya, S., Dak, A. Y., & Atefi, A. (2013). A hybrid intrusion detection system based on different machine learning algorithms.
[28] Zhao, G., Song, J., & Song, J. (2013, March). Analysis about performance of multiclass SVM applying in IDS. In International Conference on Information, Business and Education Technology ICIBIT.
[29] NERLIKAR, P., PANDEY, S., SHARMA, S., & BAGADE, S. (2020). Analysis of intrusion detection using machine learning techniques. International Journal of Computer Networks and Communications Security, 8(10), 84-93.
[30]Aburomman, A. A., & Reaz, M. B. I. (2016). A novel SVM-kNN-PSO ensemble method for intrusion detection system. Applied Soft Computing, 38, 360-372.
[31] Wang, J., Hong, X., Ren, R. R., & Li, T. H. (2009). A real-time intrusion detection system based on PSO-SVM. In Proceedings. The 2009 International Workshop on Information Security and Application (IWISA 2009) (p. 319). Academy Publisher.
[32] Lin, S. W., Ying, K. C., Lee, C. Y., & Lee, Z. J. (2012). An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection. Applied Soft Computing, 12(10), 3285-3290.
[33] Mulay, S. A., Devale, P. R., & Garje, G. V. (2010, June). Decision tree based support vector machine for intrusion detection. In 2010 International Conference on Networking and Information Technology (pp. 59-63). IEEE.
[34] Li, Y., Xia, J., Zhang, S., Yan, J., Ai, X., & Dai, K. (2012). An efficient intrusion detection system based on support vector machines and gradually feature removal method. Expert systems with applications, 39(1), 424-430.
[35] Chen, W. H., Hsu, S. H., & Shen, H. P. (2005). Application of SVM and ANN for intrusion detection. Computers & Operations Research, 32(10), 2617-2634.
[36] Mukkamala, S., Janoski, G., & Sung, A. (2002, May). Intrusion detection using neural networks and support vector machines. In Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290) (Vol. 2, pp. 1702-1707). IEEE.
[37] Ahanger, A. S., Khan, S. M., & Masoodi, F. (2021, April). An effective intrusion detection system using supervised machine learning techniques. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1639-1644). IEEE.

[38] Bhavsar, Y. B., & Waghmare, K. C. (2013). Intrusion detection system using data mining technique: Support vector machine. International Journal of Emerging Technology and Advanced Engineering, 3(3), 581-586.
[39] Bhati, B. S., & Rai, C. S. (2020). Analysis of support vector machine-based intrusion detection techniques. Arabian Journal for Science and Engineering, 45(4), 2371-2383.
[40]Hsiao-Ying, Lin., Battista, Biggio. (2021). Adversarial Machine Learning: Attacks From Laboratories to the Real World. IEEE Computer, 54(5):56-60. doi: 10.1109/MC.2021.3057686
[41] .Hamidreza, Habibollahi, Najaf, Abadi. (2021). Adversarial Machine Learning Attacks on Condition-Based Maintenance Capabilities.. arXiv: Learning,
[42] Akm, Iqtidar, Newaz., Amit, Kumar, Sikder., Mohammad, Ashiqur, Rahman., A., Selcuk, Uluagac. (2019). HealthGuard: A Machine Learning-Based Security Framework for Smart Healthcare Systems. 389-396. doi: 10.1109/SNAMS.2019.8931716
[43] Akm, Iqtidar, Newaz., Amit, Kumar, Sikder., Mohammad, Ashiqur, Rahman., A., Selcuk, Uluagac. (2019). HealthGuard: A Machine Learning-Based Security Framework for Smart
Healthcare Systems. 389-396. doi: 10.1109/SNAMS.2019.8931716
[44] Eirini, Anthi., Lowri, Williams., Malgorzata, Slowinska., George, Theodorakopoulos., Peter, Burnap. (2019). A Supervised Intrusion Detection System for Smart Home IoT Devices. IEEE Internet of Things Journal, 6(5):9042-9053. doi: 10.1109/JIOT.2019.2926365
[45] Sandy, H., Huang., Nicolas, Papernot., Ian, Goodfellow., Yan, Duan., Pieter, Abbeel. (2017). Adversarial Attacks on Neural Network Policies.
[46] Naveed, Akhtar., Ajmal, Mian. (2018). Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. IEEE Access, 6:14410-14430. doi: 10.1109/ACCESS.2018.2807385
[47] Ishai, Rosenberg., Asaf, Shabtai., Lior, Rokach., Yuval, Elovici. (2018). Generic Black-Box End-to-End Attack Against State of the Art API Call Based Malware Classifiers. 490-510. doi: 10.1007/978-3-030-00470-5_23
[48].Ishai, Rosenberg., Asaf, Shabtai., Lior, Rokach., Yuval, Elovici. (2017). Generic Black-Box End-to-End Attack Against State of the Art API Call Based Malware Classifiers. arXiv: Cryptography and Security,
[49] Eirini, Anthi., Lowri, Williams., Malgorzata, Slowinska., George, Theodorakopoulos., Peter, Burnap. (2019). A Supervised Intrusion Detection System for Smart Home IoT Devices. IEEE
Internet of Things Journal, 6(5):9042-9053. doi: 10.1109/JIOT.2019.2926365