Issues in Information Systems
Volume 22, Issue 3, pp. 224-240, 2021

DOI: https://doi.org/10.48009/3_iis_2021_242-258

Security and privacy in machine learning: A survey

Gayatri Sravanthi Kuntla, *Kennesaw State University, gkuntla@students.kennesaw.edu*
Xin Tian, *Kennesaw State University, xtian2@kennesaw.edu*
Zhigang Li*, Kennesaw State University, zli8@kennesaw.edu*

Abstract

Machine Learning has received its attention over the past decade in the security and privacy related applications. The information is enormous, and manual working on data could be difficult and makes it more complex to predict the future. Machine Learning has been introduced and applied in a wide range of technical applications in computer science to overcome this problem. Machine Learning is popularly used in real-time decision-making applications, big data processing to build an efficient time calculated and decisive solution. Machine learning model is used for classifying the data with at the most accuracy. Various kinds of machine learning algorithms and techniques are available. However, the latest research studies show that machine learning using data could be vulnerable to security attacks, malware detection, and many more. This paper highlights a survey of machine learning tools, techniques, and applications and systematically analyzes security attacks over the past few years.

Keywords: Machine Learning, security threats, survey, cybersecurity, defensive techniques

Introduction

In the latest years, technology advances in machine learning (ML) are embodied with development in computational capabilities and transformed by automation of Machine Learning as a technology platform. Machine Learning techniques impacted in daily world application and made a breakthrough in several years in various fields and altered the practices. Machine Learning models explored in the prevailing research fields by reconstruction of face biometric data from the social media and gathering sensitive medical details of the person is revealed, that may turn out to be a privacy issue of the individual and focusing on improvement in all realistic domains. However, this paper provides more insight into machine learning applications in Cybersecurity, IoT systems, network intrusion detection. In some applications, machine learning shows errors in some applications like image classification, face detection, which may exceed the errors caused by humans. These are caused due to defects in the machine learning systems, security issues and are worth attention. Understanding machine learning and security issues based on mobile devices, like android platforms and other usable equipment, makes the situation more challenging and complicated. Human intervention is required to maintain conventional software security, i.e., to analyze threats, pull out characteristics and encrypt the attributes into the software system for detection. This involves an entirely labor-demanding method and can be economical by implementing machine learning algorithms. Researchers discovered machine learning algorithms to detect incidents which is more efficient and consistent. These edited books (Barbara and Jajodia, 2002; Maloof, 2006)

were published and conducted two workshops at research conferences and in past years. (Chan et al., 2003; Brodley et al., 2004)

The outline of this survey is structured as follows: Section 1. lists out an overview of machine learning. Section 2, literature review- applications of machine learning techniques in evolving areas. Section 3 security and privacy in machine learning. Then section 4, future works, and finally, section 5, concludes the paper.

Overview of Machine Learning

Machine Learning is the field that enables computers to work on problems and decode them without any programming language required. It predicts the results based on the previous observations. The purpose of this section is to provide an overview of a high level of machine learning paradigms and classifications along with architecture. The learning technique can sum together machine learning algorithms as it varies significantly and categorizes the functions based on tasks or the depth of performing.

Machine learning is classified into three major areas: *supervised, unsupervised*, and *reinforcement learning*. Later, few more categories have emerged in more detail: *semi-supervised, active, and ensemble* learning.

Machine Learning Techniques

Based on the technique the model is trained the data, machine learning is classified into three major areas: *supervised, unsupervised* and *reinforcement learning*. Later, few more categories have emerged in more detail: *semi supervised, active and ensemble* learning.

Supervised Learning: In supervised learning, there will a pair of labeled inputs and desired outputs for the model learn from the training data. By analyzing the dataset, a mapping function generates between input(x) and output (y). The most common application tasks are Classification and Regression.

Unsupervised Learning: In unsupervised learning, there are no labels in the model for training dataset and used to find the structure or models based on the information and characteristics of data. There are no labeled inputs and desired output variables. The most common application tasks are Dimensionality Reduction, Clustering, and Association Rule Learning. Most attacks will be on unsupervised machine learning as language models.

Reinforcement Learning: Reinforcement learning, interacts with the external environment and learns itself from the actions of using trial and error. By gaining the ideas from experience, it is trained to predict future observations. There are no privacy related attacks observed in this type of learning.

Semi-Supervised Learning: It is a combined algorithm of supervised and unsupervised learning where labeled data and unlabeled data is used to achieve the desired model from the training dataset. Semi-supervised learning algorithm uses unlabeled data for superior level interpretation and then uses labeled data to simplify the downstream tasks. Tasks include Classification and Regression.

Active Learning: In active learning, training data is selected actively to reduce the large amount of labeled data with more flexibility. This influences the cost and time for collecting the labeled training data.

Ensemble Learning: In ensemble learning, multiple weak classifiers combine and are built to form strong classifiers by predicting each observation and taking individual decisions. Samples of boosting and bagging are examples of ensemble learning.

Machine learning depth is classified into shallow and deep learning to distinguish the machine learning techniques based on how deep the recognition task path is.

Shallow Learning:

Shallow Learning is a standard machine learning model in which the training dataset do not consider multiple deep layers. As a result, complexity of computation is less that develops from multiple deep layers. Thus, shallow models are limited when compared to other models and unable to detect the correlations among other models.

Deep Learning:

Deep Learning uses multiple deep layers of simple modules and reduces the complexity of the model. To achieve desired output model, this involves supervised and unsupervised learning related with the system by using labeled data and unlabeled data.

Deep Learning is used for distributed computing, analysis and learning of unlabeled, uncategorized data. Various machine learning applications uses deep learning models in contribution to speech recognition, computer vision by producing better data sets. It is also used for solving higher level of technical related problems over a large scale.

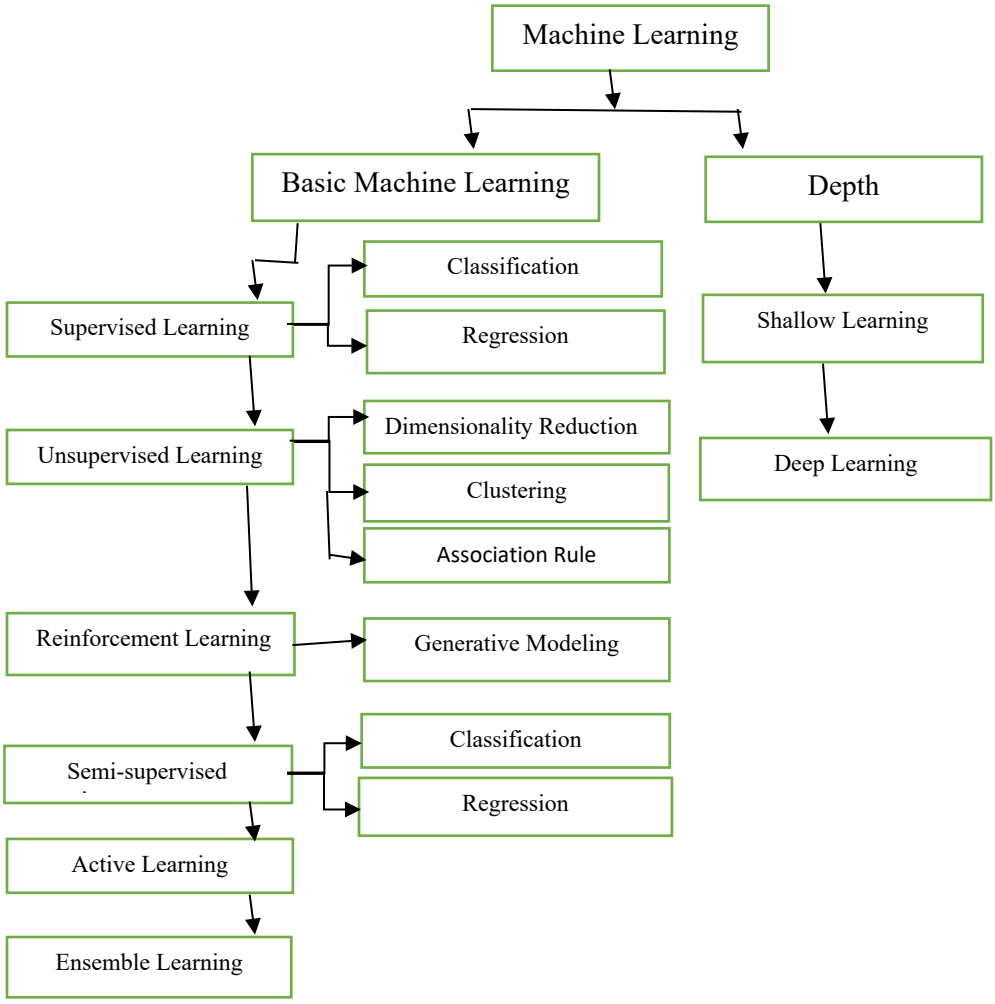

Figure 1 (Ibitoye, 2019): Classification of Machine Learning

2 Literature Review

Many companies have started using machine learning as a service. Some of the famous companies are Google and Amazon. The platform supports the ML model wherein the data holder uploads the data on a demand basis. Machine Learning has extended its application in many cloud applications and services. The details about the underlying model and the data uploaded must be maintained at the most privacy and secured. The internal information is kept secure. Owing to the increase in usage of service-related applications of machine learning in a cloud environment, there exist privacy threats towards the model and the data.

In this section, a detailed survey about the applications where machine learning is used popularly and the threat towards the machine learning model and the machine learning data is presented in detail.

References	Description	Scope			
		ML applications	**ML security**	**ML attacks**	**ML defenses**
[1]	Security matters: a survey on adversarial machine learning		+	F	F
[2]	The Threat of Adversarial Attacks Against Machine Learning in Network Security: A Survey	F		+	
[3]	A Survey of Privacy Attacks in Machine Learning			F	+
[4]	A Survey of Data Mining and Machine Learning: Methods for Cyber Security Intrusion Detection	F			
[5]	Machine Learning Security: Threats, Countermeasures, and Evaluations		F	F	F
[6]	A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data-Driven View			F	F
[7]	Machine Learning Techniques and Tools: A Survey	F			
[8]	SoK: Security and Privacy in Machine Learning		F		
[9]	When Machine Learning meets Security Issues: A Survey		F		+
[10]	A Machine Learning Security Framework for IoT Systems		F		
[11]	Machine Learning Threatens 5G Security.		F		+
[12]	The security of machine learning		F		
[13]	Applications in Security and Evasions in Machine Learning: A Survey		F	+	
[14]	Applications of Machine Learning in Cyber Security	F			
[15]	Machine Learning for Computer Security	F	+		

Figure 2 (Suomalainen et al., 2020): Surveys are classified according to their focus areas, i.e., the focus of a survey is highlighted with `F' and briefly covered areas with `+'.

The integration of the applications, security and privacy of machine learning can lead to possible security attacks and faces security issues if proper consideration is not given. Some of these evolving vulnerabilities have been identified by examining and mentioned in surveys, which are listed in fig 2.

Application of Machine Learning in Cyber Security:

Machine Learning techniques have been successfully implemented over large areas of problems of systems. Applications of machine learning in cybersecurity is categorized into two prospects:

General Application of ML in Security issues

Cyberspace security had many issues over many years before introducing machine learning. Challenges faced: 1. Due to the high volume of data, manual analysis was made impossible. 2. Security threats have evolved at a higher rate, making short use of patterns, and resulting in new threats highly.3. New threats like evasive were hard to detect and control. 4. Cost consumption has been increased thoroughly over the years. 5. Development of new algorithm and implementation involves time and cost. 5. Thus, it became more difficult for domain experts.

Finally, implementing machine learning has made a lot easier in cyberspace security. The following advantages are 1. With ML, cybersecurity systems could prevent similar attacks by analyzing the patterns and respond to altering performance. 2. Be more proactive in preventing threats and reacting to active attacks in the actual moment. 3. Distinguish normal models to anomalous ones with behavior modeling. 4.It becomes accessible and more comparable in threat domain experts. 5. Advanced Persistent Threat (APT) is well detected in real terms and future time.

Practical Applications in Cyber Security

A. Network Intrusion Detection:

Network Intrusion Detection (NID) systems identify the malignant web activity results to integrity, confidentiality, and availability breach of the computer network. It is a technology that detects threats to cybersecurity. ML uses techniques is NID for specific tasks to develop new compliance and unknown threats. Intrusion detection in cybersecurity scrutinizes network traffic and susceptible issues in the network.

The Intrusion detection process automatically executes by an Intrusion detection system (IDS). An Intrusion prevention system (IPS) can likely stop interruption on the foundation of IDS. Based on IDS, the cyber analytics method is classified into three categories: signature-based (misuse Base), anomaly-based, hybrid. (Buczak & Guven, 2016b)

Anna L. Buczak introduced clear definitions on previously mentioned scientific classification and appended immense measure of datasets: Association Rules, NetFlow Data, Public Dataset, Packet-Level Data, Bayesian Network, Clustering, Artificial Neural Networks, Inductive Learning, Naïve Bayes, Decision tree, Support Vector Machine, Ensemble Learning (Buczak & Guven, 2016b).

B. Malware Detection:

Malware Detection is used to identify all types of defects in the system and retrieve the users' information without permission. Malicious attacks are the most potent threats in cybersecurity. As business maintains a large amount of data, machine learning models are useful for such type of datasets and detects the malicious

attacks in the system. Methods for detecting malware delivery, malicious attacks, malicious web pages, ML techniques, and models are applied in the systems. A Conventional way to detect malware is by using blacklists. But these blacklists have severe limitations like lack of reliability and difficulty to detect new generated attacks. More overview of malware URLs using ML techniques is presented by Doyen Sahoo and presents more learning algorithms in this domain (Sahoo, 2017).

Mal-ID, a new methodology analysis to detect malware files, uses a segmental approach as an alternative to entire data (Tahan, 2012). The Mal-ID basic approach is more effective in real-time when compared to other alternative methods. An Alert Classification System was developed against Distributed Denial of Service (DDoS)using Support Vector Machines (SVM) and Neural Networks (NN). A virtual environment was implemented with the "Snort" technique for intrusion detection, "packit"- network packets in simulating DDoS attacks and forwarding to the target system (Saini et al., 2020). The alerts created by the snort intrusion detection tool were caught and taken care of and classifying the alerts as true-positives or false-positives by back-spread neural network and support vector machines. This process was found in a reduction of 95% in the total number of alerts, proved by the researchers.

Machine Learning in Network Security

Due to innovations, technologies in today's network, and the future, network security is a high priority as network architecture has become complicated. Machine learning techniques have implemented various networks for security and have increased its importance due to defenses in many network layers considering from start to endpoint. Here are some applications of machine learning categorized into five parts: a. ML for network protection b. ML for endpoint protection c. ML for application security d. ML for user behavior analytic. e. ML for process behavior analytic.

Applications of Network Security	Detection type	ML technique
Network Protection	Intrusion detection, Anomaly detection	KNN, Decision trees, SVM, Multi-Layer Perception
Endpoint Protection	Malware detection, Access control, Authentication	Logistic Regression, Random Forest, Naïve Bayes, Sequential Minimal Optimization (SMO)
Application Security	Malicious URL, Phishing, Spam	SVM, Neural networks, Leave one model, Biased SVM, Self-Organizing Maps
User Behavior	Keystroke dynamics, Breaking human interaction Proofs (CAPTCHA's)	Support Vector Machine (SVM)
Process Behavior	Process Anomaly, Fraud detection	Decision tree (C5.0, CandRT, CHAID) and SVM (linear and radial basis, Kernels of polynomial, sigmoid)

Figure 3: ML applications of Network Security

Security and Privacy

In recent years, the security of machine learning algorithms has attracted widespread locations, and many researchers worked on the strategies for threats and attacks. However, earlier sections explain machine learning Security in different applications- malware detection, network security, cybersecurity, and so on. This section highlights security threats, attacks, and privacy of data protection.

Mode of Adversary:

The attack behavior is focused here. The attacker will have to know a few of the information about any one of the following,

Goal: The target towards the ML is finalized. Either it will be towards the data or the model

Knowledge: The depth of the information the attacker knows about the model. It can be either a white box or a black box. In the white box, all the information is known to the attacker. While in the black box, no additional detail is known as only the target is finalized to be attacked.

Capability: The influence towards the application which uses the ML model and the data processed with it, enables to know the additional details in depth and pave the way to influence the system to get attacked

Strategy: The strategy of the attack is to embed with the ML model and optimize the same, so it behaves like optimized methods and, in turn, will create the impact

The security threat towards the Machine Learning models is based on the three major perspectives. They are security violation, influence towards the classifier, and the perspective on the attack specificity.

Causative Attack: Any modification made by the adversary towards the training data is called a causative attack.

Integrity/Availability Attack: The increase in the False Negatives [FN] and the increase in the False Positive [FP] are referred to as Integrity/Availability attacks, respectively.

Explorative Attack: The attack of the adversary that led to misclassification is termed as an explorative attack.

Privacy Violation Attack: The attack towards the privacy and the sensitive information available in the system is generally termed as a privacy violation attack.

Targeted Attack: This attack is primarily focusing on the performance of the system. The invalidation and the lack of performance are done with this targeted attack.

Indiscriminate Attack: Owing to the increase in the sample size, the attack is focused on the classifier.

Privacy Threat on model

In this section, the privacy attack towards the model is presented in detail. In (Tramèr, 2016), a model extraction attack is presented. This is a type of privacy attack. The plugin is equipped with the API to initiate the attack. The major intention behind this type of attack is to own the services in the future and the evasion sample construction by using the gradient information (Zhang et al., 2021) (Papernot, 2016). However, the model information is considered confidential information. Hence the stealing of such model information led to a privacy attack. The privacy data may include sensitive and confidential information. In (Tramèr, 2016), the authors have used decision tree, logistic regression, and Neural networks for demonstration

purposes. Here using the classification technique, the information is tried to predict. In (Papernot, 2016), The attack towards the ML model is focused. The authors have used the Jacobian matrix-based data authentication technique to predict the functionality of the underlying model. This enables the stealing of the model, which results vulnerable to the system. In (Orekondy, 2018), (J. Zhang & Li, 2019), the research focused on the usage of non-problem domain data. The unlabeled details are primarily focused here. The overview of the model that is the Black Box of the model is tried to attain here. The Black box of the ML model does not include detailed information about the individual segments. But it provides an overall detail about the entire process. In (Wang, 2018), the privacy issues were focused on the hyperparameters. The relationship between the hyperparameter, a methodology is experimented to find out. The hyperparameter is vulnerable enough to find out the function of the machine learning model. The solution of the hyperparameter was achieved by using the least squared method. In (Oh, 2017), the introduction of the model attributes to predict the model of the neural network. The dynamic input is fed into the static model and the invasion is tried to attain. (Orekondy, 2019) proposed the method of poison attacking as a mode of stealing attack. (Juuti, 2018) found that the data-stealing is done as a hierarchical process and the discrimination between the normal and the synthetic queries.

Privacy threat on data

The threat towards the data is more of the known types of knowing a fewer piece of information and lead to a pre-emptive attack mode. In (Song et al., 2017), the usage of the third-party untrusted algorithm is used for the memorization of the data. This experiments in the trusted isolation environment. The complete encoding of the training set is done without altering the built-in model. The untrusted medium will always remain a threat even after the completion of the present threat towards the data. This remains a continuous attack. In (Fredrikson et al., 2015), the model inversion attack is done by using the genotype of the information. This is done towards sensitive information. The data which is available in public is tried to be modeled out to fetch the inference out of it. The target is chosen with the publicly available sensitive information about the data. The data released out for a purpose, or a cause is multiplied with threats and ended up in a data attack. (Yang et al., 2019) presents the generally available data and the prediction vectors to reconstruct the sample through the available information. Here the partially available data were considered. Based on that the generation of the complete information is tried to achieve. In (Shokri, 2016), the usage of membership inference attack is materialized. This misleads the classification problem. The attackers construct the shadow model to perform this attack. Through it, the prediction model is tried to be framed out. (Salem, 2018) found that data transferring attack is possible even if partial data is available or only if the data is coming from the various sources. (Long, 2018) proves that whether the target record fits under the conditions but gets prevailed to the several types of attacks. (Yeom et al., 2018) explained the membership advantage, (Pyrgelis, 2017) presented the existence of the targeted individual and combined the GAN to find out the study of generative attacks.

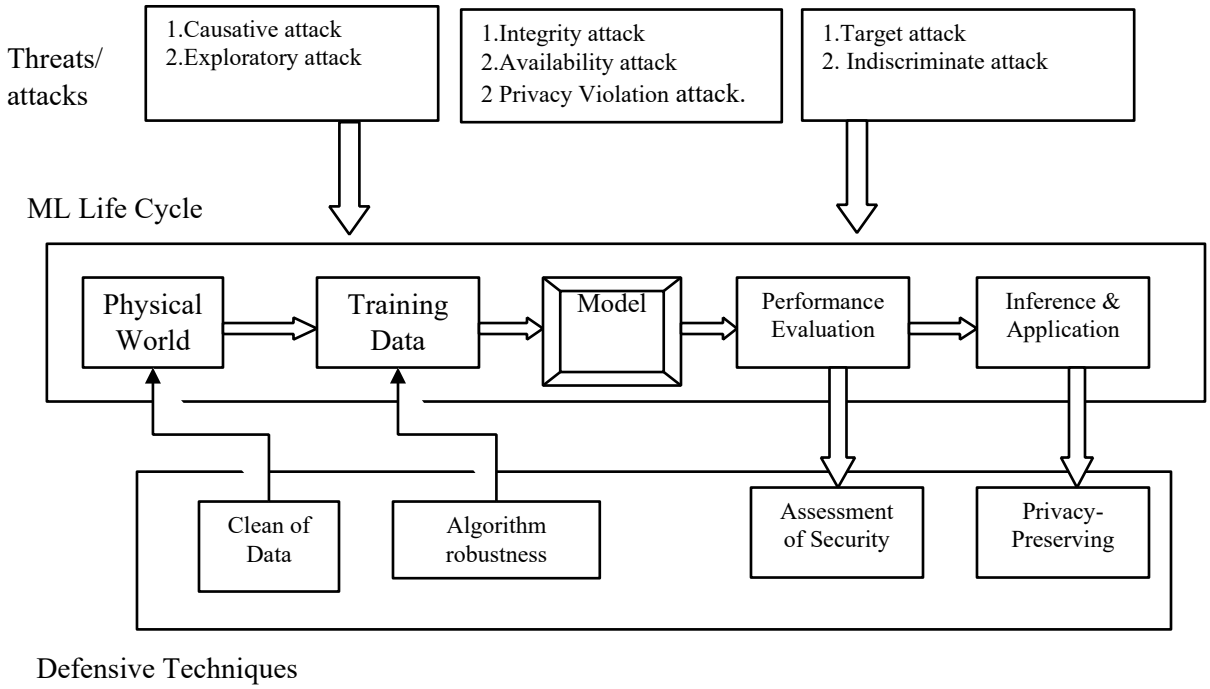

Fig 4 (Liu et al., 2018): Threats and Defensive methodologies on Machine Learning models

Privacy Protection

In this section, we present collaborative learning, Homomorphic encryption, and Differential privacy as the privacy protection techniques in machine learning applications.

Collaborative Learning:

In a centralized learning application, there arise critical situations to reveal the information to the public. The details may be sensitive information as well. To solve these kinds of issues, (Shokri & Shmatikov, 2015) presented distributed learning method to ensure privacy. The consensus of the architecture is reached by the participants of the models. Hence the privacy is preserved to a larger extent. The modeling in the local dataset is done by Stochastic Gradient Descent. The selective gradients are also shared with the server. The researchers from google (McMahan, 2016) named it federated learning. Privacy protection is enabled between the cloud and the client. The gradient parameters are uploaded, and the model is shared with the client. Hence the participants access the protected model, and therefore, the privacy is preserved in collaborative learning.

Homomorphic Encryption:

In (Rivest et al., 1978), homomorphic encryption is initially proposed. (Gentry, 2009) proposed homomorphic encryption for the Neural Network pattern. (Bansal et al., 2010) proposed a method of encryption standard to preserve the information. (Q. Zhang et al., 2016) presented another encryption method in a cloud environment. In homomorphic encryption, the primary focus is on the data. There was partial homomorphic encryption like additive encryption and multiplicative encryption presented in various

studies in state of the art. But the fully homomorphic encryption is where the data are partitioned arbitrarily based on homomorphic nature and property. Followed by which the activation function is calculated based on the portion of the partitioned information available. Piece-wise linear approximation is carried out to evaluate the activation function. Therefore, the cloud is resisted with no access to the actual data.

Differential Privacy:

This is studied in (Dwork et al., 2017) as one of the statistical methods of privacy protection methodologies. If there is any publicly available sensitive information, there arises a chance of differential attack. Differential privacy is one such method to prevent the differential and any kind of statistical attack. The statistical attack is more viable towards the data as well as the model. So, this differential privacy paves the way through dealing with statistical data. They were mainly concentrated towards handling differential attacks. Many researchers primarily used attribute correlation in the application to classify gender utilization. Meta classifier is used to infer the training set attributes. In (Phan, 2017) presented a method of Adaptive Laplace mechanism. This enables adding the essential component to fetch the resultant more secured.

Observations:

Based on the learning in the above studies and the research held in the state of the art, the following observations were made based on the security and privacy issues towards the machine learning classification techniques.

Author & Year	**Tool/Technique**	**Type**	**Method**	**Application**
Goodfellow, 2014	Gradient-Based Approaches	Adversarial attacks	Minimax Defense	ML classification
J. Kos, I. Fischer, 2018	Deep generative models	Adversarial attack	VAE- GAN	Privacy Application
S.-M. Moosavi-Dezfooli, 2016	Deep generative models	Adversarial attack	Deep Fool	NN based application.
N. Papernot,2016	Deep generative models	Adversarial attack	DNN	Privacy Application
Florian Tramèr,2016	decision tree, logistic regression and	Stealing ML model	Neural network	Confidential data Application
Xinbo Liu, 2019	Gradient-based information	Attack towards model	Any classification	All application
Nicolas Papernot, 2017	Black box attack	Attack towards model	Neural network	Confidential data Application
Tribhuvanesh Orekondy, 2019	Black box attack	Stealing ML model	Neural network	Confidential data Application
Jiliang Zhang, 2019	Unlabeled information attack	Adversarial attacks	VAE- GAN	Privacy Application
Binghui Wang, 2018	Hyperparameters Stealing	Stealing ML model	Neural network	Confidential data Application
Seong Joon Oh, 2018	Black box attack	Stealing ML model	Neural network	Confidential data Application
Tribhuvanesh Orekondy, 2019	Constrained Defenses Against Model Stealing	Stealing ML model	Neural network	Confidential data Application

Mika Juuti, 2019	Protecting model	Stealing ML model	DNN	Privacy application
Congzheng Song, 2017	Unlabeled information attack	Adversarial attacks	VAE- GAN	Privacy Application
Matt Fredrikson, 2015	model inversion attack	Model attack	Genotype of the information	Privacy Application
Ziqi Yang, 2019	Deep generative models	Adversarial attack	DNN	Privacy Application
Reza Shokri, 2017	Inference Attacks	Model attack	Genotype of the information	Privacy Application
Ahmed Salem, 2019	Data Independent Membership Inference Attacks	Data Attack	data transferring attack	Privacy application
Yunhui Long, 2018	Data Independent Membership Inference Attacks	Data Attack	data transferring attack	Privacy application
Jamie Hayes, 2019	Inference Attacks	Data Attack	Generative Models	Privacy application

Fig 5: Details regarding each of the ML tools/techniques reviewed.

In (Goodfellow, 2014), presented gradient-based approaches for handling adversarial attacks. The method used in this study is the minimax defense technique. This is considered one of the best defense methods of ML classification. The libraries of the learning are used here for computing the gradient.

Deep generative models for handling the adversarial attacks are presented in (Kos, 2017), (Moosavi-Dezfooli, 2015) and (Papernot et al., 2016). The privacy towards the neural network is focused on all these approaches, The applications are all the sensitive and confidential neural patterns.

The attacks towards the model of machine learning especially focused on handling confidential information is considered by various researchers in (Tramèr, 2016), (Zhang et al., 2021), and (Papernot, 2016). The attack is in the neural network and the classification models in (Orekondy, 2018) have also researched this attack. The black box attacks are most initiated here. Without knowing the internal details of the model, an attempt was made to attack the model with the Black Box details as generical information.

The partially labeled or the unlabeled information attack is presented in (J. Zhang & Li, 2019). With the available information, the model is tried to be hacked or attacked. It is mostly done with the application that regularly accesses the ML services. In (Wang, 2018), presented the hyperparameter stealing, the hyperparameter enables the details to know about the ML model.

The Data Independent Membership Inference Attacks, inference attacks were presented in (Shokri, 2016), (Salem, 2018), (Long, 2018), and (Hayes, 2017). These attacks are very vulnerable to the data. These are done with the ML application which is done for the public domain, sensitive application in the medical domain, and social media modeling as well. The information released public is taken for a negative cause and the privacy of the application is detained.

The above table presents the overview of the previous work done by the researchers upon the various attack towards the privacy and security issues of the machine learning models. The research towards the model as well as the data was presented.

Future Works

Based on our survey, machine learning has successful attempts carried out on data with very few security attacks on data in transit. However, applications of machine learning in cybersecurity, network security domains such as in the field of intrusion detection remain challenging for new emerging security attacks on data in transit. Hence, to understand the possible risks of security attacks and data privacy in machine learning more research is required on the possible defense techniques. Some of the major concerns in maintaining the security and privacy of data in machine learning need to be considered for the following scenarios.

- There would be constantly emerging new security attacks in the application of machine learning. This survey paper focuses on the ML-based algorithm techniques and verify them in real work.
- Privacy of data is on high priority when the use of machine learning for security and therefore, constant development in machine techniques is required. High-level cost-effective privacy technology would be challenging work to study further.
- New applications of machine learning are demanding in cybersecurity and network security. A Well-defined security assessment is needed to study for establishing high standardized techniques.

Studying further on all the above scenarios recommends a high level of security and privacy of data in machine learning.

Conclusion

The improvements of the privacy and security aspect in the machine learning models and their classification were studied in this research. The privacy and security issues towards the machine learning applications are briefly presented and the causes of the attack at the machine learning model and towards the data are researched. The existing vulnerabilities with the methods found in the literature were analyzed and briefly presented. The attacks at the training and the inference stages were presented. Owing to the application of intelligence and learning methods in every field, the proper protection methods were also researched and presented.

References

Bagaa, M., Taleb, T., Bernabe, J. B., & Skarmeta, A. (2020). A Machine Learning Security Framework for Iot Systems. *IEEE Access*, *8*, 114066–114077. https://doi.org/10.1109/access.2020.2996214

Bansal, A., Chen, T., & Zhong, S. (2010). Privacy preserving Back-propagation neural network learning over arbitrarily partitioned data. *Neural Computing and Applications*, *20*(1), 143–150. https://doi.org/10.1007/s00521-010-0346-z

Barreno, M. (2010, May 20). *The security of machine learning*. Machine Learning. https://link.springer.com/article/10.1007/s10994-010-5188-5?error=cookies_not_supported&code=22f367e9-8b0d-4422-bf37-3d3dddf98361

Buczak, A. L., & Guven, E. (2016a). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials*, *18*(2), 1153–1176. https://doi.org/10.1109/comst.2015.2494502

Buczak, A. L., & Guven, E. (2016b). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials*, *18*(2), 1153–1176. https://doi.org/10.1109/comst.2015.2494502

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2017). Calibrating Noise to Sensitivity in Private Data Analysis. *Journal of Privacy and Confidentiality*, *7*(3), 17–51. https://doi.org/10.29012/jpc.v7i3.405

Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. Published. https://doi.org/10.1145/2810103.2813677

Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. *Proceedings of the 41st Annual ACM Symposium on Symposium on Theory of Computing - STOC '09*. Published. https://doi.org/10.1145/1536414.1536440

Goodfellow, I. J. (2014, December 20). *Explaining and Harnessing Adversarial Examples*. ArXiv.Org. https://arxiv.org/abs/1412.6572v2

Guan, Z., Bian, L., Shang, T., & Liu, J. (2018). When Machine Learning meets Security Issues: A survey. *2018 IEEE International Conference on Intelligence and Safety for Robotics (ISR)*. Published. https://doi.org/10.1109/iisr.2018.8535799

Hayes, J. (2017, May 22). *[1705.07663v4] LOGAN: Membership Inference Attacks Against Generative Models*. Http://Arxiv-Export-Lb.Library.Cornell.Edu/Abs/1705.07663v4. http://arxiv-export-lb.library.cornell.edu/abs/1705.07663v4

Ibitoye, O. (2019, November 6). *The Threat of Adversarial Attacks on Machine Learning in Network. . .* ArXiv.Org. https://arxiv.org/abs/1911.02621

Juuti, M. (2018, May 7). *PRADA: Protecting against DNN Model Stealing Attacks*. ArXiv.Org. https://arxiv.org/abs/1805.02628

Kos, J. (2017, February 22). *Adversarial examples for generative models*. ArXiv.Org. https://arxiv.org/abs/1702.06832

Li, G. (2018, October 16). *Security Matters: A Survey on Adversarial Machine Learning*. ArXiv.Org. https://arxiv.org/abs/1810.07339

Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., & Leung, V. C. M. (2018). A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View. *IEEE Access*, *6*, 12103–12117. https://doi.org/10.1109/access.2018.2805680

Long, Y. (2018, February 13). *[1802.04889] Understanding Membership Inferences on Well-Generalized Learning Models*. Https://Export.Arxiv.Org/Abs/1802.04889. https://export.arxiv.org/abs/1802.04889

McMahan, B. H. (2016, February 17). *Communication-Efficient Learning of Deep Networks from Decentralized Data*. ArXiv.Org. https://arxiv.org/abs/1602.05629v2

Moosavi-Dezfooli, S. (2015, November 14). *DeepFool: a simple and accurate method to fool deep neural networks*. ArXiv.Org. https://arxiv.org/abs/1511.04599v3

Obulesu, O., Mahendra, M., & ThrilokReddy, M. (2018). Machine Learning Techniques and Tools: A Survey. *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*. Published. https://doi.org/10.1109/icirca.2018.8597302

Oh, S. J. (2017, November 6). *Towards Reverse-Engineering Black-Box Neural Networks*. ArXiv.Org. https://arxiv.org/abs/1711.01768

Orekondy, T. (2018, December 6). *Knockoff Nets: Stealing Functionality of Black-Box Models*. ArXiv.Org. https://arxiv.org/abs/1812.02766

Orekondy, T. (2019, June 26). *Prediction Poisoning: Towards Defenses Against DNN Model Stealing Attacks*. ArXiv.Org. https://arxiv.org/abs/1906.10908

Papernot, N. (2016, February 8). *Practical Black-Box Attacks against Machine Learning*. ArXiv.Org. https://arxiv.org/abs/1602.02697

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The Limitations of Deep Learning in Adversarial Settings. *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. Published. https://doi.org/10.1109/eurosp.2016.36

Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). SoK: Security and Privacy in Machine Learning. *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*. Published. https://doi.org/10.1109/eurosp.2018.00035

Phan, N. (2017, September 18). *[1709.05750v1] Adaptive Laplace Mechanism: Differential Privacy Preservation in Deep Learning*. Http://Export.Arxiv.Org/Abs/1709.05750v1. http://export.arxiv.org/abs/1709.05750v1

Pyrgelis, A. (2017, August 21). *Knock Knock, Who's There? Membership Inference on Aggregate. . .* ArXiv.Org. https://arxiv.org/abs/1708.06145v2

Rigaki, M. (2020, July 15). *A Survey of Privacy Attacks in Machine Learning*. ArXiv.Org. https://arxiv.org/abs/2007.07646

Sagar, R. (2020). *Applications in Security and Evasions in Machine Learning: A Survey*. MDPI. https://www.mdpi.com/2079-9292/9/1/97

Sahoo, D. (2017, January 25). *Malicious URL Detection using Machine Learning: A Survey*. ArXiv.Org. https://arxiv.org/abs/1701.07179

Saini, P. S., Behal, S., & Bhatia, S. (2020). Detection of DDoS Attacks using Machine Learning Algorithms. *2020 7th International Conference on Computing for Sustainable Global Development (INDIACom)*. Published. https://doi.org/10.23919/indiacom49435.2020.9083716

Salem, A. (2018, June 4). *ML-Leaks: Model and Data Independent Membership Inference Attacks. . .* ArXiv.Org. https://arxiv.org/abs/1806.01246v2

Shokri, R. (2016, October 18). *Membership Inference Attacks against Machine Learning Models*. ArXiv.Org. https://arxiv.org/abs/1610.05820v2

Shokri, R., & Shmatikov, V. (2015). Privacy-Preserving Deep Learning. *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. Published. https://doi.org/10.1145/2810103.2813687

Song, C., Ristenpart, T., & Shmatikov, V. (2017). *Machine Learning Models that Remember Too Much.* Dblp Computer Science Bibliography. https://dblp.org/rec/conf/ccs/SongRS17.html

Suomalainen, J., Juhola, A., Shahabuddin, S., Mammela, A., & Ahmad, I. (2020). Machine Learning Threatens 5G Security. *IEEE Access*, *8*, 190822–190842. https://doi.org/10.1109/access.2020.3031966

Tahan, G. (2012). *Mal-ID: Automatic Malware Detection Using Common Segment Analysis and Meta-Features*. Https://Www.Jmlr.Org/Beta/Papers/V13/Tahan12a.Html. https://www.jmlr.org/beta/papers/v13/tahan12a.html

Tramèr, F. (2016). *Stealing Machine Learning Models via Prediction APIs | USENIX*. Https://Www.Usenix.Org/Conference/Usenixsecurity16/Technical-Sessions/Presentation/Tramer. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer

Wang, B. (2018, February 14). *Stealing Hyperparameters in Machine Learning*. ArXiv.Org. https://arxiv.org/abs/1802.05351

Xue, M., Yuan, C., Wu, H., Zhang, Y., & Liu, W. (2020). Machine Learning Security: Threats, Countermeasures, and Evaluations. *IEEE Access*, *8*, 74720–74742. https://doi.org/10.1109/access.2020.2987435

Yang, Z., Zhang, J., Chang, E. C., & Liang, Z. (2019). Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. Published. https://doi.org/10.1145/3319535.3354261

Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2018). Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. Published. https://doi.org/10.1109/csf.2018.00027

Zhang, J., & Li, C. (2019). Adversarial Examples: Opportunities and Challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 1–16. https://doi.org/10.1109/tnnls.2019.2933524

Zhang, Q., Yang, L. T., & Chen, Z. (2016). Privacy Preserving Deep Computation Model on Cloud for Big Data Feature Learning. *IEEE Transactions on Computers*, *65*(5), 1351–1362. https://doi.org/10.1109/tc.2015.2470255

Zhang, Z., Yang, X., & Huang, K. (2021). Attacking Sequential Learning Models with Style Transfer Based Adversarial Examples. *Journal of Physics: Conference Series*, *1880*(1), 012021. https://doi.org/10.1088/1742-6596/1880/1/012021