Security Risk and Attacks in Artificial Intelligence (AI): A Survey of Security and Privacy

Md Mostafizur Rahman[1], Aiasha Siddika Arshi[2], Md Mehedi Hasan[3], Sumayia Farzana Mishu[4], Hossain Shahriar[5], Fan Wu[6]
[12345]Kennesaw State University, Georgia, USA [6]Tuskegee University, Alabama, USA
{md.mostafizur.rn, aiasha.arshi, hasan.ict072}@gmail.com, smishu@students.kennesaw.edu,
hshahria@kennesaw.edu, fwu@tuskegee.edu

***Abstract*—This survey paper provides an overview of the current state of Artificial Intelligence (AI) attacks and risks for AI security and privacy as artificial intelligence becomes more prevalent in various applications and services. The risks associated with AI attacks and security breaches are becoming increasingly apparent and cause many financial and social losses. This paper will categorize the different types of attacks on AI models, including adversarial attacks, model inversion attacks, poisoning attacks, data poisoning attacks, data extraction attacks, and membership inference attacks. The paper also emphasizes the importance of developing secure and robust AI models to ensure the privacy and security of sensitive data. Through a systematic literature review, this survey paper comprehensively analyzes the current state of AI attacks and risks for AI security and privacy and detection techniques.**

***Index Terms*—Artificial intelligence, Machine learning, Security, Privacy, Adversarial attacks, Secure machine learning.**

I. Introduction

Artificial Intelligence (AI) is rapidly advancing and being integrated into various applications and services, resulting in an increase in AI attacks and risks for AI security and privacy. As a result, it is crucial to investigate the current state of AI attacks and the risks they pose to AI security and privacy. Drawing from recent internet data and literature, several prevailing patterns and perils pertaining to AI security and privacy involving Deepfakes [1] that creates synthetic audiovisual representations designed to appear authentic but in reality are skillfully doctored to mislead and delude individuals, AI-driven malware attacks [2], data privacy, lack of transparency and insider threats. To mitigate these risks, establishments must adopt resolute security protocols, such as stringent access controls, advanced encryption mechanisms, and frequent security evaluations. Furthermore, they must ensure that their AI systems are transparent and subject to oversight, conforming to privacy laws and regulations and adopting frameworks like trustworthy AI [3].

This survey paper provides an in-depth review of AI attacks and risks for AI security and privacy. It categorizes the different types of attacks on AI models and the data used to train them, including adversarial attacks, model inversion attacks, poisoning attacks, data poisoning attacks, data extraction attacks, and membership inference attacks. The paper emphasizes the importance of developing secure and robust AI models to ensure the privacy and security of sensitive data. This paper aims to inform and educate researchers, practitioners, and policymakers on the potential risks and challenges in securing AI systems by examining the various types of AI attacks. [4]

II. Methodology

To conduct this survey paper on AI security and privacy, the following methodology was used:

*1) **Research question**:* The research question for this survey paper is: "What are the current trends and threats related to AI security and privacy based on real internet data and articles?"

*2) **Data collection**:* To complete the suvey paper our team done a comprehensive literature review to gather relevant research articles, papers, and reports related to AI security and privacy. We tried to cover various areas, including AI attacks and risks, and methodologies for addressing AI security and privacy concerns, also trying to identify the attack detection techniques. The search was conducted on various online databases, including IEEE Xplore, ACM Digital Library, ScienceDirect, and Google Scholar, using relevant keywords such as "AI security," "AI privacy," "AI attacks," "AI risks," "AI defence," "AI mitigation," and "AI protection." The search was limited to articles published between 2010 and 2022.

*3) **Data analysis**:* The collected data were analyzed using qualitative and quantitative methods. The qualitative data analysis identified common subjects and concepts related to AI security and privacy. The quantitative data analysis involved analyzing the frequency and distribution of the identified core concepts and patterns.

*4) **Categorization of findings**:* The findings from the data analysis were categorized based on their relevance to the research question. The categorization was done based on different aspects, including types of AI attacks and defenses, industry-specific risks, approaches to AI security and privacy, and comprehensive previous survey papers, which include various research directions.

*5) **Interpretation of results**:* The data analysis was interpreted to draw conclusions about the research question. The results were interpreted based on their significance, relevance, and implications for AI security and privacy.

*6) **Presentation of results**:* The results of the survey paper were presented in a clear and concise manner. The presentation

included tables, graphs, and other visual aids to make the findings easier to understand.

*7) **Conclusion**:* Based on the research question, data collection, data analysis, and validation of findings, the survey paper drew conclusions about the current trends and threats related to AI security and privacy based on real internet data and articles. The conclusions were based on the evidence presented in the paper.

*8) **Future research directions**:* Finally, the survey paper identified future research directions to address the gaps in the current knowledge related to AI security and privacy based on real internet data and articles. The future research directions were based on the limitations of the current study and the potential for future advancements in the field.

III. Results

Adversarial attacks, model inversion attacks, poisoning attacks, data poisoning attacks, data extraction attacks, and membership inference attacks are the major types of attacks on AI and machine learning models. These attacks have been extensively studied in recent years [5] [6] [7] [8].

*1) **Adversarial attack**:* Adversarial attacks are among the most common attacks on AI and machine learning models. Adversarial attacks are designed to add small turmoils to the input data to cause the model to misclassify the input. Adversarial attacks can be targeted or untargeted. Targeted attacks force the model to output a specific incorrect result, while untargeted attacks cause the model to output an incorrect result. Adversarial attacks can result in significant losses in many applications, such as self-driving cars and medical diagnoses. [9] [10] [11]

*2) **Model inversion attack**:* Model inversion attacks are another attack on AI and machine learning models. These attacks aim to extract information about the training data from the model. In model inversion attacks, an adversary can use the model's output to reconstruct the input data to mislead the original machine learning model [12]. These attacks can leak sensitive information, which can be used for various unethical work.

*3) **Poisoning attack**:* Poisoning attacks are another common type of attack on AI and machine learning models. In poisoning attacks, an attacker introduces malicious data into

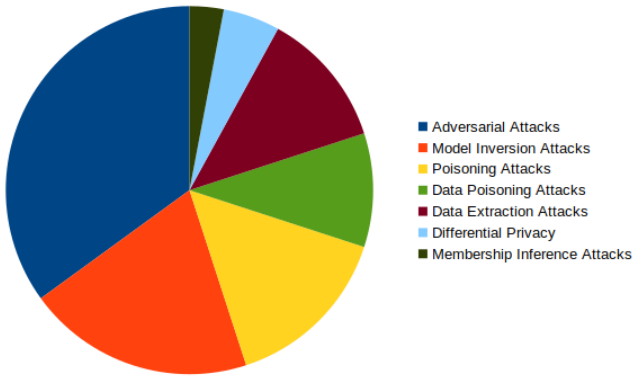

Fig. 1. Different Attacks Occurrence in Percent

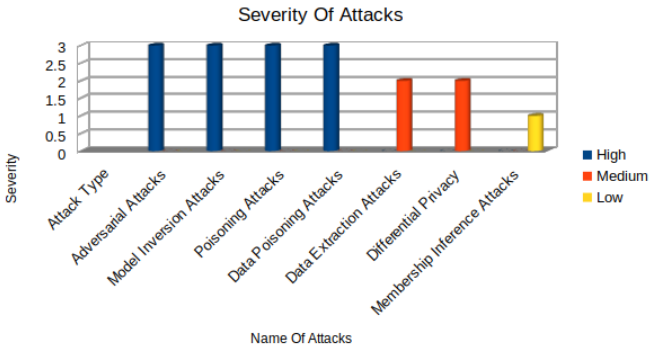

Fig. 2. Severity Of Different Attacks

the training data set to manipulate the model's behavior at test time [13]. These attacks can be challenging to detect and mitigate because the poisoned data may not be apparent during the training phase.

*4) **Data poisoning attack**:* Data poisoning attacks are similar to poisoning attacks but focus on manipulating the training dataset to affect the model's behavior. An attacker can modify the training dataset to introduce partial or incorrect data in data poisoning attacks [14]. These attacks can result in significant performance or correctness problems in the model. It significantly degrades the overall performance of prediction as well as robustness.

*5) **Data extraction attack**:* In such an attack, the adverser, who has no prior knowledge about the model, try to extract sensitive data used to train the model. Data extraction attacks are designed to extract information about the training data from the model. An attacker can use the model's output to infer information about the training data. These attacks can result in the leakage of sensitive information. [15] [16]

*6) **Membership inference attack**:* Membership inference attacks are designed to determine if a particular data point was used in the training dataset [17]. In membership inference attacks, an attacker can use the model's output to determine or infer if a particular data point was used in the training dataset. In this attack, the attacker tries to leak training data from a prediction from the model response. These attacks can result in the leakage of sensitive information. Researchers have proposed various methods to prevent these attacks, such as adversarial training, data sanitization, and model pruning [18] [19] [20]. Adversarial training involves training the model on adversarial examples to improve its robustness against adversarial attacks. Data sanitization involves filtering the training dataset to remove malicious or irrelevant data. Model pruning involves removing unnecessary features or connections from the model to reduce complexity and increase robustness. [21] [22] [23].

IV. Securing AI Models

Securing AI models has become a critical aspect of AI development due to the potential for AI systems to be exploited by attackers. The security of AI models involves protecting the models from various attacks, including adversarial attacks,

model inversion attacks, poisoning attacks, data extraction attacks, and membership inference attacks.

There are several approaches to securing AI models. One direction is to use defensive mechanisms, such as adversarial training, to make the model more robust against adversarial attacks. Adversarial training involves training the model on a combination of clean and adversarial examples to improve its accuracy and robustness against adversarial attacks [9].

Another approach is to use anomaly detection techniques to identify when the model is under attack. Anomaly detection techniques involve monitoring the input and output of the model and looking for unexpected behavior or patterns that could indicate an attack. For instance, if the model's output suddenly changes significantly from what it should be, this could mean an attack [9].

A. *Discussion On Different Approaches*

Securing AI models has become an important concern due to the increasing prevalence of AI-based applications in various industries. Several approaches have been proposed to ensure the security of AI models, including:

*1) **Adversarial training**:* This approach involves adding adversarial examples to the training data to improve the robustness of the model against adversarial attacks. Strengths of this approach include its simplicity and effectiveness in improving the model's resistance to adversarial attacks. However, it can be computationally expensive and may not provide complete protection against all types of attacks [9] [48].

*2) **Input sanitization**:* This approach involves preprocessing the input data to remove potential malicious code or inputs. Strengths of this approach include its effectiveness in preventing input-based attacks and its low computational cost. However, it may not be effective against more sophisticated attacks that can bypass input sanitization techniques [48] [49].

*3) **Model explainability**:* This approach involves improving the transparency and interpretability of the model to detect and prevent attacks. Strengths of this approach include its ability to identify potential vulnerabilities and the potential for increased trust in the model. However, it can be difficult to implement and may not be effective against attacks that exploit weaknesses in the model architecture [50] [51].

*4) **Model diversification**:* This approach involves training multiple models with different architectures or parameters to increase the overall robustness of the system. Strengths of this approach include its effectiveness against a wide range of attacks and the potential for increased accuracy. However, it can be computationally expensive and may not be practical for all applications [48] [49].

*5) x**Model diversification**:* This approach involves using hardware-based security measures to protect the model and data during run-time. Strengths of this approach include its ability to prevent attacks at the hardware level and the potential for increased security. However, it can be expensive to implement and may not be effective against all types of attacks [48] [30].

*6) **Federated learning**:* This approach involves training the model using data from multiple sources without sharing the data itself, improving privacy and reducing the risk of attacks. Strengths of this approach include its potential for increased accuracy and privacy, as well as reduced vulnerability to attacks. However, it can be computationally expensive and may require significant coordination and communication between the different sources of data.

Securing AI models also involves ensuring the privacy of the data used to train the models. This can be achieved through techniques such as differential privacy, which adds random noise to the data to prevent individual data points from being identified [9].

There are also challenges to securing AI models, such as the lack of interpretability of deep learning models, which makes it difficult to understand how the models make decisions and identify when they are under attack. Additionally, the complexity of AI models and the diversity of attack methods make it challenging to develop effective defenses.

In order to address these challenges and effectively secure AI models, best practices have been proposed, such as regularly updating and testing the defenses of AI models, using explainable AI methods to improve interpretability, and ensuring that security is integrated throughout the entire AI development process [9].

Overall, securing AI models is a complex and evolving area of research, but it is crucial for ensuring the reliability and trustworthiness of AI systems in various domains, such as healthcare, finance, and autonomous vehicles.

V. AI or ML Attack Detection Techniques

Detection techniques for AI or ML model attacks can be referred to as the methods and tools used to identify if a model has been attacked or compromised. Malicious activity detection techniques play a crucial role in enhancing the security of AI and ML models by helping to identify potential attacks and minimize their impact.

The benefits of using detection techniques include early detection of potential attacks, minimizing the damage caused by attacks, and improving the overall security and trustworthiness of AI and ML models. By using these techniques, organizations can proactively monitor their models and identify any suspicious activities, and thereby organizations can improve their ability to respond quickly and effectively to security incidents. Additionally, these techniques can help to build trust among users and stakeholders by demonstrating a commitment to securing sensitive data and ensuring the accuracy and reliability of AI and ML models. Now we will discuss the different types of detection techniques mentioned in the table: I

*1) **Defensive Distillation**:* Defensive distillation is a detection technique for detecting adversarial attacks on deep neural networks. It was introduced by Papernot et al. in 2016 as a means of defending against adversarial examples, which are inputs to a machine learning model that are specifically designed to cause it to make incorrect predictions. Defensive distillation works by training a second neural network, known

TABLE I
ATTACKES AND DETECTION TECHNIQUES

Attack Type	Detection Technique	Security Measures	References
Adversarial Attacks	Defensive Distillation	Feature Squeezing, Adversarial Training, Ensembling	[9], [24], [25], [5], [26] [27] [16] [28] [29]
Model Inversion Attacks	Regularization	Secure Multiparty Computation, Federated Learning, Differentially Private Learning	[9], [25], [30], [31] [27] [32] [33]
Poisoning Attacks	Detection of Outliers in the Data	Data Filtering, Input Validation, Regularization Techniques	[9], [34], [35] [36] [37]
Data Poisoning Attacks	Robust Statistical Methods	Dataset Verification, Input Validation, Randomization	[9], [38], [25] [39] [40] [41]
Data Extraction Attacks	Differential Privacy	Secure Multiparty Computation, Federated Learning, Data Anonymization	[9], [25], [30], [31] [42] [43] [44]
Membership Inference Attacks	Randomized Response Mechanisms	Differential Privacy, Membership Revocation	[9], [38], [30], [31] [45] [46] [47]

as the distilled model, to approximate the outputs of the original model, which is known as the teacher model.

The teacher model is first trained on a large data set, and then the distilled model is trained on the outputs of the teacher model. This process essentially distills the knowledge of the teacher model into the distilled model, resulting in a more robust model that is less susceptible to adversarial attacks. The distilled model is then used for inference rather than the teacher model.

The main advantage of defensive distillation is that it can be applied to any machine learning model, including those that were not specifically designed with security in mind. It is also effective against a wide range of attack types, including both white-box and black-box attacks. Additionally, defensive distillation is relatively easy to implement, requiring only minor modifications to the training process.

One potential weakness of defensive distillation is that it is not effective against all types of attacks. In particular, it is vulnerable to attacks that specifically target the training process, such as data poisoning attacks. It is also relatively resource-intensive, requiring the training of two separate models.

To implement defensive distillation, several tools, and techniques are required, including a deep learning framework such as TensorFlow or PyTorch, as well as access to a large dataset for training the teacher model. There are also several open-source implementations of defensive distillation available, such as CleverHans and Adversarial Robustness Toolbox.

Several research papers have evaluated the effectiveness of defensive distillation against various types of adversarial attacks. For example, Papernot et al. demonstrated that defensive distillation can improve the robustness of deep neural networks against both white-box and black-box attacks. In another study, Samangouei et al. showed that defensive distillation can also improve the accuracy of machine learning models when applied to natural image classification tasks.

Overall, defensive distillation is a promising technique for improving the security and robustness of machine learning models. While it is not a silver bullet solution to the problem of adversarial attacks, it can be an effective tool in the broader arsenal of techniques used to secure AI and ML systems. [52]

*2) **Regularization**:* Regularization is a widely used detection technique for securing machine learning models from adversarial attacks. The main goal of regularization is to limit the model's complexity and avoid overfitting, which is a common vulnerability that can be exploited by attackers to manipulate the model's output.

There are different types of regularization techniques, such as L1, L2, and dropout, each of which applies a different form of penalty to the model's parameters during training to encourage simpler models. L1 regularization, for example, adds a penalty to the absolute value of the model's parameters, while L2 regularization adds a penalty to the squared value of the parameters. Dropout regularization randomly drops out a fraction of the model's neurons during training to prevent the model from relying too much on specific features.

Regularization has been shown to be effective in improving the robustness of machine learning models against different types of attacks, including adversarial attacks and poisoning attacks. However, it may not be sufficient on its own and should be combined with other detection techniques and security measures.

The implementation of regularization requires the use of specific tools and techniques, such as TensorFlow, PyTorch, and scikit-learn libraries, to modify the machine learning model's code and add the regularization penalties.

Overall, regularization is a crucial detection technique for securing AI and ML models and can significantly improve their robustness against different types of attacks. However, it should be used in combination with other techniques and best practices for optimal security and privacy protection. [5] [53]

*3) **Detection of Outliers in Data**:* Detection of outliers in data is another important technique used to identify attacks on AI and ML models. Outliers are observations that are significantly different from other observations in the dataset, and their presence can indicate anomalies or potential attacks. Outliers can be detected using various statistical and machine

learning methods such as clustering, classification, and regression analysis.

One popular technique for detecting outliers is the Local Outlier Factor (LOF) method. LOF identifies anomalies based on the density of neighboring points in a given data set. The technique works by computing the density of points around a particular data point and comparing it to the density of points around its neighboring points. If the density of the point is significantly lower than that of its neighbors, it is considered an outlier.

Another method for detecting outliers is Principal Component Analysis (PCA), which is a statistical technique used for dimensionality reduction. PCA identifies the variables that contribute the most to the variance in the dataset and projects the data onto a lower-dimensional space. Outliers can be identified by looking for data points that are far away from the center of the projected data.

The benefit of outlier detection techniques is that they can be used to identify attacks that may not be detected by traditional security measures. By detecting outliers in the data, analysts can identify potential attacks on the AI or ML model and take appropriate measures to secure it. However, one of the challenges with outlier detection is that it can also generate false positives, which can lead to unnecessary alerts and increase the workload of analysts.

To overcome this challenge, it is important to use a combination of techniques such as defensive distillation, regularization, and outlier detection to enhance the security of AI and ML models. Additionally, continuous monitoring and updating of these techniques are necessary to keep up with the evolving nature of attacks on AI and ML models. [5] [54] [55]

*4) **Robust Statistical Methods**:* Robust statistical methods are an important approach to detecting anomalies or outliers in data that can lead to attacks on AI and ML models. These methods involve the use of statistical models that are resistant to outliers and are able to detect any deviations from expected patterns in the data accurately.

One of the main advantages of robust statistical methods is their ability to handle data that is contaminated or contains noise, which is common in many real-world applications. Some of the techniques used in this approach include the use of robust regression, robust covariance estimation, and trimmed means.

Robust statistical methods can be applied at different stages of the machine learning pipeline, including during data preprocessing, model training, and model evaluation. However, one of the challenges of using these methods is their increased computational complexity compared to traditional statistical methods.

*5) **Differential Privacy**:* Differential privacy is a technique used to protect sensitive information while processing data. It involves adding random noise to the data to obscure any individual's personal information while still allowing useful insights to be drawn from the data. In the context of AI and ML security, differential privacy can be used to prevent attackers from inferring sensitive information from a model's training data or outputs.

To implement differential privacy, several tools, and techniques can be used, including adding noise to the data or modifying the training process to ensure that the model does not learn sensitive information. This technique has been applied to various applications, such as image recognition, natural language processing, and recommendation systems.

One of the strengths of differential privacy is that it provides strong guarantees of privacy protection, even against powerful adversaries. However, it may also introduce additional noise into the data, which can impact the accuracy of the model. Therefore, finding a balance between privacy protection and model accuracy is crucial.

*6) **Randomized Response Mechanisms**:* Randomized Response Mechanisms is a detection technique for AI or ML attacks that aims to preserve the privacy of sensitive data while providing statistical information. This technique involves introducing randomness into the data to hide the true value while maintaining a distribution that is statistically similar to the original data.

Randomized Response Mechanisms use a probabilistic algorithm to introduce noise into the data. The amount of noise introduced is controlled by a parameter called the privacy budget, which determines the level of privacy protection provided.

This technique can be implemented using various tools such as the Differential Privacy Library, PySyft, and IBM's Privacy-Preserving Deep Learning Library. It can be used to detect and prevent attacks such as membership inference, model inversion, and data poisoning.

One of the main advantages of Randomized Response Mechanisms is that it provides a rigorous mathematical framework for measuring privacy guarantees. Additionally, it can be used in a variety of settings, including healthcare, finance, and social media. [56]

VI. Conclusion

AI attacks and risks for AI security and privacy are growing concerns as AI models become more prevalent. This survey paper has highlighted the different types of AI attacks and risks for AI security and privacy. It is important to develop robust and secure AI models that are resilient to attacks to ensure the privacy and security of sensitive data.
The future direction of research in AI and ML security and privacy should focus on addressing the increasing sophistication of attacks and the need for more robust defences. This includes developing new detection techniques and security measures that can keep pace with evolving threats and improving the explainability and transparency of AI models to enhance trust and accountability. Additionally, research should also explore the ethical and societal implications of AI and ML technologies, such as fairness, bias, and privacy concerns. Finally, interdisciplinary collaborations between computer science, law, ethics, and other fields will be crucial in developing comprehensive and effective solutions to AI and ML security and privacy challenges.

VII. Acknowledgements

This work is partially supported by the National Science Foundation under NSF Award #2100115, #2209638, #2209637, #2100134, and #1663350. Any opinions, findings, and recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation

References

[1] C. Campbell, K. Plangger, S. Sands, and J. Kietzmann, "Preparing for an era of deepfakes and ai-generated ads: A framework for understanding responses to manipulated advertising," *Journal of Advertising*, vol. 51, no. 1, pp. 22–38, 2022.

[2] B. Guembe, A. Azeta, S. Misra, V. C. Osamor, L. Fernandez-Sanz, and V. Pospelova, "The emerging threat of ai-driven cyber attacks: A review," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2037254, 2022.

[3] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, "Trustworthy ai: From principles to practices," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–46, 2023.

[4] H. Shahriar, M. A. I. Talukder, M. Rahman, H. Chi, S. Ahamed, and F. Wu, "Hands-on file inclusion vulnerablity and proactive control for secure software development," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2, 2019, pp. 604–609.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015.

[6] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Sok: Security and privacy in machine learning," 04 2018, pp. 399–414.

[7] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.

[8] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction apis," 2016.

[9] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Machine Learning and Knowledge Discovery in Databases*, H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 387–402.

[10] A. Raghunathan, J. Steinhardt, and P. Liang, "Certified defenses against adversarial examples," 2020.

[11] B. Kolosnjaji, A. Demontis, B. Biggio, D. Maiorca, G. Giacinto, C. Eckert, and F. Roli, "Adversarial malware binaries: Evading deep learning for malware detection in executables," 2018.

[12] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *Proceedings 2018 Network and Distributed System Security Symposium*. Internet Society, 2018.

[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[14] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," 2018.

[15] O. Suciu, S. E. Coull, and J. Johns, "Exploring adversarial examples in malware detection," 2019.

[16] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," 2019.

[17] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," 05 2019, pp. 707–723.

[18] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," 2019.

[19] J. Gilmer, L. Metz, F. Faghri, S. S. Schoenholz, M. Raghu, M. Wattenberg, and I. Goodfellow, "Adversarial spheres," 2018.

[20] W. Hu and Y. Tan, "Generating adversarial malware examples for black-box attacks based on gan," 2017.

[21] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," 2017.

[22] J. Xu, C. Zhang, X. Zheng, L. Li, C.-J. Hsieh, K.-W. Chang, and X. Huang, "Towards adversarially robust text classifiers by learning to reweight clean examples," 01 2022, pp. 1694–1707.

[23] D. Wang, W. Yao, T. Jiang, G. Tang, and X. Chen, "A survey on physical adversarial attack in computer vision," 2023.

[24] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," 2018.

[25] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroSP)*, 2016, pp. 372–387.

[26] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, "Adversarial perturbations against deep neural networks for malware classification," 2016.

[27] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, dec 2018.

[28] A. Fawzi, O. Fawzi, and P. Frossard, "Analysis of classifiers' robustness to adversarial perturbations," 2016.

[29] Y. Dong, F. Liao, T. Pang, X. Hu, and J. Zhu, "Discovering adversarial examples with momentum. arxiv preprint 1710.06081 (2017)," *arXiv preprint arXiv:1710.06081*, 2017.

[30] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel, "On the (statistical) detection of adversarial examples," 2017.

[31] N. Papernot, *Adversarial Machine Learning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2019, pp. 1–4.

[32] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1322–1333.

[33] P. Buddareddygari, T. Zhang, Y. Yang, and Y. Ren, "Targeted attack on deep rl-based autonomous driving with learned visual patterns," 2022.

[34] L. Muñoz-González, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu, and F. Roli, "Towards poisoning of deep learning algorithms with back-gradient optimization," 2017.

[35] J. Kos, I. Fischer, and D. Song, "Adversarial examples for generative models," 02 2017.

[36] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, "Parseval networks: Improving robustness to adversarial examples," 2017.

[37] B. Miller, A. Kantchelian, S. Afroz, R. Bachwani, E. Dauber, L. Huang, M. C. Tschantz, A. D. Joseph, and J. Tygar, "Adversarial active learning," ser. AISec '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 3–14.

[38] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 3–18.

[39] M. C. Tschantz, D. Kaynar, and A. Datta, "Formal verification of differential privacy for interactive systems (extended abstract)," *Electronic Notes in Theoretical Computer Science*, vol. 276, pp. 61–79, 2011, twenty-seventh Conference on the Mathematical Foundations of Programming Semantics (MFPS XXVII).

[40] C. Zhu, W. R. Huang, A. Shafahi, H. Li, G. Taylor, C. Studer, and T. Goldstein, "Transferable clean-label poisoning attacks on deep neural nets," 2019.

[41] H. A. Gulshan Kumar, "Machine learning techniques for intrusion detection systems in sdn-recent advances, challenges and future directions," *Computer Modeling in Engineering & Sciences*, vol. 134, no. 1, pp. 89–119, 2023.

[42] Z. Chen, Z. Wang, J. Huang, W. Zhao, X. Liu, and D. Guan, "Imperceptible adversarial attack via invertible neural networks," 2023.

[43] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2017.

[44] I. Seraphim, S. Palit, K. Srivastava, and P. Eswaran, "A survey on machine learning techniques in network intrusion detection system," 12 2018, pp. 1–5.

[45] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning models," 2018.

[46] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser. arxiv e-prints, page," *arXiv preprint arXiv:1712.02976*, 2017.

[47] S. Cheng, Y. Dong, T. Pang, H. Su, and J. Zhu, "Improving black-box adversarial attacks with a transfer-based prior," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[48] K. Shaukat Dar, S. Luo, V. Varadharajan, I. Hameed, and M. Xu, "A survey on machine learning techniques for cyber security in the last decade," *IEEE Access*, 11 2020.

[49] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," 2017.

[50] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016.

[51] Z. C. Lipton, "The mythos of model interpretability," 2017.

[52] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," 2016.

[53] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 06 2014.

[54] G. Li, K. Ota, M. Dong, J. Wu, and J. Li, "Desvig: Decentralized swift vigilance against adversarial attacks in industrial artificial intelligence systems," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 5, pp. 3267–3277, 2020.

[55] C. C. Aggarwal and C. K. Reddy, "Data clustering," *Algorithms and Applications*, 2016.

[56] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation*, M. Agrawal, D. Du, Z. Duan, and A. Li, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 1–19.