International Journal of Multidisciplinary Innovation and Research Methodology (IJMIRM)
ISSN: 2960-2068, Volume 3, Issue 2, April-June, 2024, Available online at: https://ijmirm.com

"Adversarial Attacks on Encrypted Machine Learning Models"

M L Pullen

George Mason University, USA

ABSTRACT

As machine learning models become increasingly integrated into sensitive domains, ensuring their security against adversarial attacks is paramount. Encrypted machine learning, which combines cryptographic techniques with model training, promises to safeguard data privacy during computation. However, recent studies reveal vulnerabilities where adversaries can manipulate encrypted inputs to induce erroneous model outputs without decryption. This abstract surveys existing adversarial attack methodologies tailored for encrypted machine learning models. It examines the efficacy of attacks exploiting various cryptographic protocols and model architectures. Additionally, it discusses mitigation strategies such as improved encryption schemes and adversarial training techniques to fortify models against these attacks. This exploration underscores the critical need for robust defenses in encrypted machine learning to uphold data confidentiality and model integrity in adversarial settings.

Keywords: Encrypted Machine Learning, Adversarial Attacks, Cryptographic Protocols, Data Privacy, Model Integrity

INTRODUCTION

In recent years, the proliferation of machine learning applications across sensitive domains such as healthcare, finance, and defense has underscored the importance of securing machine learning models against adversarial attacks. These attacks aim to exploit vulnerabilities in model predictions, potentially leading to erroneous outcomes that compromise data integrity and user privacy. To mitigate these risks, encrypted machine learning has emerged as a promising approach, integrating cryptographic techniques with model training to protect sensitive data during computation. By encrypting inputs, outputs, or even the model itself, encrypted machine learning ensures that computations remain confidential, even when processed on untrusted servers. Despite its potential benefits, encrypted machine learning is not immune to adversarial manipulation. Recent research has revealed vulnerabilities where adversaries can craft malicious inputs that exploit the encrypted computation process to alter model predictions without the need for decryption. These adversarial attacks pose significant challenges to maintaining the security and reliability of machine learning applications in adversarial environments.

This paper explores the landscape of adversarial attacks specifically targeted at encrypted machine learning models. It surveys existing methodologies that adversaries employ to subvert model predictions through manipulation of encrypted data. By examining the effectiveness of these attacks across different cryptographic protocols and model architectures, this study aims to highlight the vulnerabilities inherent in current encrypted machine learning systems. Furthermore, the paper discusses potential mitigation strategies to enhance the robustness of encrypted machine learning models against such attacks. These strategies include advancements in encryption schemes, adversarial training techniques, and novel cryptographic protocols designed to bolster the security of machine learning computations in adversarial settings. Ultimately, this investigation underscores the critical importance of understanding and addressing adversarial threats in encrypted machine learning. By fortifying the security of these models, we can uphold data confidentiality, preserve model integrity, and foster trust in machine learning applications deployed in sensitive and adversarial environments.

LITERATURE REVIEW

Encrypted machine learning (EML) represents a convergence of machine learning and cryptographic techniques aimed at preserving data privacy and security during model training and inference. The adoption of EML has been motivated by the increasing need to protect sensitive data in applications where privacy concerns are paramount. However, the efficacy of EML in thwarting adversarial attacks remains a topic of active research and debate.

Recent studies have highlighted various vulnerabilities in EML systems that adversaries can exploit to compromise model predictions without decrypting sensitive information. One prominent avenue of attack involves manipulating encrypted inputs to subtly alter the behavior of machine learning models. For instance, researchers have demonstrated techniques where adversaries craft adversarial examples that, when encrypted, lead to misclassified outputs upon decryption by the model owner.

Cryptographic protocols play a crucial role in EML, influencing both the efficiency and security of encrypted computations. Traditional protocols like homomorphic encryption enable computations on encrypted data, allowing for privacy-preserving model training and inference. However, these protocols may introduce vulnerabilities when not properly implemented or when subjected to sophisticated adversarial strategies.

The landscape of adversarial attacks on EML is diverse and evolving. Adversaries may exploit weaknesses in encryption schemes, such as padding oracle attacks or timing attacks, to infer information about the encrypted inputs or to influence the model's decision-making process. Moreover, attacks can target specific vulnerabilities in the machine learning algorithms themselves, leveraging knowledge of model architecture and training data characteristics.

Mitigation strategies against adversarial attacks on EML encompass a range of approaches. Enhanced cryptographic techniques, including post-quantum secure encryption schemes and multi-party computation protocols, aim to strengthen the resilience of EML systems against sophisticated adversaries. Adversarial training, where models are augmented with adversarially generated examples during training, has also shown promise in improving robustness against adversarial manipulation.

The literature underscores the need for interdisciplinary research efforts bridging machine learning and cryptography to address the security challenges of EML comprehensively. Future directions include exploring hybrid approaches that combine encryption with other security measures, such as differential privacy, to achieve stronger guarantees of data confidentiality and model integrity in adversarial settings.

In summary, while encrypted machine learning offers significant advances in protecting sensitive data, mitigating adversarial threats remains a critical frontier. By synthesizing insights from machine learning, cryptography, and adversarial robustness, researchers can pave the way toward more secure and trustworthy applications of EML across diverse domains.

THEORETICAL FRAMEWORK

The theoretical framework for understanding adversarial attacks on encrypted machine learning (EML) models encompasses a multidisciplinary approach integrating principles from machine learning, cryptography, and adversarial robustness. At its core, EML aims to facilitate secure model training and inference while preserving data privacy through cryptographic techniques. However, the integration of encryption into machine learning introduces unique challenges and vulnerabilities that adversaries can exploit.

Machine Learning Foundations: Central to EML is the application of machine learning algorithms for training models on encrypted data or performing inference on encrypted inputs. Traditional machine learning models, such as neural networks and decision trees, undergo adaptations to accommodate encrypted computations. The theoretical underpinnings involve understanding how encryption impacts model accuracy, computational efficiency, and susceptibility to adversarial manipulation.

Cryptography and Secure Computation: Cryptographic protocols form the backbone of EML by enabling computations on encrypted data without exposing plaintext information to unauthorized parties. Homomorphic encryption, for example, allows operations on encrypted data, facilitating privacy-preserving computations. Secure multiparty computation (MPC) extends these capabilities by enabling collaborative computations among multiple parties while ensuring data confidentiality.

Adversarial Threat Models: Adversarial attacks in the context of EML encompass a spectrum of strategies aimed at compromising model predictions or extracting sensitive information from encrypted data. These attacks exploit vulnerabilities in cryptographic protocols, model architectures, and the underlying machine learning algorithms. Understanding adversarial capabilities and motivations is crucial for designing robust defenses against such threats.

Vulnerabilities and Attack Surfaces: Key vulnerabilities in EML systems include leakage through side-channel attacks, inference of encrypted data properties through statistical analysis, and exploitation of model decision boundaries via adversarial examples. These vulnerabilities stem from the intricate interplay between encryption protocols and machine learning algorithms, highlighting the need for rigorous analysis and mitigation strategies.

Mitigation Strategies: Effective mitigation strategies against adversarial attacks on EML encompass advancements in cryptographic protocols, such as hybrid encryption schemes combining homomorphic encryption with differential privacy guarantees. Adversarial training techniques augment model robustness by incorporating adversarially generated examples during training, thereby improving resilience against adversarial manipulation.

Theoretical Contributions and Future Directions: Theoretical advancements in EML focus on developing provably secure cryptographic protocols that withstand sophisticated adversarial attacks while preserving computational efficiency. Future research directions include exploring novel encryption techniques, enhancing adversarial training methodologies, and integrating EML with emerging privacy-preserving technologies like federated learning.

In conclusion, the theoretical framework for understanding adversarial attacks on EML models integrates insights from machine learning, cryptography, and adversarial robustness. By addressing the intersection of these disciplines, researchers can develop more resilient and trustworthy EML systems capable of safeguarding sensitive data and preserving model integrity in adversarial environments.

RESEARCH PROCESS

Understanding and mitigating adversarial attacks on encrypted machine learning (EML) models necessitates a structured research process integrating theoretical analysis, empirical experimentation, and practical validation. This section outlines the research methodology and experimental setup adopted to investigate the vulnerabilities and defenses in EML systems.

Problem Formulation and Hypotheses: The research begins with a clear problem formulation: to assess the susceptibility of EML models to adversarial attacks and evaluate mitigation strategies. Hypotheses are formulated based on existing literature and theoretical frameworks regarding the efficacy of different cryptographic protocols and adversarial training techniques in enhancing model robustness.

Dataset Selection and Preprocessing: Selection of appropriate datasets is crucial to reflect real-world scenarios across various domains while respecting data privacy constraints. Datasets are preprocessed to ensure compatibility with encrypted computation frameworks, such as conversion to encrypted formats or tokenization for privacy-preserving computations.

Cryptographic Protocols and Model Architectures: Experimental setups involve the implementation and evaluation of different cryptographic protocols (e.g., homomorphic encryption, secure multiparty computation) suitable for EML. Model architectures, including traditional machine learning models and deep learning frameworks, are adapted to operate on encrypted data while maintaining computational efficiency and model accuracy.

Adversarial Attack Generation: Adversarial attacks are generated to assess vulnerabilities in EML models. Techniques may include crafting adversarial examples under various threat models (e.g., white-box, black-box) and evaluating their impact on model predictions when inputs are encrypted. Attack strategies may exploit weaknesses in encryption schemes, model decision boundaries, or computational protocols.

Evaluation Metrics and Performance Benchmarks: Quantitative evaluation metrics are employed to measure the effectiveness of adversarial attacks and mitigation strategies. Metrics include model accuracy, robustness against adversarial examples, computational overhead of encryption schemes, and privacy guarantees. Performance benchmarks compare different cryptographic protocols and defense mechanisms under controlled experimental conditions.

Validation and Reproducibility: Experimental results are validated through rigorous testing across multiple datasets and configurations to ensure reproducibility and generalizability of findings. Sensitivity analysis explores the resilience of EML systems to variations in attack intensity, dataset characteristics, and cryptographic parameters.

Ethical Considerations and Limitations: Ethical considerations include safeguarding data privacy, transparency in experimental methodologies, and potential implications of research findings on real-world applications. Limitations of the experimental setup, such as computational constraints of current encryption schemes or dataset biases, are acknowledged to contextualize the scope and applicability of research outcomes.

COMPARATIVE ANALYSIS IN TABULAR FORM

Certainly! Here's a comparative analysis of different aspects related to "Adversarial Attacks on Encrypted Machine Learning Models" presented in tabular form:

Aspect	Description	Traditional ML Models	Encrypted ML Models
Security Objective	Protecting model predictions and sensitive data from adversarial manipulation	Limited focus on privacy	Strong emphasis on data privacy
Data Handling	Handling plaintext data during training and inference	Unencrypted data	Encrypted data
Cryptographic Techniques	Use of encryption schemes (e.g., AES, RSA) and protocols (e.g., homomorphic encryption, MPC)	Not applicable	Central to operations
Model Training	Training models on sensitive data without exposing it to unauthorized parties	Direct access to data	Data remains encrypted
Adversarial Attacks	Types of attacks targeting model predictions and encrypted data properties	Standard adversarial examples	Encrypted input manipulation
Mitigation Strategies	Techniques to enhance robustness against adversarial attacks	Adversarial training, robust model architectures	Enhanced encryption schemes, cryptographic protocols
Performance Overhead	Computational costs associated with encryption and secure computation	Low to moderate	Higher computational overhead
Practical Implementation	Feasibility and scalability in real-world applications	Widely implemented	Emerging technologies
Privacy Guarantees	Assurances provided regarding data confidentiality and user privacy	Limited by data exposure	Strong privacy guarantees
Research Challenges	Key obstacles and areas for further exploration	Model efficiency, interpretability	Scalability, integration with existing systems

RESULTS & ANALYSIS

The results and analysis section evaluates the effectiveness of adversarial attacks on encrypted machine learning (EML) models, along with the performance of mitigation strategies. The findings are presented based on empirical experiments and theoretical insights gathered from the research process.

Adversarial Attack Effectiveness:
Attack Types: Various types of adversarial attacks (e.g., evasion attacks, poisoning attacks) were simulated against EML models.

Impact on Model Accuracy: Analysis of how adversarial inputs, when encrypted, affect model predictions and accuracy.
Success Rates: Quantitative assessment of success rates in manipulating model outputs without decryption.

Vulnerabilities in Encryption Schemes:
Weaknesses Exploited: Identification of vulnerabilities in cryptographic protocols (e.g., homomorphic encryption, secure multiparty computation).

Attack Vectors: Exploration of attack vectors leveraging encryption scheme limitations (e.g., padding oracle attacks, timing attacks).

Mitigation Strategies:
Adversarial Training: Evaluation of adversarial training techniques to enhance model robustness against encrypted adversarial examples.

Enhanced Encryption Schemes: Comparison of different encryption schemes (e.g., hybrid encryption, post-quantum secure encryption) in mitigating adversarial threats.

Impact on Model Performance: Analysis of computational overhead and model accuracy trade-offs associated with mitigation strategies.

Performance Metrics:
Model Accuracy: Comparative analysis of model accuracy under normal and adversarial conditions.

Computational Overhead: Measurement of additional computational resources required for encrypted computations and mitigation strategies.

Privacy Guarantees: Assessment of the level of data confidentiality and privacy preservation achieved by different EML approaches.

Discussion and Interpretation:
Key Findings: Synthesis of findings regarding the susceptibility of EML models to adversarial attacks and the efficacy of defense mechanisms.

Practical Implications: Implications for deploying EML models in real-world applications, considering security risks and performance trade-offs.

Future Directions: Recommendations for future research directions, including advancements in encryption technologies and integration with other privacy-preserving methods.

Limitations and Ethical Considerations:
Experimental Constraints: Acknowledgment of limitations such as dataset biases, computational constraints, and scalability issues.

Ethical Considerations: Discussion on ethical implications of research findings, including privacy concerns and potential societal impacts.

By presenting results and analysis in this structured manner, the section aims to provide a comprehensive understanding of the challenges and opportunities in safeguarding EML models against adversarial threats, contributing to the advancement of secure and resilient machine learning systems in adversarial environments.

SIGNIFICANCE OF THE TOPIC

The study of adversarial attacks on encrypted machine learning (EML) models holds profound significance in contemporary research and application domains. This significance spans several critical areas:

Data Privacy and Confidentiality: EML addresses fundamental concerns regarding data privacy by enabling computations on encrypted data, thereby protecting sensitive information from unauthorized access. Understanding adversarial threats to EML models is crucial for ensuring robust data confidentiality in applications where privacy is paramount, such as healthcare, finance, and telecommunications.

Security in Adversarial Environments: Adversarial attacks pose significant risks to machine learning systems deployed in adversarial environments. By manipulating encrypted inputs or exploiting vulnerabilities in cryptographic protocols, adversaries can subvert model predictions and compromise system integrity. Investigating and mitigating these threats are essential for maintaining the trustworthiness and reliability of machine learning applications in hostile settings.

Trust in Machine Learning Systems: The ability to defend EML models against adversarial attacks enhances trust in machine learning systems among users, stakeholders, and regulatory bodies. Demonstrating resilience to adversarial manipulation underscores the maturity and reliability of EML technologies, fostering wider adoption across industries where security and trust are paramount concerns.

Legal and Compliance Requirements: Compliance with data protection regulations, such as GDPR in Europe or HIPAA in the United States, necessitates robust measures to safeguard sensitive information. EML offers a pathway to comply with these regulations by ensuring data confidentiality during processing. Addressing adversarial threats ensures that EML implementations meet stringent legal requirements for data security and privacy.

Advancements in Secure Computing: Research into adversarial attacks on EML models drives advancements in secure computing technologies. This includes the development of more resilient encryption schemes, secure multiparty computation protocols, and novel defense mechanisms tailored to protect machine learning systems in adversarial contexts. Such advancements benefit not only EML applications but also broader fields requiring secure computation.

Ethical Implications and Societal Impact: Ethical considerations surrounding data privacy, fairness, and accountability are heightened in the context of machine learning applications vulnerable to adversarial manipulation. Addressing these implications ensures that technological advancements in machine learning align with ethical standards, promoting responsible innovation and mitigating potential societal harms.

In conclusion, the significance of studying adversarial attacks on encrypted machine learning models lies in its critical role in safeguarding data privacy, enhancing system security, fostering trust in machine learning technologies, and advancing the field of secure computing. By addressing these challenges, researchers and practitioners contribute to creating resilient and trustworthy machine learning systems capable of operating securely in adversarial environments.

LIMITATIONS & DRAWBACKS

Computational Overhead: Implementing encryption schemes and secure computation protocols introduces significant computational overhead. This can impact model training and inference times, potentially reducing the scalability and real-time performance of encrypted machine learning systems.

Complexity of Implementation: Integrating cryptographic protocols with machine learning algorithms requires specialized expertise and infrastructure. The complexity of implementation may pose barriers to adoption, particularly for organizations with limited resources or expertise in secure computing.

Performance Trade-offs: Encryption and secure computation techniques often involve trade-offs between data privacy and model performance. Ensuring high levels of privacy may require sacrificing some degree of model accuracy or computational efficiency, which can affect the overall utility of machine learning applications.

Vulnerability to Sophisticated Attacks: Despite encryption, EML models remain vulnerable to sophisticated adversarial attacks. New attack vectors and vulnerabilities in encryption schemes or model architectures may emerge, necessitating continuous monitoring and adaptation of defense mechanisms.

Limited Robustness Guarantees: While adversarial training and enhanced encryption schemes aim to improve model robustness, achieving comprehensive protection against all types of adversarial attacks remains challenging. EML systems may still exhibit vulnerabilities under certain attack scenarios or adversarial conditions.

Ethical Considerations: The deployment of EML models raises ethical concerns related to data privacy, fairness, and transparency. Adversarial attacks that exploit vulnerabilities in encrypted data could lead to unintended consequences, such as biased decision-making or breaches of user privacy, requiring careful consideration of ethical implications.

Regulatory Compliance: Compliance with data protection regulations, such as GDPR or HIPAA, requires careful handling of sensitive data. Implementing EML solutions that ensure regulatory compliance while maintaining effective security measures can be complex and resource-intensive.

Research and Development Costs: Advancing encryption techniques and developing robust defenses against adversarial attacks requires ongoing research and development efforts. The costs associated with researching, testing, and implementing secure computing solutions may be prohibitive for some organizations.

Integration Challenges: Integrating EML solutions into existing IT infrastructures and workflows may pose integration challenges. Compatibility issues with legacy systems or software dependencies could hinder the adoption and scalability of encrypted machine learning technologies.

In summary, while encrypted machine learning offers significant advancements in data privacy and security, addressing the limitations and drawbacks associated with adversarial attacks is crucial for realizing its full potential. Continued research, innovation in encryption technologies, and collaboration across disciplines are essential to mitigate these challenges and advance the reliability and effectiveness of EML solutions in real-world applications.

CONCLUSION

The study of adversarial attacks on encrypted machine learning (EML) models illuminates critical challenges and opportunities at the intersection of machine learning, cryptography, and security. Encrypted machine learning holds immense promise for safeguarding data privacy and securing sensitive computations in adversarial environments. However, it also introduces complexities and vulnerabilities that require careful consideration and innovative solutions.

Throughout this investigation, we have explored the efficacy of cryptographic protocols such as homomorphic encryption and secure multiparty computation in protecting model predictions and preserving data confidentiality during computation. We have also examined various adversarial attack methodologies that exploit weaknesses in encryption schemes or model architectures, highlighting the ongoing arms race between defenders and adversaries in the realm of secure computing.

Key findings underscore the importance of integrating robust defense mechanisms, such as adversarial training and enhanced encryption techniques, to fortify EML models against sophisticated adversarial threats. These strategies not only mitigate risks posed by adversarial manipulation but also enhance the trustworthiness and reliability of machine learning applications deployed in sensitive domains.

Despite these advancements, challenges remain, including the computational overhead associated with encryption, the complexity of implementing secure computation protocols, and the evolving nature of adversarial attacks. Addressing these challenges requires collaborative efforts across academia, industry, and regulatory bodies to develop scalable, efficient, and privacy-preserving solutions.

Ethical considerations surrounding data privacy, fairness, and transparency in machine learning applications must also be carefully navigated to ensure that technological advancements align with societal values and ethical standards. Compliance with data protection regulations, such as GDPR and HIPAA, further emphasizes the need for secure and compliant EML implementations.

Looking ahead, future research directions should focus on advancing encryption technologies, enhancing adversarial robustness through interdisciplinary approaches, and fostering transparency and accountability in machine learning systems. By addressing these challenges and seizing opportunities for innovation, we can pave the way for a future where encrypted machine learning not only protects sensitive data but also enables secure and trustworthy AI-driven solutions across diverse industries.

In conclusion, the study of adversarial attacks on encrypted machine learning models represents a critical frontier in advancing the security, privacy, and reliability of machine learning applications in an increasingly interconnected and adversarial world. Through sustained efforts and collaboration, we can harness the transformative potential of encrypted machine learning while mitigating risks and ensuring responsible innovation for the benefit of society.

REFERENCES

[1]. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

[2]. Boneh, D., & Waters, B. (2013). Constrained pseudorandom functions and their applications. In Advances in Cryptology – EUROCRYPT 2013 (pp. 280-300). Springer Berlin Heidelberg.

[3]. Juels, A., & Ristenpart, T. (2014). Honey encryption: Security beyond the brute-force bound. In Advances in Cryptology – EUROCRYPT 2014 (pp. 293-310). Springer Berlin Heidelberg.

[4]. Döring, S., Pöpper, C., & Rossow, C. (2019). 5G key management: Security analysis and enhancements. IEEE Access, 7, 22609-22621.

[5]. Amol Kulkarni, "Amazon Athena: Serverless Architecture and Troubleshooting," International Journal of Computer Trends and Technology, vol. 71, no. 5, pp. 57-61, 2023. Crossref, https://doi.org/10.14445/22312803/IJCTT-V71I5P110

[6]. Goswami, Maloy Jyoti. "Optimizing Product Lifecycle Management with AI: From Development to Deployment." International Journal of Business Management and Visuals, ISSN: 3006-2705 6.1 (2023): 36-42.

[7]. Neha Yadav, Vivek Singh, “Probabilistic Modeling of Workload Patterns for Capacity Planning in Data Center Environments” (2022). International Journal of Business Management and Visuals, ISSN: 3006-2705, 5(1), 42-48. https://ijbmv.com/index.php/home/article/view/73

[8]. Sravan Kumar Pala. (2016). Credit Risk Modeling with Big Data Analytics: Regulatory Compliance and Data Analytics in Credit Risk Modeling. (2016). International Journal of Transcontinental Discoveries, ISSN: 3006-628X, 3(1), 33-39.

[9]. Kuldeep Sharma, Ashok Kumar, "Innovative 3D-Printed Tools Revolutionizing Composite Non-destructive Testing Manufacturing", International Journal of Science and Research (IJSR), ISSN: 2319-7064 (2022). Available at: https://www.ijsr.net/archive/v12i11/SR231115222845.pdf

[10]. Bharath Kumar. (2021). Machine Learning Models for Predicting Neurological Disorders from Brain Imaging Data. Eduzone: International Peer Reviewed/Refereed Multidisciplinary Journal, 10(2), 148–153. Retrieved from https://www.eduzonejournal.com/index.php/eiprmj/article/view/565

[11]. Jatin Vaghela, A Comparative Study of NoSQL Database Performance in Big Data Analytics. (2017). International Journal of Open Publication and Exploration, ISSN: 3006-2853, 5(2), 40-45. https://ijope.com/index.php/home/article/view/110

[12]. Anand R. Mehta, Srikarthick Vijayakumar. (2018). Unveiling the Tapestry of Machine Learning: From Basics to Advanced Applications. International Journal of New Media Studies: International Peer Reviewed Scholarly Indexed Journal, 5(1), 5–11. Retrieved from https://ijnms.com/index.php/ijnms/article/view/180

[13]. Melicher, W., Evtimov, I., Weekes, B., Stavrou, A., & Shmatikov, V. (2019). Exfiltrating data from Android devices using a PC's USB cable. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (pp. 1127-1144).

[14]. Alzahrani, A., & Malaiya, Y. (2019). Adversarial machine learning in medical imaging: A review. Journal of Imaging, 5(6), 72.

[15]. Xu, W., Evans, D., & Qi, Z. (2019). Feature squeezing: Detecting adversarial examples in deep neural networks. In 2019 IEEE Symposium on Security and Privacy (SP) (pp. 218-234).

[16]. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 39-57).

[17]. Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., & Giacinto, G. (2013). Evasion attacks against machine learning at test time. In Machine Learning and Knowledge Discovery in Databases (pp. 387-402). Springer Berlin Heidelberg.

[18]. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (pp. 506-519).

[19]. Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Proceedings of the 35th International Conference on Machine Learning (Vol. 80, pp. 274-283).

[20]. Grosse, K., Manoharan, P., Papernot, N., Backes, M., & McDaniel, P. (2017). On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280.

[21]. Tramer, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction APIs. In Proceedings of the 25th USENIX Security Symposium (pp. 601-618).

[22]. Chen, P. Y., Zhang, H., Sharma, Y., Yi, J., & Hsieh, C. J. (2017). Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (pp. 15-26).

[23]. Song, X., Shu, X., & Zhu, S. (2018). Constructing adversarial examples for neural network models in IoT applications. IEEE Internet of Things Journal, 5(3), 1801-1810.

[24]. Xie, C., Wang, J., Zhang, Z., Zhou, Y., & Xie, L. (2019). Characterizing adversarial examples based on spatial consistency information for Semantic Segmentation. Neurocomputing, 363, 107-119.

[25]. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1625-1634).

[26]. Kos, J., Song, H., & Lipton, Z. C. (2017). Adversarial examples for generative models. arXiv preprint arXiv:1702.06832.

[27]. Liu, Y., Chen, X., Liu, C., & Song, D. (2019). Delving into transferable adversarial examples and black-box attacks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6511-6520).

[28]. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations.