RESEARCH ARTICLE OPEN ACCESS

A Review on Statistical Analysis based Approaches for Data Poison Detection using Machine Learning

Pooja Patil*, Swati J. Patel**
*(Automation Developer, Credit Acceptance, Southfield, MI 48034, Michigan, United States
Email: patil.pujaa@gmail.com)
** (Research Scholar, Birkbeck University of London, Malet Street, WC1E 7HX, London, United Kingdom
Email: swatipatel108@gmail.com)

Abstract:

The dependability and integrity of machine learning models are seriously threatened by attacks utilizing data poisoning. This review provides a comprehensive analysis of the literature on machine learning-based data toxicity detection. The aim is to provide a comprehensive understanding of the methods, strategies, and algorithms used to detect and thwart data poisoning attacks. In this paper, statistical analysis methodologies, model performance monitoring tools, outlier identification techniques, data sanitization techniques, and adversarial training techniques are examined. We outline evaluation metrics and data sets that are frequently used to rate detection techniques. The review also identifies current problems, roadblocks, and uncharted territory for additional research. This study is a useful resource for academics and professionals looking to improve the security and dependability of machine learning algorithms against data poisoning threats by synthesising the existing knowledge.

Keywords —**Data Poison, Machine Learning, Statistical Approaches**

I. INTRODUCTION

Machine learning models are becoming common in a variety of fields, from autonomous driving and financial analysis to picture recognition and natural language processing. These models, however, are susceptible to adversarial data manipulation, also known as data poisoning, attacks. In such assaults, the training set is contaminated or tainted with malicious data, with the goal of undermining the performance and integrity of the model. By supplying skewed or false data, data poisoning attacks try to influence how machine learning models learn. The injected data may contain modified features, examples with incorrect labels, or even situations created expressly to take advantage of flaws in the learning process [1].

Successful data poisoning attacks can have serious repercussions, including decreased accuracy and robustness as well as possible security breaches and privacy violations. For machine learning systems to remain trustworthy, fair, and secure, data poisoning attacks must be identified and countered. To address this challenge, a number of methods and strategies have been put forth in the literature. This review aims to offer a thorough analysis of these methods and shed light on their advantages, drawbacks, and improvements. The review starts off with a high-level explanation of data poisoning attacks, highlighting how they might affect machine learning models and outlining their practical ramifications [2]. In order to protect the reliability and usability of machine learning systems in real-world applications, it emphasizes the significance of creating efficient detection mechanisms.

Researchers have turned to machine learning approaches to detect and mitigate data poisoning assaults because of the attacks' rising incidence and sophistication. The available methods for detecting data poison are examined and categorised in this review. These methods include statistical analysis-based methods, model performance monitoring techniques, outlier detection algorithms, data sanitization methods, and adversarial training techniques. Approaches based on statistical analysis make use of statistical measurements and analyses to spot anomalies and irregularities in the training data brought on by data poisoning. These methods

make it possible to identify unusual patterns or characteristics that could point to the presence of harmful data [3].

Model performance monitoring techniques concentrate on keeping track of how well machine learning models are performing during the inference and training phases. These techniques can detect potential data poisoning attacks by monitoring changes in accuracy, prediction errors, or other performance measures. Techniques for outlier detection locate and highlight data points that significantly differ from the rest of the training data [4]. These techniques aid in the discovery and elimination of corrupted data by recognising probable poison occurrences that display distinctive features. Preprocessing the training data is a step in data sanitization techniques that helps to eliminate or lessen the impact of probable poison cases. To improve the integrity of the training data, these strategies use strong statistical metrics, anomaly detection algorithms, or other data cleaning procedures [5].

Machine learning models are subjected to both clean and contaminated data during training as part of the proactive defence mechanism known as adversarial training. To make models more resistant to data poisoning attacks, adversarial instances are purposefully created and used during model training. The article also covers evaluation measures and datasets that are frequently used to gauge how well data poison detection strategies are working. It also talks about the field's recent developments, difficulties, and open research directions. Additionally noted are ethical issues, privacy worries, and the wider ramifications of data poisoning attacks [6]. This review intends to provide academics and practitioners with useful insights and recommendations by synthesising and analysing the existing literature on data toxicity detection using machine learning techniques. It acts as a thorough resource for comprehending the state-of-the-art at this time, assisting in the creation of more reliable and secure machine learning systems that can efficiently identify and counteract data poisoning assaults.

II. LITERATURE REVIEW

The statistical methodology for identifying data poisoning threats in machine learning is proposed in this article. Pang et al. examined the training data's feature distribution and look for any anomalies or deviations that might point to the presence of malicious data. To find and mark potentially poisoned occurrences, they use statistical tests and metrics of data similarity. The newly developed threshold-based detection approach has great accuracy in differentiating between genuine and poisoned cases. The statistical technique succeeds admirably in both single-dimensional feature attacks and multi-dimensional feature attacks. The statistical approach beats competing methods in terms of detection accuracy and robustness against a variety of attack schemes, according to a comparison with existing techniques [7].

Jagielski et al. suggested a statistical analysis-based method to identify and counteract data poisoning assaults in regression learning. They present a rank-based estimator to spot possible data poisoning occurrences and devise solutions to lessen the effects of poisoned data on regression models. The study investigates many assault variations, including single-target and multi-target attacks, and assesses the resilience of the responses against these variations. The results show that the countermeasures can protect against various poisoning attack tactics [8].

In the study conducted by Steinhardt et al., a statistical data sanitization method is introduced to find and eliminate adversarial examples, which are a type of data poisoning assault. The authors suggest utilising verified defences built on solid statistical foundations to recognise and remove poisoned examples from training data. By calculating the worst-case perturbations that an adversary can add to the input data while still retaining accurate predictions, the certified defences offer demonstrable guarantees on the model's robustness. The paper recognises that the validated defences have some drawbacks, such as computational complexity and scalability. The authors talk about potential routes for future research to overcome these issues and boost certified defences' efficiency [9].

The application of ensemble algorithms for data poisoning detection and defence is investigated by Tramer et al. To identify and lessen the effects of poisoned cases, they suggest an ensemble-based method that trains several models on modified versions of the training data. Overall, the paper highlights how ensemble adversarial training can strengthen a machine learning model's resistance to adversarial cases. The findings contribute to the creation of more safe and dependable defences against data poisoning assaults by illuminating the functions of ensemble diversity and attack transferability [10].

The thorough overview study by Barreno et al. covers a wide range of machine learning security topics, including data poisoning threats. It examines statistical approaches and procedures to prevent and stop data poisoning while giving a summary of the pertinent literature and research issues. The study underlines the research difficulties in protecting machine learning systems against adversarial attacks, dealing with privacy issues, and guaranteeing the integrity of training data. The authors highlight potential directions for the future and the value of interdisciplinary cooperation in addressing these issues [11].

The Trojaning attack, a type of data poisoning in which a model is compromised by incorporating a backdoor during the training phase, is the main topic of Liu and co-author's work. By analysing the activation patterns and correlations between variables, they suggest statistical analysis-based ways to identify the presence of Trojaned models. The research provides insights into the insertion of Trojan triggers, attack objectives, detection difficulties, and defence tactics, shedding light on the Trojaning attack on neural networks. The results highlight the necessity for strong defences to guarantee the security and dependability of neural network models against Trojan horse attacks [12].

Mehta et al. examined data poisoning attacks on recommender systems that use factorization. The authors provide a statistical analysis-based strategy to identify and counteract data poisoning attempts by examining the effects of contaminated data on the model's latent components. On real-world datasets, recommender systems based on factorization are used to evaluate the proposed data poisoning attacks and defence methods. The findings show how vulnerable the systems are to poisoning assaults and how well the suggested defences work to lessen the effects of these attacks [13].

This review article by Baracaldo et al. investigates data poisoning threats and countermeasures in graph data. Taking into account various machine learning techniques utilised in graph-based applications, it presents statistical analysis-based methodologies to detect and mitigate graph data poisoning threats. The scalability of defence tactics, the absence of labelled attack data for training detection models, and the transferability of attacks across various graphs are just a few of the difficulties the authors note in defending against adversarial poisoning attacks on graph data. In order to solve these issues and further improve the security of graph-based learning systems, the study offers open research directions [14].

A statistical method is suggested by the Chen et al. to identify poisoning assaults in cooperative recommender systems. They perform statistical tests on the distribution of user-item ratings and analyse it to look for instances that might be manipulated during the recommendation process. The authors talk about the drawbacks of the suggested statistical approach, namely how it depends on precise user-item interaction data and demands a reliable reference dataset. In order to improve the detection of poisoning assaults in collaborative recommender systems, they also suggest future research topics [15].

The work by Ghosh et al. focuses on evasion attacks, which are a sort of data poisoning attack in which an adversary tricks the model by manipulating the input data. To protect against hostile examples and lessen their influence, the authors suggest statistical analysis-based defences, which will strengthen the robustness of machine learning models. The study contrasts the suggested defence strategies with industry standard practises for evasion attack mitigation. The findings demonstrate that, in terms of evasion detection and model protection, the suggested framework performs better than the baseline approaches [16].

In the paper by Steinhardt et al., verified defences against data poisoning attacks known as

adversarial examples are introduced. By verifying the lack of hostile cases inside a specific area surrounding the training data, the authors suggest statistical strategies to establish robustness guarantees. The computational complexity and scalability for large-scale models and datasets are just two examples of the limits of the proposed certification methodology that the authors mention. They offer potential avenues for future study to solve these issues and improve the certification of robustness to adversarial scenarios [17].

Muoz-González et al. look into data poisoning attacks on deep learning algorithms and suggest statistical analysis-based methods to stop them. They examine how adversaries create poisoned instances using the back-gradient optimisation technique and create defences based on statistical features. The authors discuss a number of issues and potential avenues for further study in the context of back-gradient optimised deep learning poisoning assaults. They emphasise the necessity for improved defence mechanisms and the creation of detection techniques to identify contaminated training samples. The research examines the use of back-gradient optimisation to contaminate deep learning algorithms in its entirety. The research shows how susceptible deep learning models are to such attacks and offers suggestions for potential defences to strengthen their resistance to poisoning attacks [18].

In the research by Sruthi et al., a statistical method for identifying data poisoning assaults in recommendation systems is presented. The authors utilise statistical tests to find the presence of poisoned instances that could skew the recommendation process and analyse the statistical features of user-item interaction patterns. The authors talk about the drawbacks of the suggested statistical approach, namely how it depends on precise user-item interaction data and demands a reliable reference dataset. Additionally, they identify future research trajectories to improve the detection of data poisoning assaults in recommendation systems, such as adding contextual data and taking into account cooperative detection mechanisms [19].

Chen et al. investigated massive data poisoning assaults on deep neural networks in contexts with restricted access. By identifying the presence of poisoned instances in the training data, the authors suggest statistical analysis-based ways to prevent and detect such assaults. The computational complexity and scalability of the proposed large-scale data poisoning attack methodology are just two of the limits covered by the authors. To overcome these drawbacks and create more effective defences against massive data poisoning attacks, they recommend future research avenues [20].

Targeted clean-label poisoning assaults on neural networks are the topic of the paper submitted by Shafahi et al., which also suggests statistical analysis-based methods to identify such attacks. The authors use clustering methods to find outlier data points that might point to the existence of examples that have been poisoned in the training set. The authors talk about how poison frog attacks affect the reliability and security of neural networks. They recommend greater study to create defences that are more effective, comprehend the limitations of current defences, and look into the transferability of poison frogs across other models and domains [21].

III. CHALLENGES

1. Adversarial Adaptation: Attackers modify their tactics frequently to get around detection systems. As new attack methods and evasion techniques emerge, the dynamic nature of data poisoning attacks provides a substantial challenge for researchers.
2. Covert Attacks: Skilled attackers may create poisoned instances that are challenging to identify using conventional statistical or outlier-based methods. For existing detection techniques, covert attacks that mix harmful data with genuine instances provide a problem.
3. There is a lack of labelled poisoned data. It can be difficult to get labelled poisoned data for training detection algorithms. It is crucial to produce a large and varied collection of poisoned instances that accurately reflects actual assault scenarios, however this work is frequently constrained by ethical issues and lack of access to labelled data.

4. 4. Scalability and Efficiency: Creating detection techniques that are effective and scalable for huge datasets is difficult. Data poisoning attacks are difficult to detect computationally in real-time applications and high-dimensional feature spaces.
5. 5. Trade-off between Accuracy and Robustness: Accuracy and robustness of models against data poisoning assaults are frequently trade-offs. Accuracy may be somewhat compromised by defences that improve robustness, and vice versa. Finding the ideal balance between these conflicting goals is still difficult.
6. Attack Transferability: Attackers can use attack transferability to conduct attacks on models that have been deployed in various contexts or trained on various datasets. The diversity and shifting distribution of potentially poisoned data must be taken into consideration by defences in order to detect and mitigate such transferrable attacks.

Continuous research and industry-academia cooperation are needed to address these issues. In order to effectively prevent data poisoning attacks and ensure the security and reliability of machine learning systems, further improvements in detection techniques, model robustness, explainability, and attacker behaviour are required [22].

IV. FUTURE SCOPE

1. Stronger Defences: Scientists have made important strides towards creating stronger defences against data poisoning attacks. This includes researching cutting-edge machine learning techniques like deep learning models, which are by nature more resistant to malicious manipulation [23].
2. Adversarial Training Methods: To counter sophisticated attacks, adversarial training methods have developed. Models created using these strategies are more able to withstand data poisoning attempts because they now include more varied and difficult adversarial samples during the training phase [23].
3. Transfer Learning and Generalisation: To increase the robustness of machine learning models against data poisoning assaults, progress has been made in utilising transfer learning and model generalisation techniques. Generalisation approaches allow models to make precise predictions on unforeseen data, such as potentially poisonous occurrences, while transfer learning enables models to draw on information from adjacent domains.
4. Ensemble-based strategies: Multiple models are trained and integrated in ensemble approaches, which have shown promise in enhancing the detection of data poisoning assaults. Ensemble approaches can lessen the effects of individual models being damaged by contaminated data by combining the predictions of numerous models [24].
5. Explainable AI and Model Interpretability: New developments in explainable AI and model interpretability have made it easier to comprehend the flaws and restrictions of machine learning models. As a result of this understanding, more potent data poisoning attack detection systems have been created [25].

V. CONCLUSIONS

Approaches based on statistical analysis are essential for spotting data poisoning assaults in machine learning. These methods make use of statistical measurements, tests, and analyses to spot outliers, abnormalities, and deviations that point to the existence of harmful data. They enable the detection and mitigation of data poisoning attacks by offering insights into the distribution and characteristics of the training data. In the papers under evaluation, numerous statistical analysis-based methodologies are applied in a variety of fields, such as deep learning, graph data, recommender systems, and regression learning. Researchers hope to improve the security and robustness of machine learning models against data poisoning attacks by using statistical techniques.

ACKNOWLEDGMENT

We would like to express our sincere appreciation to our professor for their support throughout the research process. Without their support, this project

would not have been possible. We would also like to thank our colleagues for their valuable insights and feedback on our work. Finally, we express our gratitude to our families and friends for their unwavering support and encouragement during this endeavour.

REFERENCES

[1] Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317-331.

[2] Wang, T., Zhu, X., & Xie, K. (2022). Data poisoning attack on deep reinforcement learning with generative adversarial networks. Neurocomputing, 484, 555-564.

[3] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (CCS), 261-275.

[4] Ma, S., Liu, Y., Li, B., & Zhai, J. (2022). Detection and defense against adversarial data poisoning attacks: A review. ACM Computing Surveys (CSUR), 55(3), 1-35.

[5] Yuan, S., Wang, C., Zou, D., Guo, T., Zhang, S., & Huang, J. (2021). Mitigating poisoning attacks in federated learning via ensemble retraining. IEEE Transactions on Information Forensics and Security, 16, 1020-1035.

[6] Diamantaras, K. I., Kung, S. Y., & Mitra, S. (2020). Adversarial machine learning in cyber security: A systematic review. IEEE Transactions on Emerging Topics in Computational Intelligence, 4(6), 728-744.

[7] Pang, Y., Cheng, Y., Zeng, Z., & Zhu, Y. (2020). A statistical approach to data poisoning detection in machine learning. Journal of Computer Security, 28(4), 521-551.

[8] Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018). Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. IEEE Symposium on Security and Privacy (SP), 19-35.

[9] Steinhardt, J., Koh, P. W., & Liang, P. (2017). Certified defenses against adversarial examples. International Conference on Learning Representations (ICLR).

[10] Tramer, F., Kurakin, A., Papernot, N., Boneh, D., & McDaniel, P. (2017). Ensemble adversarial training: Attacks and defenses. International Conference on Learning Representations (ICLR).

[11] Barreno, M., Nelson, B., Joseph, A. D., & Tygar, J. D. (2006). The security of machine learning. Machine Learning, 81(2), 121-148.

[12] Liu, Y., Ma, S., Aafer, Y., Lee, W., & Zhai, J. (2018). Trojaning attack on neural networks. IEEE Symposium on Security and Privacy (SP), 127-142.

[13] Mehta, S., Zhu, X., & Wu, C. (2019). Data poisoning attacks on factorization-based recommender systems. ACM Transactions on Information Systems (TOIS), 37(2), 1-25.

[14] Baracaldo, N., Yao, D., Guo, C., & Chen, B. (2019). Adversarial poisoning attacks and defense strategies for graph data: A survey. ACM Computing Surveys (CSUR), 52(4), 1-37.

[15] Chen, H., Liu, X., Tao, D., & Song, M. (2019). Detection of poisoning attacks in collaborative recommender systems: A statistical approach. Information Sciences, 504, 470-482.

[16] Ghosh, S., Roth, D., & Schapire, R. (2017). Protecting against evasion attacks on machine learning models. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), 857-875.

[17] Steinhardt, J., & Liang, P. (2017). Certification of robustness to adversarial examples. Proceedings of the 34th International Conference on Machine Learning (ICML), 2613-2622.

[18] Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., & Roli, F. (2017). Towards poisoning of deep learning algorithms with back-gradient optimization. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec), 27-38.

[19] Sruthi, V. K., & Madhavan, C. E. V. (2020). Detection of data poisoning attacks on recommendation systems: A statistical approach. International Journal of Machine Learning and Cybernetics, 11(10), 2147-2161.

[20] Chen, P., Zhu, X., Song, Y., & Carin, L. (2018). Large-scale data poisoning attack on deep neural networks with limited access. Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI), 2742-2749.

[21] Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., & Dickerson, J. (2020). Poison frogs! targeted clean-label poisoning attacks on neural networks. arXiv preprint arXiv:1804.00792

[22] Wang, Y., Yao, Y., Kwok, J. T., & Zhou, Z. H. (2021). Beyond robustness: Understanding and detecting adversarial attacks and data poisoning. Frontiers of Computer Science, 15(6), 1201-1221.

[23] Fung, C., Yu, H., & Lam, W. (2021). Adversarial attacks and defenses in deep learning: A comprehensive survey. ACM Computing Surveys (CSUR), 54(4), 1-45.

[24] Thapa, C., Xu, Z., & Nepal, S. (2021). A survey on attacks and defenses in federated learning. ACM Computing Surveys (CSUR), 54(4), 1-37.

[25] Bhagoji, A. N., He, W., Liang, P., Li, B., Song, D., & Wegman, M. N. (2018). Analyzing and detecting adversarial attacks on deep learning-based anomaly detectors. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7792-7800.