International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

DATA POISONING ATTACKS ON REGRESSION MODELS AND THEIR DEFENSES

Spoorti P Patgar
Student, Dept. of Cyber Security
APS College of Engineering
spoortipatgar@gmail.com

Akshata Jawade
Student, Dept. of Cyber Security
APS College of Engineering
akshatajawade12@gmail.com

Abhishek R Karanje
Student, Dept. of Cyber Security
APS College of Engineering
abhishekrk0001@gmail.com

Abstract

Machine learning models are widely deployed in critical domains such as healthcare, finance, transportation, and cyber-physical systems, making their security and reliability a paramount concern. Among the most significant threats to these systems are data poisoning attacks, which compromise model integrity by injecting malicious or manipulated data into training datasets, leading to degraded performance and unreliable predictions. While data poisoning has been extensively studied in classification tasks, its impact on regression models — which are equally critical for applications such as medication dosage management, power supply regulation, and financial forecasting — remains comparatively underexplored.

Experimental evidence consistently demonstrates that even a minimal fraction of poisoned data, as low as 2%, can drastically increase prediction error by up to 150% in mean squared error (MSE), underscoring the severity of this threat. Common attack strategies include label manipulation, feature perturbation, adversarial data injection, and optimization-based black-box attacks that operate without prior knowledge of the target model. In response to these vulnerabilities, researchers have proposed several defense mechanisms, including Trimmed Loss, Differential Privacy, and the novel Iterative Trim (I Trim) method, which effectively detects and removes poisoned samples without requiring prior knowledge of attack intensity.

Additionally, adaptive strategies such as dynamic network structure adjustment and adaptive learning weights have demonstrated strong potential in reducing the influence of poisoned data during training, thereby preserving model accuracy and robustness. Evaluations conducted across diverse datasets confirm that these defense frameworks significantly mitigate poisoning effects while maintaining practical model performance. These findings collectively emphasize the urgent need for secure data pipelines, robust learning algorithms, and adaptive defense strategies to safeguard machine learning systems against evolving adversarial threats in real-world applications.

Keywords: Data Poisoning, Regression Models, Adversarial Machine Learning, Outlier Detection, RANSAC, Robust Learning.

Problem Statement

Regression models are vulnerable to adversarial data poisoning attacks, where

malicious data points are introduced into the training dataset, resulting in degraded model performance and unreliable predictions. The challenge is to design effective defense mechanisms that can detect and mitigate such attacks while preserving model accuracy.

Introduction

Machine learning has become a foundational technology across numerous mission-critical domains, including healthcare, finance, autonomous systems, industrial quality control, and transportation. Applications such as pharmaceutical dosage prediction, stock price forecasting, traffic state estimation, and predictive maintenance now rely heavily on regression models — algorithms designed to predict continuous values with high precision. As the adoption of these systems grows rapidly, so does the urgency to address their security vulnerabilities, particularly those that arise from their dependence on large, often unverified datasets.

One of the most significant threats to machine learning systems is the **data poisoning attack** — a class of adversarial attack that occurs during the training phase, where a malicious actor deliberately injects crafted or corrupted samples into the training dataset. Unlike evasion attacks, which attempt to deceive a model at inference time, poisoning attacks are embedded at the source, causing the model to learn incorrect or manipulated patterns that persist throughout its entire lifecycle.

The consequences of such attacks range from general performance degradation and prediction bias to the creation of hidden backdoors that can be exploited on demand. Specific poisoning strategies include label flipping, feature corruption, and adversarial data insertion — all of which can cause severe disruption to model stability and reliability.

While poisoning attacks have been extensively studied in the context of classification models, regression models have received comparatively little attention from the research community. This gap is particularly alarming given that regression systems directly influence real-world, high stakes decisions. A compelling illustration of this risk is Warfarin dosage prediction in clinical settings — a domain where even a minor skew in model output due to poisoned training data can result in dangerous over- or under-medication, posing life-threatening risks to patients. Similarly, in financial systems, corrupted regression models can lead to faulty forecasts, causing significant economic harm.

Existing defense mechanisms against data poisoning primarily rely on data preprocessing, anomaly detection, and robustness optimization techniques. However, these approaches suffer from notable limitations: they frequently fail to identify well-hidden malicious samples, struggle to generalize across diverse attack strategies, and often introduce prohibitively high computational overhead.

There is therefore a pressing need for efficient, adaptive defense frameworks that not only enhance model robustness against

poisoning attacks but also maintain predictive performance without significant computational cost. This research aims to bridge this gap by systematically analyzing the impact of data poisoning on both linear and nonlinear regression learners and proposing an improved, adaptive defense mechanism capable of securing regression models in real-world deployment scenarios.

Literature Review

The growing deployment of machine learning systems in safety-critical domains has attracted significant research attention toward understanding and mitigating adversarial threats, particularly data poisoning attacks. Early investigations into adversarial machine learning focused predominantly on classification models, establishing a foundational understanding of how malicious actors could manipulate model behavior through carefully crafted inputs. Over time, this body of work has expanded to encompass regression models, defense mechanisms, and real-world application-specific vulnerabilities. [11],[17],[18].

1.Early Work on Classification Poisoning

The earliest systematic studies on poisoning attacks targeted classification models such as Support Vector Machines (SVMs), Lasso, and Ridge Regression. These works employed Karush-Kuhn-Tucker (KKT) optimality conditions to mathematically derive optimal poisoning samples that maximally disrupted model decision

boundaries. Subsequent research extended these techniques to multi-class classification scenarios and deep neural networks, leveraging back-gradient optimization to craft poisoned samples that propagated adversarial influence through multiple layers of model training. Clean label attacks — where adversaries manipulate feature representations without altering ground-truth labels — further expanded the threat surface, making detection considerably more difficult. These foundational classification-focused studies established the theoretical and computational frameworks that later researchers adapted for regression settings.[2],[3],[4],[19],[25].

2. Poisoning Attacks on Regression Models

Despite the critical importance of regression models in real-world decision making, research on regression-specific poisoning attacks has remained notably sparse. Jagielski et al. (2018) conducted one of the first and most influential systematic studies on poisoning attacks targeting linear regression models. Their work introduced gradient-based optimization strategies specifically designed to maximize model error, demonstrating that inserting as little as 2% poisoned data into a training set can increase the Mean Squared Error (MSE) by a factor of over 150 compared to a clean model. This finding underscored the extreme sensitivity of regression learners to even minimal adversarial interference.

Critically, prior to recent advances, nonlinear regression models — including Neural Networks and Kernel Support Vector

Regression (SVR) — had received virtually no attention in the poisoning attack

literature, representing a significant and dangerous research gap. [1],[5],[12].

3.Application-Specific Vulnerabilities

Beyond theoretical analysis, several studies have demonstrated the real-world consequences of regression poisoning across diverse domains. Wang et al. (2024) formulated data poisoning attacks as a sensitivity analysis problem within Traffic State Estimation (TSE) systems, showing how adversaries can deliberately flatten the slope of traffic flow prediction models to cause congestion mismanagement at scale. In smart grid environments, Zhu et al. (2023) developed online poisoning frameworks where attackers incrementally pollute incoming data streams to evade detection while continuously degrading model reliability. Perhaps most alarmingly, medical applications such as Warfarin dosage prediction have been identified as high-risk targets, where even minor prediction errors induced by poisoned training data can result in life-threatening patient outcomes. These applications specific studies collectively demonstrate that the consequences of regression poisoning extend far beyond academic metrics and into direct real-world harm. [9],[10],[20],[21],[22].

4. Defense Mechanisms and Their Limitations

In response to the growing threat of data poisoning, researchers have proposed a variety of defense strategies, each with distinct strengths and limitations. Early

geometric defenses, such as the Sphere defense, operated by removing training points that fell outside a defined spherical radius in feature space, while the Slab defense filtered samples based on directional constraints relative to the data distribution. The TRIM algorithm, considered a state-of-the-art defense for regression poisoning, is an iterative approach that fits a regressor on the subset of training data exhibiting the smallest residual errors. However, TRIM suffers from a critical practical limitation: it requires prior knowledge of the exact fraction of poisoned data in the training set, an assumption that is rarely satisfied in real deployment scenarios.

Differential Privacy (DP) has also been explored as a defense paradigm. Ma et al. (2019) formally proved that differentially private learners are inherently resistant to poisoning attacks, as the DP property mathematically limits the influence any single data point — whether clean or malicious — can exert on the final trained model. Building on this, game-theoretic frameworks have been developed to model the strategic interaction between attackers and defenders in differentially private learning environments. Baracaldo et al. (2017) proposed a data provenance approach, leveraging metadata such as data origin and sensor identifiers to segment training data and flag suspicious sources prior to model training.

On the detection front, Müller et al. (2020) proposed statistical detection-based defenses specifically tailored for regression poisoning scenarios, using anomaly scoring to isolate malicious samples. Chen et al.

(2021) introduced De-Pois, an attack agnostic defense framework that

restructures the training optimization process to minimize the influence of poisoned data irrespective of the specific attack strategy employed. More recently, Mao et al. (2025) proposed decentralized optimization techniques resilient to local data poisoning in federated settings, while Zheng et al. (2025) introduced game theoretic defense strategies designed for crowdsensing environments where data collection is inherently distributed and difficult to verify. [1],[5],[6],[7],[8],[13],[14],[16].

5. Identified Research Gaps

Despite these advances, the existing literature reveals several persistent and critical limitations. First, the majority of poisoning defenses have been designed and evaluated exclusively for classification models, leaving regression learners comparatively under protected. Second, defenses such as TRIM rely on unrealistic assumptions about attacker knowledge or poisoning rates. Third, many proposed methods incur high computational costs or suffer from incomplete detection of well-hidden malicious samples, rendering them impractical for large-scale or real-time systems. Fourth, the study of black-box poisoning attacks — where the adversary has no direct access to model internals — and bilevel optimization-based poisoning strategies remains an emerging and underdeveloped area. These gaps collectively highlight the urgent need for adaptive, lightweight, and assumption-free defense mechanisms capable of protecting nonlinear regression models in practical, real-world deployment conditions. [1],[11],[12],[23],[15].

Methodology

This study follows a structured and comprehensive experimental pipeline designed to investigate the impact of data poisoning attacks on regression models and evaluate the effectiveness of both existing and proposed defense mechanisms. The methodology is organized into six key phases: dataset preparation, model design, attack modelling, defense mechanisms, experimental setup, and performance evaluation.

1. Dataset Preparation:

The experimental framework utilizes multiple benchmark datasets to ensure the generalizability of findings across diverse domains. Datasets include regression oriented data such as housing prices and medical records, as well as classification benchmarks such as the Iris dataset — a well-established collection of 150 samples distributed across three categories with four input features each. All datasets are split into dedicated training and testing subsets to maintain evaluation integrity. Prior to model training, a thorough data preprocessing pipeline is applied. Outlier detection is performed using the k-Nearest Neighbors (k-NN) algorithm, which identifies and removes data points that deviate significantly from the expected distribution of normal samples. Following anomaly removal, all feature values are normalized to the range [0, 1] through minmax scaling, ensuring training stability, accelerating convergence, and preventing any single

feature from disproportionately influencing model learning.

2. Baseline Model Design:

To comprehensively evaluate poisoning impacts across different model architectures, both linear and nonlinear regression models are included in the experimental setup. Linear models examined include Ridge Regression, Lasso Regression, Elastic Net, and Huber Regression — each representing varying degrees of regularization and robustness to outliers. Nonlinear models include Neural Networks, Support Vector Regression (SVR), and Kernel Ridge Regression, which capture complex, high-dimensional relationships in data. For the adaptive defense experiments, a Multi-Layer Perceptron (MLP) architecture is employed as the primary model, configured with an input layer of 4 neurons, one or more hidden layers initialized with 10 neurons using RELU activation functions, and an output layer sized according to the prediction task. This diverse model selection ensures that findings are not limited to a single architecture and reflect the broader vulnerability landscape of regression learners.

3. Attack Modelling:

The study considers a realistic black-box attack scenario, wherein the adversary has no direct knowledge of the target model's architecture, hyperparameters, or training dataset. Instead, the attacker operates with access to a substitute dataset and the limited ability to inject a controlled number of malicious samples into the training pipeline.

Three primary poisoning strategies are investigated to simulate varied adversarial behaviors:

- Label Flipping: The adversary modifies the continuous target values of selected training samples, pushing them to extreme ends of the feasible value range while keeping input features unchanged. This directly contradicts the true data distribution and forces the model to learn erroneous mappings.

- Feature Manipulation: Input feature values of selected samples are altered to create misleading data representations, causing the model to associate incorrect patterns with given outputs.
- Adversarial Sample Insertion: Carefully optimized data points are crafted and inserted into the training set to maximally disrupt model learning, guided by a bilevel optimization objective that seeks to maximize prediction error on a clean validation set.

A novel attack strategy, termed the **Flip Attack**, is also proposed in this work. The Flip Attack identifies training points in the substitute dataset whose target values lie closest to the boundaries of the feasible domain and flips those values to the opposite extreme. Crucially, this attack is model-agnostic — it operates independently of the specific regressor being targeted — making it a practical and widely applicable black-box threat. Poisoning levels are varied systematically across experiments (ranging from 2% to 75% of the training set)

to observe the progressive degradation in model performance and identify critical vulnerability thresholds.

4. Defense Mechanisms:

(a) TRIM Defense: The TRIM algorithm is an iterative defense that fits the regression model on a subset of training data characterized by the smallest residual errors, effectively isolating and discarding suspected poisoned samples. While TRIM represents a strong baseline, its practical utility is constrained by its requirement for an accurate prior estimate of the poisoning rate — a piece of information rarely available to defenders in real-world deployments.

(b) I Trim — Iterative Trim (Proposed): To overcome the limitations of TRIM, this study introduces the Iterative Trim (I Trim) defense. I Trim eliminates the need for a known poisoning rate by systematically testing multiple candidates estimates and selecting the optimal value based on observed training loss behavior. Specifically, the algorithm monitors the training loss curve across different poisoning rate assumptions and identifies a characteristic inflection point — a "kink" — at which the loss stabilizes, indicating that all poisoned samples have been successfully removed. This data-driven, assumption-free approach significantly improves upon TRIM in practical scenarios.

(c) Adaptive MLP Defense: Additionally, an adaptive defense mechanism integrated directly into the MLP architecture is proposed. The model dynamically adjusts its network structure every 20 training epochs based on observed accuracy: if accuracy falls below 85%, the number of hidden neurons is halved to reduce model complexity and limit the influence of noisy data; if accuracy meets or exceeds 85%, the neuron count is increased to enhance learning capacity. Simultaneously, the model assigns reduced learning weights to training samples identified.

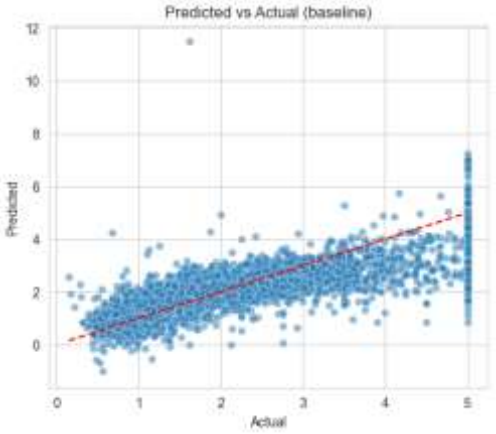
5. Experimental Setup:

A total of 26 datasets and seven regressor configurations — four linear and three nonlinear — are used across experiments, resulting in over 2,000 individual experimental runs to ensure statistical reproducibility. Hyperparameter tuning is performed using Cross-Validated Grid Search. Three experimental conditions are compared for each model: a clean baseline model trained on unpoisoned data, a poisoned model subjected to adversarial data injection, and a defended model incorporating the proposed I Trim or adaptive MLP defense.

6. Evaluation Metrics:

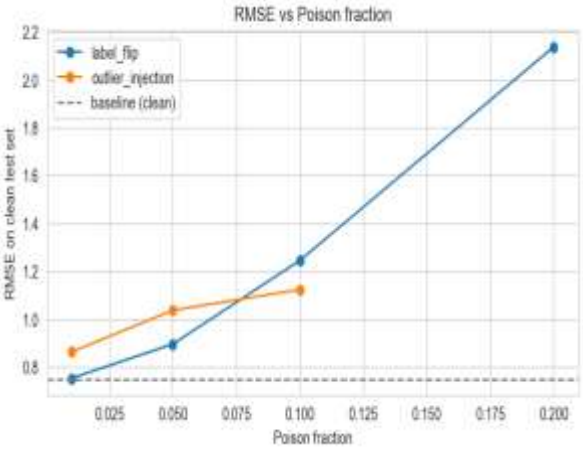
Model performance is assessed using the following quantitative metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R^2 Score, and overall classification or prediction accuracy. These metrics collectively capture both the magnitude of prediction error and the proportion of variance explained by the model, providing a holistic picture of performance degradation under attack and recovery under defense.

Experimental Results and Analysis



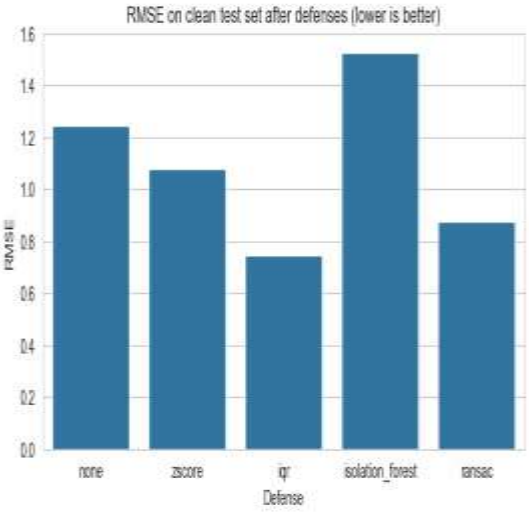
The scatter plot illustrates the relationship between the predicted and actual values obtained from the regression model trained on clean data. The red dashed line represents the ideal scenario where predicted values exactly match the actual values. Most data points are clustered around this line, indicating that the model performs reasonably well under normal conditions. However, some dispersion is observed, suggesting the presence of inherent noise and minor prediction errors in the dataset.

Figure 1 : Predicted vs Actual values for baseline regression model.



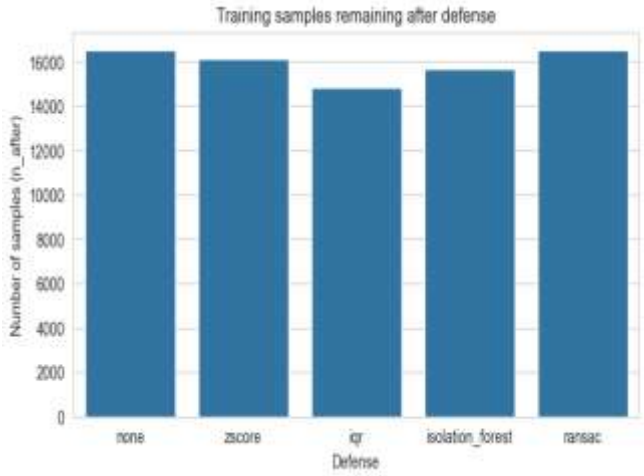
This graph shows how model performance degrades as the proportion of poisoned data increases. Both label flipping and outlier injection attacks significantly increase RMSE, indicating reduced prediction accuracy. The label flipping attack exhibits a sharper increase in error at higher poison fractions, highlighting its effectiveness in misleading the model. The baseline (clean data) RMSE is shown as a reference line, emphasizing the extent of degradation caused by adversarial manipulation.

Figure 2 : Effect of data poisoning on model performance



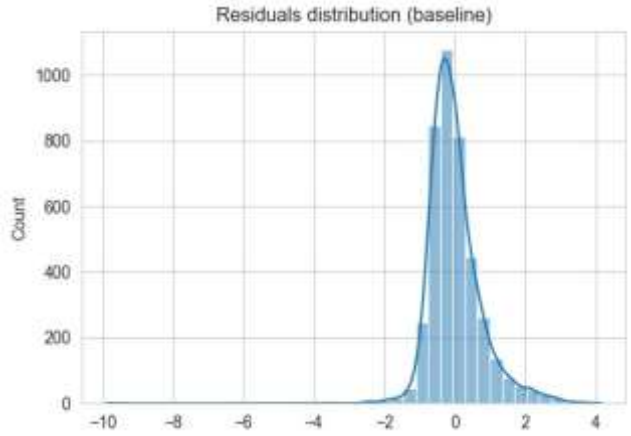
This figure compares the effectiveness of different defense mechanisms applied to poisoned datasets. The results indicate that IQR filtering achieves the lowest RMSE, making it the most effective defense in this scenario. RANSAC regression also performs well by maintaining robustness without removing data points. In contrast, Isolation Forest shows poor performance, likely due to excessive removal of valid data points. This highlights the importance of selecting appropriate defense strategies based on attack characteristics.

Figure 3 : Comparison of defense mechanisms based on



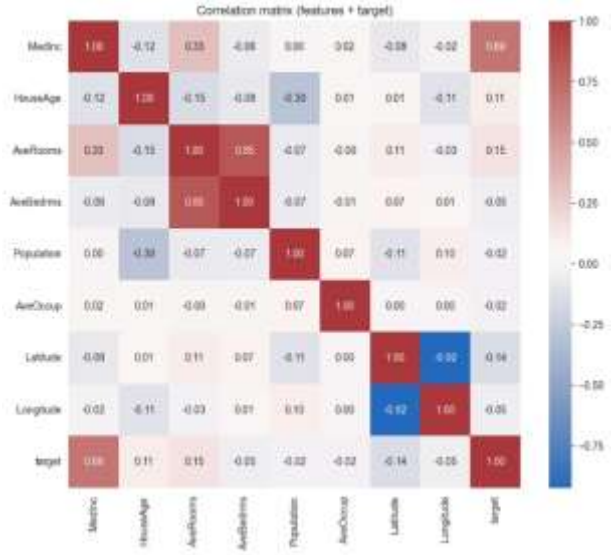
This graph illustrates how different defense mechanisms impact the size of the training dataset. IQR and Z-score filtering remove a portion of data points identified as outliers, reducing dataset size. Isolation Forest removes a larger number of samples, which may lead to loss of important information. RANSAC, on the other hand, retains all samples while internally handling outliers, demonstrating its advantage in preserving data while maintaining robustness.

Figure 4 : Number of training samples remaining after applying defenses.



The residual distribution shows the difference between actual and predicted values. The distribution is centered around zero, indicating that the model does not exhibit significant bias. However, the spread of residuals suggests the presence of prediction errors, which can be amplified under poisoning conditions. This analysis helps in understanding model accuracy and identifying deviations from ideal performance.

Figure 5 : Distribution of residual errors for baseline model.



The correlation heatmap illustrates relationships between input features and the target variable. Strong positive and negative correlations can be observed among certain features, which influence the regression model's predictions. For example, features with higher correlation to the target contribute more significantly to prediction accuracy. Understanding these relationships is essential for interpreting model behavior and identifying potential vulnerabilities to data poisoning.

Figure 6 : Correlation matrix of features and target variable.

Conclusion

This study analyzed the impact of data poisoning attacks on regression models and evaluated the effectiveness of various defense mechanisms. The experimental results demonstrate that even a small proportion of poisoned data can significantly degrade model performance, as observed through increased RMSE and reduced R^2 scores.

Among the evaluated attacks, label flipping and outlier injection were found to be effective in misleading the model. To counter these effects, multiple defense techniques were implemented and compared. The results indicate that IQR filtering provided the best performance improvement by effectively removing extreme outliers, while RANSAC regression demonstrated strong robustness without discarding data points. In contrast, Isolation Forest showed inconsistent performance due to excessive filtering of valid samples.

Overall, this study highlights the importance of incorporating robust defense mechanisms in machine learning pipelines to ensure reliability in adversarial environments. Future work can extend this research to classification models and deep learning systems, as well as explore more advanced adaptive defense strategies.

References

[1] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *Proc. IEEE Symposium on Security and Privacy (S&P)*, San Francisco, CA, USA, 2018, pp. 19–35. https://doi.org/10.1109/sp.2018.00057

[2] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proc. 29th International Conference on Machine Learning (ICML)*, Edinburgh, Scotland, 2012, pp. 1807–1814.

[3] L. Muñoz-González, B. Biggio, A. Demonte's, A. Paudie, V. Congressmen, E. C. Lupu, and F. Roli, "Towards poisoning of deep learning algorithms with back gradient optimization," in *Proc. ACM Workshop on Artificial Intelligence and Security (Asie)*, Dallas, TX, USA, 2017, pp. 27–38.

[4] J. Steinhardt, P. W. Koh, and P. Liang, "Certified defenses for data poisoning attacks," in *Proc. Advances in Neural Information Processing Systems (Nauris)*, Long Beach, CA, USA, 2017, pp. 3517–3529.

[5] N. Müller, D. Kowatsch, and K. Böttinger, "Data poisoning attacks on regression learning and corresponding defenses," in *Proc. IEEE Pacific Rim International Symposium on Dependable Computing (PRDC)*, Perth, Australia, 2020, pp. 80–89.

[6] Y. Ma, X. Zhu, and J. Hsu, "Data poisoning against differentially-private learners: Attacks and defenses," in *Proc. 28th International Joint Conference on Artificial Intelligence (IJCAI)*, Macao, China, 2019, pp. 4732–4738. https://doi.org/10.48550/arxiv.1903.09860

[7] J. Chen, X. Zhang, R. Zhang, C. Wang, and L. Liu, "De-Pois: An attack-agnostic defense against data poisoning attacks," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 16, pp. 3412–3425, 2021.

[8] N. Baracaldo, B. Chen, H. Ludwig, and J. A. Safavi, "Mitigating poisoning attacks on machine learning models: A data provenance-based approach," in *Proc. 10th ACM Workshop on Artificial Intelligence and Security*, Dallas, TX, USA, 2017, pp. 103–110. https://doi.org/10.1145/3128572.3140450

[9] F. Wang, X. Wang, Y. Hong, R. T. Rockefeller, and X. Ban, "Data poisoning attacks on traffic state estimation and prediction," *Transportation Research Part C: Emerging Technologies*, vol. 168, p. 104577, 2024. https://doi.org/10.1016/j.trc.2024.104577

[10] Y. Zhu, H. Wen, R. Zhao, Y. Jiang, Q. Liu, and P. Zhang, "Research on data poisoning attack against smart grid cyber–physical system based on edge computing," *Sensors*, vol. 23, no. 9, p. 4509, 2023. https://doi.org/10.3390/s23094509

[11] H. I. Kure, P. Sarkar, A. B. Ndamase, and A. O. Najrana, "Detecting and preventing data poisoning attacks on AI models," *arrive preprint*, arXiv:2503.09302, 2025. https://doi.org/10.48550/arxiv.2503.09302

[12] A. Paracha, et al., "Deep behavioral analysis of machine learning algorithms under adversarial conditions," *International Journal of Information Security*, 2025.

[13] Y. Mao, et al., "Decentralized optimization resilient against local data poisoning," *IEEE Transactions on Automatic Control (TAC)*, 2025.

[14] Z. Zheng, et al., "Defending data poisoning attacks in differentially private based crowdsensing," *IEEE Transactions on Mobile Computing (TMC)*, 2025.

[15] S. S. Ullah, et al., "Mitigating content poisoning attacks in machine learning systems," *Artificial Intelligence Review*, 2025.

[16] M. Alad, et al., "Preventing data poisoning attacks by using generative models," in *Proc. IEEE Signal Processing and Communications Applications Conference (UBMYK)*, 2019.

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

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[19] H. Xiao, B. Biggio, G. Brown, G. Fumera, C. Eckert, and F. Roli, "Is feature selection secure against training data poisoning?" in *Proc. 32nd International Conference on Machine Learning (ICML)*, Lille, France, 2015, pp. 1689–1698.

[20] M. A. Ma, et al., "Machine learning for warfarin dose prediction: A systematic review," *PLOS ONE*, 2018.

[21] International Warfarin Pharmacogenetics Consortium (IWPC), "Estimation of the warfarin dose with clinical and pharmacogenetic data," *New England Journal of Medicine*, vol. 360, no. 8, pp. 753–764, 2009.

[22] M. A. Sharabi ani, et al., "Revisiting machine learning-based warfarin dosing in a new clinical setting," *Pharmacogenomics Journal*, vol. 15, no. 6, pp. 509–514, 2015.

[23] "Provably effective detection of data poisoning attacks," *arrive preprint*, Corr, 2025.

[24] "Sponge attack against multi-exit networks," *IEEE Access*, vol. 12, 2024.

[25] A. Paudie, L. Muñoz-González, and E. C. Lupu, "Label sanitization against label flipping poisoning attacks," in *Proc. ECML PKDD Workshop on Machine Learning and Data Mining for Cybersecurity*, Dublin, Ireland, 2018.