International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 02 | Feb - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Detecting Intentional Dataset Poisoning Through Training-Time Gradient Behavior and Internal Data Contradictions

Chahat R. Bakhtar
Diploma
Dept. of Computer Engineering
DR. PDGP, Amravati

Saket R. Bobade
Assistant Professor
Dept. of Computer Engineering
DR. PDGP, Amravati

Sumit M. Dhopte
H.O.D
Dept. of Computer Engineering
DR. PDGP, Amravati

Abstract - Machine learning models are highly dependent on the integrity of training data, making them vulnerable to dataset poisoning and backdoor attacks. Such manipulations can subtly alter model behavior while maintaining normal validation performance, making detection challenging. Existing approaches largely focus on post-training inspection or trigger reconstruction, which may fail to provide early-stage protection. This paper proposes a structured training-time framework for detecting intentional dataset poisoning by monitoring behavioral learning signals during model optimization. The methodology integrates gradient anomaly analysis, internal contradiction detection, signal aggregation, and intent-based inference to evaluate dataset integrity. By distinguishing structured adversarial patterns from natural training instability, the framework provides a probabilistic risk assessment prior to deployment. A conceptual experimental design demonstrates how persistent gradient bias and class-level instability can indicate potential malicious manipulation. The proposed approach introduces an intent-aware detection mechanism that enhances early-stage adversarial resilience in machine learning systems.

Key Words: Dataset Poisoning, Backdoor Attacks, Training-Time Detection, Gradient Monitoring, Intent-Based Inference, Adversarial Machine Learning, Data Integrity Monitoring.

1. INTRODUCTION

Machine learning systems are increasingly deployed in critical domains such as healthcare, finance, autonomous systems, and cybersecurity. The reliability of these systems fundamentally depends on the integrity of the data used during training. However, the open and distributed nature of modern data collection processes exposes machine learning models to adversarial risks, particularly dataset poisoning and backdoor attacks. In such scenarios, an adversary intentionally manipulates a subset of training samples to influence model behavior while maintaining acceptable performance on standard validation data.

Unlike traditional software vulnerabilities, data-level attacks do not require modification of model architecture or deployment environments. Instead, they exploit the learning process itself. Carefully crafted poisoning samples can manipulate gradient updates during optimization, steering the model toward adversarial objectives. Backdoor attacks further complicate detection by embedding hidden trigger behaviors

that remain dormant until specific input patterns are encountered. As a result, conventional evaluation metrics such
as overall accuracy are often insufficient to reveal malicious manipulation.

Existing detection approaches primarily focus on post-training analysis, including trigger reconstruction, activation clustering, or model auditing. While such techniques provide valuable insights, they typically evaluate a fully trained model rather than monitoring behavioral signals during the learning process. Consequently, detection may occur only after model compromise has already taken place.

This paper proposes a structured training-time framework for detecting intentional dataset poisoning by analyzing behavioral learning signals generated during optimization. The methodology integrates gradient monitoring, internal contradiction detection, signal aggregation, and intent-based inference to provide a

probabilistic risk assessment of dataset integrity. By shifting detection from post-training inspection to training-time behavioral analysis, the proposed approach aims to strengthen early-stage resilience against adversarial manipulation.

2. BACKGROUND AND RELATED WORK

Research in adversarial machine learning has established that machine learning systems are vulnerable to manipulation at both training and inference stages [1][2]. Data poisoning attacks specifically target the training phase by injecting carefully crafted samples that influence model optimization. Early work demonstrated how adversarial training samples could compromise classical models such as support vector machines [3], while more recent studies introduced gradient-based poisoning techniques capable of steering model updates at scale [4]. These findings highlight the importance of understanding training dynamics when addressing dataset-level threats.

Backdoor attacks represent a more covert form of poisoning in which malicious behavior is embedded into a model without significantly affecting validation accuracy. Studies have shown that hidden trigger patterns can cause targeted misclassification while maintaining overall performance [5]. Detection approaches such as spectral analysis and trigger reconstruction have been proposed to identify such attacks [6][7][8], yet these methods often operate after training has concluded, limiting early-stage protection.
The role of gradients in adversarial manipulation has been extensively examined in optimization-based poisoning research. Back-gradient techniques demonstrate how adversarial objectives can exploit parameter updates during training [9], and influence-based analysis has shown that individual samples can disproportionately affect model predictions [10]. These works indicate that monitoring gradient behavior provides meaningful insight into training-time irregularities.

Finally, distinguishing malicious manipulation from natural dataset noise remains a significant challenge. Research on label noise suggests that random errors typically produce dispersed and inconsistent instability patterns [11], while statistical approaches attempt to identify mislabeled samples within datasets [12]. However, limited work integrates gradient monitoring, contradiction analysis, and intent reasoning into a unified training-time detection framework, motivating the proposed approach.

3. PROBLEM STATEMENT

The reliability of machine learning systems is fundamentally dependent on the integrity of the data used during training. In dataset poisoning attacks, adversaries intentionally modify a portion of the training data to influence model behavior, either by degrading its overall performance or by embedding hidden backdoor functionality. Such manipulations are often subtle and may not significantly affect validation accuracy, making them difficult to detect using conventional evaluation techniques. Additionally, distinguishing between unintentional data noise and deliberate malicious interference remains a complex challenge. Therefore, there is a clear need for a structured training-time analysis framework capable of monitoring learning behavior—particularly gradient dynamics and internal data inconsistencies—to identify potential indicators of intentional dataset poisoning prior to model deployment.

4. PROPOSED METHODOLOGY

Machine learning systems are vulnerable to manipulation at the training stage, as established in early adversarial learning research [1]. Subsequent studies have shown that poisoning attacks can directly influence model optimization processes, including gradient updates during training [3][4]. In addition, backdoor attacks demonstrate that models can behave normally on clean data while containing hidden malicious behavior embedded within the training set [5]. These findings indicate that validation accuracy alone is insufficient to ensure dataset integrity.

Furthermore, prior work has highlighted that malicious manipulation differs from accidental label noise, particularly in terms of consistency and targeted impact [11]. Based on these observations, this work proposes a structured training-time detection framework that monitors gradient behavior and internal data contradictions to infer potential intentional poisoning before model deployment. Rather than reconstructing triggers after training [7], the proposed approach evaluates behavioral learning signals during optimization and produces a dataset-level risk assessment.

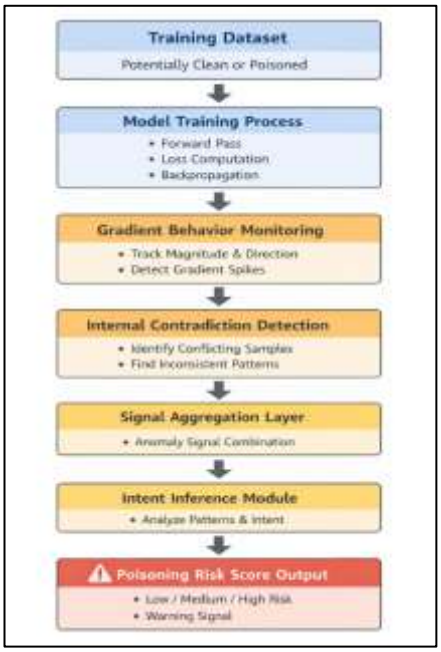

Fig – 1: Proposed Training-Time Dataset Poisoning Detection Framework

4.1 Training Dataset

The proposed framework assumes a supervised learning setting in which a labeled training dataset is provided as input to the learning system. Although the dataset may appear structurally valid, prior research has demonstrated that even a limited number of strategically manipulated samples can significantly influence model behavior during optimization [3][5]. In poisoning and backdoor scenarios, adversarial modifications are often subtle and remain undetectable through simple inspection or validation accuracy alone.

Standard preprocessing operations such as normalization, feature scaling, encoding, or resizing are applied based on the data modality. The framework does not perform early filtering or manual inspection of samples. Instead, it treats the dataset as operational input and defers integrity assessment to behavioral monitoring during the training process. This design ensures that detection relies on learning dynamics rather than direct examination of raw data features.

4.2 Model Training Process

Model training proceeds using conventional optimization procedures, including forward propagation, loss computation, and backward propagation to update parameters through gradient-based learning. During each iteration, gradients determine how the model adapts to the training data. Prior studies have shown that poisoning attacks can intentionally manipulate these gradient updates to steer learning in adversarial directions [4][9]. As a result, the optimization process itself becomes a valuable source of behavioral information.

The proposed framework does not interfere with or modify the training mechanism. Instead, it integrates a passive monitoring layer that records gradient-related signals generated during backpropagation. By embedding the detection mechanism within the training loop, the framework enables real-time observation of learning behavior rather than relying on post-training inspection methods such as trigger reconstruction or model auditing [7].

This stage establishes the operational environment within which behavioral anomaly detection modules are activated in subsequent sections.

4.3 Training-Time Gradient Monitoring

Gradient monitoring refers to the process of observing how model parameters are updated during optimization. In gradient-based learning, each training sample contributes to parameter adjustments through computed gradients derived from the loss function. These gradients determine both the direction and magnitude of learning. Because gradients directly reflect how the model responds to the training data, analyzing their behavior can provide insight into the integrity of the dataset.

Previous research has shown that poisoning attacks exploit this mechanism by carefully crafting training samples that manipulate gradient updates to influence model convergence [3][4][9]. In targeted scenarios such as backdoor attacks, adversarial samples are optimized to consistently steer the model toward specific outputs while maintaining normal validation performance [5]. Therefore, abnormal gradient behavior may reveal hidden malicious influence during the training process.

The proposed framework incorporates a dedicated gradient monitoring module that passively observes optimization signals without altering parameter updates. This module evaluates three structured indicators:

- **Gradient Magnitude Variability** – Large or irregular fluctuations in update magnitude across batches may indicate disproportionate influence from certain samples.

- **Directional Consistency Bias** – Persistent alignment of gradients toward a specific class or output target over multiple epochs may suggest structured adversarial steering.

- **Repeated Influence Contribution** – Samples that consistently generate strong gradients across iterations may indicate targeted manipulation rather than natural variation.

Under clean training conditions, gradient magnitude and direction typically stabilize as the model converges. In contrast, optimization-driven poisoning strategies often introduce repeatable and structured deviations in gradient dynamics [4]. Unlike random annotation noise, which tends to produce inconsistent and non-persistent disturbances [11], intentional poisoning is more likely to generate consistent behavioral patterns over time.

The module computes a quantitative **Gradient Anomaly Score (GAS)** representing deviation from expected learning behavior. This score serves as an early behavioral indicator of potential dataset-level poisoning without attempting to directly identify specific adversarial samples.

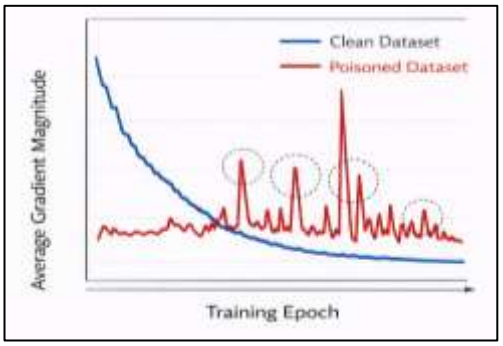

Fig - 2: Conceptual Gradient Behavior Under Clean and Poisoned Training

4.4 Internal Contradiction Detection

While gradient monitoring analyzes how the model updates its parameters, it does not fully capture structural inconsistencies within the dataset. A poisoned dataset may contain samples that appear valid individually but collectively introduce instability or conflicting learning behavior. Research on label noise and dataset corruption indicates that accidental errors tend to be randomly distributed, whereas malicious manipulation often produces structured and targeted inconsistencies [11][12].

The Internal Contradiction Detection module focuses on identifying such structured instability during training. Rather than examining raw input features or performing spectral analysis [6], this module evaluates behavioral inconsistencies that emerge during optimization. It operates by analyzing three indicators:

- **Conflicting Gradient Contributions** – Samples that are semantically similar but consistently push model parameters in opposing directions may indicate embedded contradictions within the dataset.

- **Class-Level Instability Patterns** – If a particular class repeatedly exhibits unstable loss behavior across epochs, it may suggest targeted poisoning toward that class, especially in backdoor-style attacks [5].

- **Persistent High-Loss Samples** – Training samples that repeatedly generate high loss values and resist convergence may indicate adversarial manipulation rather than natural variation.

Under normal conditions, datasets exhibit gradual stabilization in loss distribution across classes. In contrast, targeted poisoning strategies can introduce concentrated instability that persists across training iterations [4]. Unlike random noise, which produces scattered irregularities, structured contradictions tend to cluster around specific labels or influence patterns.

This module computes a quantitative **Contradiction Score (CS)** that reflects the degree of structured instability observed during training. The score does not isolate specific malicious samples but evaluates whether the dataset exhibits behavioral patterns consistent with intentional manipulation.

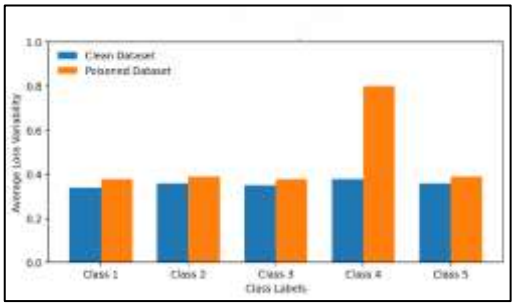

Fig - 3: Conceptual Illustration of Internal Dataset Instability

4.5 Signal Aggregation Layer

The Gradient Anomaly Score (GAS) and the Contradiction Score (CS) represent two independent behavioral indicators extracted during training. While each score provides useful information individually, relying on a single metric may lead to incomplete interpretation. For example, gradient irregularities may arise temporarily during early training phases, and isolated contradictions may occur due to natural class imbalance. Therefore, a combined evaluation mechanism is necessary.

The Signal Aggregation Layer integrates outputs from the gradient monitoring module and the internal contradiction detection module into a unified anomaly representation. Each score is first normalized to a comparable scale to prevent dominance by magnitude differences. The framework then applies weighted combination principles to produce a consolidated behavioral anomaly value.

This aggregation step ensures that detection decisions are not based on isolated signals but rather on consistent interaction between multiple behavioral indicators. For instance, a high gradient anomaly combined with concentrated class instability may indicate stronger evidence of malicious manipulation than either signal independently.

The output of this stage is an **Aggregated Anomaly Index (AAI)**, which serves as structured input to the subsequent intent inference mechanism.

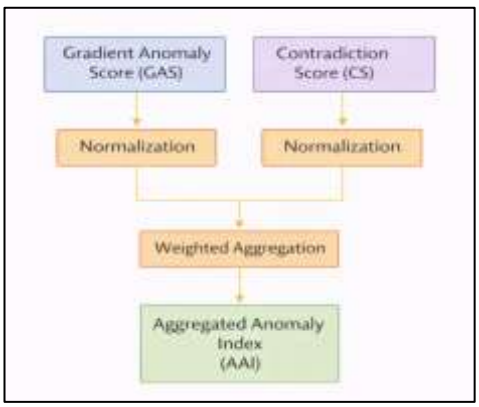

Fig - 4: Signal Aggregation Process

4.6 Intent Inference Mechanism

The primary objective of the proposed framework is not merely to detect anomalous training behavior, but to determine whether such behavior reflects accidental data irregularities or intentional adversarial manipulation. This distinction is critical, as machine learning systems frequently encounter minor instability due to natural class imbalance, noisy annotations, or early-stage optimization dynamics. Therefore, identifying anomalies alone is insufficient; interpreting their intent is essential.

Research in adversarial machine learning has demonstrated that poisoning attacks are typically designed with targeted objectives, such as influencing specific classes or embedding hidden backdoor functionality while maintaining overall performance [4][5]. Unlike random noise, which tends to produce irregular and distributed disturbances, intentional poisoning often introduces structured and repeatable patterns across training iterations [11]. These characteristics form the basis for intent-level reasoning in the proposed framework.

The Intent Inference Mechanism operates on the Aggregated Anomaly Index (AAI) produced by the previous stage. Rather than evaluating isolated anomalies, it analyzes behavioral patterns over time and across classes. This temporal and structural evaluation ensures that inference is based on persistence and concentration rather than momentary deviations.

The mechanism evaluates three core dimensions:

I. Temporal Consistency of Anomalies

If anomaly scores remain elevated across multiple consecutive epochs, this suggests systematic influence rather than random fluctuation. Intentional poisoning strategies are typically optimized to maintain consistent effect throughout training [4].

II. Targeted Concentration of Instability

Backdoor and targeted poisoning attacks are often class-specific, affecting particular labels disproportionately [5]. If behavioral instability is consistently concentrated within a single class or a narrow subset of samples, this increases the likelihood of deliberate manipulation.

III. Repetition of Structured Patterns

Intentional attacks frequently rely on optimization-based objectives that repeatedly reinforce a specific gradient direction or influence pattern [4][9]. Recurrent behavioral

signatures across epochs strengthen evidence of adversarial intent.

By jointly analyzing these dimensions, the system classifies the observed behavior into three levels:

- **Low Intent Confidence** – Patterns resemble natural noise or transient instability.
- **Moderate Intent Confidence** – Structured anomalies are present but lack strong persistence.
- **High Intent Confidence** – Persistent, concentrated, and repeatable patterns consistent with targeted poisoning.

This mechanism represents the conceptual core of the proposed framework. While earlier modules detect behavioral irregularities, this stage interprets those signals within an adversarial reasoning context. It transforms raw anomaly indicators into structured risk intelligence, enabling informed decision-making prior to model deployment.

Importantly, the mechanism does not claim definitive attribution of attack or identification of adversarial actors. Instead, it provides a principled inference layer that distinguishes between benign irregularities and patterns consistent with intentional dataset compromise.

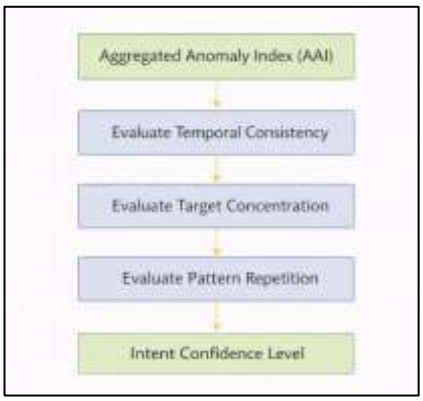

Fig - 5: Intent-Based Decision Mechanism

4.7 Poisoning Risk Score and Alert Layer

The final stage of the proposed framework translates the inferred intent confidence into an operational dataset-level risk assessment. While the previous modules extract behavioral signals and evaluate their structured persistence, this stage converts analytical interpretation into a clear and actionable outcome.

The Poisoning Risk Score is derived from three structured inputs:

- Gradient Anomaly Score (GAS)
- Contradiction Score (CS)
- Intent Confidence Level

Rather than relying on a single indicator, the framework integrates these components into a consolidated evaluation model. The final score reflects both the intensity of behavioral anomalies and the inferred likelihood of intentional manipulation. This layered assessment reduces the probability of false alarms caused by temporary training instability or naturally occurring data noise.

The risk scoring mechanism categorizes the dataset into three operational states:

- **Low Risk** – Minor anomalies consistent with normal training fluctuations.
- **Moderate Risk** – Noticeable structured deviations requiring further inspection.
- **High Risk** – Persistent, concentrated, and repeatable behavioral patterns consistent with intentional dataset poisoning.

This output functions as an early-warning system prior to model deployment. The framework does not automatically terminate training or modify data; instead, it provides interpretable evidence that enables informed decision-making. Such structured reporting aligns with adversarial machine learning principles emphasizing proactive monitoring rather than post-deployment recovery [1][15].

By transforming behavioral monitoring into a risk-oriented evaluation, the framework completes the transition from signal extraction to actionable intelligence. This stage represents the final operational layer of the proposed methodology.

5. EXPERIMENTAL DESIGN

To conceptually evaluate the proposed framework, a controlled comparison between a clean dataset and a deliberately manipulated dataset is considered.

In the clean training scenario, a supervised classification model is trained on an unaltered five-class dataset. During optimization, gradient magnitudes gradually stabilize, directional updates fluctuate naturally without persistent bias, and class-wise loss variability remains balanced across all labels. Consequently, the Gradient Anomaly Score (GAS) and Contradiction Score (CS) remain low. The aggregated anomaly index does not exhibit sustained elevation across epochs, resulting in a Low Intent Confidence and Low Risk classification.

In the manipulated scenario, a small subset of samples associated with one target class is intentionally modified to simulate backdoor-style poisoning [5]. During training, the framework observes repeated directional bias in gradient updates linked to the target class. Additionally, the Internal Contradiction Detection module identifies elevated loss variability concentrated within the same class, indicating structured instability rather than random noise [11]. These behavioral signals cause both GAS and CS to increase.

The Signal Aggregation Layer combines these indicators into an Aggregated Anomaly Index that remains consistently elevated across multiple epochs. The Intent Inference Mechanism evaluates the persistence and concentration of these patterns and assigns a High Intent Confidence level. As a result, the Risk Scoring Layer classifies the dataset as High Risk prior to deployment.

This example demonstrates how the proposed methodology integrates gradient behavior monitoring, contradiction detection, signal aggregation, and intent reasoning into a unified detection process capable of distinguishing structured poisoning from natural training variability.

6. RESEARCH GAP COVERAGE

Extensive research has been conducted on adversarial machine learning, particularly in the areas of data poisoning and backdoor attacks [1][3][5]. Several studies have demonstrated how carefully crafted training samples can manipulate gradient updates and influence model behavior [4][9]. Additionally, backdoor research has shown that models can maintain high validation accuracy while embedding hidden malicious functionality [5].

While detection approaches such as trigger reconstruction and activation analysis have been proposed [7], many existing methods focus primarily on post-training inspection. These techniques often evaluate a fully trained model rather than monitoring behavioral signals during optimization. As a result, detection may occur only after model training is complete.

Furthermore, prior work on label noise emphasizes the challenge of distinguishing between accidental data corruption and structured malicious manipulation [11][12]. However, limited research has integrated gradient dynamics, dataset contradiction patterns, and intent-level reasoning into a unified training-time detection framework.

The proposed methodology addresses this gap by combining three behavioral perspectives: gradient anomaly analysis, internal contradiction detection, and structured intent inference. By embedding monitoring directly within the training process, the framework enables early-stage risk assessment before deployment. This integrated approach extends beyond isolated anomaly detection and introduces an intent-aware evaluation mechanism for dataset integrity.

7. LIMITATIONS AND FUTURE SCOPE

The proposed framework presents a structured training-time approach for detecting potential dataset poisoning through gradient monitoring, contradiction analysis, and intent-based reasoning. By integrating multiple behavioral indicators, the system aims to provide an early-stage risk assessment before model deployment.

However, the methodology is subject to certain limitations:

- It requires access to internal training signals such as gradients and loss behavior, which may not be available in strict black-box settings.
- Highly adaptive adversaries may design poisoning strategies that minimize observable behavioral deviations.
- The framework provides probabilistic risk assessment rather than definitive attribution of malicious intent.

Despite these limitations, the proposed methodology establishes a practical foundation for behavior-based dataset integrity evaluation. It demonstrates that structured monitoring during optimization can enhance early detection capability beyond traditional validation-based approaches.

The future scope of this work includes:

- Developing adaptive threshold calibration mechanisms to reduce false positive detections.
- Incorporating influence-based sample tracing techniques for deeper anomaly investigation [10].
- Extending the framework to distributed and federated learning environments.

8. CONCLUSION

This research proposed a structured training-time framework for detecting potential dataset poisoning in machine learning systems. By integrating gradient behavior monitoring, internal contradiction analysis, and intent-based reasoning, the methodology enables behavioral evaluation of dataset integrity during model optimization rather than relying solely on post-training inspection. The framework emphasizes early detection through structured signal analysis and risk-based interpretation.

The proposed approach demonstrates that persistent gradient anomalies, concentrated class instability, and repeated behavioral patterns can serve as meaningful indicators of intentional manipulation. By transforming these signals into an aggregated anomaly index and intent confidence level, the system provides a practical dataset-level risk assessment mechanism prior to deployment.

Overall, this work contributes a unified, behavior-driven methodology for identifying structured dataset manipulation and establishes a foundation for further exploration in adversarial machine learning defense strategies.

9. REFERENCES

1. B. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine Learning*, vol. 81, no. 2, pp. 121–148, 2010.

2. B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, arXiv preprint arXiv:1712.03141v2, 2018.

3. B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proceedings of the 29th International Conference on Machine Learning (ICML),* arXiv preprint arXiv:1206.6389v3, 2012.

4. J. Geiping, L. Fowl, W. R. Huang, M. Czaja, G. Taylor, M. Moeller, and T. Goldstein, "Witches' brew: Industrial scale data poisoning via gradient matching," in *International Conference on Learning Representations (ICLR)*, 2020.

5. T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying vulnerabilities in the machine learning model supply chain," arXiv preprint arXiv:1708.06733, 2017.

6. B. Tran, J. Li, and A. Madry, "Spectral signatures in backdoor attacks," in *Advances in Neural Information Processing Systems (NeurIPS)*, arXiv preprint arXiv: 1811.00636v1, 2018.

7. B. Wang *et al*., "Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks," *2019 IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, 2019, pp. 707-723, doi: 10.1109/SP.2019.00031.

8. B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Molloy, B. Srivastava, "Detecting backdoor attacks on deep neural networks by activation clustering," in *AAAI Conference on Artificial Intelligence*, 2018.

9. 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," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, arXiv preprint arXiv: 1708.08689v1, 2017.

10. P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

11. B. Frenay and M. Verleysen, "Classification in the Presence of Label Noise: A Survey," in *IEEE*

Transactions on Neural Networks and Learning Systems, vol. 25, no. 5, pp. 845-869, May 2014, doi: 10.1109/TNNLS.2013.2292894.

12. C. G. Northcutt, L. Jiang, and I. L. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *Journal of Artificial Intelligence Research*, 2021.

13. A. Saha, A. Subramanya, and H. Pirsiavash, "Hidden Trigger Backdoor Attacks", *AAAI*, vol. 34, no. 07, pp. 11957-11965, Apr. 2020.

14. K. Doan, Y. Lao, and P. Li, "LIRA: Learnable, Imperceptible and Robust Backdoor Attacks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

15. M. Goldblum *et al.*, "Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1563-1580, 1 Feb. 2023, doi: 10.1109/TPAMI.2022.3162397.