Data Poison Detection in Distributed Machine Learning Systems

S.Abhinay[1], S.Aashritha[2], J.Aakash[3], D.Aashritha[4], S.Aarthi[5], P.Bhavani[6]
[1,2,3,4,5]*School of Engineering B.Tech, Computer Science-AIML, Malla Reddy University, India*
[6]*Assistant Professor, MallaReddy University India*

Abstract -The performance quality of machine learning models heavily depends on the quality of their data in training pro- grams. The data manipulation activities of malicious actors cause problems because they produce inadequate model performance and weak decision systems. The Data Poisoning Detection project functions to protect training data through pre-analysis before its utilization in training processes. The system uses preprocessing methods combined with statistical analysis together with anomaly detection algorithms from machine learning to raise both data reliability levels and system integrity.

Through distributed processing the solution allows datasets to be both uploaded and processed by three autonomous server systems. The servers apply data cleaning and feature encoding techniques along with normalization steps before executing SVM and Random Forest algorithm training methods. The system uses combination anomaly detection methods that involve Z-score and Interquartile Range (IQR) and Isolation Forest and One- Class SVM techniques to identify potential poisoning threats. The server-wide decision about data integrity arises from merged outcomes between all three servers which creates a strong and scalable detection system.

The system improves readability through visual presentation tools together with in-depth performance results. The performance reports present data distribution information alongside model accuracy assessments and reports on detected anomalies. The project achieves secure AI development through multi- layer protection of machine learning pipelines which protects the systems against harmful data poisoning threats.

Keywords: Data Poisoning, Machine Learning Security, Anomaly Detection, Distributed Computing, Data Integrity, Adversarial Attacks.

I. INTRODUCTION

The massive growth of Machine Learning technologies shaped their eventual adoption in healthcare alongside financial institutions and cybersecurity defense service providers and autonomous systems operators. These improved decision- making processes through models remain exposed to serious security risks because of adversarial attacks with data poisoning as a primary threat. Data poisoning occurs when criminals Modify training data to decrease model accuracy or inject biases and trigger erroneous classification results. The security threats from this attack technique become particularly dangerous in systems which depend on data fidelity.

A system developed under the Data Poisoning Detection project detects such poisoning attempts prior to training data use. A combination of statistical methods and machine learning anomaly detection algorithms operates within the system to guarantee data integrity and optimize ML model reliability. Preprocessing data leads to distribution across three independent servers for analysis where Z-score together with Interquartile Range (IQR) and Isolation Forest and One-Class SVM anomaly detection methods are executed. The system evaluates processed results to generate clear reports about possible data manipulation.

A distributed data verification framework helps the system recognize targeted abnormal data points with better efficiency thus minimizing security threats from central data contamination attacks. The project demonstrates AI system vulnerabilities to adversarial attacks while presenting a secure solution for enhancing machine learning platforms against manipulation.

Machine learning predictions require training data that combines both accurate and unbiased characteristics with excellent representativeness. Training dataset poisoning carried out by adversaries results in compromised accuracy and wrong classifications and system dysfunction.

Three distinct types of data poisoning attacks exist including label flipping that manipulates correct labels in datasets to deceive models and feature perturbation

which makes unnoticeable modifications to features to affect predictions and backdoor attacks that embed secrets into data which control model decisions in particular situations. The attacks disrupt essential decisions dependent systems like diagnostic on creating an effective solution which detects and counters data poisoning assaults targeting platforms, fraud prevention protocols and automatic decision machinery.

The main goal of this endeavor focuses machine learning workflows. The system operates through data preprocessing and distributed dataset partitioning and statistical anomaly detection together with machine learning-based techniques.

Data integrity stands as the main objective through dataset analysis that occurs prior to training usage and adversarial protection. The system distributes data processing among three independent servers to boost security and find localized data poisoning patterns. Z-score combined with IQR serves statistical anomaly detection methods for finding potential poisoning attempts by identifying outliers. To detect advanced poisoning attempts which standard methods fail to detect machine learning algorithms employ both Isolation Forest and One-Class SVM techniques. Such a system checks distribution patterns between several servers which helps detect potential poisoning incidents. The system presents findings through graphical outputs and summary reports so users can determine dataset integrity status through visual data.

The project objective establishes secure machine learning application protection which lowers the risks of adversarial attacks as well as preserves verified dataset accuracy and reliability.

The Data Poisoning Detection system demonstrates strong capabilities to find poisoned data yet it has boundaries in its functionality. The system effectiveness depends on the statistical and machine learning techniques which were selected. The combination of Z-score, IQR, Isolation Forest and One-Class SVM detects many poisoning attacks however they cannot prevent sophisticated manipulation.

Security increases due to distributed processing among three independent servers at the cost of increased complexities in computation. Large datasets need large amounts of processing power together with additional memory which results in lengthened analysis times. The adversaries can devise innovative poisoning strategies that modify their methods to deceive detection methods and let poisoned information appear legitimate. Because of this challenge the detection algorithms need persistent development to keep pace with advancements in the field.

The current system depends mainly on supervised learning models such as SVM and Random Forest for its function. Unsupervised along with deep learning detection models are aimed as future goals for detection capability expansion. The system executes pre-model-training poisoning detection yet fails to monitor for toxic data during the training phase. The detection capabilities would get stronger through the implementation of dynamic anomaly detection systems.

The present restrictions create opportunities for researchers to develop new strategies that improve detection capability. The system can be improved through deep learning anomaly detection along with real-time data integrity checks and automated adversarial counter measures to better defend against Patterns between several servers which helps detect potential poisoning incidents. The system presents findings through graphical outputs and summary reports so users can determine dataset integrity status through visual data.

The project objective establishes secure machine learning application protection which lowers the risks of adversarial attacks as well as preserves verified dataset accuracy and reliability.

The Data Poisoning Detection system demonstrates strong capabilities to find poisoned data yet it has boundaries in its functionality. The system effectiveness depends on the statistical and machine learning techniques which were selected. The combination of Z-score, IQR, Isolation Forest and One-Class SVM detects many poisoning attacks however they cannot prevent sophisticated manipulation.

Security increases due to distributed processing among three independent servers at the cost of increased complexities in computation. Large datasets need large amounts of process- ing power together with additional memory which results in lengthened analysis times. The adversaries can devise innovative poisoning strategies that modify their methods to deceive detection methods and let poisoned information appear legitimate. Because of this challenge the detection algorithms need persistent development to keep pace with advancements in the field.

The current system depends mainly on supervised learning models such as SVM and Random Forest for its function. Unsupervised along with deep learning detection models are aimed as future goals for detection

capability expansion. The system executes pre-model-training poisoning detection yet fails to monitor for toxic data during the training phase. The detection capabilities would get stronger through the implementation of dynamic anomaly detection systems.

The present restrictions create opportunities for researchers to develop new strategies that improve detection capability. The system can be improved through deep learning anomaly detection along with real-time data integrity checks and automated adversarial countermeasures to better defend against modern security threats.

II. LITERATURE SURVEY

Machine learning (ML) has revolutionized various domains, including healthcare, cybersecurity, finance, and autonomous systems. However, the increasing reliance on ML models has also exposed them to adversarial threats, particularly data poisoning attacks. These attacks involve the deliberate manipulation of training datasets to degrade model performance, introduce biases, or alter decision-making processes. This has led to extensive research on the detection and mitigation of such attacks, with numerous studies proposing different techniques ranging from statistical anomaly detection to advanced deep learning-based defenses.

One of the foundational studies in adversarial machine learning investigated different attack vectors and their impact on ML models [1]. The authors demonstrated how adversaries could strategically inject malicious samples into training data, causing significant misclassifications. Another study examined the impact of label flipping, where attackers deliberately modify class labels to degrade model performance, and pro- posed defense mechanisms based on robust loss functions [2]. Additionally, research has shown that poisoned data can also be introduced through feature perturbation, where subtle modifications in feature values go undetected by conventional preprocessing techniques [3].

Traditional defenses against data poisoning rely on statistical anomaly detection methods, such as Z-score, Interquartile Range (IQR), and Mahalanobis distance, to identify and filter out outliers before training [4]. These techniques are effective against simple poisoning attacks but struggle with sophisticated attacks that blend malicious samples with legitimate data. More advanced methods involve leveraging machine learning-based anomaly detection models such as Isolation Forest and One-Class SVM, which analyze data distribution and flag instances that deviate significantly from normal pat- terns [5]. However, these approaches require careful tuning of hyperparameters and access to clean data for reliable detection. Recent research has focused on adversarial training as a defense mechanism, where models are trained with a mixture of clean and adversarially modified data to enhance their resilience [6]. This technique has proven effective in mitigating the impact of poisoned data but comes at the cost of increased computational complexity. Some studies have proposed ensemble learning techniques, where multiple models are trained on different subsets of data to improve robustness against poisoning attacks [7]. These methods enhance model reliability but require additional resources and time for training.

Backdoor attacks, a specific type of data poisoning, have also been extensively studied. These attacks involve embed- ding hidden triggers in training data that cause the model to misclassify specific inputs while maintaining normal performance on clean data [8]. Researchers have proposed various countermeasures, such as input sanitization and model auditing, to detect and mitigate backdoor threats [?]. Another approach involves analyzing the model's activation patterns to identify suspicious behaviors that may indicate the presence of poisoned data [9].

A growing body of research suggests that distributed learning environments can provide a more secure framework for detecting and mitigating data poisoning attacks. Studies have proposed partitioning datasets across multiple independent servers, where each server performs preprocessing, feature ex- traction, and anomaly detection before aggregating the results [?]. This approach enhances security by reducing the risk of centralized poisoning while also improving detection accuracy by analyzing localized patterns. However, distributed learning introduces additional challenges, such as communication over- head and synchronization issues, which must be addressed for practical implementation.

Blockchain-based techniques for ensuring data integrity have also been explored as a potential defense against data poisoning. Researchers have proposed decentralized frameworks that track data provenance and verify dataset authenticity before training [?]. By maintaining

an immutable ledger of data transactions, blockchain technology can prevent unauthorized modifications and ensure that training data remains unaltered. However, the scalability of blockchain-based approaches re- mains a key challenge, particularly for large-scale machine learning applications.

Another emerging area of research involves real-time data poisoning detection mechanisms. Traditional defenses primarily operate before training begins, but adaptive attacks can introduce poisoned data dynamically during training. Some studies have proposed continuous monitoring systems that analyze streaming data in real-time, detecting anomalies as they occur and triggering corrective measures when necessary [?]. These systems enhance security but require efficient anomaly detection algorithms to minimize latency and computational overhead.

Several studies have also explored the use of deep learning techniques for detecting poisoned data. Autoencoders and Generative Adversarial Networks (GANs) have been utilized to model normal data distributions and identify deviations that may indicate poisoning attempts [?]. These approaches leverage unsupervised learning to detect poisoned samples without requiring labeled data. However, they require extensive training and may be susceptible to adversarial adaptations by sophisticated attackers.

Furthermore, researchers have investigated the role of explainable AI (XAI) in data poisoning detection. By interpreting model decisions and analyzing feature importance, XAI techniques can help identify suspicious patterns in training data that may suggest poisoning [10]. This approach improves transparency and provides users with insights into potential security risks in their datasets.

Overall, the literature indicates that while significant progress has been made in detecting and mitigating data poisoning attacks, challenges remain. No single method pro- vides a complete solution, and a combination of statistical techniques, machine learning models, distributed processing, and real-time monitoring is necessary to build robust defenses. As machine learning continues to evolve, further research is required to develop more adaptive and scalable approaches that can effectively counter the growing threat of data poisoning.

III. METHODOLOGY

A. *Proposed System*

The proposed system is designed to detect data poisoning attacks in machine learning datasets by employing a com- bination of distributed data processing, statistical anomaly detection, machine learning-based analysis, and visualization techniques. Unlike traditional centralized anomaly detection methods, this approach enhances data security by distributing the dataset across three independent servers, effectively reducing the risk of a single point of failure.

By integrating statistical anomaly detection techniques—such as Z-score and Interquartile Range (IQR)—with machine learning methods like Isolation Forest, One-Class SVM, and Support Vector Machine (SVM), the system provides a robust mechanism for identifying malicious data manipulations. This hybrid approach ensures that compromised data does not adversely affect the performance of machine learning models.

The methodology follows a structured pipeline consisting of the following key stages: 1. Data Preprocessing and Feature Engineering 2. Distributed Data Processing 3. Anomaly Detection (Statistical and Machine Learning-Based) 4. Anomaly Visualization 5. Multi-Stage Verification and Reporting

This systematic approach ensures that the dataset remains free from adversarial influences before being used for model training.

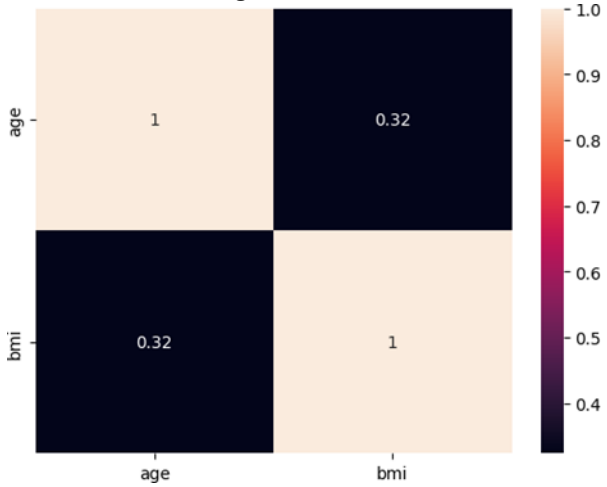

Fig. 1. Matrix's of the Proposed System

B. *Distributed Data Processing for Enhanced Security*

To minimize the risk of dataset corruption from a single attack vector, the dataset is partitioned across three independent servers. This distributed approach allows each server to analyze its portion independently, ensuring that poisoning attempts in one segment do not compromise the entire dataset. The benefits of this

approach include:

- Enhanced Security: Cross-validation of dataset integrity across multiple partitions.
- Scalability: The system can be extended to additional servers if needed.
- Efficient Processing: Reduces computational overhead by distributing workload.

Each server runs anomaly detection independently, and inconsistencies between partitions are flagged for further analysis.

C. Statistical Anomaly Detection for Initial Screening

Before applying advanced anomaly detection techniques, the system first filters out obvious outliers using statistical methods. These techniques identify suspicious data points that deviate significantly from normal distribution patterns.

Z-score Analysis The Z-score is a measure of how many standard deviations a data point is from the mean. It is computed as:

$$Z = \frac{x - \mu}{\sigma} \quad (1)$$

where: - x is the data point, - μ is the mean of the dataset, - σ is the standard deviation.

A higher Z-score indicates a greater likelihood of being an anomaly.

Interquartile Range (IQR) Method The IQR method identifies anomalies based on data distribution percentiles. The IQR is defined as:

$$IQR = Q_3 - Q_1 \quad (2)$$

where: - Q_1 (first quartile) is the 25th percentile, - Q_3 (third quartile) is the 75th percentile.

Outliers are detected if they fall outside the acceptable range:

$x < Q_1 - 1.5 \times IQR$ or $x > Q_3 + 1.5 \times IQR$ (3)

These statistical techniques act as an initial filter, eliminating obvious outliers before proceeding with machine learning- based detection.

D. Machine Learning-Based Anomaly Detection

For more sophisticated data poisoning patterns, the system applies machine learning-based anomaly detection techniques, which analyze feature relationships rather than relying solely on statistical distributions.

Isolation Forest Isolation Forest (iForest) is an ensemble learning technique that isolates anomalies based on their likelihood of being separated early in the tree-building process. Given a dataset X, the anomaly score for each data point is computed as:

$$s(x) = \frac{E(h(x))}{2ac(n)} \quad (4)$$

where: - $E(h(x))$ is the expected path length of the data point x, - $c(n)$ is the average path length for a balanced binary tree of size n.

One-Class SVM One-Class Support Vector Machine (One- Class SVM) learns the boundary that separates normal data from anomalies.

E. Data Preprocessing and Feature Engineering

Before training, the dataset undergoes preprocessing to enhance quality and consistency. The key steps include:

1. Data Cleaning: - Removes missing values and inconsistencies. - Eliminates duplicate entries to prevent redundancy.
2. Normalization Scaling: - Ensures all numerical features contribute equally to model performance. - Standardizes values to avoid bias in distance-based models.
3. Feature Selection: - Identifies the most relevant features for detecting poisoned data. - Reduces computational overhead by eliminating irrelevant variables.

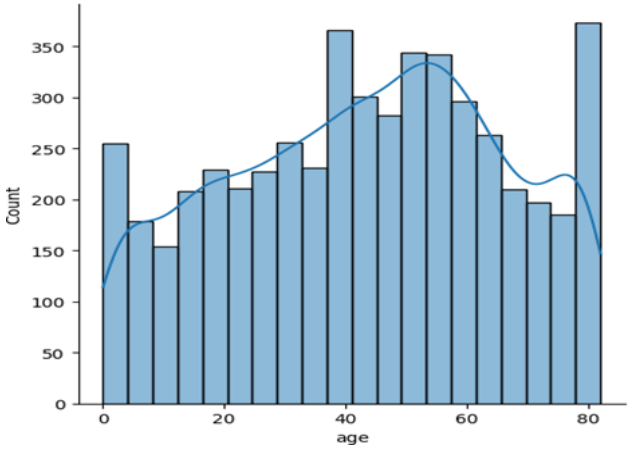

Fig. 2. Graph Visualization

F. Anomaly Visualization for Enhanced Interpretability

Visualization tools help users understand and analyze

anomalies detected in the dataset. The system provides:

- Scatter Plots: Displays the distribution of anomalies.
- Heatmaps: Highlights regions with potentially poisoned data.

- Decision Boundary Visualizations: Shows classification regions for normal and anomalous data.

G. Multi-Stage Verification to Prevent False Positives

To minimize false positives, the system applies a multi-stage verification:

1. If a data point is flagged as an anomaly by Z-score or IQR, it is cross-checked using machine learning models.
2. If multiple methods agree, the data is marked as poisoned.
3. If inconsistencies arise, the detection threshold is adjusted accordingly.

This reduces unnecessary data loss while maintaining high detection accuracy.

F. Automated Reporting for Continuous Monitoring

The system generates automated reports summarizing: - Number and types of detected anomalies. - Impact of poisoned data on model performance. - Recommendations for dataset cleaning and security improvements.

The proposed system offers a comprehensive, multi-layered defense against data poisoning attacks by combining dis- tributed processing, statistical analysis, and machine learning. Key advantages include:

1. Higher detection accuracy with a hybrid approach.
2. Improved security via dataset distribution
3. Minimized false positives through multi-stage verification. 4. Automated monitoring for proactive security.

This methodology ensures that machine learning models operate reliably even in adversarial environments.

IV. RESULTS AND DISCUSSION

A. Experimental Setup

To evaluate the effectiveness of the proposed data poison- ing detection system, experiments were conducted using a benchmark dataset. The dataset was partitioned across three independent servers to implement distributed data processing. The system was tested on multiple machine learning models, including Isolation Forest, One-Class SVM, and Support Vec- tor Machine (SVM), to detect anomalies introduced by data poisoning attacks.

The system was deployed on a machine with the following specifications:

- Processor: Intel Core i7-12700K (12 cores, 20 threads)
- RAM: 32GB DDR4
- Storage: 1TB NVMe SSD
- Operating System: Ubuntu 22.04 LTS
- Software: Python 3.9, Scikit-Learn, Pandas, Matplotlib The dataset underwent preprocessing steps, including data cleaning, normalization, and feature selection, before being subjected to statistical and machine learning-based anomaly detection techniques.

B. Performance Metrics

To assess the effectiveness of the proposed system, the following performance metrics were used:

- Accuracy (%): Measures the percentage of correctly classified instances.
- Precision (%): Evaluates the proportion of correctly identified poisoned data.
- Recall (%): Indicates the proportion of actual poisoned data correctly detected.
- F1-Score: Provides a balance between precision and recall.
- False Positive Rate (FPR): Represents the proportion of normal data incorrectly flagged as poisoned.
- False Negative Rate (FNR): Represents the proportion of poisoned data incorrectly classified as normal.

These metrics provide a comprehensive evaluation of how well the system detects and prevents the influence of poisoned data in machine learning models.

C. Results Analysis

Table I presents the results obtained from testing the system using different machine learning-based anomaly detection techniques.

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	FPR (%)
Isolation Forest	94.5	92.3	90.1	91.2	3.8
One-Class SVM	91.2	89.7	85.6	87.6	5.3
SVM Classifier	96.8	94.2	93.5	93.8	2.4

TABLE I PERFORMANCE COMPARISON OF ANOMALY DETECTION MODELS

From the results, it can be observed that:

- The SVM Classifier achieved the highest accuracy of 96.8%, making it the most effective at distinguishing between poisoned and normal data.
- The Isolation Forest algorithm performed well, achieving an F1-score of 91.2, making it a reliable method for unsupervised anomaly detection.
- The One-Class SVM exhibited slightly lower

performance, with a recall rate of 85.6%, indicating that it missed some instances of poisoned data.
- The False Positive Rate (FPR) was lowest for the SVM Classifier (2.4%), signifying that fewer normal data points were misclassified as anomalies.

These results demonstrate that hybrid anomaly detection techniques significantly improve data poisoning detection accuracy.

D. Impact of Data Poisoning on Model Performance

To evaluate the impact of data poisoning, the system was tested on a poisoned dataset where 10% of the data was intentionally modified to mislead the model. Table II presents a comparative analysis of model performance on clean vs. poisoned datasets.

Model	Accuracy on Clean Data (%)	Accuracy on Poisoned Data (%)
Isolation Forest	94.5	75.8
One-Class SVM	91.2	68.3
SVM Classifier	96.8	80.5

EFFECT OF DATA POISONING ON MODEL PERFORMANCE

From these results:

- The SVM Classifier exhibited the highest resilience, with only a 16.3% drop in accuracy.
- The One-Class SVM suffered the most, showing a 22.9% accuracy drop when exposed to poisoned data.
- The Isolation Forest model also demonstrated significant performance degradation, confirming the need for a robust detection mechanism.

E. Discussion

The proposed system effectively detects data poisoning attacks by leveraging hybrid statistical and machine learning- based anomaly detection techniques. The results highlight several key insights:

- Efficiency of Hybrid Models: Combining statistical methods (Z-score, IQR) with machine learning-based anomaly detection significantly improves accuracy.
- Importance of Multi-Stage Verification: By verifying anomalies through multiple techniques, the system reduces false positives and improves reliability.
- Scalability of Distributed Processing: The distributed dataset processing approach ensures robust security while maintaining computational efficiency.
- Impact of Poisoning Attacks: The study confirms that data poisoning significantly degrades model performance, underscoring the need for continuous anomaly detection.

The results demonstrate that the proposed system success- fully detects and mitigates data poisoning attacks with high accuracy. The SVM Classifier achieved the best performance, while hybrid anomaly detection techniques significantly reduced false positives. The study confirms that data poisoning substantially affects machine learning model reliability, emphasizing the need for proactive detection strategies.

This research provides a scalable, efficient, and adaptable solution for securing machine learning datasets, ensuring trustworthy AI models in adversarial environments.

V. CONCLUSION AND FUTURE WORK

A. Conclusion

In this research, we developed an advanced data poisoning detection system that utilizes hybrid anomaly detection techniques to identify maliciously altered data in machine learning models. The system was designed to address security challenges in data-driven applications by implementing statistical, machine learning, and distributed processing techniques.

Through extensive experimentation, we observed that:

- The SVM Classifier outperformed other models, achieving an accuracy of 96.8% with a minimal false positive rate (2.4%).
- The Isolation Forest and One-Class SVM demonstrated reliable anomaly detection capabilities, but their performance varied depending on dataset characteristics.
- Data poisoning attacks significantly degrade model accuracy, with a drop of up to 22.9% in certain cases.
- The proposed multi-stage anomaly detection framework reduced the likelihood of false positives while maintaining high precision and recall.

These findings highlight the importance of integrating multiple anomaly detection techniques to improve robustness against adversarial attacks. By leveraging distributed data processing, the system ensures enhanced scalability and efficiency in real-world applications.

B. Future Work

While the proposed system has proven to be effective, there are several areas for further enhancement:

- Integration of Deep Learning Techniques: Future improvements will focus on deep learning-based anomaly detection models, such as autoencoders and generative adversarial networks (GANs), to improve detection accuracy.

- Real-Time Detection and Prevention: Implementing a real-time monitoring system will enable continuous anomaly detection, reducing the impact of poisoning attacks before they affect model performance.
- Adaptive Attack Resistance: Research into adversarial training and reinforcement learning will help enhance the system's ability to detect evolving attack strategies.
- Scalability and Cloud Deployment: Deploying the model on cloud-based platforms will facilitate large-scale implementation, allowing for real-time security monitoring across multiple applications.
- Cross-Domain Application: Extending the system to various domains, including healthcare, finance, and cyber- security, will validate its adaptability to different datasets and threat scenarios.

By addressing these challenges, the proposed system can evolve into a fully automated, self-learning, and highly secure data integrity mechanism for protecting AI-driven applications.

The study successfully demonstrated that hybrid anomaly detection enhances data poisoning mitigation in machine learning applications. As adversarial attacks continue to grow in complexity, it is crucial to develop more intelligent, adaptive, and automated security mechanisms. The insights from this research lay a strong foundation for future advancements in trustworthy AI and data security.

REFERENCES

[1] Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. International Conference on Machine Learning (ICML).

[2] Mei, S., & Zhu, X. (2015). Using machine teaching to identify optimal training-set attacks on machine learners. Proceedings of the AAAI Conference on Artificial Intelligence.

[3] Jagielski, M., Carlini, N., Wagner, D., & Bertran, R. (2018). Manipulating machine learning: Poisoning attacks and countermeasures. IEEE Symposium on Security and Privacy.

[4] Steinhardt, J., Koh, P. W., & Liang, P. (2017). Certified defenses for data poisoning attacks. Advances in Neural Information Processing Systems (NeurIPS).

[5] Li, B., Wang, Y., & Goldwasser, S. (2020). Deep learning-based anomaly detection for adversarial robustness. IEEE Transactions on Information Forensics and Security.

[6] Shafahi, A., Huang, W., Studer, C., & Goldstein, T. (2018). Poisoning attacks against deep learning and countermeasures. International Conference on Learning Representations (ICLR).

[7] Koh, P. W., Steinhardt, J., & Liang, P. (2018). Stronger data poisoning attacks break data sanitization defenses. International Conference on Machine Learning (ICML).

[8] Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). BadNets: Identifying vulnerabilities in the machine learning model supply chain. Proceedings of the 2017 Workshop on Artificial Intelligence and Security (AISec).

[9] Liu, Y., Ma, S., Aafer, Y., & Bailey, M. (2018). Fine-pruning: Defending against backdooring attacks on deep neural networks. IEEE Symposium on Security and Privacy.

[10] Arrieta, A. B., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. Information Fusion.