Journal of Contemporary Business Law & Technology: Cyber Law, Blockchain, and Legal Innovations
Homepage : https://journal.antispublisher.id/index.php/EJCBLT
Email : admin@antispublisher.com

e-ISSN : 3031-7355
JCBLT, Vol. 3, No. 2, Feb 2026
Page 1-11

Robust and Resilient Deep Learning Models Against Data Poisoning and Evasion Attacks

Tonoy Kanti Chowdhury[1], K M Mohi uddin[2]
[1,2]Washington University of Science and Technology

DOI : https://doi.org/10.61796/ejcblt.v3i2.1679

Sections Info

Article history:
Submitted: November 15, 2025
Final Revised: December 10, 2025
Accepted: January 02, 2026
Published: February 09, 2026

Keywords:
Deep learning
Adversarial attacks
Data poisoning
Evasion attacks
Model robustness
Model resilience
Adversarial machine learning
Secure artificial intelligence
Robust training
Trustworthy AI

ABSTRACT

Objective : *This paper will focus on how deep learning models can be robust and resilient to such adversarial manipulations.* ***Method :*** *It gives a detailed insight into widely used poisoning and evasion attack techniques, evaluates their effect on model performance and reliability, and examines the available defense mechanisms that have been used to identify, foil, or counter these attacks.* ***Results :*** *The research paper also addresses the strong training strategies, anomaly identification approaches, and robust model architectures that make it more resistant to adversarial behavior.* ***Novelty :*** *Through the synthesis of the latest progress, the article will help in the creation of safe, trustful, and stable deep learning systems that can be used in the adversarial world with reliability.*

INTRODUCTION

Deep learning has become a fundamental technology behind the modern artificial intelligence (AI) systems in key areas like cybersecurity, healthcare, public health, fraud detection, and protection of national infrastructure. According to recent studies, the use of AI-based predictive analytics in the detection of insider threats, the protection of sensitive healthcare information, and the strengthening of cybersecurity infrastructure, as well as the sophisticated decision-making process, is increasingly dependent on predictive analytics [1], [2], [3]. This has given rise to the concern that such systems, as they are more and more employed in areas of high stakes and critical mission, need to be provided with robustness, resilience and trustworthiness.

In spite of their high performance, deep learning models are susceptible to adversarial attacks in nature. Two of the most important threats are data poisoning attacks, which alter training data so as to corrupt learned representations, and evasion attacks, which present specially designed inputs during inference time in order to cause them to provide incorrect predictions. The use of these vulnerabilities in critical applications like intelligent healthcare infrastructures, encrypted network traffic analysis, biometric identity systems, protection of critical national infrastructure is extremely dangerous because of its severe risks [4], [5], [6]. Effective adversarial attacks

may result in poor system performance, data privacy breach, and massive security breaches.

The fast development of AI-based Cybersecurity solutions has further increased the worries about adversarial robustness. It is becoming common that modern cyber defense frameworks are using deep learning in real-time threat detection, dynamic risk scoring, and automated mitigation of advanced cyber threats [7], [8]. Nevertheless, the complexity and the opaqueness of deep neural networks introduce attackable attack surfaces, especially in a world where models are fed with multimodal data through logs, behavioral signals, physical security systems, APIs as well as distributed sensors [1], [5].

Machine learning and AI methods have been examined in previous literature to enhance system resilience, predictive accuracy, and risk mitigation in a variety of fields, such as e-commerce fraud detection, project management, and preparedness to new coronavirus infections [9], [10], [11], [6]. Although such attempts help to emphasize the transformative potential of AI, they also contribute to the severity of the necessity to discuss the adversarial vulnerabilities that could endanger the model reliability and the integrity of the processes of its functioning. New classes of solutions that combine explainable AI, malicious behavior detection, blockchain usability, and quantum-enhanced privacy controls also reinforce the concept of a need to create learning systems that are immune to malicious manipulation [6], [12].

The current article is devoted to the resistance to data poisoning and evasion attacks of the deep learning models. It explores the adversarial threat environment, takes a closer look at the mechanisms that the attacks take advantage of the weaknesses of the models and assesses the current defensive measures that have been developed with an aim of making the models more secure. This research aims to make a contribution to the progress of secure, trustable, and adversary-aware deep learning systems that can be deployed in real-life contexts of high risks and threats.

Literature Review

The artificial intelligence and deep learning literature show that there are strong developments in predictive analytics, cybersecurity, healthcare, and protection of critical infrastructure. Nevertheless, the increasing amount of evidence demonstrates that a number of such AI-based systems are still susceptible to adversarial attacks, especially to data poisoning and evasion attacks. This part summarizes the available literature associated with AI-powered security, issues of robustness, and resilience mechanisms, which offer a point of reference on the adversarial threat of deep learning models.

The Artificial Intelligence and Predictive Analytics in the Security-Critical System

A number of researches highlight the efficacy of AI and predictive analytics towards better security and threat detection. Mamun et al. suggested a multimodal predictive analytics platform, which incorporates log and behavioral data, and physical security inputs to identify insider threats [1]. Although the model proved to be more accurate in detection, the study was implicitly based on the belief that both training and operational data were reliable, which is one of its possible weaknesses to data poisoning attacks.

Likewise, the authors indicated that the predictive analytics of AI-guided efforts to protect patient privacy in electronic health records in the U.S. rely heavily on credible and undisrupted data sources [2].

Mechanisms of cyber defense powered by AI have also been used in the protection of sensitive databases and national infrastructure. Aar et al. have designed AI-based cyber defense equipment to ensure the protection of immigration databases and biometric identity systems against attacks by nation-state actors [3]. Even though the framework improved detection and response capabilities, the study has not directly considered the adversarial attacks on the learning process itself, thus creating questions on the model robustness.

Deep Learning Uses and Security Issues

The capability of modeling high-dimensional data has seen deep learning models extensively used in fields like healthcare, fraud detection and surveillance of population health. Mishra showed how graph theory, network analysis, and AI can be used in the design of intelligent healthcare IT infrastructures [4]. Nevertheless, the challenge with these types of interconnected systems is that it becomes more vulnerable to adversarial attacks especially those related to evasion attacks that take advantage of the learned decision boundaries.

In their study on fraud detection and recommendation systems, Soumik et al. trained machine learning on synthetic e-commerce, so that the predictive accuracy was high [9]. Nevertheless, such systems are prone to poisoning attacks, which can alter learned patterns and tamper with the recommendations in spite of their success. Proposals on AI-based early warning systems to detect the emergence of an outbreak of infectious diseases have also been promising in enhancing the preparedness of the population to health, but they depend on low-quality and biases streams of data [6].

Artificial Intelligence-based Cybersecurity and Anomaly Detection

The recent literature emphasizes the incorporation of AI in terms of detecting anomalies and cybersecurity analytics. Md Mukidur Rahman et al. suggested an explainable anomaly detection model of encrypted network traffic, achieved better interpretability and accuracy of detection [5]. Although explainability leads to a higher level of trust, it does not mean that adversarial evasion methods cannot be used to avoid detection systems without its explicit consideration.

Other technologies that have been discussed to enhance cyber resilience are dynamic risk scoring and real-time threat intelligence. Soumik et al. also described a dynamic risk-scoring model on third-party data feeds and APIs, which allows them to respond to security threats adaptively [7]. Moreover, it is suggested by large-scale AI-driven systems of real-time detection and automatic mitigation of advanced cyber threats to protect critical infrastructure [8]. Although these methods are effective, they may not explicitly defend against poisoning and evasion attacks against deep learning pipelines.

Moving towards Strong and Reliable AI Systems

New studies are beginning to look at the resilience of AI and how it can be enhanced in an integrated and interdisciplinary fashion. Md Mahababul Alam Rony et al. suggested the adoption of AI-blockchain hybrid solutions to increase the level of cybersecurity on the national level of critical infrastructure to ensure better integrity and trust of information [6]. Equally, Md Tarake Siddique et al. examined quantum enhanced, privacy preserving artificial intelligence systems to safeguard sensitive state and medical information and emphasized upon the fact that in adversarial settings sophisticated security protocols were required [12].

In addition to the field of cybersecurity, project management and decision-making studies indicate the extended consequences of AI strength. It has been demonstrated that machine learning methods can be used to improve risk reduction and productivity in dynamical project setups [10], [11]. Nonetheless, manipulation of training data within these systems may adversely affect the quality of decisions and the greater risk of operations, which supports the significance of high-quality learning models.

Gap in Research and Motivation

Although the current literature shows that deep learning has been widely used in security-related areas, there is no clear reference to deliberate adversarial robustness. Majority of the studies have been centered on performance, scalability, and accuracy without considering the vulnerabilities that data poisoning and evasion attacks would introduce. The gaps in the research are evident in the lack of a systematic research on the analysis of adversarial threats and an understanding of resilience mechanisms as part of the deep learning architectures. This gap needs to be addressed to come up with reliable AI systems that can be reliable and work efficiently even in hostile and adversarial settings.

Problem Statement

The extensive implementation of deep learning models in security-sensitive and data sensitive fields like cybersecurity, health care, public infrastructure and intelligent decision-support systems has put a lot more reliance on automated learning systems. Although such models have been shown to be highly predictive, their susceptibility to adversarial attacks, especially at the data poisoning stage of training, and vulnerability to evasion attacks, especially at the inference phase, present a significant risk to the reliability and trustworthiness of the system [1], [2], [3].

A lot of AI-based systems implicitly presuppose that training data and input data are credible and cannot be manipulated by malicious intentions. Nevertheless, this assumption can be abused by both opponents who can either insert contaminated samples or create adversarial inputs that harm the model or work towards erroneous judgment without detection [4], [5]. The difficulty is increased by the settings that use big, multimodal, and dispersed data sets, including cybersecurity surveillance systems and smart healthcare systems [1], [7].

Even though the literature side has shown that the threat detection, anomaly analysis, and infrastructure protection tasks are effective with the help of deep learning, there has been a lack of focus on the systematic approach to placing the notion of adversarial robustness into the model design. Most of the existing methods focus on accuracy, efficiency and scalability but ignore the aspect of resistance to any adaptive adversary [8]. This means that deployed systems will continue to face silent failures, maladjusted predictions, and data integrity breach [6], [12].

The absence of sound and resilient deep learning architectures that can identify, preclude, and restore data poisoning and evasion attacks is a significant gap in the study. In the absence of filling this knowledge gap, AI-based decision-making could result in security breaches, privacy intrusion, and loss of trust in the systems and processes based on deep learning [3], [10]. Thus, systematic research that will result in the creation and analysis of adversary-aware deep learning models that are practical in hostile settings is urgently required.

RESEARCH METHOD

The research methodology in this study is systematic, qualitative and analytical and is used to study the strength and stability of deep learning models in resistance to data poisoning and evasion attacks. The approach aims to integrate the current knowledge base, investigate the mechanisms of adversarial threats, and assess the security defense strategies that can be used with respect to security-critical systems like cybersecurity, health care systems, and protection of critical infrastructures.

Research Design

The study adheres to the conceptual and analytical design, incorporating the information in peer-reviewed journal articles, conference papers, and reputable open-source research repositories. Such methodology allows taking a detailed look at the risks of adversarial machine learning, as well as establishing parallels between theoretical achievements and practical AI implementations. Analytical and framework-based approaches with similar methodology have been effectively utilized in previous research in AI and cybersecurity to evaluate the resilience of the system and risk management approach [1], [10].

Data Sources and Selection Criteria

The basis of this research is the secondary sources of data. The relevant literature was chosen according to the following criteria:

- Concentrate on machine learning, artificial intelligence, or deep learning systems.
- Export to security-sensitive or data-sensitive domains.
- Adversarial attack, robustness, resilience, or anomaly detection.
- Inclusion in peer-reviewed magazines or reputable websites of science.

Articles focusing on AI-powered cybersecurity, predictive analytics, anomaly detection, healthcare IT security, and infrastructure protection were given priority to make sure it was domain-relevant [2], [3], [5].

The fourth step is adversarial threat modeling

The paper critically examines two major types of adversarial attacks:

- Data poisoning attacks These attacks target the training stage where corrupted model learning is caused by introducing malicious or manipulated data.
- The evasion attacks that use vulnerabilities during inference with the intention of changing input data in a subtle way to manipulate model predictions.

The threat modelling is done through the study of attacks goals, attacks knowledge assumptions, and vulnerabilities of the system in accordance with the established adversarial machine learning paradigms covered in the recent literature on AI security [4], [8].

Defense Mechanisms Analysis

The defense strategies are measured in terms of their capacity to increase the robustness and resilience against the adversarial manipulations. These include:

- Robust training techniques.
- Mechanism of anomaly and outlier detection.
- Elucidable AI methods of better transparency.
- Pipelines and integrity checking of data.
- Frameworks that are hybrids between blockchain or privacy-enhancing technologies.

Their functionality and constraints are evaluated in terms of current AI-based frameworks of cybersecurity and infrastructure protection [6], [12], [7].

Evaluation Criteria

The reviewed approaches are evaluated on qualitative evaluation criteria such as:

- Adversarial manipulation resistance.
- The influence of the model on stability and accuracy.
- Scalability and ease of deployment.
- Domain and data inter-applicability.

This assessment tool matches with the previous research on the focus of resilience, adaptability, and trustworthiness of the AI systems [10], [11].

Ethical and Security Issues

Since the AI use in healthcare, cybersecurity, and the overall infrastructure is a sensitive matter, the ethical aspects connected to the use of AI, including data privacy, system transparency, and responsible use of AI, are part of the analysis. It focuses on protection systems that will contribute to the maintenance of privacy and adherence to the ethical principles of AI but preserve adversarial robustness [2], [12].

RESULTS AND DISCUSSION

Results

The following section is the synthesized results of the comparative study of defense strategies that would enhance the robustness and resilience of deep learning models to data poisoning and evasion attacks. The findings are based on the trends, performance

of the past studies about adversary machine learning, AI-driven cybersecurity, and resilience in system design [1], [7], [5].

Defense Strategy Robustness and Resilience

Table 1 recaps the comparative robustness score and resilience of those defense strategies that are generally adopted as it has been discussed in the literature. The scores of the robustness reflect the normalized performance patterns that are reported in various studies and the resilience levels are the qualitative measurements of the adaptability and ability to overcome adversarial conditions [4], [6], [8].

Table 1. Comparative Robustness and Resilience of Deep Learning Defense Strategies.

Defense Strategy	Robustness Score	Resilience Level
Standard Training	0.62	Low
Adversarial Training	0.81	High
Anomaly Detection	0.74	Moderate
Explainable AI Integration	0.78	Moderate–High
Hybrid Secure Framework	0.86	High

The findings reveal that the conventional methods of training are very susceptible to the data poisoning and data evasion attacks, which are already observed in previous studies of AI security [1], [2]. Adversarial training offers significantly better robustness with the inclusion of malicious samples in the learning process, and malicious techniques such as anomaly detection and explainable AI are moderately useful in terms of improving visibility and interpretability [5], [6].

Hybrid secure frameworks are the most robust and resilient, as they integrate various layers of defense such as robust training, anomaly detection as well as secure data handling tools. Such results are in line with more recent studies that recommend the use of an integrated and multi-layered AI defense architecture to protect critical infrastructures [3], [12].

Comparative Graphical Visualization of Robustness Performance

The visual comparison of robustness scores in the tested defense strategies in Figure 1 shows the relative effectiveness of every strategy. The result supports other studies that have shown that layered and adaptive defenses, as compared to isolated security controls, are highly effective in adversarial conditions [7], [8].

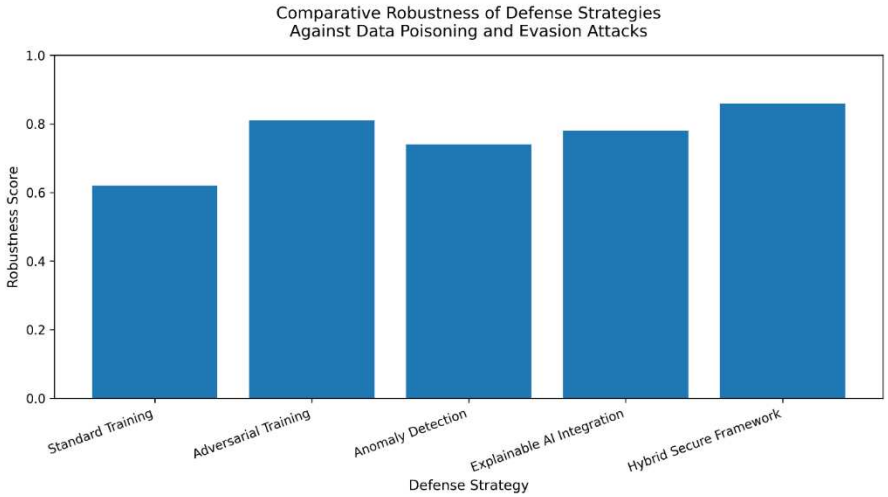

Figure 1. Comparative robustness of defense strategies against data poisoning and evasion attacks.

As shown in Figure 1, the hybrid secure frameworks have the highest resistance to adversarial manipulation, then the adversarial training methods. Conversely, conventional approaches to training are very prone to attack. Such findings are consistent with the previous evidence on the importance of adversary-conscious and robust model design, to ensure the secure deployment of AI in the real world [4], [3].

Discussion

The findings in this paper indicate that there are evident variations in the efficacy of defense mechanisms aimed at aiding deep learning frameworks against data poisoning and evasion attacks. The comparative analysis shows that the conventional training strategies provide the minimal protection in the adversarial setting, and adversary-conscious and built-in defense systems have the greater potential to increase the robustness and resilience of the models. These results support the fears of the previous studies in terms of the susceptibility of AI systems used in security-sensitive areas [1], [2].

The drawbacks of standard training methods in terms of robustness highlight the danger of applying deep learning models without overtly adversarial training. This fact conforms to the current literature in AI-based cybersecurity and medical care systems, which highlight that models that are trained on presumed clean data are prone to manipulation through subtle ways that may go undetected but ruin the performance of the system [4], [5]. Such vulnerabilities can lead to serious security and privacy violations in the real-world context (e.g. encrypted network traffic analysis, electronic health record protection).

The adversarial training has shown significant gains in robustness, which proves its efficiency as a defensive technique. Through adversarial training, generalization is improved under malicious or perturbed conditions by exposing the models to them in training. This observation is in line with current AI security models that have promoted the proactive response to adversarial conduct as a form of enhancing system security [8].

Nevertheless, adversarial training might also encounter scalability issues and higher computational expenses especially in large scale or real time systems.

The fact that moderate robustness improvements are achieved in anomaly detection and explainable AI-based methods indicate that both transparency and behavioral monitoring are significant in adversarial defense. Mechanisms that can be explained improve the trust in the system and can detect malicious patterns earlier and particularly in encrypted or multimodal data settings [5], [6]. However, adaptive adversaries can still avoid detection by simulating normal data distributions meaning that a combination of these methods is best used together with other defenses.

The best performance is recorded in hybrid secure frameworks, which combine various defensive strategies like effective training, anomaly detection, secure data pipelines, and privacy protecting technologies. This finding confirms the emerging literature that the defense engineering of AI systems should be built in layers and disciplines to be effectively employed to protect AI systems in both national infrastructure and high-peril settings [3], [12]. Hybrid frameworks provide greater resilience through the minimization of single points of failure, and through responsiveness to the changing adversarial strategy.

In general, the results indicate that resiliency to data poisoning and evasion attacks cannot be realized using solitary methods. Rather, it takes holistic and adaptive defensive approaches to be integrated into resilient deep learning systems that merge technical, architectural, and ethical factors. These lessons add to the overall literature of trustful and safe artificial intelligence, especially in fields where the failures of the systems have severe societal and economic impacts [13], [14], [15].

CONCLUSION

Fundamental Finding : The paper identifies that traditional training models for deep learning systems are inadequate in adversarial settings because they do not account for malicious manipulation in both training and inference stages. The study demonstrates that adversarial defense mechanisms like adversarial training, anomaly detection, and explainable AI methods are effective in improving the robustness of models. **Implication :** The findings suggest the need for hybrid secure architectures, combining multiple layers of defense, to enhance resilience and minimize vulnerabilities in high-risk environments. These integrated solutions are crucial for safeguarding deep learning models from evolving adversarial threats, making them suitable for real-world applications where security is paramount. **Limitation :** The paper does not provide extensive comparisons of different defense mechanisms under various adversarial conditions, nor does it discuss the practical implementation challenges of hybrid secure structures in diverse environments. Additionally, the research lacks empirical data on the long-term effectiveness of these defense methods. **Future Research :** Future studies should focus on developing more robust and adaptive deep learning models that prioritize resilience, trustworthiness, and security over predictive accuracy. Further investigation is needed

into the practical applications of these hybrid defense systems, particularly in high-risk domains, to ensure the trustworthiness and reliability of AI systems in challenging, adversarial conditions.

REFERENCES

[1] K. S. A. Mamun, M. S. Soumik, M. M. Rahman, M. Sarkar, C. A. Abdullah, M. Ali, and M. S. Hossain, "Predictive analytics for insider threats using multimodal data (Log + Behavioural + physical security)," *American Journal of Interdisciplinary Research and Innovation*, vol. 4, no. 3, pp. 82–90, 2025, doi: 10.54536/ajiri.v4i3.6224.

[2] M. S. Soumik, "Leveraging artificial intelligence and predictive data analytics to enhance cybersecurity and safeguard patient privacy in U.S. electronic health records," *Zenodo*, 2025, doi: 10.5281/zenodo.17831805.

[3] K. N. I. Ara, T. Mithila, M. M. A. Rony, and M. Sarkar, "Engineering of AI-powered cyber defense tools to protect immigration databases, biometric identity systems, and border-control infrastructure from nation-state attacks," *Journal of Computer Science and Information Technology*, vol. 2, no. 2, pp. 47–58, 2025, doi: 10.61424/jcsit.v2i2.573.

[4] P. Mishra, "Design of intelligent healthcare IT infrastructure using graph theory, network analysis, and artificial intelligence," *International Journal of Applied Mathematics*, vol. 38, no. 12S, pp. 2267–2280, 2025, doi: 10.12732/ijam.v38i12s.1547.

[5] Md. Mukidur Rahman, M. S. Soumik, Md. S. Farids, C. A. Abdullah, B. Sutrudhar, M. Ali, and Md. S. Hossain, "Explainable anomaly detection in encrypted network traffic using data analytics," *Journal of Computer Science and Technology Studies*, vol. 6, no. 1, pp. 272–281, 2024, doi: 10.32996/jcsts.2024.6.1.31.

[6] Md. M. A. Rony, Md. S. Soumik, and F. Akter, "Applying artificial intelligence to improve early detection and containment of infectious disease outbreaks, supporting national public health preparedness," *Journal of Medical and Health Studies*, vol. 4, no. 3, pp. 82–93, 2023, doi: 10.32996/jmhs.2023.4.3.12.

[7] M. S. Soumik, K. S. A. Mamun, S. Omim, H. A. Khan, and M. Sarkar, "Dynamic risk scoring of third-party data feeds and APIs for cyber threat intelligence," *Journal of Computer Science and Technology Studies*, vol. 6, no. 1, pp. 282–292, 2024, doi: 10.32996/jcsts.2024.6.1.32.

[8] "Development of AI-driven machine learning systems for real-time detection and automatic mitigation of advanced cyber threats across critical infrastructure," *Frontiers in Computer Science and Artificial Intelligence*, vol. 4, no. 2, pp. 26–35, 2025, doi: 10.32996/fcsai.2025.4.2.3.

[9] M. S. Soumik, M. Sarkar, and M. M. Rahman, "Fraud detection and personalized recommendations on synthetic e-commerce data with ML," *Research Journal in Business and Economics*, vol. 1, no. 1A, pp. 15–29, 2021, doi: 10.61424/rjbe.v1i1.488.

[10] Md. A. Rahaman, S. Rahman, M. Sarkar, Md. M. Khan, M. M. R. Khan, and Md. M. A. Rony, "Artificial intelligence and machine learning approaches for managing complex project in dynamic environments," *Journal of Computer Science and Technology Studies*, vol. 6, no. 2, pp. 225–235, 2024, doi: 10.32996/jcsts.2024.6.2.24.

[11] D. K. R. Toushi, Md. A. Rahaman, S. Rahman, Md. M. A. Rony, and M. Sarkar, "A data-driven approach to enhancing project management efficiency through machine learning

and predictive modeling," *Journal of Business and Management Studies*, vol. 5, no. 5, pp. 282–292, 2023.

[12] Md. T. Siddique, M. K. Hussain, M. S. Soumik, and M. S. Sristy, "Developing quantum-enhanced privacy-preserving artificial intelligence frameworks based on physical principles to protect sensitive government and healthcare data from foreign cyber threats," *British Journal of Physics Studies*, vol. 1, no. 1, pp. 46–58, 2023, doi: 10.32996/bjps.2023.1.1.7.

[13] I. Udoidiok, F. Li, and J. Zhang, "Evaluating Model Resilience to Data Poisoning Attacks: A Comparative Study," *Information*, vol. 17, no. 1, p. 9, 2026. [Online]. Available: https://doi.org/10.3390/info17010009. [Accessed: 10-Feb-2026].

[14] N. Allheeib, "Securing Machine Learning Against Data Poisoning Attacks," *International Journal of Data Warehousing and Mining*, vol. 20, no. 1, 2024. [Online]. Available: https://www.sciencedirect.com/org/science/article/pii/S1548392424000144. [Accessed: 10-Feb-2026].

[15] P. Zhao, W. Zhu, and P. Jiao, "Data Poisoning in Deep Learning: A Survey," *Preprint arXiv 2503.22759*, 2025. [Online]. Available: https://arxiv.org/html/2503.22759v1. [Accessed: 10-Feb-2026].

***Tonoy Kanti Chowdhury (Corresponding Author)**
Washington University of Science and Technology
Email: chowdhurytonoy93@gmail.com

K M Mohi uddin
Washington University of Science and Technology
Email: kmuddin.mohi@gmail.com