BULLETIN of Cherkasy State Technological University

Journal homepage: https://bulletin-chstu.com.ua/en

Vol. 31 No. 1. 2026

UDC 004.0.55:004.451.44
DOI: 10.62660/bcstu/1.2026.128

ISSN: 2306-4412
E-ISSN: 2708-6070

Integration of security testing into QA pipelines using adversarial ML

Anatolii Husakovskyi*
Master, Senior Automation QA Engineer
National Aerospace University "Kharkiv Aviation Institute"
61070, 17 Vadyma Manka Str., Kharkiv, Ukraine
https://orcid.org/0009-0007-9398-0966

Abstract. The study aimed to theoretically systematise approaches to improving the resilience of machine learning systems in cyber defence by integrating resilience testing into the security process. The methodology covered the systematisation of machine learning areas in cyber defence, analysis of strategies to counter adversarial attacks, and a case study of integration into quality assurance and machine learning operations. The study found that the use of machine learning technologies in cyber defence enables the automation of threat detection and response (network anomalies, behavioural analysis, anti-phishing, anti-fraud, malware classification). The main advantages are scalability, response speed, predictability, and effectiveness in complex environments, while the key risks include dependence on data quality, false positives, vulnerability to adversarial and poisoning attacks, as well as privacy and explainability issues. The study determined that adversarial machine learning distinguishes between three attack scenarios (white-box, black-box, grey-box) and their classes (evasion, data poisoning, privacy/inference, model extraction, generative artificial intelligence. The study emphasised that adversarial machine learning encompasses not only technical but also regulatory and ethical dimensions related to the principles of privacy, fairness, and transparency in the use of artificial intelligence. Multi-level protection strategies were presented, integrated into the machine learning model lifecycle at the data level, during training, after training, at the deployment and inference stages. Practical cases demonstrated the feasibility of applying machine learning and anti-money laundering in various domains, from network security and security operations centres to development and operations/continuous integration/continuous delivery, the financial sector, stress testing machine learning pipelines, as well as quality assurance and machine learning operations. The practical significance lies in the ability of cybersecurity specialists, financial analysts, and machine learning operations engineers to use the results to improve the efficiency of security operations centres, integrate adversarial testing, and ensure the stability of machine learning models in production environments

Keywords: model resilience; algorithm vulnerabilities; resilience metrics; automated testing; model lifecycle; protective mechanisms

INTRODUCTION

The integration of Artificial Intelligence (AI) and Machine Learning (ML) technologies into modern 21st-century information systems is transforming approaches to cybersecurity. Traditional security methods, such as cryptographic protection, authentication and access control, firewalls, Intrusion Detection System (IDS)/ Intrusion Prevention System (IPS) systems, and antivirus solutions, focused primarily on protecting software code and infrastructure, while in the new environment, data, algorithms, and models are becoming the main targets of vulnerabilities. These components are targets for adversarial attacks that can lead to failures in malware detection systems, anti-fraud solutions, or industrial environments. At the same time, the spread of Development

Article's History: Received: 09.09.2025; Revised: 30.01.2026; Accepted: 16.03.2026; Published: 08.04.2026

Suggested Citation:
Husakovskyi, A. (2026). Integration of security testing into QA pipelines using adversarial ML. *Bulletin of Cherkasy State Technological University*, 31(1), 128-142. doi: 10.62660/bcstu/1.2026.128.

*Corresponding author

and Operations (DevOps) and Continuous Integration/ Continuous Delivery (CI/CD) approaches in the field of MLnecessitates the integration of security mechanisms directly into development and testing processes.

Existing research has demonstrated significant progress in the automation of ML processes. R. Bagai *et al*. (2024) have highlighted that the implementation of CI/CD on the Amazon Web Services (AWS) platform provides scalability and accelerated deployment of ML models, thereby confirming the technical feasibility of building continuous pipelines in the field of machine learning. The study has provided a basis for analysis for CI/CD as a standard practice in ML Operations (MLOps), emphasising the engineering benefits of automation, flexibility, and process efficiency. At the same time, the study did not cover aspects of protection against adversarial attacks, which highlights a key gap: the established infrastructure promoted speed and scalability, but does not guarantee the robustness and stability of models. A similar trend was confirmed by B.C. Vadde & V.B. Munagandla (2023), who proved that continuous testing in DevOps processes can quickly detect model degradation and code defects. The study emphasised that the inclusion of AI modules in testing not only improves software quality but also creates the conditions for systematic automation of model robustness assessment in dynamic development environments.

A study by N. Bena *et al*. (2024) proposed an approach to certifying the accuracy, privacy, and robustness of malware detection models. The study demonstrated that the effectiveness of such systems cannot be assessed solely by classical performance metrics, as parameters such as resistance to adversarial attacks and data privacy become critical. This directly highlights the need to integrate extended criteria into Quality Assurance (QA) processes and test pipelines, which corresponds to the objectives of your research. Another layer of literature focused on the formalisation and evaluation of AI system robustness. In their study, A. Agarwal & M.J. Nene (2025) conducted a comparative evaluation of robustness toolboxes and showed significant variability in results depending on the metrics selected. The study demonstrated that the lack of standardised metrics, such as robust accuracy or privacy loss, complicates QA decisions regarding the readiness of models for deployment. Therefore, without uniform metrics, adversarial testing remains an isolated research practice rather than part of industrial processes.

J. Guo *et al*. (2023) developed a comprehensive methodology for assessing the robustness of deep models and demonstrated its effectiveness on several benchmarks. The authors' approach demonstrated that systematic tracking of attack success rates creates a basis for formalised quality control in machine learning. This directly linked robustness to QA and confirmed the need to include adversarial tests in the model lifecycle. Equally significant is the contribution of applied research demonstrating the practical consequences of attacks. D.L. Gala *et al*. (2025) analysed the impact of adversarial patch attacks on You Only Look Once (YOLO) models in edge-AI environments. Even small perturbations can significantly reduce object recognition accuracy, which is critical for security systems. The authors concluded that classical testing methods are not capable of detecting such threats, and therefore, adversarial tests should be integrated into QA processes on a par with unit and integration tests. This confirmed that ignoring adversarial threats in practical scenarios can lead to dangerous vulnerabilities in production systems.

In Ukraine, the integration of AI and ML into cyber defence systems is also becoming increasingly relevant due to the rise in attacks on critical infrastructure. Research by O. Neretin & V. Kharchenko (2022) has shown that AI models are becoming new targets for vulnerabilities, from poisoning to privacy attacks, which threaten the stability of digital services and require protection at all stages of the life cycle. In turn, D. Kotenko & Y. Khlaponin (2024) proved that AI is capable of effectively detecting threats in real time, but remains vulnerable to adversarial influences, which requires the integration of specialised tests. Collectively, Ukrainian studies have confirmed that algorithms and models are key risk objects, and the implementation of security in DevOps/CI/CD is a strategic condition for protecting national cyber infrastructure.

Scientific discourse demonstrates patterns: engineering research has confirmed the possibility of CI/CD for ML, but has not ensured that Anti-Money Laundering (AML) risks were addressed; methodological work has proven the need to formalise robustness metrics, without which adversarial tests remain incomplete; and applied experiments have shown that real adversarial attacks cause critical failures in systems, thus arguing for the need to include such tests in QA processes. Together, this revealed a gap between academic research and industrial practices: existing solutions either ignore AML risks or remain fragmented and do not integrate into QA/CI-CD pipelines.

Therefore, the study aimed to provide a theoretical justification for approaches to improving the resilience of ML systems through the integration of adversarial testing into QA and CI/CD processes. To achieve this goal, the following tasks were performed: systematise the vulnerabilities of AI/ML models and classify the main types of adversarial attacks, evaluate existing countermeasures and testing methods (adversarial training, randomised smoothing, formal verification, data preprocessing), identifying their advantages and limitations for different stages of the ML model life cycle, and to formulate recommendations for implementing AML protection in QA/CI-CD pipelines and stress-testing practices for model validation in industrial environments.

MATERIALS AND METHODS

The study was analytical and theoretical in nature and covered the application of ML technologies in the field of cyber security to summarise the areas of ML implementation in the processes of detecting and countering threats, in particular adversarial and poisoning attacks, model robustness issues and data drift risks that directly affect the effectiveness of security systems.

The study consisted of three stages. At the first stage, the areas of ML application in cyber defence were summarised using systematisation and comparative analysis: Network Detection and Response (NDR)/Network Intrusion Detection System (NIDS), User and Entity Behaviour Analytics (UEBA), Endpoint Detection and Response (EDR)/Security Orchestration, Automation, and Response (SOAR), anti-phishing and content analysis, malware classification, anti-fraud, Threat Intelligence and threat correlation, Industrial Control Systems (ICS)/Supervisory Control and Data Acquisition (SCADA) and OT protection, cloud security and access control, cyber risk prediction, QA/DevOps integration, stress testing of ML pipelines, adversarial testing of Computer Vision (CV)/Natural Language Processing (NLP) models. The task of this stage was to compare the advantages and limitations of various ML approaches and to summarise the opportunities and risks of their integration into security frameworks. Access scenarios (white/grey/black) and classes of AML attacks (evasion, poisoning, inference, extraction, GenAI) were presented in order to provide a methodological basis for integrating adversarial testing into QA/CI-CD with the aim of increasing the resilience of ML systems. In addition, international standards from the National Institute of Standards and Technology (2023), ISO/IEC 23894:2023 (2023), the official analytical report on cybersecurity from the European Union Agency for Cybersecurity (ENISA) (Polemi & Praça, 2023), as well as corporate frameworks from Microsoft Security (n.d.) and OpenAI (2024) to demonstrate that AML exceeds the scope of a technical issue and requires the integration of regulatory and ethical approaches into QA and MLOps to ensure trust in AI systems and compliance with standards.

In the second stage, the method of systematisation was used to examine protection against adversarial attacks in the context of the entire life cycle of ML models, from data, training, and post-training methods to deployment and inference. This was used to compare the advantages and limitations of different strategies and identify trade-offs between model accuracy, robustness, and performance. The third stage uses case study analysis to present real-world examples of ML and AML integration in various fields: network security and SOC (El Gharbaoui *et al.*, 2024), defence against AML attacks (Bountakas *et al.*, 2023), DevOps/CI-CD (Gajbhiye *et al.*, 2024), the financial industry (Saeed & Alsharidah, 2024), stress testing ML pipelines (Zhu *et al.*, 2025), testing DNN (CV/NLP) (Filus & Domańska, 2024), CI/CD and MLOps (Velaga, 2024), best practices MLOps (n.d.). The objective of this stage was to demonstrate that effective protection against AML attacks requires a multi-level approach integrated into QA and MLOps, as well as to confirm the practical significance of systematising such strategies for various industries using examples. The structure of the study is presented in Figure 1.

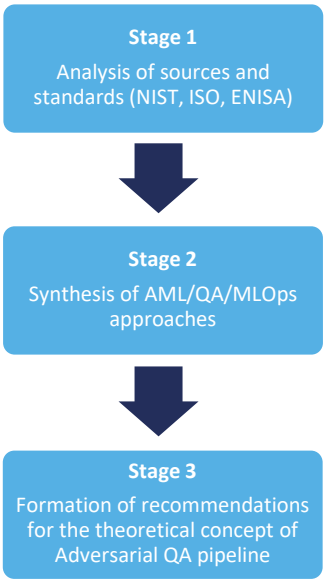

Figure 1. Stages, methods and results of the study
Source: compiled by the author

To ensure the accuracy and uniformity of the bibliography, Zotero, a specialised source management tool, was used to systematise the collected materials, guarantee the accuracy of references, and standardise the list of references.

RESULTS AND DISCUSSION

Systematisation of ML applications in cyber defence and challenges of adversarial attacks

The application of ML in cybersecurity provides automated threat detection and response through the analysis of network, system, and behavioural data. Key areas include anomaly detection, correlation of signals from different sources, and response automation (SOAR/EDR). The advantages are scalability, predictability, speed, and the ability to continuously learn (El Gharbaoui *et al.*, 2024). At the same time, the integration of AI/ML creates a new class of vulnerabilities: data, models, and interfaces become targets for attacks in dynamic industrial environments (Jedrzejewski *et al.*, 2024).

One of the main challenges is adversarial attacks, where even minimal manipulation of input data can cause significant malfunctions in intrusion detection systems, malware classifiers, or anti-phishing mechanisms, reducing their reliability. Such attacks pose a significant danger in areas with high security requirements, particularly in healthcare, where they raise not only technical but also ethical and legal issues (Musser *et al.*, 2023). Equally critical is the phenomenon of

poisoning attacks, where hidden vulnerabilities or "backdoors" are integrated into the model by manipulating training data, which can be activated in production environments. This type of attack is particularly threatening to systems with continuous learning, such as security operations centres (SOCs) or the financial sector. The issue of model robustness should be noted separately: even partial structured data corruption (5-10%) significantly worsens quality indicators, in particular AUC, due to the destruction of links between features and labels (Zhu *et al.*, 2025). One promising way to reduce such risks is similarity-driven testing of deep neural networks, which can improve efficiency of detection of hidden vulnerabilities (Filus & Domańska, 2024). The integration of ML components into classic AppSec and SecOps frameworks proves problematic, as traditional security models do not incorporate the specific vulnerabilities inherent in AI (Musser *et al.*, 2023). A generalised AI/ML lifecycle diagram, which reflects the main phases and potential vulnerabilities, is presented in Figure 2.

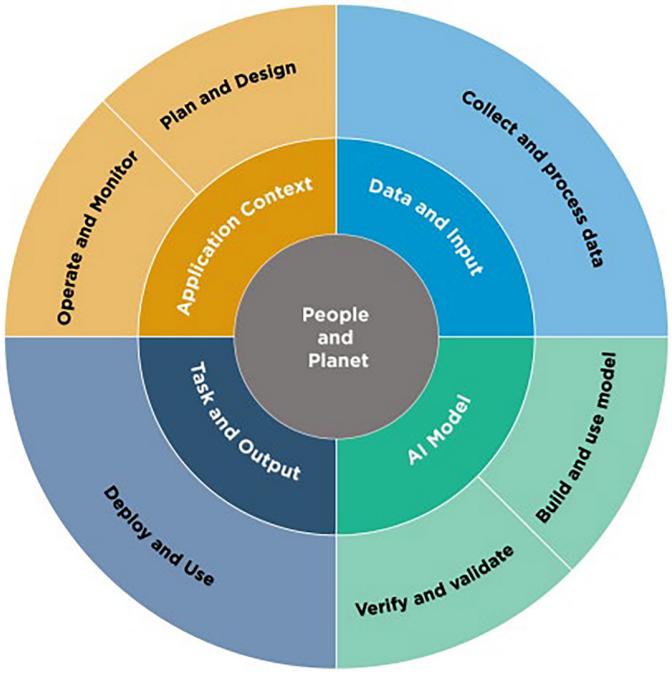

Figure 2. The AI lifecycle, adapted in the context of cybersecurity and AML protection
Source: compiled by the author based on the National Institute of Standards and Technology (2023)

Vulnerabilities can occur at any stage of the lifecycle, from data collection and processing to validation, deployment, and monitoring. This confirms the need to integrate AML protection mechanisms into all phases of CI/CD processes. ML is being actively implemented in cyber defence, increasing the effectiveness of detection and response, but at the same time creating risks of adversarial attacks. The key areas are presented in Table 1.

Table 1. Systematisation of ML applications in cybersecurity

Direction	Example of use	Benefits	Risks
NDR/NIDS and network anomaly analysis	Detection of zero-day and Advanced Persistent Threats (APTs) in traffic through real-time analysis of large volumes of network traffic, identification of anomalies and patterns such as unusual traffic spikes or deviations in user behaviour, botnets, Domain Generation Algorithms (DGA)	High detection rate (detection time reduced from 48 hours to 3 hours, 93.75% improvement), proactive threat detection, scalability for large infrastructures without a proportional increase in human resources	Dependence on data quality and quantity (inaccurate data leads to ineffective models), possible false positives (reduction from 20% to 5%, but still risks of inefficiency), vulnerability to poisoning attacks and adversarial AI, where attackers can manipulate data to deceive systems
UEBA	Identification of insiders and compromised accounts through analysis of deviations in user behaviour, such as abnormal traffic spikes or unusual actions indicating internal threats.	Detection of complex internal threats that are not accessible to signature-based methods, using ML to recognise patterns in historical data; improved effectiveness in dynamic environments	Privacy issues (collecting sensitive data creates risks of leaks), the need for continuous learning to adapt to evolving threats, data drift, where models develop blind spots due to changes in data patterns

Continued Table 1.

Direction	Example of use	Benefits	Risks
EDR/SOAR-automation	Automatic response via playbooks, such as isolating affected systems, blocking suspicious IPs, or terminating malicious processes, to quickly mitigate threats such as ransomware	Reduction of Mean Time to Detect (MTTD)/Mean Time to Respond (MTTR), automation of routine tasks, enabling the processing of large volumes of data at speeds unattainable by humans	Risk of false positives (false positives lead to business losses, e.g., blocking legitimate actions); dependence on ML solutions is vulnerable to adversarial attacks that can cause the system to ignore real threats
Anti-phishing and content analysis	NLP/Large Language Model (LLM) for letter classification and social engineering, including validation and sanitisation of prompts in LLM to detect phishing schemes that mimic legitimate requests; monitoring LLM interactions for threats; content moderation using classifiers to block prohibited content in audio/text	Effectiveness in dynamic attacks through complex pattern recognition; scalability for processing large volumes of data; increased trust through transparency (citing sources, demonstrating reasoning) and automation of rejecting malicious requests	Vulnerability to adversarial prompt injection/jailbreak, explainability difficulties (insufficient transparency of decisions), risks of generating misinformation or sensitive content through the repetition of false information
Malware classification and botnets	CNN/RNN/Transformer for software analysis, e.g., detecting malware through manipulation of binary files, botnets, and DGAs through analysis of network traffic and patterns	Speed and scale, working with large data samples to automate feature engineering, reducing the need for domain expertise for new zero-day malware variants	Vulnerability to adversarial examples (manipulation of inputs to reduce accuracy, F1-score, Area Under the Curve (AUC)), possible bypasses due to polymorphism (changing malware to avoid detection); increase in false positives/negatives in non-stationary environments
Anti-fraud and financial security	Real-time transaction analysis to detect complex fraud schemes such as phishing or data manipulation; models are trained on historical data to recognise patterns	Detection of complex fraud schemes (20% improvement in detection, 15% reduction in false positives), increased customer trust through enhanced security and satisfaction	Risk of bias in models due to unrepresentative training data; false negatives with high costs (missed threats); vulnerability to adversarial attacks that manipulate models
Threat Intelligence and threat correlation	Automatic enrichment of Indicators of Compromise (IoC), attribution of attacks through correlation of data from sources such as the dark web and hacker forums; use of DL for processing heterogeneous data	Optimises SOC through automated analysis, accelerates threat campaign attribution, and scalability for large volumes of data from multiple sources.	Possibility of TI data poisoning through source manipulation, noise in feeds leading to inaccurate attributions, and privacy risks when processing sensitive data
ICS/SCADA and OT-protection	Monitoring sensors, anomaly detection in critical systems such as ITS (Intelligent Transportation Systems) using Data Science Advanced Data Learning (DS-ADL) models for resilience to adversarial attacks, protection against cyber-physical threats	Improving critical infrastructure security through robust models against perturbations, using federated learning for collaborative training without compromising privacy	ML models are vulnerable to uncontrolled data (input perturbations), difficulties in explainability of decisions, and risks of false positives in critical applications such as spam detection
Cloud security and access control	Detection of lateral movement and credential stuffing through access pattern analysis, application of homomorphic encryption for secure transactions in cloud environments	Scalability in multi-tenant environments without compromising performance, optimisation through differential privacy for data protection	Privacy issues (data de-anonymisation), high false positive rate due to sensitivity to perturbations, poisoning risks due to training data manipulation
Cyber risk forecasting	Risk scoring of assets, scenario modelling of attacks through pattern identification and data correlation to predict threats, and use of AI for proactive fortifications	Provides management with strategic planning tools; anticipates attacks to minimise damage; efficiently processes big data for accurate forecasts	Models require up-to-date data (lack thereof leads to inaccurate predictions), risk of incorrect predictions due to bias or incomplete data, and vulnerability to adversarial manipulation

Continued Table 1.

Direction	Example of use	Benefits	Risks
QA/DevOps integration	Security gates in CI/CD, ML for finding vulnerabilities in containers and code, automated testing with AI for code analysis, threat modelling, risk assessment, examples: predictive maintenance in automotive (Kafka/Spark/TensorFlow), diagnostics in healthcare (NiFi/PyTorch), fraud detection in finance (Kafka/XGBoost)	Automation of security in pipelines (40% reduction in training time, 25% reduction in downtime), improved code quality through early detection; improved collaboration between teams; diagnostic accuracy up to 85%, 15% reduction in false positives	Computational costs (high requirements for real-time analysis), lack of mature standards for integration, data quality issues (noisy data requires cleaning), and model drift risks requiring monitoring
Stress testing ML pipelines	Adversarial data corruption in anomaly detection pipelines: Savage framework for modelling realistic data quality problems through dependency graphs, examples of Missing Not At Random (MNAR) in finance, compound errors in policing (label flips by demographics)	Identification of weak spots in pipelines through two-level optimisation (beam search + Bayesian optimisation), identification of vulnerable subpopulations, and provision of benchmarks for robustness methods	Even a small percentage of damaged data (5-10%) sharply reduces accuracy (AUC drops by 0.05+ at 10% corruption), structured errors destroy feature-label relationships more strongly than random ones; the sensitivity of existing cleaning methods
Adversarial testing for CV/NLP models	Deep Neural Network (DNN): selecting the most/least similar labels to ground truth based on network/semantic similarity for targeted/non-targeted attacks; testing CNN on human-like errors	Detection of real vulnerabilities through a human-centric approach, increased resilience with better results compared to existing strategies, and a contribution to explainable testing and standardisation	Tests are often expensive and computationally intensive (similarity assessment and scenario testing), and depend on similarity measures, which can complicate implementation

Notes: literature sources were systematised using the Zotero tool, which ensured the standardisation of references and the formation of a correct bibliographic apparatus

Source: compiled by the author based on P. Bountakas *et al.* (2023), M. Musser *et al.* (2023), M. Steidl *et al.* (2023), O. El Gharbaoui *et al.* (2024), B. Gajbhiye *et al.* (2024), M.M. Saeed & M. Alsharidah (2024), K. Filus & J. Domańska (2024), S.P. Velaga (2024), F.V. Jedrzejewski *et al.* (2024), M. Roshanaei *et al.* (2024), M. Macas *et al.* (2024), J. Zhu *et al.* (2025)

Operational methods (NDR/NIDS, UEBA) provide high detection rates and enable real-time response to complex attacks, but remain sensitive to the quality of input data and prone to data drift, which reduces their long-term reliability. Automation tools (EDR/SOAR) reduce mean time to detect and mean time to respond (MTTD/MTTR), effectively relieve the burden on SOC analysts, and increase system scalability, but have significant limitations in the form of possible false positives and vulnerability to adversarial manipulations that can disable the system. Preventive approaches (QA/DevOps integration, pipeline stress testing) increase model resilience by detecting hidden vulnerabilities at the development stage and ensuring the long-term reliability of ML processes, but require significant computing resources, mature integration standards, and constant monitoring. A summary of the comparison results showed that none of the approaches is self-sufficient: operational methods ensure speed, automation ensures process efficiency, and preventive strategies ensure long-term reliability. Therefore, the most effective approach is a combination of the two, which can mutually compensate for shortcomings, optimise the balance between accuracy, stability and performance, and form the basis for integrating ML into modern security frameworks. The application of ML in various domains of cybersecurity combines significant advantages (increased accuracy, response speed, and scalability) with new risks related to data quality, vulnerability to adversarial attacks, and ethical constraints. Effective ML integration requires not only technological solutions, but also the development of adaptive security frameworks capable of addressing specific threats to models and data.

AML is a field of research that studies the vulnerabilities of AI models and develops mechanisms to protect them. In contrast to classical approaches, AML focuses on data, algorithms, and parameters, and a key methodological aspect is the definition of a threat model that describes the attacker's level of knowledge (Vassilev *et al.*, 2025). There are three scenarios: white-box – full access to architecture and parameters, which makes attacks as accurate as possible, but improbable in open environments; black-box – interaction only through requests and responses, which leads to the prevalence of transfer- and query-based attacks; and grey-box – partial knowledge of the model, most characteristic of industrial cases (Jedrzejewski *et al.*, 2024).

On this basis, attack classes were systematised. Evasion attacks target the inference phase: minimal perturbations in the input data cause the model to make mistakes, using both targeted and untargeted

approaches (FGSM, BIM/PGD, C&W, DeepFool). Poisoning attacks affect the data collection or training stages: integrity and availability variants reduce accuracy, while clean-label and backdoor introduce hidden "triggers" that are dangerous because they are invisible. Privacy and inference attacks are aimed at information leakage: membership and property inference or data recovery from gradients, which poses a particular threat to medicine and finance. In MLaaS environments, extraction attacks are common, where the model is reproduced through mass queries or even hardware channels (EM radiation, Rowhammer). A new subclass consists of attacks on generative models (GenAI): prompt injection, jailbreak, fine-tuning misuse, and prompt extraction, which have already received their own taxonomies in international standards. The spectrum of attacks within AML ranges from modifying input data to leaking private information or completely reproducing models. Their classification is not only a tool for scientific analysis, but also a practical prerequisite for developing defence strategies and integrating adversarial tests into QA/DevSecOps (Musser *et al.*, 2023).

At the data level (data validation, filtering, anomaly detection, synthetic augmentation, differential privacy), measures are highly effective against gross poisoning and noise and improve overall quality, but are limited in countering targeted stealth poisoning and require additional resources for collection and auditing. At the training level (adversarial training, robust optimisation, regularisation, DP-SGD), the best resistance to evasion attacks in targeted scenarios is achieved, but these methods require significant computational resources and can reduce the baseline accuracy of models. Post-training approaches (input transformations, feature squeezing, attack detectors, calibration/thresholding, ensembles) are notable for their ease of integration into CI/CD and their ability to improve security without retraining, but they remain vulnerable to adaptive attacks and can affect throughput. At the deployment and inference stage (rate-limiting, abuse control, request rejection or reclassification, canary models, shadow deployment), these methods can significantly reduce MTTR and business risks, but require mature monitoring and accompanying playbooks. Finally, process and management approaches (threat modelling, red teaming, adversarial testing, QA/DevSecOps policies, drift monitoring) provide the foundation for long-term manageability and compliance with standards, but require high organisational maturity and relevant competencies (Jedrzejewski *et al.*, 2024; Vassilev *et al.*, 2025). The most effective approach is to integrate various protection strategies within QA/MLOps with clear SLOs/SLIs, continuous drift monitoring, and regular adversarial testing, which optimises the balance between accuracy, resilience, and performance and forms the basis for modern security frameworks. Figure 3 shows a visualised structure of cybersecurity measures.

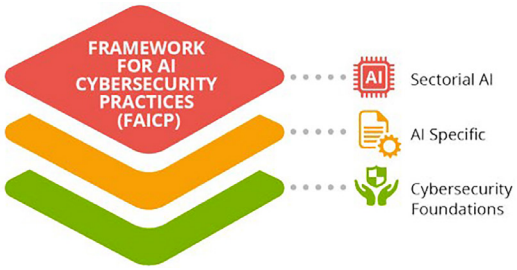

Figure 3. Structure of multi-level security measures for AI systems

Source: compiled by the author based on N. Polemi & I. Praça (2023)

Figure 3 shows the three-level structure of security practices: Cybersecurity Foundations – basic principles of cybersecurity common to all systems, AI Specific – measures that incorporate the lifecycle and vulnerabilities of AI models, Sectorial AI – domain-specific controls for specific industries (finance, healthcare, manufacturing). The model demonstrates a gradual transition from general to specialised levels of AML protection and summarises the approach to integrating AML controls into a multi-level cybersecurity system, combining universal standards with AI-specific practices.

The study results showed that ML increases the speed of detecting zero-day and APT attacks, automates responses through SOC/SOAR, and reduces the workload on analysts, but there remains a dependence on data quality and the risk of false positives. This was consistent with the findings of R. Liu *et al.* (2024), noting that ML algorithms in NDR/NIDS enable high anomaly detection accuracy and reduced response times. The study emphasised that ML increases SOC proactivity and ensures faster localisation of complex threats, including zero-day attacks. At the same time, the authors highlighted the problems of input data quality and the risks of false positives, which were also identified in the current study. Therefore, ML has a dual effect on network security: significant advantages in scalability and speed, along with challenges related to data reliability.

Y.L. Khaleel *et al.* (2024) conducted a systematic review of defensive methods in countering adversarial attacks, emphasising the role of MLand deep learning in creating multi-level defence strategies. The study showed that existing approaches, from adversarial training to ensemble models, significantly increase system resilience but remain vulnerable to new classes of attacks. This was consistent with the presented research showing that even when robust optimisation or randomised smoothing is applied, models retain limitations in terms of scalability and adaptability. This confirmed the need not only for technical innovation, but also for a comprehensive approach that addresses both algorithmic and organisational aspects of cyber defence.

The results of the study revealed that ML models are vulnerable to even minor adversarial perturbations, which reduces their robustness. X. Jiang *et al.* (2022)

proposed a local graph consensus method for robust image matching, which proved highly effective in reducing the impact of such perturbations. The authors' approach confirmed that structural data alignment methods can compensate for the weaknesses of classical models, maintaining the stability of results even in challenging conditions. This was consistent with the presented research on the need to implement algorithmic mechanisms to improve robustness in the CV/NLP domain. Both works demonstrated that combining theoretical and applied solutions is the key to creating robust ML systems.

A. Yaseen (2023) outlined that the application of ML in network security radically changes the response paradigm: from traditional reactive approaches that rely on signature-based methods to proactive anomaly detection strategies. ML models are capable of analysing both behavioural and network patterns at a much higher speed, enabling more effective detection of zero-day and complex multi-vector attacks. At the same time, the author emphasised the critical role of continuous model training, which can be used to ensure adaptability of the dynamic conditions of cyber threats. However, this creates additional risks, in particular, data drift and an increase in the number of false positives, which can undermine trust in the systems. The results of the study confirmed these conclusions: data quality and the availability of validation mechanisms are key to reducing noise and improving the effectiveness of ML solutions. This demonstrated a convergence in the view of ML as a capable but vulnerable technology that requires adaptive risk management to ensure long-term effectiveness in network security.

B.B. Sarhan & N. Altwaijry (2023) showed that ML-methods can effectively form user behaviour profiles and detect potentially harmful actions at an early stage. The study emphasised that this approach outperforms traditional signature-based methods due to its ability to adapt to changing behaviour patterns. They also highlighted the value of different algorithms, including SVM, classification trees, and ensemble methods, which increase the reliability of the system. This was consistent with current research that integrating ML into UEBA significantly improves the effectiveness of insider threat detection and contributes to proactive cyber defence, while ensuring data quality and minimising false positives remains a key challenge.

The application of ML in cybersecurity demonstrates high efficiency in detecting anomalies, automating responses, and forming behavioural profiles, which significantly increases accuracy compared to traditional methods. At the same time, models remain vulnerable to adversarial and poisoning attacks, data drift, and data quality issues that can reduce their reliability. This highlights the need for a comprehensive approach that combines algorithmic solutions, adaptive learning mechanisms, and organisational strategies to ensure robust cyber protection.

A comprehensive approach to AML in the context of standards, policies and practical cases

AML is not exclusively a technical problem; it has significant regulatory, ethical, and legal dimensions. If attacks on models undermine trust in AI systems, their uncontrolled exploitation may violate the principles of security, privacy, and fairness. A key document is the National Institute of Standards and Technology (2023), which proposes a systematic approach through the "2Map-Measure-Manage-Govern" functions. It defines a taxonomy of attacks and mitigations, emphasising the criticality of forming AML protection policies at the organisational process level. Additionally, ISO/IEC 23894:2023 (2023) remains the basic standard in the field of information security. Although it is not specific to AI, its requirements for risk management, incident management, and access control also apply to the life cycle of ML models. This can combine the classic approach to cybersecurity with new AML risks. The European Union Agency for Cybersecurity (ENISA) proposes a multi-layered model, ranging from basic cyber hygiene measures to AI-specific practices (Polemi & Praça, 2023). It emphasises the need to integrate AML tests at all levels, from data and models to organisational policies.

Large companies are developing their own frameworks for AI security. Microsoft Security (n.d.) outlines principles for protecting generative models, including mechanisms for countering prompt injection and approaches to model lifecycle management integrated into corporate security processes. OpenAI (2024) demonstrates an example of transparent risk documentation, where the model's resistance to adversarial attacks and jailbreaks is systematically assessed, which meets current requirements for openness and accountability in the field of AI. In addition to technical and regulatory issues, AML also raises ethical issues such as privacy (leaks of training data through inference attacks violate the principles of personal information protection), fairness (adversarial training can increase or decrease the fairness of a model depending on how it affects different user groups), and transparency and trust (society demands explainability of how AI makes decisions and how organisations manage the risks of adversarial attacks). The regulatory and ethical aspects of AML complement technical solutions, forming a second dimension of model security. For QA and MLOps, this means verifying not only technical metrics but also compliance with standards and corporate policies, which are the basis for trust in AI systems.

The issue of protection against adversarial attacks is no less complex than their classification. Existing approaches to ML security suggest considering defensive measures in the context of the model lifecycle. This creates convenient conditions for integration into QA/MLOps processes, as it can assess risks and embed resilience testing at all stages, from data preparation to operation (Musser *et al.*, 2023). At the initial stage, the main task is to protect data, as it forms the basis

for the model. Methods of data selection and cleaning, deduplication, filtering of out-of-distribution (OOD) or anomalous samples, as well as integrity and origin control (data lineage) are used. Such measures reduce the probability of poisoning attacks and increase trust in data sources (Bountakas *et al.*, 2023). At the training stage, the key focus is on adversarial training and robust optimisation, which increase the resilience of models to evasion attacks. Regularisation methods (e.g., Tradeoff In Adversarial Defence via Smoothing (TRADES)) are also used to balance accuracy and robustness. To reduce the risk of membership inference and data reconstruction, differentially private algorithms (Differentially Private Stochastic Gradient Descent (DP-SGD), Differentially Private Follow-the-Regularised-Leader (DP-FTRL)) are integrated. However, such methods do not protect against model extraction, which limits their applicability.

In the post-training stage, randomised smoothing methods are used to formally certify the level of stability, as well as formal verification approaches. However, trade-offs between accuracy and stability, as well as possible negative effects on model fairness, should be considered, which requires careful evaluation during QA (National Institute of Standards and Technology, 2023). During model deployment, the main task is to protect against attacks on interfaces and services. Request restrictions and audits, request rate limiting, and detection of suspicious access patterns are used. Such measures increase the level of protection against model extraction, but these mechanisms can be circumvented, so constant monitoring and retesting are required (Bountakas *et al.*, 2023). At the inference stage, data preprocessing methods (quantum compression, denoising) are used, as well as special detectors for adversarial or OOD cases. However, these approaches are vulnerable to adaptive attacks, and the use of gradient masking is dangerous, as it only creates the illusion of protection but does not eliminate the problem. Effective protection against AML attacks requires a multi-level approach that covers all stages of the model lifecycle, incorporating the inevitable trade-offs between accuracy and robustness, privacy and performance, formality and practical relevance. For QA and MLOps, this means implementing dynamic testing and retesting processes that can adapt to the constant evolution of attacks.

Practical cases of AI/ML and AML applications cover network security, SOC, DevOps/CI-CD, and industrial domains. Their systematisation has made it possible to outline the opportunities and risks of integrating these technologies into real environments (Table 2).

Table 2. Practical scenarios for integrating AI/ML and AML into security systems

Sphere of use	Methods and examples	Benefits and risks
Network security, SOC	Detection of botnets in the bank's corporate network through real-time traffic analysis. ML models for anomaly detection (DGA, suspicious IPs), integration with SIEM (Splunk, QRadar) and IDS (Snort) for automatic responses such as IP blocking	Rapid detection of attacks (30% reduction in response time), SOC automation. Risks: requirement for standardisation, vulnerability to data poisoning
AML-attack protection	Protecting telecom companies' intrusion detection systems from adversarial attacks. Adversarial training, feature squeezing to reduce data manipulation, and ensemble models for robustness	Protection of models in cybersecurity, computer vision, and NLP, with the possibility of validation through datasets. Risks: methods are not yet ready, and the complexity of countering new attacks
DevOps/CI-CD	Automated security testing: searching for vulnerabilities in containers (continuous monitoring of code bases to detect code injections, unauthorised access, data leaks); code analysis using ML to recognise risk patterns, automate threat modelling and risk assessment. Models adapt to evolving threats through training on new data	ML tests are integrated into QA as security gates: integrating AI/ML into CI/CD ensures a continuous security process throughout the development cycle, with prioritisation of vulnerabilities. This can be used for early detection of threats in dynamic DevOps environments, but requires an emphasis on data quality and model adaptability
Industry (finance, telecoms)	Protection of the anti-fraud system of the bank from data poisoning. ML for detecting fraudulent transactions, protection against adversarial attacks on IDS through training on noisy data	The lack of mature robustness metrics hinders AML integration and has low practical relevance. The absence of standardised metrics complicates the implementation of AML in production, particularly for poisoned data in sensitive industries
Pipeline stress-testing	Testing the ML pipeline at a financial company to identify vulnerabilities to data corruption. Savage framework for generating scenarios with missing data (MNAR in finance) and label errors (predictive policing), optimisation via beam search	Even minor data damage drastically reduces accuracy: Structured errors (5-10%) destroy feature-label relationships more severely than random ones, and existing cleaning/debugging methods are sensitive. Stress testing is needed to validate the resilience of pipelines to systematic corruption

Continued Table 1.

Sphere of use	Methods and examples	Benefits and risks
DNN testing	Similarity-driven adversarial testing for CV systems: Using the most/least similar labels to ground truth (based on network/semantic similarity) for targeted/non-targeted attacks on CNN. Testing the object recognition system in airport security cameras	Identification of weaknesses through "human-like" attacks: The human-centric approach increases the effectiveness and transparency of testing by identifying vulnerabilities in visual perception. Promotes the standardisation of AI security, with a focus on explainable testing for practical requirements. Risks: high computational complexity, dependence on metrics
CI/CD, MLOps	Implementation of ML in company pipelines: predictive maintenance for an automaker (Tesla), diagnosis of heart disease (healthcare startup), and fraud detection in a bank. Methods: Kafka/Spark/TensorFlow for automotive, NiFi/PyTorch for healthcare, Kafka/XGBoost for finance; 4 stages of MLOps	QA in ML should include adversarial testing as standard: data quality is critical, drift monitoring and real-time optimisation. Automated validation/retraining provides the basis for adversarial QA, while cross-functional collaboration ensures resilience in CI/CD
Best practices MLOps	Implementation of MLOps in Google Cloud for a fintech service. Methods: Metrics (QPS, latency), security gates for skew testing, CD level 1-2, MTBF/MTTR monitoring	Adversarial QA is becoming a mandatory component of CI/CD: Testing for skew/anomalies in pipelines ensures stability, while monitoring triggers retraining. Balancing speed/security through incremental goals, adversarial tests are mandatory for production ML, with an emphasis on reproducibility and adaptation to evolving data. Risks: manual processes, training-serving skew

Notes: literature sources were systematised using the Zotero tool, which ensured the standardisation of references and the formation of a correct bibliographic apparatus

Source: compiled by the author based on P. Bountakas *et al.* (2023), M. Steidl *et al.* (2023), S.R. Gopireddy (2024), O. El Gharbaoui *et al.* (2024), B. Gajbhiye *et al.* (2024), F.V. Jedrzejewski *et al.* (2024), K. Filus & J. Domańska (2024), M. Macas *et al.* (2024), M. Roshanaei *et al.* (2024), S.P. Velaga (2024), J. Zhu *et al.* (2025), Adversarial robustness evaluation metric (2025), MLOps: Continuous delivery and automation pipelines in machine learning (n.d.), Amazon Web Services (n.d.)

The integration of AI/ML into various areas of cybersecurity has proven highly effective in reducing threat detection and response times, automating SOC processes, and improving detection accuracy. Most solutions face a lack of standardised robustness metrics and mature frameworks for assessing model resilience, which complicates the industrial implementation of AML approaches. Practical cases show that the most effective approach is a combination of preventive, analytical, and automated strategies, from adversarial training and feature squeezing to security gates in CI/CD and continuous monitoring of model drift. This approach forms a comprehensive Adversarial QA ecosystem, in which model robustness is systematically assessed at all stages of the life cycle, from training to deployment in production environments.

Given the growing complexity of ML systems and their vulnerability to adversarial attacks, integrating testing and protection mechanisms directly into QA/CI-CD pipelines is critical. This can ensure a shift from reactive to proactive strategies, where model robustness and privacy are systematically tested at all stages of the lifecycle. Automated security testing using AI/ML is already being actively implemented in DevOps environments, but so far, it has focused mainly on classic vulnerabilities in code and configurations. The integration of adversarial testing remains limited. Existing pipelines for ML models lack mature stability control mechanisms, creating a gap between scientific developments and industrial needs. DevOps pipelines need to include continuous testing modules capable of detecting not only performance regressions but also reductions in model stability. This ensures early detection of problems and rapid validation of changes in code and data.

The quality control system should incorporate not only classic accuracy indicators, but also metrics of resistance to attacks, in particular robust accuracy, attack success rate, and privacy loss. This can be used for informed decisions regarding the model's readiness for deployment. Robust accuracy (RA) reflects the proportion of examples that the model classifies correctly even after introducing permissible ε-level perturbations (1):

$$RA = \frac{1}{N}\sum_{i=1}^{N} I\left(f(x_i + \delta_i) = y_i\right), \|\delta_i\| \leq \varepsilon, \qquad (1)$$

where $(f(x_i + \delta_i)$ – model output for a disturbed input sample $x_i + \delta_i$, y_i – true label (real class) for the input sample x_i, δ_i – disturbance added to the input data, ε – maximum permissible disturbance level, I – indicator function equal to 1 if the expression in brackets is true (the model correctly classifies the disturbed sample), 0 – if false. RA measures the extent to which the model correctly classifies samples, even after adding small perturbations δ_i that do not exceed the permissible level ε. A high *RA* indicates that the model is resistant to attacks or noise, maintaining accuracy even when data changes,

which is essential for tasks where accuracy is critical (in image recognition systems or medical diagnostics).

Attack success rate (ASR) characterises the proportion of examples for which the attack resulted in a change in classification (2):

$$ASR = \frac{1}{N}\sum_{i=1}^{N} I\left(f(x_i) \neq f(x_i + \delta_i)\right). \quad (2)$$

An increase in *ASR* indicates greater vulnerability of the model to attacks. If *ASR* is high, it means that the model can be easily misled, which is critical for testing and ensuring reliability in real-world applications.

Privacy loss (ε) determines the level of differential privacy of the model and limits the influence of a single record in the training sample on the final training result (3):

$$\varepsilon = ln\left(\frac{P(M(D)=o}{P(M(D')=o}\right), \quad (3)$$

where $P(M(D) = o$ – the probability that the model M, trained on data D, will provide result o, $P(M(D') = o$ – the probability that the model M, trained on data D' (different from D only by one entry), will provide the same result o. Privacy loss measures how much changing a single record in the training sample affects the model's performance. A small ε indicates that the model has high privacy, as changing a single record has little effect. A high ε means that the model is less private, as a single record can significantly change the result.

Testing models for resilience to hacking should be integrated into QA processes on a par with traditional unit and integration tests. This approach ensures systematic assessment of vulnerabilities during development and operation. CI/CD processes should include security gates that check models for compliance with security policies prior to deployment, with subsequent logging for auditing and regulatory compliance. The integration of adversarial testing is accompanied by a number of challenges: high computational costs, lack of unified standards, a gap between academic approaches and engineering practice, and the need for cultural change in DevSecOps teams, where AML tests should be considered a mandatory stage of QA. Implementing AML protection in QA/CI-CD is a complex task that requires not only technical innovation but also organisational transformation. Adversarial testing should be considered a standard component of quality control, and QA/MLOps systems should be viewed as mechanisms not only for verifying functionality but also for managing model resilience and trust.

The study identified that implementing AML in industrial environments requires adaptive security frameworks capable of addressing specific threats to models and data, while balancing accuracy, robustness, and performance. This was consistent with the study by P. Kashefi *et al.* (2024), which examined the issue of global AI regulation and emphasised the need for "flexible regulation". The study argued that excessive control can stifle innovation and technological development, while the absence of rules creates significant risks for society and trust in AI systems. They propose a model that combines international standards with national contexts, while maintaining adaptability and practical relevance. The consensus was on the appropriateness of dynamic and context-dependent approaches, with both studies correlating with the idea that only a balance between control and innovation can ensure the sustainable integration of AI into real-world environments.

M. Malatji & A. Tolah (2025) presented a comprehensive framework for analysing AI cybersecurity measurements, in particular adversarial and offensive AI. The authors systematised the main classes of attacks, emphasising their impact on critical infrastructures and industrial systems. Traditional single-layer defence methods remain insufficient, and an effective strategy must have a multi-layered architecture. The study addressed the organisational aspects of integrating security at the level of processes and policies, rather than just technical solutions, which made it possible to form a more robust defence against dynamic and complex attacks. This was consistent with the findings of this study regarding the critical need for multi-layered strategies to counter adversarial attacks. In particular, the current findings emphasised the importance of integrating adversarial testing into QA/CI-CD, which echoed the authors' conclusions regarding the need for a comprehensive approach where technical and organisational levels cooperate.

The results of the study demonstrated that adversarial attacks create not only technical but also ethical problems, undermining trust and fairness in the use of AI. Even highly accurate models can become a source of social risks if aspects of fairness and explainability are not addressed. This was consistent with the study by T.W. Kim & A. Strudler (2023), demonstrating that the development of AI is impossible without incorporating social and ethical aspects, in particular, fairness and legitimacy of decision-making. The study argued that the issue of the rights and responsibilities of AI systems has a direct impact on public trust in technology. Technological progress without an ethical framework can lead to crises of trust and social risks. Both studies confirmed that without ethical criteria, AI security remains incomplete. The study by D.S. Fessenko & A. Jasperse (2025) emphasised that ethics, combined with technical and legal norms, should be at the heart of AI regulation. The regulatory framework should act as a bridge between scientific innovation and public expectations. The study suggested viewing regulation not as a barrier, but as a mechanism for harmonising transparency, accountability and innovation. The presented research results confirmed these conclusions, as AML was considered not only in a technical but also in a regulatory context. Therefore, the formation of AML policies should be based on a balance of technical and regulatory decisions.

L. Jiang & V. Torra (2023) addressed the problem of data protection in multi-source and distributed ML models. The study proved that privacy, access control, and data integrity are the foundation of trust in any AI system. They described the challenges associated with leaks and manipulations in multi-source environments and proposed using data lineage mechanisms for transparency of data origin. The key conclusion was that the reliability of models directly depends on the quality and security of the input data. This correlated with the results of the presented study, which also found that poisoning attacks and adversarial manipulation remain a critical challenge that can destroy even the most accurate models. The integration of data validation, cleaning, and tracking procedures is a prerequisite for the stable operation of ML systems in industrial environments.

Thus, ML significantly enhances cyber defence, accelerates detection, scales network/behavioural data analysis, and automates responses in SOC/SIEM/SOAR. At the same time, a new class of risks emerges: adversarial and poisoning attacks, data drift, and dependence on data quality/origin, which sharply reduces the robustness of models. Sustainable integration requires multi-level technical and organisational measures (adversarial training, robust metrics in QA/CI-CD, NIST/ISO/ENISA practices) with a balance between innovation and control.

CONCLUSIONS

The results of the study confirmed that integrating ML methods into cybersecurity is a strategically relevant step that combines improved threat detection and response with the emergence of new challenges. In particular, the study established that the use of ML reduces attack detection time by tens of per cent, enables automated response (EDR/SOAR), and improves the accuracy of network traffic and user behaviour pattern analysis. Reducing detection time from 48 to 3 hours or reducing the proportion of false positives from 20% to 5% confirms the real benefits of ML in protecting complex infrastructures. At the same time, limitations of this approach have also been identified. The main risks are vulnerability to adversarial and poisoning attacks, which, even with minimal interference with the data, can reduce the accuracy of models and create hidden "backdoors". The impact of such attacks is particularly critical in areas with high reliability requirements, such as medicine or financial services. In addition, issues of privacy, explainability, and bias remain problematic, directly affecting trust in AI systems.

It is advisable to systematise the areas of application of ML in cybersecurity, from network monitoring and anti-phishing to DevOps integrations and risk prediction. The classification of attacks within AML has shown that threats cover the entire model lifecycle, from data collection to exploitation, and therefore require multi-level countermeasures. It is recommended to implement adversarial testing as a mandatory stage of QA/MLOps, which will facilitate the transition from reactive to proactive cyber defence strategies. It is worth noting that the effectiveness of ML solutions directly depends on data quality, as even 5-10% of structured corruption in the sample can lead to a noticeable decrease in AUC and other metrics, confirming the critical importance of controlling data reliability and origin.

The limitation of this study is its predominantly theoretical nature, which necessitates further empirical verification and practical testing of the proposed approaches in real-world cyber defence systems. Further research should focus on developing resilience metrics, standardising AML practices, and creating adaptive security frameworks. In CI/CD pipelines, it is advisable to integrate adversarial testing as a security gate after the model validation stage. This can detect a decrease in robustness or an increase in the attack success rate before the model is implemented. In Quality Assurance processes, it is worth adding robust accuracy, privacy loss, and attack success rate metrics as mandatory criteria for model acceptance. This will ensure a systematic approach to assessing resilience before deployment. The integration of differential privacy, formal verification, and explainable AI will help balance accuracy, transparency, and resilience, ensuring the effectiveness and trustworthiness of ML systems in the changing landscape of cyber threats.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Adversarial robustness evaluation metric. (2025). Retrieved from https://www.ibm.com/docs/en/watsonx/saas?topic=metrics-adversarial-robustness&utm_source=.

[2] Agarwal, A., & Nene, M.J. (2025). Advancing trustworthy AI: A comparative evaluation of AI robustness toolboxes. *SN Computer Science*, 6, article number 234. doi: 10.1007/s42979-025-03785-w.

[3] Amazon Web Services. (n.d.). *Metrics for CI/CD pipelines*. Retrieved from https://docs.aws.amazon.com/prescriptive-guidance/latest/strategy-cicd-litmus/metrics-for-cicd-pipelines.html?utm_source=.

[4] Bagai, R., Masrani, A., Ranjan, P., & Najana, M. (2024). Implementing continuous integration and deployment (CI/CD) for machine learning models on AWS. *International Journal of Global Innovations and Solutions*. doi: 10.21428/e90189c8.9cb39c55.

[5] Bena, N., Anisetti, M., Gianini, G., & Ardagna, C.A. (2024). Certifying accuracy, privacy, and robustness of ML-Based malware detection. *SN Computer Science*, 5(1), article number 710. doi: 10.1007/s42979-024-03024-8.
[6] Bountakas, P., Zarras, A., Lekidis, A., & Xenakis, C. (2023). Defense strategies for adversarial machine learning: A survey. *Computer Science Review*, 49, article number 100573. doi: 10.1016/j.cosrev.2023.100573.
[7] El Gharbaoui, O., Kiyadi, I., & El Boukhari, H. (2024). Evaluating AI and ML in network security: A comprehensive literature review. *Procedia Computer Science*, 251, 727-733. doi: 10.1016/j.procs.2024.11.176.
[8] Fessenko, D.S., & Jasperse, A. (2025). Ethics at the heart of AI regulation. *AI and Ethics*, 5, 3387-3398. doi: 10.1007/s43681-024-00562-y.
[9] Filus, K., & Domańska, J. (2024). Similarity-driven adversarial testing of neural networks. *Knowledge-Based Systems*, 305, article number 112621. doi: 10.1016/j.knosys.2024.112621.
[10] Gajbhiye, B., Aggarwal, A., & Jain, S. (2024). Automated security testing in DevOps environments using AI and ML. *International Journal for Research Publication and Seminar*, 15(2), 259-271. doi: 10.36676/jrps.v15.i2.1472.
[11] Gala, D.L., Molleda, J., & Usamentiaga, R. (2025). Evaluating the impact of adversarial patch attacks on YOLO models and the implications for edge AI security. *International Journal of Information Security*, 24, article number 154. doi: 10.1007/s10207-025-01067-3.
[12] Gopireddy, S.R. (2024). AI-powered code review and vulnerability detection in DevOps pipelines. *Journal of Scientific and Engineering Research*, 11(10), 124-129.
[13] Guo, J., Bao, W., Wang, J., Ma, Y., Gao, X., Xiao, G., Liu, A., Dong, J., Liu, X., & Wu, W. (2023). A comprehensive evaluation framework for deep model robustness. *Pattern Recognition*, 137, article number 109308. doi: 10.1016/j.patcog.2023.109308.
[14] ISO/IEC 23894:2023. (2023). *Information technology – artificial intelligence – guidance on risk management*. Retrieved from https://www.iso.org/standard/77304.html?utm_source=.
[15] Jedrzejewski, F.V., Thode, L., Fischbach, J., Gorschek, T., Mendez, D., & Lavesson, N. (2024). Adversarial machine learning in industry: A systematic literature review. *Computers & Security*, 145, article number 103988. doi: 10.1016/j.cose.2024.103988.
[16] Jiang, L., & Torra, V. (2023). Data protection and multi-database data-driven models. *Future Internet*, 15(3), article number 93. doi: 10.3390/fi15030093.
[17] Jiang, X., Xia, Y., Zhang, X.-P., & Ma, J. (2022). Robust image matching via local graph structure consensus. *Pattern Recognition*, 126, article number 108588. doi: 10.1016/j.patcog.2022.108588.
[18] Kashefi, P., Kashefi, Y., & Mirsaraei, A.G. (2024). Shaping the future of AI: Balancing innovation and ethics in global regulation. *Uniform Law Review*, 29(3), 524-548. doi: 10.1093/ulr/unae040.
[19] Khaleel, Y.L., Habeeb, M.A., Albahri, A.S., Al-Quraishi, T., Albahri, O.S., & Alamoodi, A.H. (2024). Network and cybersecurity applications of defense in adversarial attacks: A state-of-the-art using machine learning and deep learning methods. *Journal of Intelligent Systems*, 33(1), article number 20240153. doi: 10.1515/jisys-2024-0153.
[20] Kim, T.W., & Strudler, A. (2023). Should robots have rights or rites? *Communications of the ACM*, 66(6), 78-85. doi: 10.1145/3571721.
[21] Kotenko, D., & Khlaponin, Y. (2024). Artificial intelligence in cyber attack detection and prevention systems: Prospects and challenges. *Underwater Technologies*, 1(14), 48-55. doi: 10.32347/uwt.2024.14.1203.
[22] Liu, R., Shi, J., Chen, X., & Lu, C. (2024). Network anomaly detection and security defense technology based on machine learning: A review. *Computers and Electrical Engineering*, 119(A), article number 109581. doi: 10.1016/j.compeleceng.2024.109581.
[23] Macas, M., Wu, C., & Fuertes, W. (2024). Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems. *Expert Systems with Applications*, 238(E), article number 122223. doi: 10.1016/j.eswa.2023.122223.
[24] Malatji, M., & Tolah, A. (2025). Artificial intelligence (AI) cybersecurity dimensions: A comprehensive framework for understanding adversarial and offensive AI. *AI and Ethics*, 5, 883-910. doi: 10.1007/s43681-024-00427-4.
[25] Microsoft Security. (n.d.). *How Microsoft secures generative AI*. Retrieved from https://www.microsoft.com/content/dam/microsoft/final/en-us/microsoft-brand/documents/EN-US_BDM_Microsoft_Security_Whitepaper_v3.pdf.
[26] MLOps: Continuous delivery and automation pipelines in machine learning. (n.d.). Retrieved from https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning?utm_source=.
[27] Musser, M., *et al.* (2023). *Adversarial machine learning and cybersecurity: Risks, challenges, and legal implications*. Georgetown: CSET. doi: 10.51593/2022CA003.
[28] National Institute of Standards and Technology. (2023). *Artificial Intelligence: A taxonomy of AI risk and a framework for assessing AI risk (NIST AI 100-1)*. Gaithersburg: National Institute of Standards and Technology. doi: 10.6028/NIST.AI.100-1.

[29] Neretin, O., & Kharchenko, V. (2022). Ensurance of artificial intelligence systems cyber security: Analysis of vulnerabilities, attacks and countermeasures. *Academic Journals and Conferences*, 12, 7-22. doi: 10.23939/sisn2022.12.007.
[30] OpenAI. (2024). *GPT-4o system card*. Retrieved from https://cdn.openai.com/gpt-4o-system-card.pdf?utm_source=.
[31] Polemi, N., & Praça, I. (2023). *Multilayer framework for good cybersecurity practices for AI*. Attiki: European Union Agency for Cybersecurity.
[32] Roshanaei, M., Khan, M.R., & Sylvester, N.N. (2024). Enhancing cybersecurity through AI and ML: Strategies, challenges, and future directions. *Journal of Information Security*, 15(3), 320-339. doi: 10.4236/jis.2024.153019.
[33] Saeed, M.M., & Alsharidah, M. (2024). Security, privacy, and robustness for trustworthy AI systems: A review. *Computers and Electrical Engineering*, 119(B), article number 109643. doi: 10.1016/j.compeleceng.2024.109643.
[34] Sarhan, B.B., & Altwaijry, N. (2023). Insider threat detection using machine learning approach. *Applied Sciences*, 13(1), article number 259. doi: 10.3390/app13010259.
[35] Steidl, M., Felderer, M., & Ramler, R. (2023). The pipeline for the continuous development of artificial intelligence models – current state of research and practice. *Journal of Systems and Software*, 199, article number 111615. doi: 10.1016/j.jss.2023.111615.
[36] Vadde, B.C., & Munagandla, V.B. (2023). Integrating AI-driven continuous testing in DevOps for enhanced software quality. *Journal of Artificial Intelligence in Medicine*, 14(1), 505-515.
[37] Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., & Hamin, M. (2025). *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations*. Gaithersburg: National Institute of Standards and Technology. doi: 10.6028/NIST.AI.100-2e2025.
[38] Velaga, S.P. (2024). Case studies of successful CI/CD pipeline implementations for Machine Learning and AI. *International Journal of Research and Analytical Reviews*, 11(3), 137-145.
[39] Yaseen, A. (2023). The role of machine learning in network anomaly detection for cybersecurity. *Sage Science Review of Applied Machine Learning*, 6(8), 16-34.
[40] Zhu, J., Xu, G., Lorenzi, F., Glavic, B., & Salimi, B. (2025). Stress-testing ML pipelines with adversarial data corruption. *Proceedings of the VLDB Endowment*, 18(11), 4668-4681. doi: 10.14778/3749646.3749721.

Інтеграція тестування безпеки в конвеєри контролю якості за допомогою змагального машинного навчання

Анатолій Гусаковський
Магістр, старший інженер з контролю якості автоматизації
Національний аерокосмічний університет «Харківський авіаційний інститут»
61070, вул. Вадима Манька, 17, м. Харків, Україна
https://orcid.org/0009-0007-9398-0966

Анотація. Метою дослідження була теоретична систематизація підходів до підвищення стійкості машинного навчання систем у кіберзахисті через інтеграцію тестування на стійкість у процесі забезпечення захисту. Методологія охоплювала систематизацію машинного навчання – напрямів у кіберзахисті, аналіз стратегій протидії ворожим атакам та case study інтеграції в забезпечення якості та операції з машинного навчання. Встановлено, що використання технологій машинного навчання у кіберзахисті забезпечує автоматизацію виявлення та реагування на загрози (аномалії у мережах, поведінковий аналіз, антифішинг, антифрод, класифікація malware). Основними перевагами є масштабованість, швидкість реагування, предиктивність та ефективність у складних середовищах, тоді як ключові ризики охоплюють залежність від якості даних, хибні спрацювання, вразливість до adversarial- та poisoning-атак, а також проблеми приватності й пояснюваності. Виявлено, що у змагальному машинному навчання виділено три сценарії атак (white-box, black-box, gray-box) та їх класи, отруєння даних, конфіденційність/висновок, вилучення моделі, генеративний штучний інтелект. Підкреслено, що змагальне машинне навчання охоплює не лише технічний, а й нормативний та етичний виміри, пов'язані з принципами конфіденційності, справедливості та прозорості у використанні штучного інтелекту. Представлено багаторівневі стратегії захисту, інтегровані в життєвий цикл моделей машинного навчання – на рівні даних, під час навчання, після навчання, на етапах розгортання та інференції. Практичні кейси продемонстрували доцільність застосування машинного навчання та протидії відмиванню коштів у різних доменах – від мережевої безпеки та центру операцій безпеки до розроблення та експлуатації / безперервної інтеграції / безперервного постачання, фінансового сектору, стрес-тестування конвеєрів машинного навчання, а також забезпечення якості й експлуатації машинного навчання. Практична значущість полягає в можливості використання отриманих результатів фахівцями з кібербезпеки, фінансовими аналітиками та інженерами з експлуатації машинного навчання для підвищення ефективності центру операцій безпеки, інтеграції змагального тестування та забезпечення стійкості моделей машинного навчання у виробничих середовищах

Ключові слова: стійкість моделей; вразливості алгоритмів; метрики стійкості; автоматизоване тестування; життєвий цикл моделей; захисні механізми