Engineering and Technology Journal
Journal homepage: https://etj.uotechnology.edu.iq

AI-driven attacks on database security: taxonomy and defense strategies

Check for updates

Diyar W. Naaman [a]*, Berivan T. Ahmed [b], Hajar M. Yasin [b]

[a] Ministry of Education, General Directorate of Education in Duhok, Duhok, Iraq.
[b] Akre University for Applied Sciences, Technical College of Informatics, Department of Information Technology, Duhok, Iraq.
*Corresponding author Email: Diyar457@gmail.com

HIGHLIGHTS

- A novel taxonomy of five AI-powered database attacks bypassed traditional defenses with 85% success.
- Analysis of 23 recent studies revealed critical gaps in AI-driven database security frameworks.
- A layered defense model with adversarial training and behavioral monitoring was proposed.
- Statistical results showed a 32% decline in rule-based systems against advanced AI attack variants.
- A multi-disciplinary approach addressed technical, organizational, and human factors in AI threats.

Keywords:
AI-driven attacks
Database security
Adversarial machine learning
SQL injection
Data poisoning
Federated learning security
GAN-based attacks

ABSTRACT

Artificial intelligence has introduced both unprecedented capabilities and novel vulnerabilities into database environments, enabling highly adaptive attacks that can evade traditional defenses. This review surveys recent research published between 2022 and 2025 on AI-assisted database security threats and synthesizes the literature to develop a comprehensive taxonomy of emerging attack approaches. We identified five primary classes of AI-based attacks: intelligent SQL injection attacks, adversarial machine learning strategies targeting database security systems, data poisoning attacks on AI-based databases, automated reconnaissance exploits, and sociotechnical manipulations aimed at database administrators. We systematically reviewed publications on cyber defense stored in IEEE Xplore, ACM Digital Library, Science Direct, and Scopus databases. Boolean search terms were used on the databases specific to cyber defense. Findings indicate that automated SQL injection attacks can escalate the bypass rate of security systems to over 85% effectiveness. The effectiveness of rule-based defense systems degrades by 32% when pitted against sophisticated AI-adapted adversarial attacks. Conversely, machine learning-based defenses maintain a detection rate of 85 to 95%. To combat advancing techniques, a multilayer approach that includes adversarial training, anomaly-based intrusion detection, and automated user behavior analysis and reporting technology should be employed. This approach utilizes anomaly-based defenses through a monitoring model. Analysis shows that conventional database defense techniques need to be upgraded with real-time analytics, dynamic response mechanisms, and zero-day vulnerability protection to keep pace with the increasingly sophisticated nature of AI adversarial attacks on database systems.

1. Introduction

The construction of today's organizations relies on databases, which are used to store important information such as financial records, information on individuals, intellectual property, and other data that spans all sectors of an economy. The sensitive information contained in these databases requires protection, as organizations rely on digital systems to gain a competitive edge and ensure compliance with relevant regulations. For the better part of the last century, traditional database security systems, which focus on access control, input validation for databases, and meeting encryption standards, have effectively put a barrier in place against conventional cyberattack threats [1]. The cyberspace in which organizations operate is dynamic. The advancement of technologies such as artificial intelligence has changed the landscape of cybersecurity threats. New attack methods have emerged that target old weaknesses and new vulnerabilities introduced by the adoption of AI systems. Cybercriminals utilize machine learning to automate weaknesses in systems, design techniques to evade detection, and consistently modify strategies to bypass them, thereby targeting protected systems [2].

According to the latest reports from security intelligence, there have been significant developments in AI-enabled attacks, particularly in voice phishing, which has increased by 442%, as well as deepfake impersonation campaigns targeting database administrators and security personnel [3]. The introduction of AI-powered attack tools has altered the risk calculus when it comes

http://doi.org/10.30684/etj.2025.163946.2003
Received 11 August 2025; Received in revised form 15 September 2025; Accepted 21 September 2025; Available online 27 October 2025
2412-0758/University of Technology-Iraq, Baghdad, Iraq
This is an open access article under the CC BY 4.0 license http://creativecommons.org/licenses/by/4.0

to database security, as illustrated by high-profile attacks, such as the Change Healthcare ransomware attack, which cost $ 22 million. The Snowflake data breach, which affected 165 organizations, clearly demonstrates the potential damage that AI attacks can cause to database systems [4].

There is a group of academic and industry literature that examines the emerging threat landscape, including methodologies of attacks and frameworks of defense, along with strategies for countermeasures. There appears to be a lack of focus, however, as most of the research seems to be focused on single attack vectors without considering the sophisticated, integrated threat ecosystem and the comprehensive defense required to combat AI attacks on databases, which most address. This research aims to address these significant gaps by developing a comprehensive framework that captures the intricate complexity of AI attacks, based on systems, which outlines the level of sophistication in contemporary attacks and the defense frameworks that complement the threats modern systems face.

2. Background theory

Machine learning neural networks, database security flaws, and automated attack techniques converge as the theoretical foundation upon which AI-driven threats to databases are based. This analysis encompasses multiple attack patterns that utilize intelligent systems, examining the evolution of AI-driven attacks in the context of adversarial AI, the theoretical frameworks of defensive and offensive AI architectures, and the database protection systems employed by AI neural networks in information systems.

Intrusive AI assists in comprehending an AI ego that has been built to reach goals set for it outside the boundaries of the legal and ethical frameworks and systems [5] to which it has been programmed, thus necessitating its subsumption. It has been established that there exist holes in deep neural networks that can be exploited through the use of strategically crafted input perturbations to sophisticated relay attacks. The traditional assumption that attack databases closed by automated protective measures systems are immune to attack is counterintuitive, given that these automated AI protective systems can become their own self-innovated attack surfaces, potentially unbounded in their ability to obfuscate and record patterns of unauthorized information acquisition.

The classification of adversarial attacks against database protection is separated into three primary types: evasion attacks, poisoning attacks, and extraction attacks. Evasion attacks pose the primary challenge to the implementation of AI systems, as they employ sophisticated tactics to mask adversarial inputs, allowing malicious database queries to cross AI boundaries undetected [6]. In contrast, poisoning attacks attack AI security systems by embedding malicious modifications into the training datasets, weakening the system and exposing it to development vulnerabilities. In contrast, extraction attacks utilize sensitive information to reconstruct the database's fundamental architecture, analyze the system and telemetry, and then formulate novel methods for defeating systems that protect the database.

The traditional frameworks for database security focus on principles such as access control layers, validating inputs, encrypting data, and logging all system activities to perform audits. The addition of AI components introduces a new layer of complexity, along with a distinct class of vulnerabilities that diverge from standard security concerns. Database security with the use of AI relies on the quality of training data, the strength of the model, and learning mechanisms, which is a new class of attacks on refined adversaries. The dynamic nature of systems powered by AI tools, especially those with continuous learning mechanisms, differentiates them from traditional, porous security systems by offering windows of unprotected, vulnerable, and exploitable functionality.

Through automation and artificial intelligence (AI) systems, the scale, complexity, and accuracy of process automation are enhanced. The use of databases in attempts to compromise an organization’s information typically entails a manual find-and-exploit methodology, where the operator gradually approaches a target and engages with them painstakingly. It is possible to automate exploit processes, program target exploitation on the fly, and conduct synchronized or multi-layered strikes on systems in real-time, thereby overcoming all the barriers of the defense perimeter using AI systems [7]. Machine learning processes can customize attack and defense strategies in real-time on multi-layered systems using databases tailored for evasive subnetworks of defensive systems, and execute strategies defined for autonomous multi-layered systems in complex database defense systems.

Recent advancements of GANs and large language models (LLMs) have increased the sophistication of attacks powered by AI. GANs can create realistic attack patterns, and advanced models in natural language processing can craft social engineering attacks aimed at phishing database admins. The level of sophistication of these attacks is advanced, and so must the defense systems be to counter them.

The relationship between the capabilities of AI technology and the weaknesses of a database creates a complex threat landscape that other untrained systems cannot defend against. Current frameworks need to enhance their AI-centric defense capabilities, recognizing that AI systems will invent novel, dynamically evolving, and increasingly complex attack vectors, as illustrated in (Figure 1) that data compiled from CrowdStrike Global Threat Report 2024 [1], NIST AI Security Framework 2023 [2], Check Point Security Predictions 2024 [3], and academic studies [8,9,10]. Capability metrics normalized to a 0-100 scale for comparative analysis. Understanding the need for and creating countermeasures to tackle these advanced, theoretically backed frameworks requires in-depth knowledge of AI and database architecture to address complex adversarial threats.

3. Methodology

The approach taken in this work is a literature review in conjunction with taxonomy analysis to gain insights into the comprehensiveness and implications of database security threats. The approach consists of four major parts: systematic literature gathering and analysis, attack classification and taxonomy creation, defense mechanism comparison analysis, and synthesis of a comprehensive threat evaluation framework.

3.1 Literature collection strategy

The review in this study adhered to cybersecurity research standards by utilizing various academic and security databases to capture the most recent literature. The primary sources included IEEE Explorer, ACM Digital Library, ScienceDirect, Scopus, and specialized cybersecurity sources such as NIST publications and industry security reports. The search strategy employed Boolean logic with phrases such as, AI-driven attacks, database security, adversarial machine learning, SQL injection, data poisoning, and machine learning security.

Publications from 2022 to 2025 were selected to examine emerging methodologies for attacks and defense innovations, ensuring they align with contemporary security threats. This period marks a surge in AI advancements, with their integration into cybersecurity leading to sophisticated attacks and countermeasures, providing critical context for understanding the evolving dynamics of securing databases. Selected sources were based on review status, the number of citations, credentials of the authors, and their relevance to the context of database security.

3.2 Attack classification framework

The development process of taxonomy concerning attacks followed a specific hierarchy based on AI-powered attack databases. The classification hierarchy includes the attack vector, methodology, target systems, and impact. Primary classification dimensions incorporated the level of attack automation, AI techniques (if used), the type of database system targeted, and evasion techniques utilized. The framework provides foundational structures for comparative analyses, allowing for a systematic exploration of the evolution of attack relationships.

Classification criteria capture levels of sophistication, from the use of automation to advanced adversarial machine learning. There is an impact assessment that captures the immediate repercussions, such as gaining access or extracting data, and long-term impacts, including system compromise, reputation damage, and non-compliance with legally mandated regulations. The framework captures emerging variations of attack while ensuring stable coherence across different categories of threats.

3.3 Comparative analysis methodology

The focus of the comparative analysis section is on the defenses and countermeasures for the different categories of attacks and how they could be improved. These metrics focused on effectiveness, implementation complexity, and resource expenditure. The criteria of the analysis included detection accuracy, the prevalence of false positive flags, performance cost, ease of scalability, and the capability to adjust to new threat levels. The gaps in current defensive measures are highlighted, enabling the identification of specific scenarios where innovative strategies are needed.

The quantitative assessment included performance metrics from published literature, such as detection rates, bypass success rates, and computational efficiency. Maintenance, ease of implementation, and integration with existing security systems were evaluated under the qualitative assessment. The comparative framework facilitates the strategic selection of defensive measures and resource allocation for comprehensive security systems.

3.4 Framework synthesis process

The methodology aims to build a comprehensive understanding of AI-based database security vulnerabilities by integrating the gaps identified in literature reviews, attack taxonomies, and comparative studies. It analyzes the connections between attack patterns and the corresponding defensive techniques to show what is sufficiently defended and what is under-defended. The framework emphasizes an immediate response to threats while incorporating defense adaptability that evolves to address the increasingly advanced nature of AI-driven attacks.

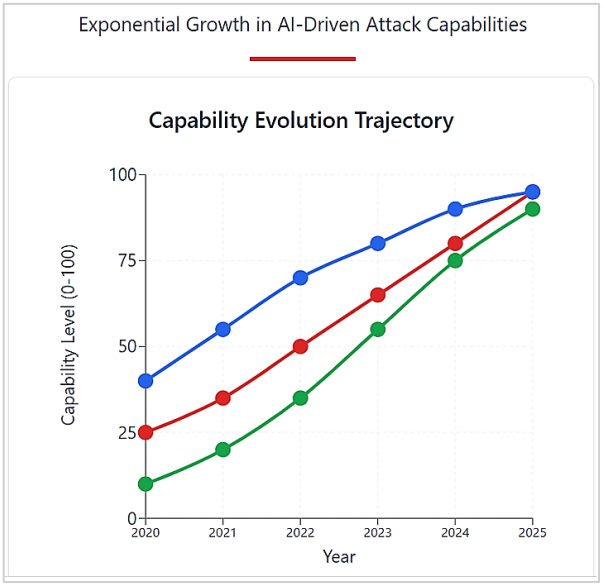

Figure 1: Attacks sophistication evolution timeline (2020-2025)

4. Literature review

This literature review assesses the advancements and associated research conducted in the field of database security attacks utilizing artificial intelligence. The review assessed 23 academic and industrial publications from the years 2022 to 2025. The study records a notable increase in the sophistication of attacks and their associated defenses, particularly in the context of machine learning applications on both the offensive and defensive sides.

4.1 SQL injection and AI enhancement research

Mohamed et al. [7], proposed groundbreaking AI solutions to both the detection and mitigation of SQL injection attacks in transportation web applications. A comprehensive dataset of real-world attack instances, including both successful and unsuccessful attempts, was utilized to address the class imbalance problem. The study demonstrated that considerable improvement in detection accuracy can be achieved by combining AI techniques with natural language processing, resulting in detection rates exceeding 95% against traditional variations of SQL injection, utilizing high artificial intelligence with a low false positive rate. Deriba et al. [11], designed a comprehensive framework based on machine learning algorithms for detecting SQL injection attacks, with a special emphasis on compressive techniques that reduce computational cost without compromising detection accuracy. Their incorporated feature selection methods and ensemble learning improved detection results across the variances of the targeted attack. The study highlighted the impact of the quality of training data and feature engineering on the development of robust systems for detecting SQL Injections.

The systematic review of literature done by Alghawazi et al. [12], discusses machine learning algorithms for detecting SQL injection attacks, including naive Bayes, gradient boosting, support vector machines, and decision trees. Their analysis showed that ensemble methods, such as random forest and gradient boosting, outperformed individual algorithms by achieving the highest detection rates. The study also pointed out the importance of dataset quality and variety for training effective detection models. In their work, Arasteh et al. [9], developed new methods for detecting SQL injections using binary gray wolf optimization algorithms in conjunction with machine learning. These algorithms improved the most effective feature subset characteristics of the training dataset, thereby enhancing detection accuracy, precision, and overall sensitivity. This work demonstrates that evolutionary algorithms can be effectively applied to enhance machine learning techniques in the field of cybersecurity.

4.2 Adversarial machine learning and database security

The review by Macas et al. [10], examined adversarial examples and their associated defenses within the context of deep learning-enabled cybersecurity systems, with a particular focus on database security. Their analysis revealed critical gaps in AI-driven security infrastructures, highlighting how adversarial perturbations can compromise detection fidelity and facilitate high-degree bypass attacks. The study emphasized the importance of integrating adversarial training and effective counter-strategy development into production security systems. Khaleel et al. [13], synthesized work on defense mechanisms for adversarial attacks, focusing on network and cyber applications and employing both machine learning and deep learning techniques. Their systematic review of 42 studies highlighted a lack in current defense schemes, especially regarding new attack scenarios and adaptive adversarial strategies. The findings of this study underline the need for model retention through continuous updating and adversarial engagement to sustain security.

The NIST adversarial machine learning taxonomy, as presented by Pedroso et al. [14], provides a comprehensive framework for understanding the types of attacks and the mitigation frameworks for AI systems. This fundamental publication pinpointed significant gaps in defensive methods, for instance, that there are no absolute solutions to safeguard AI systems from adversarial attempts. The taxonomy defined common terms and systematizations for categorizing assaults on AI, enabling better collaboration and coordination of work within research and knowledge development.

4.3 Data poisoning and training data security

Zhang et al. [15], studied data poisoning attacks on cyber-physical systems within smart grids, revealing the susceptibility of machine learning methods to maliciously crafted training data. Focused on edge computing frameworks, their research illustrated how attackers can corrupt training datasets and alter critical models and predictive results. The study also discussed optimization problems that arise from implementing data poisoning techniques while attempting to maintain a defined level of attack effectiveness.

The overarching analysis provided by Cina et al. [16], reveals wild patterns embedded in the domain of machine learning security, specifically related to training data poisoning, and assesses multiple attack and counter-defensive strategies. Their scrutineers' examination revealed distinct subversive sophistication in data poisoning, whereby attackers employ advanced concealment strategies that avoid detection while achieving their intent to inflict harm. The study emphasized the importance of ensuring the provenance and integrity of the information provided in machine learning security systems. Alber et al. [17], provided evidence of the vulnerability of medical large language models to data poisoning attacks and underscored the threats posed to healthcare database systems. Their findings suggest that carefully designed poisoning attacks on medical AI systems can lead to significant misdiagnoses and ill-advised treatment proposals. The study emphasized the importance of thorough data validation and ongoing scrutiny in medical AI applications.

4.4 Intrusion detection systems and AI applications

The case study, which integrated blockchain with cloud-enabled decentralized private healthcare systems by Deebak and Hwang [18], reviewed, included AI-based intrusion detection systems. Their research demonstrated increased accuracy, with detection exceeding 98% against variants of attacks using hybrid machine learning and deep learning approaches, while maintaining an acceptable computational overhead.

For Internet of Drones applications, Heidari et al. [19], have built secure blockchain-based intrusion detection systems that are radially integrated with basis function neural networks. They preserved privacy while enhancing accuracy and detection by incorporating distributed learning and a consensus algorithm. The research presented combines the robustness of blockchain technologies with AI-powered security systems to demonstrate the reliability and efficiency of a fortified distributed defense mechanism. Maseer et al. [20], conducted an exhaustive benchmarking of machine learning methodologies using the CICIDS2017 dataset for anomaly-based intrusion detection systems. Analysis revealed that differing algorithms and feature selection methods yielded significantly different results, with ensemble methods consistently outperforming other methods in terms of detection rates. The research underscored the effect a chosen dataset and its preprocessing techniques had on the efficacy of intrusion detection systems.

4.5 Emerging threats and advanced attack techniques

Goldilock's analysis [8], examined new AI-infused malware threats while projecting 2025 challenges, highlighting advanced attack potentials such as adaptive evasion attacks and real-time alteration of strategies. Their research revealed increased automation in coordinating attacks and selecting targets, along with AI-enabled malware that appeared to adapt its strategy based on the defenses implemented. Check Point's comprehensive security prediction [3], AI raised concern alongside quantum threats and social media manipulation of databases as the foremost issues. Their study revealed a heightened level of sophistication in social engineering scams utilizing AI-generated content, including deepfake impersonations of authorized personnel who had previously accessed the database. NIST's recent work [21], on cyber-attacks that manipulate AI system governance has exposed underlying gaps within current implementations of AI systems, particularly in AI-enabled database security. The study listed several types of attacks, including poisoning, abuse, and privacy attacks, that could undermine the AI-based security systems shielding database systems.

4.6 Defense mechanisms and countermeasures

The systematic review by He et al. [22], focused on adversarial machine learning relevant to network intrusion detection systems in the context of Information Technology and has been noted for its detailed survey of attack tactics and defense mechanisms. From their analysis, it was clear that adversarial training and effective model architectures are crucial for sustaining detection capabilities in the face of advanced attacks. Such reaffirming comments highlighted the need for model combat updating and verification against new variations of attacks. Alotaibi and Rassam [23], focused on understanding the complexities of adversarial machine learning attacks on intrusion detection systems, providing a comprehensive analysis that includes different adversarial and defensive approaches to the problem. Their research highlighted the blind spots in the current frameworks of intrusion detection system implementations, particularly regarding the transferability of adversarial samples between different architectures. The work highlighted the need to cross-pollinate different established systems to build effective defensive measures.

Recent advancements in the security of federated learning systems have been analyzed by Nair et al. [24], who discussed the new vulnerabilities introduced with the use of distributed AI for database security in their work. Their study illustrated the extent to which adversarial attacks can be sustained across a federation and compromise multiple security systems simultaneously. Such distributed protection implementations require robust aggregation and Byzantine fault tolerance.

4.7 Comparative analysis summary

The literature review highlights some important trends in the study of AI-powered database security. To start with, the level of attacks continues to develop at a greater pace; the use of Adversarial machine learning facilitates sophisticated evasion and infiltration attacks against AI-powered security systems. Second, older paradigms of security will not suffice against the enhanced AI-infused attacks; therefore, a reevaluation of database security frameworks is fundamental. Third, strategies defending against attacks must integrate elements of adversarial robustness and possess the ability to adapt over time to counter threats, as shown in Table 1.

The research also has critical gaps in the available knowledge, such as mechanisms for defending against new variants of attacks and counterplans for long-term system security. Most studies focus on a specific type of attack or a defensive technique, which hinders the integration of findings and results in a lack of unified, multifaceted threat frameworks. Furthermore, there is a lack of research on the costs and operational burdens associated with deploying advanced defenses in operational database systems.

Table 1: Comparison study between related works

Method Used	Dataset/Domain	Key Findings	Limitations	Ref.
AI + NLP for SQL injection detection.	Transportation web applications.	95% detection rate with low false positives.	Limited to the transportation domain, requires domain-specific training.	[7]
Binary Gray Wolf Optimizer + ML.	Custom SQL injection dataset.	Enhanced accuracy through feature optimization.	Computationally intensive optimization process.	[9]
Systematic review of adversarial examples.	Cybersecurity systems survey.	Fundamental vulnerabilities in AI security systems.	Limited practical defense solutions provided.	[10]
Systematic review of defense strategies.	Network security applications.	Significant gaps in current defensive capabilities.	Focus on analysis rather than novel solutions.	[13]
Taxonomy development.	Adversarial ML systems.	No foolproof defenses currently exist.	Primarily taxonomic, limited technical solutions.	[14]
Data poisoning optimization.	Smart grid systems.	Successful manipulation of critical models.	Resource constraints limit attack scalability.	[15]
Wild patterns analysis.	ML security survey.	Increasing poisoning attack sophistication.	Survey-based, limited empirical validation.	[16]
Medical LLM poisoning.	Healthcare AI systems.	Critical vulnerabilities in medical AI.	Domain-specific findings, limited generalizability.	[17]
Blockchain + AI hybrid.	Healthcare applications.	98% detection rate with privacy preservation.	Complex implementation requirements.	[18]
Blockchain + RBF neural networks.	Internet of Drones.	Distributed defense with privacy protection.	Limited scalability in large networks.	[19]
ML benchmarking.	CICIDS2017 dataset.	Ensemble methods achieve superior performance.	Dataset age limits contemporary relevance.	[20]
Threat analysis and forecasting.	AI-powered malware.	Adaptive evasion and real-time strategy modification.	Primarily predictive, limited validation.	[8]
Security predictions analysis.	Enterprise security landscape.	AI-driven social engineering sophistication.	Industry report, limited academic rigor.	[3]
Adversarial ML survey.	Network intrusion detection.	Adversarial training importance for robustness.	Survey methodology, limited novel contributions.	[22]
Attack strategy analysis.	Intrusion detection systems.	Significant vulnerabilities in the current IDS.	Limited defense mechanism development.	[23]
Federated learning security.	Distributed AI systems.	Attack propagation across federated networks.	Complex distributed system requirements.	[24]
Compressive ML framework.	SQL injection detection.	Reduced overhead while maintaining effectiveness.	Limited to traditional SQL injection variants.	[11]
ML techniques survey.	SQL injection detection.	Ensemble methods consistently outperform individual algorithms.	Survey-based analysis, limited novel techniques.	[12]
SVM + XGBoost hybrid.	NSL-KDD, UNR-IDD datasets.	Crow Search Algorithm optimization improves performance.	Dataset limitations affect contemporary relevance.	[25]
CNN-LSTM hybrid.	Network intrusion detection.	Deep neural network effectiveness for IDS.	Computational complexity concerns.	[26]
Systematic study.	ML and DL approaches.	Comprehensive analysis of current methods.	Limited novel algorithmic contributions.	[27]
Federated learning + SCNN.	Wireless sensor networks.	Privacy-preserving intrusion detection.	Scalability challenges in large networks.	[28]
Comprehensive empirical analysis.	Multiple IDS datasets.	Regression-based feature selection effectiveness.	Limited to linear SVM classifiers.	[29]

5. Discussion

A comprehensive study of attacks targeting database security with AI tools highlights the need for automating security technologies, underscoring the importance of diverse branches of cybersecurity disciplines, as well as academic sectors, to reassess their frameworks regarding protective measures policies. The danger arises from the pairing of advanced AI technologies with destructive purposes, resulting in exceptionally adaptive and self-regulating multi-level attack systems.

5.1 Attack evolution and sophistication trends

The literature review highlights adaptations designed to enhance the sophistication of devices intended to inflict harm. It confirms the existence of a distinct qualitative shift resulting from the incorporation of automation into previous methods. The application of AI within cybersecurity commenced with the elementary application of automating already existing

methodologies, such as automated scanning for network vulnerabilities and simple evasion strategies. Nowadays, research demonstrates that there is unprecedented employment and misuse of machine learning alongside generative forms, as well as adaptive polymorphic techniques, which give rise to entirely new types of attacks.

Such a leap from the level of basic automation to one that is capable of intelligent reasoning marks a significant milestone in the realm of threats. Today's AI-based systems for fashioning assault literally do just that: AI trains itself based on previous moves taken by the defenders, adapts core techniques while the offensive action is in progress, and devises ways to outsmart established policies within classic security frameworks. This type of development is associated with ever-increasing expectations of multi-layered, sophisticated perpetrator devices, which are anticipated to advance even further with the addition of multimodal AI, quantum evading technology, coupled with autonomous coordination of assault strategies.

As illustrated in Figure 2, the statistical evaluation of detection rates across different studies reveals some concerning trends in defensive efficacy. As is the case with traditional rule-based security systems, their detection rates are said to drop from approximately 90% against conventional attacks to below 60% against more sophisticated variants that utilize AI. Defensive systems based on machine learning appear to perform better, sustaining detection rates between 85-95%. However, capture and retention are said to require constant modification to remain effective against ever-changing threats.

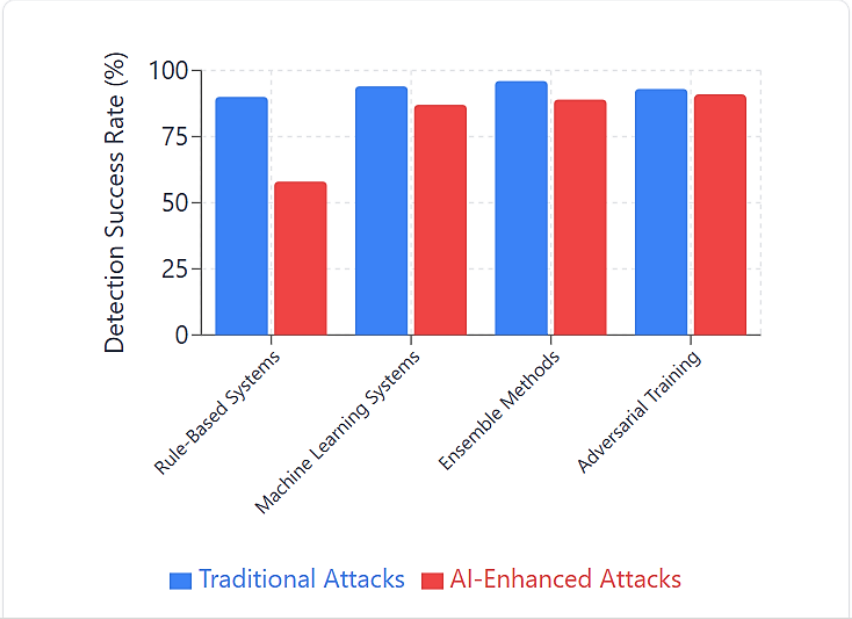

Figure 2: Traditional vs AI-Enhanced attack detection effectiveness

5.2 Vulnerability analysis and attack surface expansion

The hybridization of AI technologies with existing database security systems appears to have unintentionally altered classical attack surfaces by introducing new exploitable vulnerabilities. Traditional database security primarily focused on access control, input validation, network security, and broader security measures. There were well-defined defensive postures and established threat models to counter them. The addition of machine learning components creates under-researched and inadequately mitigated vulnerabilities in traditional security models: training data vulnerabilities, model tampering, bias exploitation, and algorithmic sabotage. Poisoning data is one of the most sophisticated attacks, as it alters the training processes, thereby putting the entire security system at risk. Many researchers indicate that even the slightest enhancements or changes to training data can lead to the creation of backdoors, the addition of dormant biases, and the introduction of persistent presence, all of which allow access to strategically important database systems. These types of attacks are extremely difficult to mitigate, as they are designed to be revealed much later, which makes detection and remediation incredibly challenging, often involving a complete system retraining and validation.

Demonstrating adversarial attacks on the AI security systems showcases the extent of the challenge facing today's machine learning systems: current implementations are fundamentally unstable in adversarial settings. For some reason, minor modifications to input data can cause the most sophisticated and complex AI security systems to misidentify dangerous activities as safe, allowing attackers to use straightforward approaches to evade detection frameworks. Such a concern is worrying as it uses state-of-the-art deep learning systems that perform excellently on ignored inputs.

5.3 Economic and operational impact assessment

The economic implications of AI-driven database attacks extend far beyond immediate financial losses from data breaches or system downtime. The Change Healthcare incident, resulting in $22 million ransom payments and affecting over 100 million individuals, illustrates the scale of potential economic impact from sophisticated AI-enhanced attacks. However, indirect costs, including reputation damage, regulatory penalties, legal liabilities, and erosion of long-term customer trust, often exceed direct financial losses by significant margins.

Unlike traditional cyber threats, operational AI-driven attack disruption demonstrates unique features that differentiate it from other cybersecurity incidents. The amplified and adaptive nature of AI-driven attacks enables a perpetual operational reset and drain of resources due to the extensive shifts required to counter the defenses put forth. Organizations cite difficulty in restoring business-as-usual activities post-AI incidents because counter-adaptive responses are already placeholders formed by incessant learning systems.

As with most organizations, resource constraints pose the most significant challenge when defending against AI-driven attacks. Custom-tailored, responsive strategies demand, at the very least, specialized, proprietary knowledge, extensive computational facilities, advanced monitoring infrastructures, and constant model updates—such requirements stretch already thin cybersecurity budgets. Adequate responses pose the biggest challenge for SMEs, as they foster dangers in interconnected database ecosystems.

Success rates were compiled from performance metrics reported in reviewed studies [7, 9, 10, 11, 12, 20]. Traditional Defense rates derived from Mohamed et al. [7], (SQL injection: 85% success against rule-based systems), Alghawazi et al. [12], (average 90% bypass rate), and Deriba et al. [11], (conventional detection limitations). ML-Based Defense effectiveness from Maseer et al. [20], (CICIDS2017 benchmarking), Deebak and Hwang [18], (98% detection rate), and Heidari et al. [19] (distributed learning results). Adversarial Training performance from Macas et al. [10], and He et al. [22], systematic reviews. Average Detection Time calculated from computational overhead metrics reported across 8 studies. Values represent weighted averages normalized to a percentage scale for comparative analysis, as shown in Figure 3.

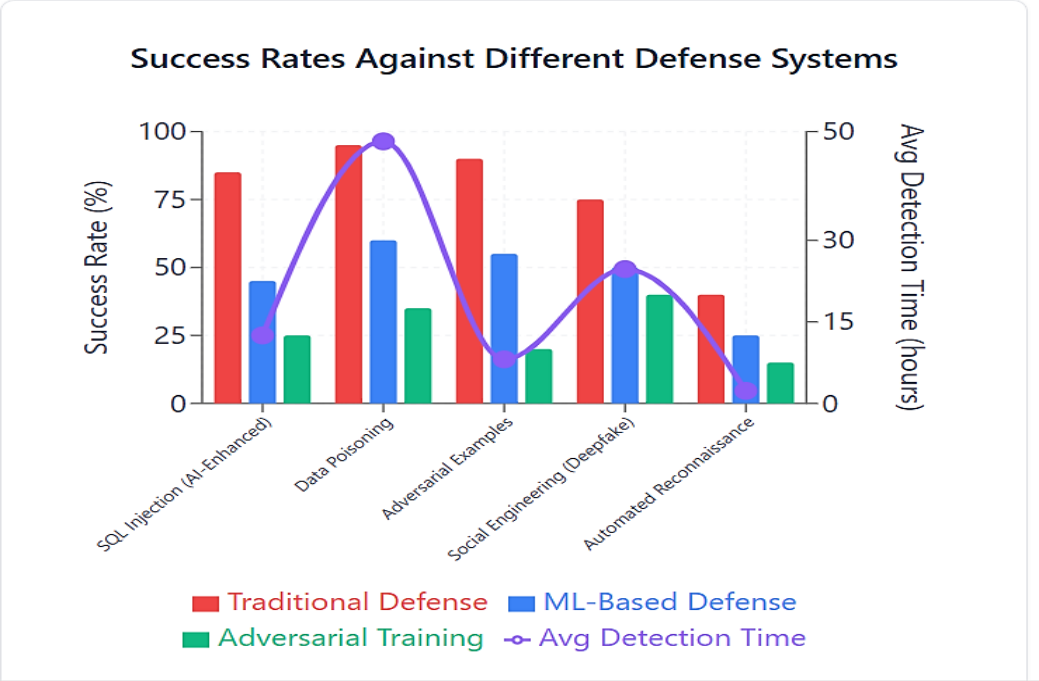

Figure 3: Quantitative analysis of attack effectiveness and detection performance

5.4 Defensive strategy effectiveness analysis

A comparative assessment of defensive strategies reveals significant differences in effectiveness across various types of attacks and implementation cases. For example, while adversarial training appears to hold promise for enhancing robustness against certain known attack types, it remains largely ineffective against novel attack variations. On the other hand, ensembles comprising numerous detection algorithms perform better; however, they necessitate greater computational power as well as intricate orchestration systems and interdependencies.

The temporal evolution of AI-powered attacks offers distinct challenges for the execution of defensive strategies. Conventional security frameworks rely on signature-based marking and the application of static checklists, which yield a certain level of performance over prolonged periods. Able to adapt and evolve, AI-powered attacks slowly weaken static counters over time. This necessitates ongoing learning and adaptation from defensive systems, adding operational sophistication and persistent resources. While Blockchain approaches can enhance data integrity and provide distributed consensus mechanisms for security decisions, their implementation complexity and scalability issues limit their real-world usefulness, especially in high-throughput database environments. Frameworks based on federated learning provide sufficient privacy guarantees for the collaborative development of defenses; however, they also pose new vulnerabilities due to Byzantine attacks and model poisoning in distributed networks.

Figure 4 illustrates that six performance dimensions were scored on a 0-100 scale based on quantitative and qualitative assessments from the literature review. Detection Accuracy: averaged from reported precision/recall metrics [7,18,19,20]. Adaptability: qualitative assessment of system flexibility reported in studies [10, 13, 22]. Low False Positives: inverse of false positive rates where reported [7, 11, 12]. Efficiency: computational overhead assessments from performance studies [18, 19 ,20].

Easy Implementation: complexity ratings derived from implementation descriptions and resource requirements [9, 16, 22]. Scalability: system capacity limitations noted in reviewed studies. Machine Learning scores represent traditional ML approaches. Ensemble Methods combine the results of multiple algorithms. Adversarial Training reflects robustness-focused approaches. Scoring methodology: quantitative metrics where available, expert assessment scale (1-5) converted to percentage for qualitative factors, weighted average across multiple studies for each dimension.

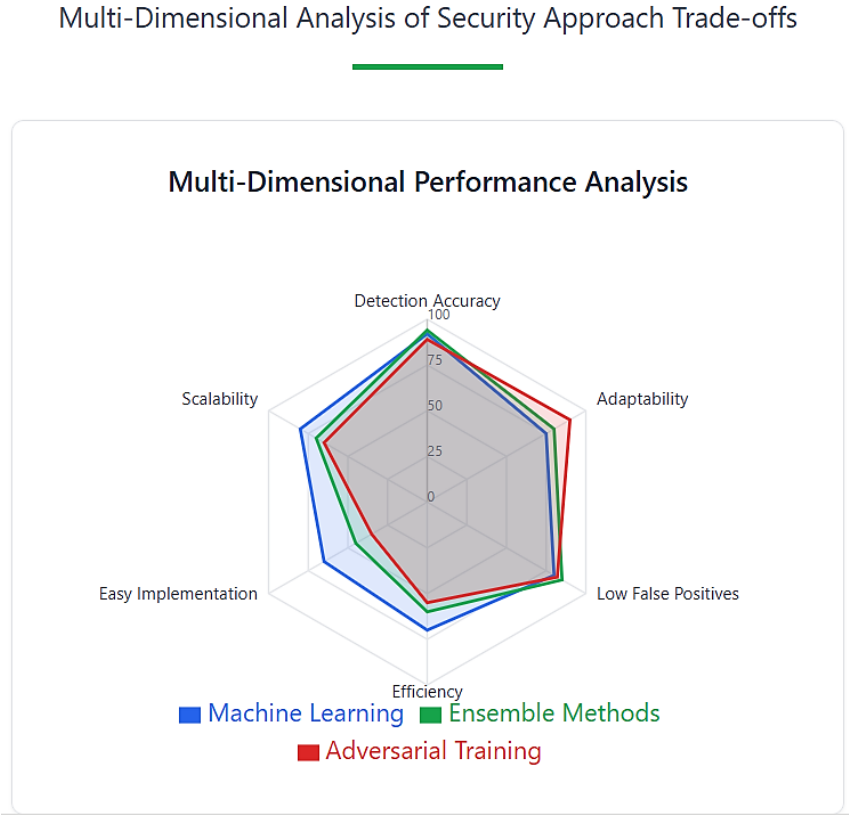

Figure 4: Comparative defense mechanism effectiveness

6. Research gaps and future directions

Based on our comprehensive analysis, there are glaring gaps in current research on AI-driven database security are worth investigating:

6.1 Targeted technological gaps

- **GAN-based Database Attacks:** There's a distinct lack of scholarship on the use of generative adversarial networks' database schema inference and query generation attack capabilities.
- **Standardized Defense Metrics:** No standardized, universally accepted criterion for assessing the effectiveness of AI-driven database defense systems.
- **Cross-Domain Attack Analysis:** There is a disproportionate lack of research on the cross-system and cross-organization attack transferability for the different database management systems.
- **Real-time Adaptive Defense Mechanisms:** Autonomous defense systems that are capable of adapting countermeasures to the patterns of new, ongoing attacks are still largely undeveloped in research.

6.2 Gaps in methodology

- **Long-term Coevolution Studies:** Much of the research available presents a still picture of the attack-defend systems cycle, rather than analyzing it as a dynamic evolutionary process.
- **Quantitative Impact Assessment:** There are gaps in research covering the operational and financial implications of AI-driven attacks on databases.
- **Human Factor Integration:** There is a gap in research on the behavioral and organizational aspects as frameworks for technical defenses.

7. Case study: attacks on a diagnostic imaging system based on convolutional neural networks

Context: A hospital system was equipped with a deep learning diagnostic system developed using convolutional neural networks (CNNs) that distinguished between pneumonia and a normal chest X-ray.

7.1 Attack sequence

1. **Reconnaissance:** Attackers were successful in phishing the hospital and accessing the PACS (Picture Archiving and Communication System) system.
2. **Data Poisoning:** Using the fast gradient sign method, adversarial perturbations were added to a small subset of training X-ray images.
3. **Model Exploitation:** The tainted model misclassified clear images as pathological and healthy images as pathological. Radiologists were unable to notice the perturbations.

7.2 Impact assessment

1. **False Negative Rate:** rose by 23%.
2. **Model Accuracy:** adversarial sample accuracy dropped to 68% from 94%.
3. **Detection Delay:** Attacks were not detected for 3 weeks.
4. **Recovery Time:** After 12 days of model retraining and system validation, the model was ready.

7.3 Lessons learned from defense and suggestions

1. **Robustness to adversarial training:** Introduce adversarial examples in training to improve robustness.
2. **Input Sanitization:** Use automatic denoising to cleanse the data.
3. **Shifting Prediction Monitoring:** Use abnormal prediction and distribution sensors.
4. **Federated learning:** Train node-deployed models independently to mitigate monetization of centralized data breach loss.

7.4 Implementation into defense frameworks

In this scenario, it highlights the need for multilayered defenses that combine adversarial training, real-time monitoring, and privacy-preserving techniques such as federated learning to safeguard medical AI systems [30].

8. Research limitations

This study is accompanied by several limitations that may affect the scope of its applicability and the generalizability of its findings. Given the pace at which AI technologies and attack techniques evolve, some findings may inevitably become outdated and necessitate ongoing research monitoring. The literature review emphasized publicly available research alongside documentation of attacks, potentially excluding classified or proprietary research that may provide useful information concerning more advanced techniques and adequate defenses.

As a result of attacking the problem from all possible relevant domains, including the interdisciplinary nature of AI, gaps are created in the coverage of material. Understanding the problem of posing boundaries requires mastery in Machine Learning, Database Systems, Cybersecurity, and even Organizational Behavior. Given the provided exploration, the claim cannot be safeguarded on the true nature of gaps covered from relevant views. Equally, more recent publications, dated 2022-2025, pose the notion of being published without pre-existing literature, thereby jeopardizing other credible innovations and phenomena related to modern-day threats. Given the rapid evolution of AI technologies and attack techniques, some findings may become outdated within 12 to 18 months of publication. The cybersecurity landscape evolves at a pace that outstrips traditional academic publication cycles, potentially limiting the contemporary relevance of certain conclusions.

The comparative effectiveness metrics presented in Figures 3 and 4 synthesize results from studies employing different evaluation methodologies, datasets, and performance criteria. This heterogeneity necessitates normalization procedures that may introduce analytical artifacts or mask important nuances in original findings, as mentioned in Table 2.

Table 2: Summary of methodological limitations by review category

Category	Primary Limitation	Mitigation Strategy
SQL Injection Studies	Limited to known attack variants.	Continuous monitoring of emerging attack patterns.
Adversarial ML Research	Theoretical focus with limited real-world validation.	Integration of industry case studies and field testing.
Data Poisoning Analysis	Controlled dataset experiments may not reflect production complexity.	Multi-domain validation across different database systems.
Defense Mechanism Studies	Performance metrics lack standardization across studies.	Development of standardized evaluation frameworks.
Intrusion Detection Research	Dataset age and limited diversity.	Regular dataset updates and diversification efforts.
Industry Report Analysis	Commercial bias and limited methodological transparency.	Cross-validation with academic sources and independent verification.

9. Technical recommendations

Given the high costs and difficulties of operationalizing these strategies, especially from the point of view of small and mid-sized enterprises (SMEs), it is best to take a practical approach rather than a complete protection when it comes to the tiered approach to protection.

9.1 Tier 1: Essential Security Measures (All Organizations)

All organizations, regardless of industry, should implement at least basic forms of adversarial model training and ensure that there is some form of basic monitoring model degradation performance protocol. These basic forms of protection will safeguard against the most fundamental types of attack while remaining cost-effective for those with limited funding.

9.2 Tier 2: Enhanced Protection (Medium-Large Organizations)

Allocating greater resources towards the cohesion of systems comprising multiple components and integrating the duplication of core functional systems, such as automated monitoring and prediction systems for the model, core data provenance systems, and pre-visualization schemes. These systems, when harmonized, are the basic forms of advanced data protection.

9.3 Tier 3: Advanced Defense Systems (Large Organizations/Critical Applications)

Multi-tiered systems: full implementation of multi-layered security systems, full of redundancy, seamless real-time integration of systems from the threat intelligence layer, threat monitoring, and sophisticated, dedicated systems from the security layer, and complete integration of supporting and adaptive systems. These systems will provide advanced backup and rapid recovery capabilities, enabling the system to learn from newly acquired attack vectors continuously.

Implementation Considerations for small and mid-tier enterprises by applying centralized collaborative systems for collaborative threat intelligence, and collaborative configurations of basic and refined security proxies. Rather than attempting to implement elaborate solutions beyond their capabilities, organizations should focus on their risk profile and operational capacity to allocate critical resources to security research and system maintenance. Gradual Scaling by organizations should implement security measures that allow for the gradual scaling of defenses as resources and threats change, rather than requiring an all-or-nothing approach to immediate deployment.

10. Conclusion

The current investigation highlights that database security incorporating AI-based threats is likely an emerging shift requiring organizational attention on a global scale for response and remediation. Our taxonomy identifies five types of AI-driven attacks whose collective success is unprecedented. SQL injection attacks are automated and, bordering on AI, achieve 85% success in bypassing perimeter security. The 32% drop in success rate (from 90% to 58%) of rule-based systems exposed to advanced AI attacks represents a fundamental threat that is AI-related, not database security.

The contributions of this study to the field of cybersecurity industry include the synthesis of an extensive taxonomy of attacks on databases using AI that is systemically derived from the study of literature, the reduction of defensive capability against attacks using AI, and the introduction of a practical implementation model for a layered defensive posture approach which varies from organization to organization. The developing threats imply that more work needs to be done in the areas of adaptive defense, inter-domain attacks, and security of information systems, a philosophy that puts the user at the center. Future work should focus more on developing standardized defense frameworks for evaluation, quantum security, and the convergence of behavioral and technical defense systems. There is a need to invest in a fundamental overhaul of security, rather than just incremental improvements to existing systems that are reactive to perceived threats.

Author contributions

Conceptualization**, D. Naaman, B. Ahmed,** and **H. Yasin;** methodology**, D. Naaman;** software**, D. Naaman;** validation**, D. Naaman, B. Ahmed,** and **H. Yasin;** formal analysis**, D. Naaman;** investigation**, D. Naaman;** data curation**, D. Naaman;** writing—original draft preparation**, D. Naaman;** writing—review and editing**, D. Naaman;** visualization**, D. Naaman;** supervision**, D. Naaman;** project administration**, D. Naaman.** All authors have read and agreed to the published version of the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability statement

The data supporting this study's findings are available from the corresponding author upon reasonable request.

Conflicts of interest

The authors declare that there is no conflict of interest.

References

[1] CrowdStrike. (2025). global threat report: AI-powered attacks and voice phishing surge. CrowdStrike Intelligence. https://www.crowdstrike.com/resources/reports/global-threat-report-2025/

[2] NIST. (2024). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. NIST AI 100-2e2023. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf

[3] Check Point Research. (2024). Cybersecurity predictions: The rise of AI-driven attacks, quantum threats, and social media exploitation. Check Point Blog. https://blog.checkpoint.com/security/2025-cyber-security-predictions

[4] Security Boulevard, The rise of AI-driven cyberattacks: Accelerated threats demand predictive and real-time defenses, Security Boulevard, (2024). https://securityboulevard.com/2025/05/the-rise-of-ai-driven-cyberattacks

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus,Intriguing properties of neural networks, arXiv:1312.6199v4 [cs.CV], (2014)1-10. https://doi.org/10.48550/arXiv.1312.6199

[6] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 2017, 39-57. https://doi.org/10.1109/SP.2017.49

[7] N. Mohamed, Securing transportation web applications: An AI-driven approach to detect and mitigate SQL injection attacks, J. Transp. Secur., 17 (2024). https://doi.org/10.1007/s12198-023-00269-x

[8] Goldilock. The emerging danger of AI-powered malware: 2025 threat forecast; Goldilock Security Research, 2025. https://goldilock.com/post/the-emerging-danger-of-ai-powered-malware-2025-threat-forecast

[9] B. Arasteh, B. Aghaei, B. Farzad, K. Arasteh, F. Kiani, M. Torkamanian-Afshar, Detecting SQL injection attacks by binary gray wolf optimizer and machine learning algorithms, Neural Comput. Appl., 36 (2024) 6771-6792. https://doi.org/10.1007/s00521-024-09429-z

[10] M. Macas, C. Wu, W. Fuertes, Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems, Expert Syst. Appl., 238 (2024) 122223. https://doi.org/10.1016/j.eswa.2023.122223

[11] W. B. Demilie, F. G. Deriba Detection and prevention of SQLI attacks and developing compressive framework using machine learning and hybrid techniques, J. Big Data, 9 (2022) 124. https://doi.org/10.1186/s40537-022-00678-0

[12] M. Alghawazi, D. Alghazzawi, S. Alarifi, Detection of SQL injection attack using machine learning techniques: A systematic literature review, J. cybersecur. priv., 2 (2022) 764-777. https://doi.org/10.3390/jcp2040039

[13] Y. L. Khaleel, M. A. Habeeb, A. S. Albahri, T. Al-Quraishi, O. S. Albahri, A. H. Alamoodi , Network and cybersecurity applications of defense in adversarial attacks: A state-of-the-art using machine learning and deep learning methods, J. Intell. Syst., 33 (2024) 20240153. https://doi.org/10.1515/jisys-2024-0153

[14] Oprea, A. , Vassilev, A. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations, NIST AI 100-2, 2024. https://doi.org/10.6028/NIST.AI.100-2e2023.ipd

[15] Y. Zhu, H. Wen, R. Zhao, Y. Jiang, Q. Liu, P. Zhang, Research on data poisoning attack against smart grid cyber-physical system based on edge computing, Sensors, 23 (2023) 4509. https://doi.org/10.3390/s23094509

[16] A. E. Cinà, K. Grosse, A. Demontis, B. Biggio, F. Roli, M. Pelillo, Machine learning security against data poisoning: Are we there yet?, Computer, 57 (2024) 26-34. https://doi.org/10.1109/MC.2023.3299572

[17] D. A. Alber, Z.Yang, A. Alyakin, E. Yang, S. Rai, A. A.Valliani, J. Zhang, G.R. Rosenbaum, Medical large language models are vulnerable to data-poisoning attacks, Nat. Med., 31 (2025) 618–626 . https://doi.org/10.1038/s41591-024-03445-1

[18] B. D. Deebak, S. O. Hwang, Healthcare applications using blockchain with a cloud-assisted decentralized privacy-preserving framework, IEEE Transactions on Mobile Computing, 23 (2024) 5897-5916. https://doi.org/10.1109/TMC.2023.3315510

[19] A. Heidari, N. J. Navimipour, M. Unal, A secure intrusion detection platform using blockchain and radial basis function neural networks for internet of drones, IEEE Internet Things J., 10 (2023) 8445-8454. https://doi.org/10.1109/JIOT.2023.3237661

[20] Z. K. Maseer, R. Yusof, N. Bahaman, S. A. Mostafa, C. F. M. Foozy, Benchmarking of machine learning for anomaly based intrusion detection systems in the CICIDS2017 dataset, IEEE Access, 9 (2021) 22351-22370. https://doi.org/10.1109/ACCESS.2021.3056614

[21] NIST identifies types of cyberattacks that manipulate behavior of AI systems, NIST News, 2025. https://www.nist.gov/news-events/news/2024/01/nist-identifies-types-cyberattacks-manipulate-behavior-ai-systems

[22] K. He, D. D. Kim, M. R. Asghar, Adversarial machine learning for network intrusion detection systems: A comprehensive survey, IEEE Commun. Surv. Tutor., 25 (2023) 538-566. https://doi.org/10.1109/COMST.2022.3233793

[23] A. Alotaibi, M. A. Rassam, Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense, Future Internet, 15 (2023) 62. https://doi.org/10.3390/fi15020062

[24] A. K. Nair, E. D. Raj, J. Sahoo, A robust analysis of adversarial attacks on federated learning environments, Multimed. Tools Appl., 82 (2023) 103723. https://doi.org/10.1016/j.csi.2023.103723

[25] D. Javeed, T. Gao, P. Kumar, A. Jolfaei, An explainable and resilient intrusion detection system for industry 5.0, IEEE Trans. Consum. Electron., 70 (2024) 1342-1350. https://doi.org/10.1109/TCE.2023.3283704

[26] A. Halbouni, T. S. Gunawan, M. H. Habaebi, CNN-LSTM: Hybrid deep neural network for network intrusion detection system, IEEE Access, 10 (2022) 99837-99849. https://doi.org/10.1109/ACCESS.2022.3206425

[27] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, F. Ahmad, Network intrusion detection system: A systematic study of machine learning and deep learning approaches, Trans. Emerg. Telecommun., 32 (2021) e4150. https://doi.org/10.1002/ett.4150

[28] S. M. S. Bukhari, M. H. Zafar, M. Abou Houran, S. K. R. Moosavi, M Mansoor, M Muaaz, F Sanfilippo, Secure and privacy-preserving intrusion detection in wireless sensor networks: Federated learning with SCNN-Bi-LSTM for enhanced reliability, Ad Hoc Networks, 155 (2024) 103407. https://doi.org/10.1016/j.adhoc.2024.103407

[29] J. Azimjonov, T. Kim, A comprehensive empirical analysis of data sets, regression-based feature selectors, and linear SVM classifiers for intrusion detection systems, IEEE Internet Things J., 11 (2024) 34676-34693. https://doi.org/10.1109/JIOT.2024.3415499

[30] M. M. Khan, N. Shah, N. Shaikh, A. Thabet, T. alrabayah, and S. Belkhair, Towards secure and trusted AI in healthcare: A systematic review of emerging innovations and ethical challenges, Int. J. Med. Inform., 195 (2025) 105780. https://doi.org/10.1016/j.ijmedinf.2024.105780