BG Pulse Journal USA
ISSN:1947-2536 http://www.bgpulseusa.com/index.php/bg

# ARTIFICIAL INTELLIGENCE IN CYBERSECURITY: MODERN METHODS OF CYBER-ATTACK DETECTION AND PREVENTION

**Temurov Sardorxon Firdavsxonovich**
High school Student of 1st General secondary school with in-depth classes in some
**J.R. Zaynalov**
Doctor of Economic Sciences, Professor
Department of Finance, Samarkand Institute of Economics and Service
Doi: https://doi.org/10.5281/zenodo.23566362

**Annotation:** This article presents an overview of modern approaches to the application of artificial intelligence in cybersecurity. The relevance of this research is driven by the increasing number of digital threats, the increasing sophistication of cyberattack methods, and the limitations of traditional information security methods. It examines the potential of machine learning, deep learning, and large-scale language models for detecting network intrusions, identifying malware, recognizing phishing messages, and automating responses to information security incidents. Particular attention is paid to the comparative characteristics of intelligent methods, issues of training data quality, false positives, and the resilience of algorithms to targeted manipulation. The prospects for the application of federated learning and explainable artificial intelligence are analyzed, as well as the need to integrate automated systems into a multi-layered information security architecture. It is argued that the effectiveness of modern solutions is determined not only by the technical characteristics of algorithms but also by the quality of organizational processes, risk management, and the participation of specialists in critical decision-making.

**Keywords:** cybersecurity, artificial intelligence, machine learning, deep learning, information security, cyberattacks, phishing, intrusion detection, large language models, data protection.

## 1. Introduction

The digital transformation of society is accompanied by the active implementation of information technology in public administration, industry, finance, education, and healthcare. Electronic document management, cloud computing, mobile applications, and remote services accelerate information processing and improve service availability. At the same time, the expansion of digital infrastructure increases the potential attack surface and creates additional risks to the confidentiality, integrity, and availability of data.

Modern cyberthreats include malware, phishing attacks, unauthorized access, exploitation of software vulnerabilities, distributed denial-of-service attacks, and ransomware attacks. A particular danger comes from multi-stage incidents, in which an attacker successively gains initial access, escalates privileges, moves within the network, and attempts to extract confidential data.

Traditional information security systems rely on signatures of known threats, predefined rules, lists of blocked resources, and access control mechanisms. These methods remain important components of protection, but their effectiveness can be reduced by the emergence of unknown threats, rapidly evolving malware, and attackers using legitimate operating system tools.

In this regard, artificial intelligence methods are attracting particular attention from researchers. Machine learning enables the identification of statistical patterns in network traffic, user behavior, and software object characteristics. Deep learning expands the capabilities of

analyzing complex data, including event sequences and natural language. Large language models can be used to analyze messages, structure incident data, and support specialists in investigating suspicious activity.

However, the use of artificial intelligence does not guarantee absolute protection. Intelligent models are subject to error, can become ineffective when the nature of the data changes, and can be subject to targeted attacks. Furthermore, attackers themselves use automated technologies to create convincing phishing messages and scale up their malicious activity.

**The aim of the study** is to systematize modern areas of application of artificial intelligence in cybersecurity and identify the main advantages, limitations, and prospects of intelligent methods for detecting and preventing cyberattacks.

To achieve this goal, the following tasks have been defined:

1. Consider the main types of modern cyber threats and the limitations of traditional means of protection.
2. Describe machine and deep learning methods used in information security.
3. To analyze the application of artificial intelligence to detect network intrusions, malware, and phishing.
4. Consider the risks associated with the security of the intelligent systems themselves.
5. To identify promising areas for improving comprehensive information security.

## 2. Materials and research methods

The study is a review study. Its methodological basis consists of an analysis of scientific publications, a systematization of existing approaches, and a comparative analysis of artificial intelligence technologies used in cybersecurity.

The information base consisted of scientific reviews on intrusion detection and phishing attack detection, as well as documents from the US National Institute of Standards and Technology (NIST) dedicated to artificial intelligence risk management and countering attacks on machine learning models.

In particular, the work of Muneer et al., published in 2024 and dedicated to a critical analysis of intelligent intrusion detection methods, a review by Kavya and Sumathi on modern phishing detection methods, as well as publications on machine learning security [1–4] were taken into account.

The comparison of technologies was carried out according to the following criteria: the nature of the analyzed data, the ability to identify previously unknown threats, requirements for training samples, computational costs, interpretability of results and resilience to changing operating conditions.

This paper does not involve conducting an independent experiment on network data. Therefore, it does not provide fictitious metrics for the accuracy, sensitivity, or efficiency of the algorithms. The conclusions are based on a qualitative analysis of published research and regulatory and methodological materials.

## 3. Modern cyber threats and limitations of traditional protection methods

**3.1. Main categories of cyberattacks.** Cyberattacks vary in their objectives, the technical means used, and the impact on an information system. Selecting an adequate defense strategy requires considering both the technical characteristics of the threat and the specifics of the infrastructure being protected. The most significant categories include:

**Malware.** Includes programs designed to gain unauthorized access, disrupt systems, steal information, or encrypt data for extortion. Some malware alters its characteristics, making it more difficult to detect using static signatures.

**Phishing attacks.** These rely on social engineering and are aimed at obtaining passwords, financial data, and other confidential information. Attackers can impersonate messages from banks, government agencies, universities, and corporate executives.

**Network intrusions.** These involve exploiting vulnerabilities, using illegally obtained credentials, and performing unauthorized actions on the network.

**Denial of service attacks** are designed to overload servers, network devices, or applications, causing legitimate users to lose access to services.

**Account attacks.** These include password guessing, using previously compromised combinations, and attempts to bypass authentication mechanisms.

**Supply chain attacks.** These involve the compromise of software components, updates, or external vendors trusted by the organization.

These threats can be combined. For example, a phishing email can be used to obtain credentials, after which the attacker gains access to the corporate network and attempts to extract confidential information.

### 3.2. Limitations of signature-based and rule-based approaches

Signature analysis involves comparing the object under investigation with known indicators of malicious activity. Its advantages include relative ease of interpretation, high effectiveness in detecting known threats, and the ability to make rapid decisions.

The main limitation of this approach is the need for appropriate signatures. If the characteristics of a malicious object change or a previously unknown attack variant emerges, the system may fail to recognize the threat.

Rule-based systems detect events that meet predefined conditions. However, a large number of rules require constant updating and tuning. Overly strict conditions can miss threats, while overly sensitive ones can increase the number of false positives.

Artificial intelligence enhances the capabilities of traditional security tools by analyzing patterns and deviations from normal behavior. However, it should not completely replace signature-based mechanisms, as their combined use can provide more robust protection.

## 4. Basic methods of artificial intelligence in cybersecurity

**4.1 Machine learning.** Machine learning is a set of techniques that enable computer systems to identify patterns in data and use them to predict or classify new observations. In cybersecurity, machine learning algorithms are used to determine whether a network connection is suspicious, whether a file is malware, and whether an email contains signs of phishing.

Depending on the formulation of the problem, three main approaches are used.

**Supervised learning.** The model is trained on labeled examples for which the correct categories are known. For example, network connections can be classified as normal or malicious. The quality of the model depends on the reliability of the labeling and the relevance of the training data to real-world operating conditions.

**Unsupervised learning.** Used when there is insufficient labeled data. Algorithms identify groups of objects and statistically unusual observations. This approach is useful for detecting anomalies, but deviations from the norm do not always indicate a cyberattack.

**Reinforcement learning.** This involves learning a behavioral strategy based on environmental feedback. In cybersecurity, this approach is being explored to support the dynamic

management of defense mechanisms and the selection of action sequences during incident response. Its practical application requires strict control, as erroneous actions can impact system availability.

Common machine learning algorithms include logistic regression, decision trees, random forests, support vector machines, and gradient boosting. The choice of a specific method is determined by data characteristics, computational requirements, and acceptable error rates.

**4.2. Deep Learning.** Deep learning is based on multilayer neural networks capable of forming complex representations of input data. These models are used to analyze sequences of events, network traffic, software objects, and text messages.

Convolutional neural networks are used to extract spatial and structural features. Recurrent neural networks, including LSTMs, are focused on analyzing sequences and temporal dependencies. Transformer-based architectures provide context processing and are used in natural language and event sequence analysis.

The advantage of deep learning lies in its ability to automatically extract features from complex data. However, training such models can require significant computational resources, large labeled samples, and additional tuning.

The difficulty of interpreting results also remains a significant issue. If a system blocks a network connection or flags a file as malicious, a specialist must understand the rationale behind the decision and be able to verify its accuracy.

**4.3. Anomaly Detection.** Anomaly detection is one of the most promising areas of application of artificial intelligence in information security.

An anomaly is an observation that significantly deviates from expected system behavior. For example, such events might include an unusually large volume of data transfers, logins at unusual times, a sharp increase in requests, or a sequence of actions that is not typical for a particular account.

Anomaly detection methods are particularly useful in situations where it's impossible to predict all potential threats in advance. However, the normal behavior of information systems varies depending on workload, employee schedules, software updates, and other factors.

Therefore, the algorithm must take into account the context of events and be regularly evaluated against new data. Otherwise, changes in normal activity may be mistakenly interpreted as attacks.

**4.4 Ensemble methods.** Ensemble methods combine the results of multiple algorithms to improve classification robustness. For example, tree-based, linear-model, and neural network solutions can be combined when analyzing network traffic. The basic idea is that different algorithms can detect different threat characteristics. Combining their results can sometimes improve detection quality compared to using a single model.

However, ensemble systems can increase computational costs and complicate interpretation. Their effectiveness should be confirmed by comparative tests on independent data, rather than assumed solely based on architectural complexity.

## 5. Using Artificial Intelligence to Detect Network Intrusions

Intrusion detection systems are designed to identify suspicious activity within a network infrastructure or on individual devices. They can analyze network packets, aggregated connection characteristics, event logs, and user activity patterns.

Traditional intrusion detection systems use signatures and rules, while intelligent solutions additionally analyze statistical patterns and deviations from normal behavior.

The process of building a machine learning system for intrusion detection typically involves several steps:

1. Collect network data and event logs.
2. Cleaning data and eliminating measurement errors.
3. Selection of features characterizing network connections and user actions.
4. Formation of training and independent test samples.
5. Training the model and tuning its parameters.
6. Evaluation of detection quality and error analysis.
7. Integration of the model into the monitoring infrastructure.
8. Quality control of the model after implementation.

The quality of a system depends on how well the data reflects real-world operating conditions. If the training set contains only a limited set of attacks, the model may not be able to recognize new scenarios effectively. Several metrics are used to evaluate the results.

**Classification accuracy** shows the proportion of correctly classified observations among all observations.

**Recall** characterizes the proportion of truly harmful events that the model was able to detect.

**The accuracy of positive predictions (precision)** reflects the proportion of truly harmful events among all events classified as harmful.

**The F1 measure** is the harmonic mean of precision and recall. For intrusion detection systems, it's important to consider class imbalance. If healthy connections significantly outnumber malicious ones, the model can demonstrate high overall accuracy while simultaneously missing a significant portion of attacks.

Therefore, the evaluation should include not only the overall classification accuracy, but also the threat sensitivity, false positive rate, detection time, and computational load.

Validating the model on data not used for training is especially important. If the test set does not reflect new operating conditions, the resulting estimates may overestimate the system's true performance.

## 6. Artificial Intelligence in Malware Detection

Malware continues to evolve, employing various methods to hide, modify code, and bypass security mechanisms. In this context, artificial intelligence can be used to analyze both the static characteristics of files and their behavior during execution.

Static analysis involves examining a software object without directly executing it. The model can take into account the file structure, libraries used, metadata, and other characteristics.

Dynamic analysis is based on observing program behavior in a controlled environment. File system access, network connections, changes to system parameters, and other actions are assessed.

Combining static and dynamic analysis allows for a more complete understanding of the object being analyzed. For example, a file that appears to be a legitimate application may exhibit suspicious activity when launched.

However, intelligent classification should not be considered an infallible mechanism. Legitimate programs can perform actions that superficially resemble malicious ones, and some malicious objects can remain inactive during analysis.

Therefore, practical systems must provide for additional checks, isolation of suspicious objects, event logging and the possibility of expert review of results.

## 7. Using artificial intelligence to counter phishing

Phishing attacks aim to obtain confidential information by deceiving users. They can be spread through email, instant messaging, social media, and fake websites.

Traditional phishing detection tools use lists of known malicious URLs, heuristics, and domain analysis. However, new fraudulent resources can appear faster than the databases can be updated.

Machine learning allows us to analyze a set of message characteristics: text content, link structure, domain characteristics, design features, and other parameters.

Natural language processing methods can detect suspicious speech patterns, signs of impersonation of official correspondence, and attempts to entice the user to take immediate action. However, a single indicator is not sufficient proof of phishing. For example, the urgent nature of a message can also be found in legitimate business communications.

A study by Kavya and Sumathi examines modern approaches to phishing detection, including machine and deep learning methods [2]. Such reviews highlight the importance of a comprehensive evaluation of algorithms, taking into account the changing characteristics of threats.

The development of generative artificial intelligence is particularly important. On the one hand, large language models make it possible to analyze suspicious messages and explain potential risks to users. On the other hand, attackers can use generative technologies to create grammatically correct and personalized fraudulent messages.

Therefore, protection against phishing should include several complementary measures: email filtering, domain verification, multi-factor authentication, access rights restrictions, and regular employee training.

## 8. Large Language Models in Cybersecurity

Large-scale language models are a class of artificial intelligence systems capable of processing and generating natural language. Their use in cybersecurity is developing in several directions.

First, such models can help specialists analyze event logs, generate brief descriptions of incidents, and identify related events from large arrays of text data.

Secondly, they are able to support the processing of suspicious activity reports, the preparation of technical documentation, and the explanation of analysis results to less experienced staff.

Third, language models can be used as an interface to access monitoring tools, incident management systems, and knowledge bases, provided that permitted actions are strictly controlled.

However, the use of such technologies is associated with particular risks. The model may generate factually incorrect explanations, misinterpret event logs, or disclose information that should not be shared with external services.

An additional danger comes from attacks via instructions contained in analyzed documents or messages. If the system doesn't separate untrusted data from control instructions, an attacker can attempt to influence the model's behavior.

Therefore, language models should not be granted unrestricted access to critical systems. Their use should be accompanied by delineation of privileges, input data validation, action logging, and human confirmation of operations that could impact the security or availability of the infrastructure.

## 9. Comparative characteristics of threat detection methods

Artificial intelligence methods differ in their requirements for data, computing resources, and the ability to explain results.

| Method | Main application | Advantages | Restrictions |
|---|---|---|---|
| Signature | Detection of | Quick and easy verification | Dependence on the |

| Method | Main application | Advantages | Restrictions |
|---|---|---|---|
| analysis | known threats | | relevance of signatures |
| Decision trees | Classification of events | Relative interpretability | Possible retraining |
| Random Forest | Network connection and file analysis | Robustness to a fraction of the errors of individual trees | Greater complexity than a single tree |
| Support vector machine | Data classification | Efficiency on specific types of features | Sensitivity to data settings and scale |
| Deep neural networks | Complex sequences and large data sets | Automatic feature extraction | High data and resource requirements |
| Anomaly detection methods | Search for unusual behavior | Possibility of identifying previously undescribed deviations | False positives for normal changes |
| Large language models | Text analysis and investigation support | Working with context and unstructured information | Erroneous conclusions and risks of data disclosure |

The table reflects the general properties of the approaches under consideration. It is not the result of an experimental comparison and does not establish a universal ranking of algorithms.

The choice of technology should be determined by the specific task. Signature-based methods remain valuable for detecting known malicious objects. Neural networks can be useful for analyzing complex event sequences, and anomaly detection algorithms can be used to identify deviations from normal behavior.

In real-life situations, the most justified approach is a combined one, in which several mechanisms complement each other.

**10. Risks and limitations of artificial intelligence in information security**

**10.1 Quality of training data.** The effectiveness of machine learning depends on the quality of the input data. Labeling errors, incomplete data, and underrepresentation of certain threat types can lead to systematic errors.

Furthermore, the behavior of information systems changes over time. New applications emerge, protocols are updated, user habits and network load patterns change. This can lead to a decrease in the quality of the model after its implementation.

To reduce risk, it is necessary to control the distribution of input data, regularly evaluate quality metrics, and revise the model when operating conditions change significantly.

**10.2 False positive and false negative results.** A false positive occurs when a legitimate action is classified as an attack. A false negative occurs when malicious activity goes undetected.

Both types of errors have practical consequences. An excessive number of false alarms increases the workload of employees and can lead to warnings being ignored. Missing a real attack creates the risk of data leakage, service disruption, and financial losses.

Therefore, system configuration must take into account the consequences of different types of errors. In critical scenarios, it is necessary not only to optimize statistical indicators but also to provide backup protection mechanisms.

**10.3. Targeted attacks on models.** Intelligent systems themselves can become targets for cyberattacks. Specifically, attackers may attempt to alter input data so that the model misclassifies a malicious object as benign.

Another risk is training data poisoning, where corrupted examples are intentionally added to the sample that can degrade the quality of the model or change its behavior.

The threats of extracting sensitive information from the model and using specially crafted queries to obtain unwanted responses are also considered.

The NIST document on machine learning attacks systematizes relevant threat types and mitigation approaches [3]. It emphasizes the need to consider model security throughout the entire lifecycle—from data preparation to exploitation.

Security should include verification of training data provenance, access control, resilience testing, and monitoring for unusual requests.

**10.4. Insufficient Explainability.** In information security systems, it is important not only to detect a threat, but also to understand why it was classified in a certain way.

Explainable AI aims to increase the transparency of results and provide the specialist with information about the factors that influenced the model's decision.

For example, the system may indicate an unusual pattern of network connections, suspicious account activity, or a combination of characteristics of a phishing message.

However, a model's explanation is not always a complete or completely reliable reflection of its internal logic. Therefore, such tools should be viewed as a means of supporting analysis, complementing the verification of underlying events.

**10.5. Confidentiality and data protection.** Network logs, user activity data, message texts, and other data containing confidential information may be used to train and operate intelligent systems.

Transferring such materials to external services creates additional risks. Organizations must define acceptable data categories, control access rights, establish storage rules, and ensure information deletion in accordance with applicable requirements.

One possible approach is federated learning. In this model, multiple participants collaborate to train an algorithm without transmitting the source data to a single, centralized repository. This approach can reduce the need to share sensitive data, but it does not eliminate all risks of model leakage or manipulation.

## 11. Integration of artificial intelligence into a comprehensive security system

Artificial intelligence is most effective when integrated into existing information security infrastructure.

One of the key components of such an infrastructure is security information and event management (SIEM) systems. They collect and analyze event logs from various sources, enabling the identification of interconnected incidents.

Intelligent models can complement such systems by identifying unusual patterns of events and helping to prioritize alerts.

Another area of focus is integration with automated incident response (SOAR) systems. These solutions enable the organization of a sequence of actions when a threat occurs: collecting additional information, generating notifications, transferring the incident to the responsible specialist, and executing pre-approved procedures.

Automation should be commensurate with the risk level. For example, automatically generating a review request is usually less dangerous than immediately blocking a critical server. The more serious the potential consequences of an action, the higher the degree of control and the requirement for specialist confirmation.

To organize comprehensive protection, it is advisable to provide for the following stages of implementation:

1. Identifying critical information assets and potential threats.
2. Selection of tasks for which the application of artificial intelligence has practical value.
3. Checking the quality, availability and admissibility of data use.
4. Testing the model on an independent sample.
5. Comparison of results with existing protective measures.
6. Pilot implementation with limited powers.
7. Performance monitoring and error logging.
8. Regular robustness testing and model revision.

This approach helps reduce the risk of unfounded trust in automated solutions and ensure the gradual integration of new technologies.

## 12. Practical significance for healthcare organizations and institutions

Cybersecurity is especially important for organizations that process sensitive information, including healthcare institutions, educational institutions, and research centers.

In healthcare, information systems are used to store electronic medical records, laboratory results, diagnostic images, prescription information, and administrative documentation. A breach of confidentiality or accessibility of such data can have not only financial but also organizational consequences.

Artificial intelligence can be used to detect unusual access to medical databases, identify suspicious logins to information systems, and analyze emails sent to employees.

For example, a sudden, massive influx of access to medical records from an account that typically handles a limited amount of data may warrant additional investigation. However, such activity doesn't necessarily indicate an attack: it could be related to a legitimate research task or a change in the employee's job responsibilities.

Therefore, the system must consider the work context, access rights, and the justification for actions. Automatically restricting access without verification can disrupt the work of medical staff and impede the provision of services.

In educational institutions, similar technologies can help detect compromised accounts, suspicious access to information resources, and fraudulent messages imitating official correspondence.

Practical implementation should be accompanied by access control, backups, multi-factor authentication, employee training, and the development of incident response procedures.

Thus, artificial intelligence should be considered as an additional tool for enhancing security, which does not replace the need to comply with organizational and technical requirements.

## 13. Prospects for the development of intelligent cybersecurity methods

The further development of artificial intelligence in information security is associated with several directions.

**The first area is improving the robustness of models.** Methods are being developed to reduce the impact of specially prepared input data and maintain detection quality as the nature of threats changes.

**The second area is explainable artificial intelligence.** Improving the interpretability of decisions can facilitate incident investigations and improve interactions between automated systems and human experts.

**The third area is federated learning.** Joint training of models without necessarily pooling source data is of interest to organizations that must comply with privacy requirements.

**The fourth area is the integration of large language models.** Such technologies can expand capabilities for analyzing text data, generating reports, and supporting specialists during incident investigations. However, their authority and access to data must be strictly controlled.

**The fifth area is automated event correlation.** Combining information from multiple sources can facilitate the earlier detection of complex, multi-stage attacks.

**The sixth area is improving performance evaluation.** Comparing methods requires reproducible experiments, independent test samples, and evaluation under conditions similar to real-world use.

Standardization of risk management processes is a key requirement for the industry's development. The NIST AI RMF 1.0 document proposes a risk management framework for artificial intelligence, aimed at increasing reliability and accountability in the design and use of intelligent systems [4].

It should be noted that risk management framework documents do not replace technical testing and specialized requirements for the protection of specific information systems.

## 14. Discussion of the review results

The approaches considered show that artificial intelligence expands the information security toolkit, but does not eliminate the fundamental uncertainty associated with the variability of cyber threats.

Machine learning can improve event classification and pattern detection, while deep learning provides the ability to analyze complex data and sequences. Large language models can assist in text processing and structuring incident information. However, neither class of technology guarantees the detection of all attacks.

A comparison of methods reveals the lack of a universal solution equally effective for all tasks. An algorithm that demonstrates good results in classifying known network connection types will not necessarily be effective in identifying new attacks or analyzing natural language.

An additional complication is the discrepancy between laboratory conditions and real-world use. Models may demonstrate strong performance on a test set, but lose effectiveness when the data distribution changes, new software components are introduced, or user behavior changes.

Therefore, the evaluation of intelligent systems must include validation on independent data, error analysis, computational cost control, and an assessment of the consequences of erroneous decisions. Critical systems require backup protection mechanisms and the ability to manually intervene.

The practical value of artificial intelligence is determined not only by the quality of the algorithm but also by how well it is integrated into security management processes. If the model's alerts are not accompanied by clear verification and response procedures, an increase in the number of detected anomalies will not necessarily lead to improved organizational security.

Thus, a promising direction is the combination of intelligent analytics, traditional protection mechanisms, organizational measures and expert control.

## 15. Conclusion

Artificial intelligence is a promising area for the development of modern cybersecurity systems. Machine and deep learning methods make it possible to analyze large amounts of network data, detect unusual activity, classify malware, and recognize signs of phishing emails.

The conducted review allows us to formulate the following conclusions.

1. Traditional signature-based and rule-based defenses remain important, but they need to be supplemented by methods that can identify statistical patterns and deviations from normal behavior.
2. The choice of an artificial intelligence algorithm should be determined by the characteristics of a specific task, data availability, acceptable error rate, and computing resource requirements.
3. The effectiveness of intelligent systems depends on the quality of training data, the correctness of testing, and the ability of the model to maintain operability under changing operating conditions.
4. False positives, lack of explainability, and targeted attacks on models remain significant limitations that must be considered when implementing automated protection.
5. The most feasible approach seems to be an integrated one that combines artificial intelligence with traditional security tools, access control, event monitoring, and expert control.
6. In organizations that process sensitive information, including healthcare and educational institutions, the implementation of intelligent systems must be accompanied by compliance with data protection requirements and an assessment of the potential consequences of erroneous decisions.

Future research opportunities include developing robust and interpretable models, improving performance evaluation methods, and exploring the potential for securely integrating large language models into information security infrastructure.

In general, artificial intelligence should not be viewed as a universal replacement for existing security mechanisms, but rather as an additional tool whose effectiveness is determined by the quality of technical implementation, process organization, and responsible risk management.

## Bibliography

1. Muneer A. et al. A Critical Review of Artificial Intelligence Based Approaches in Intrusion Detection: A Comprehensive Analysis. *Journal of Engineering* . 2024. DOI: https://doi.org/10.1155/2024/3909173
2. Kavya S., Sumathi D. Staying ahead of phishers: a review of recent advances and emerging methodologies in phishing detection. *Artificial Intelligence Review* . 2025. Vol. 58, Article 50. DOI: https://doi.org/10.1007/s10462-024-11055-z
3. Vassilev A., Oprea A., Fordyce A., Andersen H., Davies X., Hamin M. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. NIST AI 100-2e2025. National Institute of Standards and Technology, 2025. DOI: https://doi.org/10.6028/NIST.AI.100-2e2025
4. Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. National Institute of Standards and Technology, 2023. DOI: https://doi.org/10.6028/NIST.AI.100-1
5. Pinto Neto E.C., Iqbal S., Buffett S., Sultana M., Taylor A. et al. Deep learning for intrusion detection in emerging technologies: a comprehensive survey and new perspectives. *Artificial*

*Intelligence Review* . 2025. Vol. 58, Article 340. DOI: https://doi.org/10.1007/s10462-025-11346-z
6. Kytidou E., Tsikriki T., Drosatos G., Rantos K. Machine learning techniques for phishing detection: A review of methods, challenges, and future directions. 2025. DOI: https://doi.org/10.1177/18724981251366763
7. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. 2024. URL: https://www.nist.gov/itl/ai-risk-management-framework
8. MITER. MITER ATT&CK: Adversarial Tactics, Techniques, and Common Knowledge. URL: https://attack.mitre.org/
9. ISO/IEC 27001:2022. Information security, cybersecurity and privacy protection - Information security management systems - Requirements. International Organization for Standardization, 2022.
10. European Union Agency for Cybersecurity (ENISA). ENISA Threat Landscape 2024. European Union Agency for Cybersecurity, 2024.