

Improving the artificial intelligence model for real-time detection of cyberattacks and automated response based on digital technologies

T.Sh.Jumaniyozov

University of Public Safety of Republic of Uzbekistan

Abstract: *The article analyzes, from the standpoint of artificial intelligence models, the problem of detecting cyberattacks in real time and responding to them in an automated manner. Drawing on the 2025-2026 global reports of IBM and the Ponemon Institute, the regulatory and legal acts of the Republic of Uzbekistan in the field of cybersecurity, and national and foreign scientific sources, the gap between the propagation speed of an attack and the response speed of an organization is assessed quantitatively. The novelty of the model lies in combining, within a single architecture, a cost-sensitive reward function, a concept drift monitoring mechanism, an adversarial robustness module, and a federated learning scheme among sectoral SOCs. The stages of implementing the proposed solution under the conditions of Uzbekistan are aligned with the deadlines of the Cybersecurity Strategy for 2026-2030.*

Keywords: *cybersecurity, artificial intelligence, real-time mode, attack detection, automated response, deep reinforcement learning, explainable artificial intelligence, cyber resilience, critical information infrastructure*

Introduction

The deep penetration of digital services into public administration, the banking and financial system, energy, and transport has turned cybersecurity from a technical issue into a matter of national security. The defining feature of today's attacks is manifested not in their growing number but in their acceleration: the interval between the stages of reconnaissance, exploitation of a vulnerability, and data exfiltration has contracted to the point where the traditional response scheme - the one driven by the operator's hand - can no longer physically keep up.

The quantitative measure of this gap is clearly visible in international reports. According to the report published by IBM and the Ponemon Institute in July 2026, based on data studied at 602 organizations across 16 countries and 17 industries, the average cost of a single breach reached USD 4.99 million, rising by 12 percent within a year; in the United States this figure amounted to USD 11.5 million. The number of breaches in which the adversary employed artificial intelligence grew by 56 percent over the year and exceeded one quarter of all targeted attacks, while their average cost stood at around USD 6 million - approximately USD 1 million higher than that of conventional attacks. At the same time, organizations that made extensive use of artificial intelligence and automation in their security operations saved an average of USD 1.93 million and shortened the attack life cycle by 65 days.

It is precisely these two figures - the 56 percent growth on the attacker's side and the USD 1.93 million in savings on the defender's side - that define the scientific essence of the topic: today artificial intelligence is a weapon of both sides in cyberspace, and therefore improving detection and response models is not merely a technological upgrade but a condition for keeping pace in this competition.

For Uzbekistan, the relevance of the issue arises from two aspects. The first is statistical. According to data from the Ministry of Internal Affairs and the Cybersecurity Center, the number of cybercrimes recorded in the country rose from 863 in 2019 to 58,800 in 2024, that is, a 68-fold

increase over five years; the damage to citizens exceeded 1.9 trillion soums, 98 percent of which relates to fraud involving bank cards. The load at the level of the national segment is also considerable: during the first five months of 2024, more than 6,615,900 cyberattacks were observed against websites in the national segment of the internet, 183 vulnerabilities were identified and eliminated on 60 web resources of state bodies and organizations, and cybersecurity incidents were recorded on 74 websites in the «UZ» domain. This amounts to an average of more than 43,000 attack attempts per day (author's calculation), and analyzing such a volume manually is practically impossible.

The second is institutional. While the Law «On Cybersecurity» (15 April 2022, No. LRU-764) established the fundamental concepts and principles of the field, Presidential Resolution No. PR-153 of 30 April 2025 strengthened the institutional system for combating crimes committed with the use of information technologies. The Cybersecurity Strategy for 2026–2030, approved by Presidential Decree No. PD-38 of 10 March 2026, brings the issue directly into the field of the present research: the Strategy defines as separate tasks the creation of a national system for monitoring and responding to cybersecurity incidents - together with central and sectoral SOC's - the establishment of systems for the early detection of cyber threats and for the rapid exchange of information about them, and the introduction of artificial intelligence technologies among the priority areas of research in the field of cybersecurity.

1. The evolution of detection paradigms and their limitations

The evolution of attack detection tools can be divided, conditionally, into three generations. The first generation comprises signature-based IDS/IPS systems, which operate by matching traffic against known attack patterns and therefore remain almost ineffective against previously unrecorded attacks. The second generation consists of statistical anomaly detection systems; these build a profile of the «normal» behavior of the network and flag deviations from it, but they give rise to the problem of a high rate of false positives. The third generation comprises adaptive systems based on machine learning and deep learning algorithms, which are able to retrain the model on new data and adapt to a changing environment.

Table 1.

Comparative analysis of cyberattack detection paradigms

Paradigm	Core mechanism	Sensitivity to zero-day attacks	False positive rate	Response mode
Signature-based IDS/IPS	Matching against known patterns	Very low	Low	Manual or simple rule-based
Statistical anomaly	Measuring deviation from the «normal» profile	Medium	High	Alerting
Classical machine learning (RF, SVM, GB)	Classification by predefined features	Medium	Medium	Semi-automated
Deep learning (CNN-LSTM, autoencoder)	Learning spatio-temporal dependencies	High	Below medium	Semi-automated, with recommendation

DRL + SOAR (proposed approach)	Optimizing the decision policy on the basis of a reward	High	Low (owing to the cost-sensitive reward)	Conditionally autonomous, tiered
-----------------------------------	---	------	--	----------------------------------

2. Architecture of the improved model

The proposed model consists of five layers, and its central idea is to design detection and response not as separate modules but as a single closed control loop.

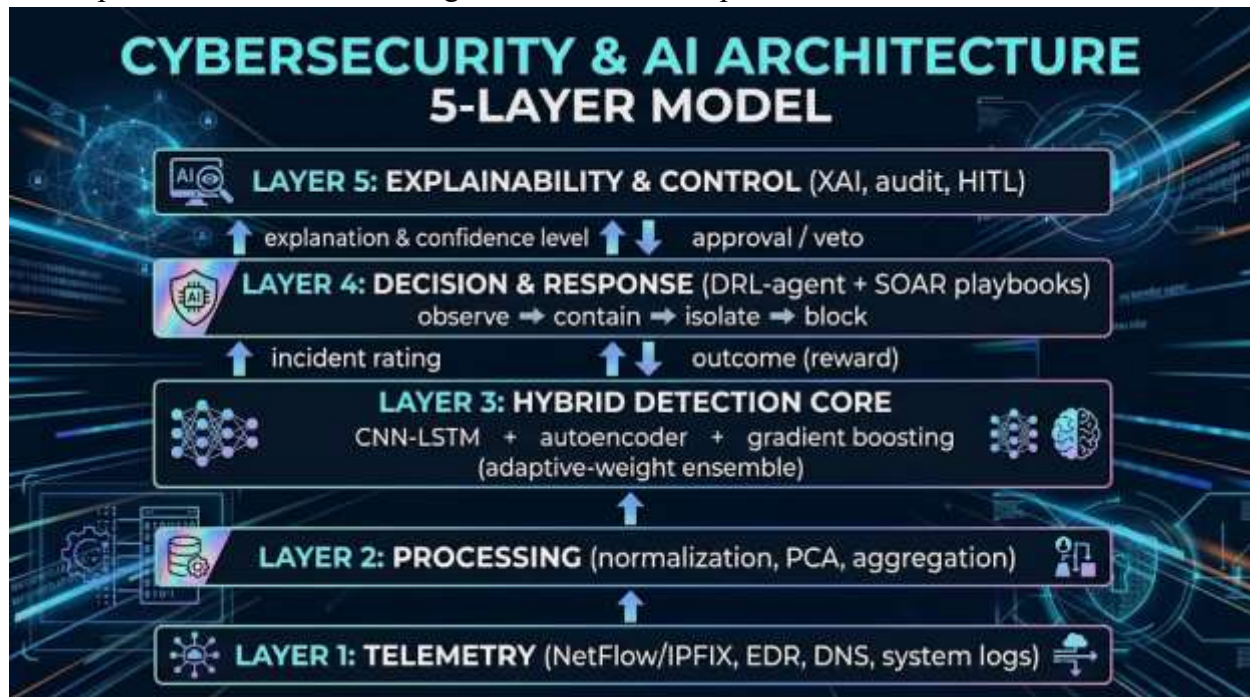


Figure 1. General block diagram of the model

Table 2.

Technological content of the model layers

Layer	Technological solution	Input data	Output result
1. Telemetry	NetFlow/IPFIX collection, EDR agents, DNS and authentication logs	Stream of network and host events	Unified event stream with timestamps
2. Processing	Normalization, feature extraction via PCA, windowed aggregation	Raw telemetry	Feature vector
3. Detection core	CNN-LSTM (spatio-temporal analysis), autoencoder (zero-day anomalies), gradient boosting (known classes)	Feature vector	Event class and confidence level
4. Decision and response	DRL agent (DDPG type) + SOAR playbooks, tiered action space	Event class, asset criticality	Automated or approval-based action
5. Oversight	XAI explanations, audit trail, human oversight (human-in-the-loop)	Decision and its rationale	Explained decision, reward signal

3. Levels of response automation

In practice, the notion of an «automated response» is understood in different ways, and the response levels are therefore classified explicitly in the model. The classification rests on the criteria of the irreversibility of an action and the criticality of the asset.

Table 3.

Levels of response automation

Level	Content	Role of the human	Recommended area of application
A0	Manual analysis and manual measures	Full control	Rare, previously unseen incidents
A1	Automated detection and alerting	Decision-maker	Low-confidence signals in critical segments
A2	A measure is recommended, the operator approves it	Approver	Assets of medium criticality
A3	Low-impact measures automated, high-impact measures subject to approval	Supervisor holding the right of veto	Mass phishing, brute force, botnet activity
A4	Fully autonomous within defined scenarios	Subsequent audit	DDoS filtering, blocking of malicious IPs and domains

The point of the tiered approach is that the level of automation need not be uniform across an organization: the same type of incident may be handled at level A4 in a test environment and at level A2 in a payment system. It is precisely this differentiation that increases trust in automation and reduces the organizational resistance expressed in the question «what happens if the automation makes a mistake?»

4. Performance criteria and training data

Two groups of criteria are applied in evaluating the quality of the model. The first group concerns classification quality: precision, recall, the F1-score, ROC-AUC, and the share of false positives. The second group comprises operational criteria: the mean time to detect an attack (MTTD), the mean time to respond to it (MTTR), the time required to contain an incident fully (MTTC), the share of automatically closed incidents, and the number of incidents per analyst. The criteria of the second group are no less important than those of the first, because the economic effect materializes precisely through the time indicators.

As the training data, it is advisable to use open benchmark datasets - including CICIOT2023, with a volume of 8.94 GB collected from 105 real IoT devices, and the flow-based CIC IoT-DIAD 2024 dataset - together with anonymized telemetry from the national segment. The problem of class imbalance (the scarcity of attack samples) is mitigated through a weighted loss function and the generation of synthetic samples. An important methodological requirement is not to confine the training of the model to open datasets alone: the traffic profile of the national infrastructure, the set of systems in use, and the attack vectors differ from those of international benchmark datasets.

5. Quantitative analysis of global dynamics

Two consecutive reports by IBM and the Ponemon Institute clearly show the direction in which the economics of cybersecurity is changing in the age of artificial intelligence.

Table 4.

Dynamics of global indicators in 2025 and 2026

Indicator	2025 report	2026 report	Change
Average cost of a single breach, USD million	4.44	4.99	+12%
Average cost in the United States, USD million	10.22	11.5	+12.5%*
Share of breaches in which the attacker used AI	16%	more than 25%	+56% growth
Incidents targeting an organization's own AI models	13%	21%	+8 p.p. (relative +61%)*
Incidents related to ungoverned AI («shadow AI»)	20%	43%	2.15-fold*
Savings owing to AI and automation, USD million	1.9	1.93	consistently high
Reduction of the attack life cycle (when AI is used)	80 days	65 days	–15 days
Absence of measures controlling access to AI	97%	92%	–5 p.p.

* *Author's calculation.*

Two conclusions follow from the table. The first is that the advantage of the attacking side is still growing: attacks employing artificial intelligence increased by 56 percent within a year and added approximately USD 1 million to the cost of an average breach; this means that the cost is nearly 21 percent higher than that of a conventional targeted attack (author's calculation: the ratio of USD 6.04 million to USD 4.99 million). The second is that the effect of automation on the defensive side is stable: savings of USD 1.9 million were recorded in 2025 and USD 1.93 million in 2026. At the same time, the fact that the reduction of the attack life cycle fell from 80 days to 65 days shows that the reserve of «easy wins» from automation is being exhausted and that further gains can come only from qualitative improvements to the models.

The composition of threat sources also deserves attention. In the 2025 report, 37 percent of the attacks carried out with the help of artificial intelligence involved phishing messages generated by a generative model, while 35 percent involved impersonation through deepfakes. In the 2026 report, impersonation through deepfakes accounted for the largest share, followed by malware enhanced with artificial intelligence and by generative phishing campaigns; phishing, for its part, remained the most widespread initial attack vector for the fourth consecutive year. This means that, in designing the model, particular attention must be paid to modules for content analysis and for monitoring identification flows.

6. Comparative assessment of detection quality indicators

Among the quantitative results reported in the literature, the indicators of the IMO-DDPG model (accuracy 94.5%, recall 94.0%, precision 92.5%, F1-score 93.5%) may be taken as baseline values. These values can be represented in percentage terms in the form of a diagram as follows:

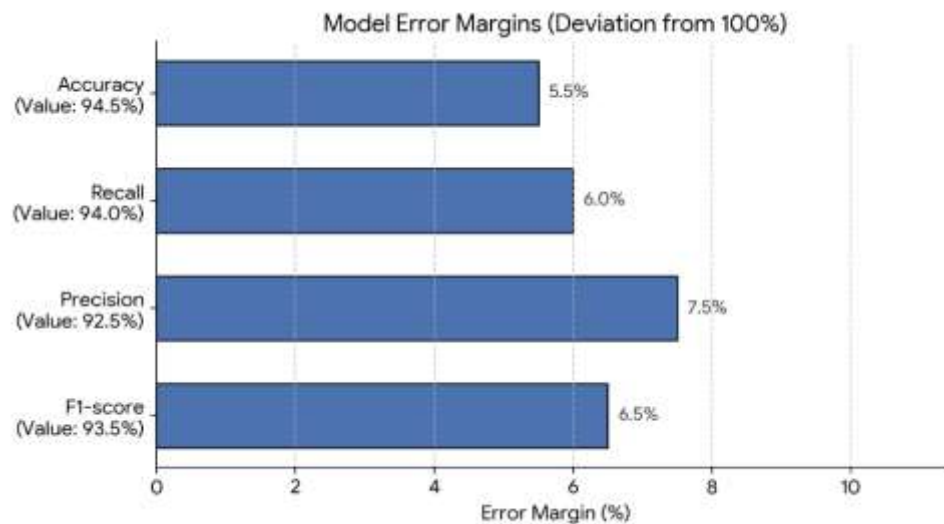


Figure 2. Indicators of the IMO-DDPG model and the corresponding error margins

Two aspects must be distinguished in assessing these results. First, an accuracy of 94 percent is regarded as a high indicator in academic testing; in practice, however, under the conditions of an imbalanced stream, an error share of 5–6 percent translates into thousands of false signals in an environment with tens of thousands of events per day. Given that, on the basis of the national data cited above, an average of more than 43,000 attack attempts per day is observed, even a 1 percent false positive rate gives rise to hundreds of futile investigations each day. For this reason, the main direction for improving the model should not be raising overall accuracy but increasing precision while maintaining high recall.

Second, detection quality indicators cannot be equated directly with operational effect. Even if an organization runs the model at 94 percent accuracy, the attack life cycle will not shorten while the decision-making process remains in human hands. The 2025 report recorded that the global average time to identify and contain a breach was 241 days, the lowest figure in nine years. If a reduction of 65 days achieved through artificial intelligence and automation is applied to that baseline, the duration falls to around 176 days, that is, by approximately 27 percent. This provides a realistic benchmark showing the true order of magnitude of the contribution made by automation.

7. The model itself as an object of threat

As artificial-intelligence-based protection is deployed, a new attack surface emerges. The data in the 2026 report confirm this thesis: 92 percent of the organizations that suffered an AI-related breach lacked appropriate measures for controlling access to the model; the share of incidents associated with the ungoverned use of artificial intelligence («shadow AI») rose within a year from 20 percent to 43 percent, that is, by a factor of 2.15. In addition, 53 percent of the organizations did not apply encryption when storing and transmitting confidential data, while the share of those using artificial intelligence agents for vulnerability management tasks amounted to only 18 percent.

The practical conclusion is as follows: deploying a detection model requires, at the same time, a threefold line of defense - role-based control of access to the model and to its training data, logging and auditing of the model outputs, and testing of the model against targeted compromise attempts (data poisoning, evasion, model inversion, prompt injection). Otherwise, the protection tool itself may turn into the most trusted point of entry.

8. Prospects for implementation under national conditions

The Cybersecurity Strategy for 2026–2030 and the «Road Map» for its implementation create a ready institutional framework for deploying the proposed model. In the table below, the stages of the model are linked to the deadlines of the Strategy.

Table 5.

Correspondence of the model implementation stages to the deadlines of the Strategy

Stage	Content	Related strategic measure	Deadline
1	Deploying the telemetry collection and data processing layers at sectoral centers	Defining the structure of central and sectoral SOC's	December 2026
2	Training the detection core on benchmark datasets and adapting it to national telemetry	Formulating research topics in the fields of AI, quantum technologies, and blockchain	September 2026
3	Testing the module for the automated detection of phishing and fraudulent resources	Anti-phishing platform and information system for blocking fraud	September 2026
4	Implementing the response layer at levels A1–A2 and testing the playbooks	Setting requirements for the national monitoring and response system	March 2027
5	Transition to level A3, launch of the federated learning scheme	Research and development program for 2026–2030	January 2027
6	Full integration within the national system and a continuous retraining cycle	Implementation of the national monitoring and response system	December 2029

The Strategy provides that a dedicated electronic platform for competitions involving independent experts in the discovery of vulnerabilities is to be created by 1 October 2026, that a cyber laboratory intended to provide information and technical support for operational activities is to be established within the Ministry of Internal Affairs, and that an online cooperation information system is to be introduced among state bodies, law enforcement agencies, banks, and telecommunications operators for the immediate identification and blocking of fraudsters. These measures form the data source and the deployment field for precisely the first and the third layers of the proposed model.

At the same time, several serious limitations must be acknowledged.

The first is data quality: however perfect the model may be, it will not work in an environment where telemetry is not collected in full, where timestamps are not synchronized, or where logs are not retained;

The second is the question of personnel: automation at levels A3–A4 demands separate qualifications in designing playbooks and monitoring the model, and the Strategy envisages meeting this need through an attestation system and international certification mechanisms;

The third is a legal question: a decision to restrict a service or to block an account automatically entails legal consequences, and the limits of liability and the appeal procedure for every autonomous action must therefore be defined in advance in departmental regulations;

The fourth limitation is methodological. The high indicators reported in the literature are, for the most part, results obtained on benchmark datasets; reproducing them in the national infrastructure requires re-evaluating the model on local traffic, conducting A/B comparisons in a controlled test environment (test-bed), and measuring the cost of false signals separately. Otherwise, claims about the effectiveness of the system remain unsubstantiated.

Conclusion

The analysis conducted shows that the principal feature of modern cyberattacks is determined by their speed and by the active use of artificial intelligence rather than by their volume. Traditional protection mechanisms therefore need to be improved with automated and intelligent tools. The improved model combines a cost-sensitive reward function, concept drift monitoring, adversarial robustness, and federated learning mechanisms within a single architecture. Response automation should be differentiated across levels A0–A4, taking into account the criticality of the asset and the degree of irreversibility of the action. At the same time, a protection system based on artificial intelligence must itself be protected against attacks on the model, with access control and auditing of its outputs. Uzbekistan has the requisite legal and institutional basis for the development of cybersecurity, and a national SOC, the early detection of threats, and the development of artificial intelligence technologies are important directions in this regard.

References

1. Law of the Republic of Uzbekistan «On Cybersecurity», 15 April 2022, No. LRU-764 // National Database of Legislation. – URL: <https://lex.uz/uz/docs/-5960604> (accessed: 18.09.2026).
2. Resolution of the President of the Republic of Uzbekistan «On measures to further strengthen activities to combat crimes committed with the use of information technologies», 30 April 2025, No. PR-153 // National Database of Legislation. – URL: <https://lex.uz/uz/docs/-7511145> (accessed: 18.09.2026).
3. Decree of the President of the Republic of Uzbekistan «On establishing the Cybersecurity Strategy of the Republic of Uzbekistan and improving the system for preventing cybercrime», 10 March 2026, No. PD-38 // National Database of Legislation, 11.03.2026, No. 06/26/38/0219. – URL: <https://lex.uz/uz/docs/-8079286> (accessed: 18.09.2026).
4. Axmedov A. M. The operating structure and algorithms of the tools currently in wide use for the detection of and protection against network attacks. – 2025. – P. 1368–1381. – DOI: 10.5281/zenodo.15280797. (In Uzbek).
5. Suyumov J. Yu., Mamarasulov H. Z., Xabibullayev Z. M. Methods of enhancing network security on the basis of artificial intelligence // SCHOLAR. – 2025. – Vol. 3, No. 15. – P. 26–31. – ISSN 2181-4147. – DOI: 10.5281/zenodo.17940097. (In Uzbek).
6. Xakimov M. I. Artificial intelligence in cybersecurity: real-time detection of phishing and DDoS attacks for the banking and government websites of Uzbekistan // International Conference on Artificial Intelligence and Applications (ICAIA). – 2026. – P. 184–187. – DOI: 10.5281/zenodo.20319625. (In Uzbek).
7. IBM Security, Ponemon Institute. Cost of a Data Breach Report 2026. – Armonk, NY: IBM Corporation, 2026. – URL: <https://www.ibm.com/reports/data-breach> (accessed: 18.09.2026).
8. Ma Y., Li C., Wang Y., Wang Y. Application of deep reinforcement learning algorithms for automatic threat detection and response in dynamic network environments to improve cybersecurity // Journal of Computational Methods in Sciences and Engineering. – 2025. – Vol. 25, Iss. 3. – P. 2112–2125. – DOI: 10.1177/14727978241309550.
9. Jumaniyozov T. S. O. G. L. OLIY HARBIY TA'LIM MUASSASALARIDA O'QUV JARAYONINI YANADA TAKOMILLASHTIRISHDA DARS JADVALINING O'RNI // Oriental renaissance: Innovative, educational, natural and social sciences. – 2022. – T. 2. – №. Special Issue 26. – C. 83-87.