

Guarding the Guardian: AI-Enabled Threat Detection and Prevention with Safeguards for Agentic and Adversarial Risks

Rohit Laxman Mankar¹, Shvet Dayavansh Saravade², Aarti Sanjay Gawai³

^{1,2}Students, Department of Information Technology and Computer Science, D.G. Ruparel College

³Assistant Professor, Department of Information Technology and Computer Science, D.G. Ruparel College
rohitmankar1324@gmail.com¹, saravadeshvet@gmail.com², aartigawai1611@gmail.com³

Abstract - Cyberattacks' scale, speed, and diversity have increased the dependency of traditional rules-based approaches on human analysis. Artificial Intelligence (AI) technologies such as machine learning (ML), deep learning (DL), anomaly detection, natural language processing (NLP), and large language models (LLMs) could help defenders in pattern detection, event correlation, prioritization, and response. This paper will explore how artificial intelligence can help in cybersecurity specifically in threat detection and prevention. A review of literature on network intrusion detection, malware and phishing analysis, identity and behavior analysis, security operations, and AI-driven incident response will be conducted. Also, the security of AI-driven defense systems will be explored in terms of adversarial manipulation, data poisoning, prompt injection, sensitive information leakage, over-agency, and attack on AI or its supply chain. Narrative literature review and conceptual synthesis was adopted based on scientific papers and authoritative guidelines issued by NIST, OWASP, ENISA, MITRE, and India government departments. A framework of layers consisting of telemetry, secure data preparation, AI analytics, risk scoring, validation by humans, controlled prevention, and continuous assurance was proposed. In the Indian context, data-protection, cyber-incident reporting, and emerging AI-safety measures should be considered alongside technical controls.

Key Words: artificial intelligence, cybersecurity, threat detection, machine learning, intrusion detection, threat prevention.

1. INTRODUCTION

The cybersecurity landscape has shifted from a relatively static framework of perimeter defenses to a continuously monitored system of identities, endpoints, cloud workload, applications, APIs, mobile devices, and interconnected services. Events are generated by each of these elements, and threats include multi-stage attacks involving credential theft, phishing, malware, exploits, lateral movement, and exfiltration of data. Thus, there are two challenges associated with security: detection of signals among large event streams and fast response to minimize damages inflicted by an intrusion.

AI is one approach to solving the issue of scalability. Machine learning can be used to classify known patterns, anomaly detection will help to detect unusual activity, deep learning to develop complex models, and natural language

processing to analyze security reports, emails, tickets, and other unstructured data. A systematic literature review conducted by Kaur, Gabrijelčič, and Klobučar on applications of AI in the Identify, Protect, Detect, Respond, and Recover activities of cybersecurity revealed 236 primary research papers among all reviewed publications [1].

Nevertheless, AI must not be viewed as the only means of providing security controls. Network segmentation, access control, secure configuration, endpoint protection, encryption, patching, backup, and response procedures are essential. AI is at its most useful where it brings a level of scale to the analytics or aids security experts in correlating pieces of evidence which would be hard to connect otherwise. Prior work on machine learning applications for intrusion detection has stressed on the challenges involved in deploying lab results in real-world operational environments due to the fact that security data are non-stationary, labels are inaccurate, and false positives are costly [2].

The problem has gained another dimension. AI-enabled defense mechanisms can themselves be attacked. Adversarial input, training data poisoning, model compromise, prompt injection, sensitive information disclosure, and excessive agent permissions may compromise AI-enabled security mechanisms. NIST has classified attack types and mitigations under various aspects of the adversarial machine learning life cycle [3], while OWASP has identified prompt injection, sensitive information disclosure, supply chain risks, data/model poisoning, improper output handling, excessive agency, and other risks associated with LLM applications by 2025 [4].

In this paper, we thus address two interrelated problems: how can AI enhance cybersecurity threat detection and prevention, and how should AI-enabled security mechanisms be designed in order to mitigate their own risks? Goals are to analyze prominent AI methodologies, apply them to popular cybersecurity tasks, discuss their pros and cons, incorporate new issues related to agentic AI cybersecurity, and develop a workable framework. This research is based on literature sources and does not have experimental data.

2. BACKGROUND AND RELATED WORK

Machine learning in cybersecurity has a rich research background. Buczak and Guven gave an overview of intrusion

detection with data-mining and machine-learning algorithms, and talked about the issues of algorithm choice, dataset, feature generation, and computation [5]. Sommer and Paxson observed that network intrusion detection was a rather challenging application of machine learning due to the indefinite and evolving operational setting [2]. That explains why the good benchmark performance should not be regarded as the indication of good production performance.

The application of deep learning widened the range of possible features that could be extracted from the network and host telemetry. Shone et al. offered a method based on representation learning and designed for deep learning network intrusion detection [6]. Vinayakumar et al. studied deep neural networks for scalable intrusion detection and focused on both network- and host-based analysis [7]. Ferrag et al. reviewed further the deep-learning techniques and datasets used for cybersecurity intrusion detection and demonstrated that the evaluation results were very much dataset-dependent [8].

The recent literature has expanded the range of application of AI technologies from the traditional intrusion detection. Kaur et al. classified applications of AI for different cybersecurity purposes [1]. The work of Apruzzese et al. explored the gap between the machine learning research community and cybersecurity practitioners [9]. The LLM research field also expanded to include the tasks of threat detection and helping the analysts. A 2024 survey analyzed LLMs applications in cyber-threat detection and their potential role in various detection scenarios [10].

On the other hand, modern AI creates new attack surfaces. The taxonomy of 2025 from NIST on adversarial machine learning includes evasion, poisoning, privacy, and misuse-related attacks for predictive and generative AI [3]. The OWASP 2025 LLM Top 10 is devoted to security risks and mitigations of LLMs for different applications [4]. MITRE ATLAS can be used alongside the above mentioned references for classification of adversarial tactics and techniques applicable to AI technologies [11].

3. RESEARCH METHODOLOGY

The methodology employed in the research is a narrative literature review and synthesis. The research synthesizes findings from peer-reviewed academic studies, standards, cybersecurity frameworks, and Indian government policy documents concerning the application of artificial intelligence for cyber threats detection and prevention. Specific focus is made on detection methods based on AI technology, limitations in the operation of AI technology, attacks on AI systems, risks of agentic AI, and governance issues.

External validation was done for current standards and Indian policy information. The priority was given to peer-reviewed publications and primary and/or authoritative sources including NIST, OWASP, MITRE, MeitY, India Code, and CERT-In. Findings in the provided draft which relied on seemingly private, proprietary or insufficiently verifiable data

sets were not considered proven quantitative findings. Notably, the numerical findings provided in the draft about agentic exfiltration, percentage of the supply chain, or breach data set have not been included in the final paper as quantitative research findings.

The paper was written based on the following research questions: (1) What kinds of AI-based tools are being used to detect potential cyber threats?; (2) What kinds of cyber threats can be detected and prevented by using artificial intelligence; (3) How the results obtained through artificial intelligence can be turned into control responses; (4) What kind of technical limitations exist; and (5) What kind of additional measures should be considered to secure a cyber-defensive system that employs AI.

4. AI TECHNIQUES FOR THREAT DETECTION

Threat detection utilizing AI can be depicted as a workflow consisting of collecting telemetry, data preparation and contextualization, modeling to detect patterns using one or several algorithms, correlation with security context, risk and confidence assessment, and making an actionable decision for a security analyst or enforcement systems. The methodology should be chosen depending on the threat model and operational requirements and not on an individual preference for certain algorithms.

4.1 Supervised Machine Learning

Supervised learning trains a model to learn a mapping between the feature set and target labels using labeled samples. Algorithms like decision tree, random forest, support-vector machine, logistic regression, gradient-boosting, and neural networks can be used for URL detection, phishing, malware family, network flows and other security problems.

The key strength of supervised learning is that the learning objective can be measured via labels. Weaknesses include the need for training data representation, class imbalance, label noise and vulnerability to shifts in distributions. Practically speaking, accuracy is not necessarily the only valuable metric.

4.2 Unsupervised and Anomaly Detection

Unsupervised and semi-supervised learning approaches are useful in finding unusual patterns where there is a lack of complete labels on attacks. Clustering, density-based methods, isolation methods, and autoencoders are useful in identifying abnormalities such as unusual patterns in network behavior, authentication behavior, insider risk, or new event patterns.

An anomaly is not synonymous with maliciousness. New versions of software, administration-related activities, travels, unusual business activities, and even legitimate behavior changes in network traffic can lead to anomalies. Therefore, AI-generated anomaly scores should always be supplemented with identity, asset, threat intelligence, and business context before any serious steps are taken.

4.3 Deep Learning

Deep learning can learn representations from high-dimensional or sequential security data. Architectures such as convolutional neural networks, recurrent neural networks, transformer-based models, and autoencoders have been explored in traffic, host event, malware, and intrusion detection [6]–[8].

Though deep learning algorithms are powerful, they could be resource-intensive and complex to understand. An intelligence analyst has to understand why a certain activity was flagged as being suspicious in order to justify why that particular solution is being recommended, especially when the proposed solution involves isolating a server or disabling an account.

4.4 NLP and Large Language Models

Cybersecurity operations involve lots of text: emails, threat reports, tickets, indicators, incident timelines, vulnerabilities, and analyst notes. NLP can identify entities, detect phishing in content, cluster reports, and normalize threat intelligence data. LLMs, on their part, are capable of alert summarization, evidence explanation, investigation queries support, generating structured incident reports, and helping analysts navigate knowledge bases [10].

The use of LLMs implies a unique security consideration. The input from outside sources may be crafted with the purpose of manipulating the model and leading to prompt injection attacks. When the model is allowed to access external tools, databases, and perform actions, a small mistake of the model turns into an operational security incident. That is why prompt injection, sensitive information exposure, wrong output management, excessive agency, and supply chain attacks are the most critical issues with regard to LLM applications [4].

4.5 Behavioral and Identity Analytics

Identity-based attacks frequently make use of valid credentials; thus, relying only on malware detection will not be enough. AI can identify behavioral profiles based on authentication time, characteristics of devices used, access patterns, privileges, session history, and resource use. A significant deviation from those behavioral profiles could lead to step-up authentication, session investigation, etc.

It is critical to take into account variance and privacy when implementing behavioral solutions. It is not about flagging any deviation from normal behavior as a potential attack, but rather integrating behavioral detection with risk and taking corresponding actions.

5. AI APPLICATIONS IN THREAT DETECTION AND PREVENTION

There are many places where AI can help with security, yet detection and prevention are tightly related. Any detection becomes practically valuable only if it can be connected to a certain context and translated into an action. When implementing the reaction path, one needs to use policy gates

and least privileges to prevent a single mistake from the AI model to disrupt operations.

Phishing and social engineering: message content analysis, URL, sender behavior, attachment analysis, and more contextual factors. This detection could be connected to any reaction: quarantine, banner warnings, URL blocking, etc.

Malware: models for malware detection can examine file attributes, API call sequencing, metadata, sandbox traces, and behavioral signals. A risk score can help to block, isolate, or quarantine malicious software. It requires continuous learning because malware creators can change attributes to bypass security tools.

Network intrusions: machine learning systems can detect intrusions by analyzing traffic flow, sessions, packets, or features based on their analysis. Confirmed alerts could help with controlling network access, implementing segmentation, firewalls, and incident response procedures [2], [5]–[8].

Account compromise: using behavioral signals, security systems can spot anomalous account logins and privilege elevations. The remediation might involve step-up authentication, token revocation, session termination, or temporary access control based on corporate policies.

Cloud and application security: machine learning can analyze identity, API, configuration, workload, and application telemetry to spot anomalies in privilege usage, API callouts, and workload behavior.

Security operations: machine learning can assist with correlating incidents, deduplication, summarizing incidents, extracting indicators, and recommending further actions to investigate.

6. PROPOSED AI-ENABLED CYBERSECURITY FRAMEWORK

This framework brings together the defensive AI pipeline of the initial document and the security guardrails highlighted in the supplied research document. This framework is based on controlled augmentation, where AI does the heavy-lifting analysis and recommendation while policy, deterministic control, and humans make the consequential decisions.

AI-enabled cybersecurity: detection → contextual decision → controlled prevention

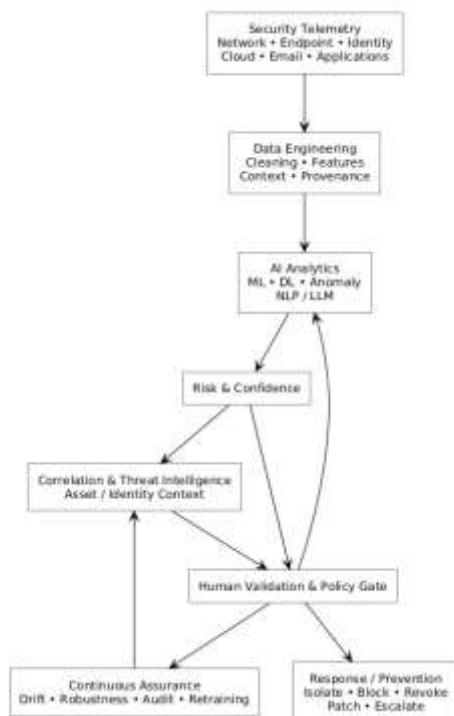


Fig -1: Proposed layered AI-enabled cybersecurity framework for detection, contextual decision-making, prevention, and continuous assurance.

6.1 Telemetry, Data Preparation and Context

The first layer will aggregate network flows, endpoint events, authentication events, cloud activity, application logs, emails, vulnerabilities, and threat intelligence. Data provenance, timestamping, asset ownership, identity, and criticality of the asset needs to be maintained. The second layer includes normalization, deduplication, feature engineering, filling in the gaps, and validating schemas. Training and production data pipelines must be isolated since training data could be tampered with in order to change future model behavior.

6.2 AI Analytics and Contextual Risk Scoring

The analytics layer is not limited to one model but can leverage multiple models including supervised models for identifying known threats, anomalies for finding deviations, deep models for handling complex patterns, and natural language processing/large language models for assisting with unstructured security information. Outputs from all those components can be combined with deterministic rules, threat intelligence, asset criticality, identity, prior incidents, and model confidence. This avoids the possibility that one model score alone determines security action.

6.3 Human Validation and Controlled Prevention

High-impact actions need to go through policy gates. Simple low-impact actions like adding an item to the review queue could be automated, but high-impact actions like disabling a privileged account, altering cloud permissions, or isolating a vital service need more validation steps or approval

by human intervention. Possible response actions would include blocking, quarantine, network isolation, token revocation, patch escalation, and incident creation. Audit logs and rollback functionality must be ensured where feasible.

6.4 Security of AI Models and Agentic Components

Modern AI-based defense mechanisms may be based on retrieval-augmented generation, tool invocation, automated investigation, or agentic workflow. The provided draft appropriately captures that the attack surface increases for an AI system if it is able to run commands or access external systems. Thus, the framework includes runtime sandboxes, clear separation of trusted instruction sets from untrusted retrieval content, least-privilege principle, short-term credentials, tool whitelisting, rate-limiting, and circuit breaking.

The AI agent must not get access to broad administration simply because the activity is easy for operation. The invocation of each tool must be authorized based on the specific task, while the irreversible activity should be covered by policy. Such an approach will reduce the impact of the attack when the attacker successfully influences the model input data.

6.5 Continuous Assurance

The last layer is responsible for monitoring model drift, false positives, false negatives, data quality, latency, model integrity, and change in attackers' behavior. The National Institute of Standards and Technology defines the AI risk management framework through the Govern, Map, Measure, and Manage functions [12]. The continuous assurance process must cover adversarial testing, versioning of models and datasets, dependency check, and revalidation.

7. COMPARATIVE ANALYSIS OF AI APPROACHES

Table -1: Qualitative comparison of AI approaches for cybersecurity threat detection and prevention.

Approach	Typical use	Strength	Limitation	Operational role
Supervised ML	Known attack classification; phishing; malware	Clear objective and measurable labels	Needs representative labeled data; drift	Detection and prioritization

Anomaly detection	Unknown behavior; insider risk; unusual traffic	Works with limited labels	Outlier may be legitimate; threshold tuning	Early warning
Deep learning	Complex traffic; sequences; host/malware behavior	Learns high-dimensional patterns	Data/compute cost; explainability	Advanced detection
NLP / LLM	Threat reports; phishing; alert summaries; analyst support	Handles unstructured language and knowledge tasks	Prompt injection; leakage; hallucination; excessive agency	Investigation support
Hybrid / ensemble	Multi-signal correlation and risk scoring	Combines complementary evidence	Higher system complexity	Decision support and orchestration

The comparison is qualitative in nature since the outcomes depend on the particular dataset, threat category, operating environment, and assessment criteria.

8. AI-RELATED THREATS TO CYBERSECURITY

Even though AI is being used for defense, this does not guarantee that there would be no attacks on the AI system itself. On the contrary, this introduces one more security layer that should be modeled and analyzed.

An attack of adversarial evasion implies manipulation of input to produce a wrong answer from the model. The 2025 taxonomy of NIST states that evasion is one of the key ideas in the context of adversarial machine learning and places it side by side with poisoning, privacy, and misuse attacks [3]. For cybersecurity, evasion implies alterations of network parameters, malware properties, URLs, and other visible aspects without changing the malicious nature of the attack.

The data poisoning affects the training/adaptation process. In case of allowing untrusted inputs/feedback to influence the model update, the adversary can affect future decisions.

Prompt injection is very pertinent in relation to security applications based on LLMs. Injurious prompts could be delivered via emails, documents, web pages, tickets, or retrieved information. If a model cannot differentiate between

information and prompts, it would act according to the injected prompts. Input isolation, trusted channel isolation, retrieval filtering, output validation, and restricted permissions for the tool are key mitigations [4].

Information exposure could happen when security data, prompts, logs, or retrieved information contains sensitive data. Security architectures need to make use of data minimization, encryption, access control, data retention policy, secret scrubbing, and logging.

Agency refers to the problem where a security application has too much power to execute prompts or system changes and does not have enough controls around it. Simply disabling automation is not the answer.

9. INDIAN TECHNO-LEGAL AND GOVERNANCE CONTEXT

The AI-security landscape in India is evolving through a blend of data-protection legislation, cyber-incident reporting mandates, technology regulation, governmental guidance, and nascent AI safety measures. The Digital Personal Data Protection Act, 2023 is a legal framework providing for the processing of digital personal data [13]. The Act and the Digital Personal Data Protection Rules, 2025 have staggered dates for coming into operation; hence, one has to differentiate between the provisions already in effect from those that will come into effect in future [14].

Responsible AI policy efforts in India also include safety, privacy, security, accountability, and responsible deployment considerations. The National Strategy for Artificial Intelligence by NITI Aayog and subsequent responsible AI policy efforts offer policy-oriented guidelines as opposed to an AI security act [16], [17]. The Ministry of Electronics and IT is working towards an IndiaAI Safety Institute that would consider AI evaluation, red-teaming, risk assessment, and AI-safety research specific to India [18].

India's responsible-AI policy work also emphasizes safety, privacy, security, accountability, and responsible deployment. NITI Aayog's National Strategy for Artificial Intelligence and subsequent responsible-AI work provide policy-oriented guidance rather than a single AI-security statute [16], [17]. MeitY has also advanced an IndiaAI Safety Institute initiative focused on AI evaluation, red-teaming, risk assessment, and India-specific AI-safety research [18].

Issues involving responsibility for the harm caused by an autonomous AI agent are still open questions for legal and policy development and are not concluded in this paper. Thus, issues like an AI Bill of Materials, higher accountability for impactful AI applications, and model-level reporting of incidents are discussed in this paper as proposals and not statutory requirements.

9.1 Proposed Techno-Legal Guardrails

AI Bill of Materials (AIBOM): Organizations implementing highly impactful AI systems may keep a record of their base model, datasets, fine-tuning artifacts, retrieval

store, third party APIs, libraries, and model version as part of their supply chain investigation and incident response process.

Runtime containment: AI agents capable of running code need to be isolated to run in well-contained environments with limited network access, job-based permissions, and a well-specified list of tools.

Context and instruction isolation: Trusted system instructions need to be separated from untrusted retrieved content, and untrusted content cannot become trusted instruction without explicit permission.

Ephemeral credentials: API tokens and credentials used by AI agents need to be temporary and narrowly defined for the job at hand. Administrative credentials need to be avoided for autonomous agents.

Human intervention for irreversible actions: Actions such as destructive database modifications, privilege change, financial transactions, and infrastructure changes require explicit policy authorization, and human intervention if needed.

Incident preparedness for AI: Security teams need to log model version, input/prompt source, tool invocations, timestamps, outputs, and actions taken so that AI incidents can be investigated using the cyber-incident process.

10. BENEFIT AND LIMITATIONS

AI can bring scaling, quick detection of patterns, prioritization of alerts, automation of correlation and assistance in processing unstructured security information. This is especially useful in scenarios where there is an abundance of events and security information or when the security team needs to discover weak signals through several different systems.

However, performance of AI heavily depends on the quality of data and circumstances. False positives will increase the workload of analysts, whereas false negatives will let attacks proceed. The shift of distribution can decrease the performance when circumstances change or attackers' behavior is altered. Complicated models may also be hard to understand, and the infrastructure for AI adds more costs and dependencies.

Therefore, the literature proves that a middle ground should be considered. That is to say, AI can improve cybersecurity, yet the quality of its work depends on architecture around it.

11. FUTURE SCOPE

The future research needs to test cybersecurity tools powered by AI under realistic distribution shifts and adversarial settings and not just on benchmark datasets. Continual learning, privacy-preserving collaborations, federated learning, graph-based threat detection, and model evaluation are some other aspects that need to be researched further.

Assistance of LLMs for cybersecurity operations is expected to increase in threat intelligence summarization,

planning investigations, knowledge lookup, explanation of alerts, and documenting incidents. Future tools should focus on provenance, secure usage of tools, permission limitations, output validation, and decrease in analyst workloads.

Specifically for India, research may be done in multilingual threat detection, AI in digital public infrastructure, region-specific phishing, and fraud, and standardized red team evaluation of systems being used for critical public services. Such research can link AI research with specific cybersecurity needs of the country without having to assume that policy issues have already been sorted out in the process.

12. DISCUSSION

The synthesized literature indicates that there are two distinct functions of AI in cybersecurity. It is used both as a defense mechanism and as an additional security asset. Focusing on the former function leads to an incomplete picture of the threat model.

The most promising design principle, derived from the analysis, is controlled augmentation, which involves the use of AI in cases when pattern recognition and scaling are clearly beneficial, while the policies, restrictions and judgment will be applied by humans when making decisions of higher impact. This also corresponds to the risk management approach in NIST, which focuses on governance and management throughout the entire lifecycle of AI [12].

The focus of the provided draft on agentic systems is an important addition to the standard research related to AI and cybersecurity. Once some AI component is able to carry out commands, access credentials, interact with APIs and change the state of a system, an error in a model will result in an incident. The solution lies in the architecture, which needs to include such components as a sandbox for the agent, isolation of trusted instructions from untrusted inputs, restrictions on credentials and validation of the output.

In the Indian scenario, AI security must be viewed along with data security, CERT-In reporting, responsible AI policies, and nascent AI safety institutions. It is not about creating novel categories of threats but looking into whether existing frameworks can be improved or new frameworks will be required for new categories of autonomous action.

13. CONCLUSION

There is a possibility to improve the process of detecting and preventing cybersecurity threats by means of the analysis of large security data sets, pattern recognition, anomalies detection, unstructured threat processing, prioritization, and incident response with the help of artificial intelligence technology. There are different capabilities of different approaches, like supervised learning, anomaly detection, deep learning, NLP, and LLM.

Also, the article demonstrates that AI cannot be used as an independent security solution per se. Model errors, limitations of data, adversarial attacks, prompt injection, data

poisoning, information leakage, and agency problems could lead to failures. It is necessary to protect both systems that are protected from threats and AI elements that perform the protection.

The suggested framework is based on telemetry, secure data engineering, multi-model analysis, contextual risk scoring, human/policy validation, controlled prevention, and continuous assurance. There is also a necessity to use the additional controls for AI agents, like sandboxing, instruction isolation, limited privileges, rate limiting, and approval gates.

On the whole, the most sensible way forward is to have AI as a tool for enhancing the cybersecurity operations in a controlled way. By integrating AI into the concept of defense-in-depth, ensuring a secure lifecycle of models and proper governance along with the right kind of human oversight, AI can be used effectively in the realm of security.

ACKNOWLEDGEMENT

Thanks is extended to the faculty and department for the academic guidance and for creating an academic environment needed to conduct this review study. Thanks is also extended to the researchers, standard-setting bodies, and public agencies that guided this research through their publications and guidelines.

REFERENCES

1. Kaur, R., Gabrijelčič, D., Klobučar, T.: Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, 97, 101804 (2023). <https://doi.org/10.1016/j.inffus.2023.101804>
2. Sommer, R., Paxson, V.: Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. *IEEE Symposium on Security and Privacy*, 305–316 (2010). <https://doi.org/10.1109/SP.2010.25>
3. Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., Hamlin, M.: Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. *NIST AI 100-2 E2025* (2025). <https://doi.org/10.6028/NIST.AI.100-2e2025>
4. OWASP Gen AI Security Project: OWASP Top 10 for Large Language Model Applications 2025. Open Worldwide Application Security Project (2025). <https://genai.owasp.org/llm-top-10/>
5. Buczak, A.L., Guven, E.: A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials*, 18(2), 1153–1176 (2016). <https://doi.org/10.1109/COMST.2015.2494502>
6. Shone, N., Nguyen, T.N., Phai, V.D., Shi, Q.: A Deep Learning Approach to Network Intrusion Detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(1), 41–50 (2018). <https://doi.org/10.1109/TETCI.2017.2772792>
7. Vinayakumar, R., Alazab, M., Soman, K.P., Poornachandran, P., Al-Nemrat, A., Venkatraman, S.: Deep Learning Approach for Intelligent Intrusion Detection System. *IEEE Access*, 7, 41525–41550(2019). <https://doi.org/10.1109/ACCESS.2019.2895334>
8. Ferrag, M.A., Maglaras, L., Moschoyiannis, S., Janicke, H.: Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study. *Journal of Information Security and Applications*, 50, 102419 (2020). <https://doi.org/10.1016/j.jisa.2019.102419>
9. Apruzzese, G., Laskov, P., Montes de Oca, E., et al.: The Role of Machine Learning in Cybersecurity. *Digital Threats: Research and Practice*, 4(1), Article 8 (2023). <https://doi.org/10.1145/3545574>
10. Chen, Y., Cui, M., Wang, D., Cao, Y., Yang, P., Jiang, B., Lu, Z., Liu, B.: A Survey of Large Language Models for Cyber Threat Detection. *Computers & Security*, 145, 104016 (2024). <https://doi.org/10.1016/j.cose.2024.104016>
11. MITRE: Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS). MITRE Corporation, continuously maintained a knowledge base. <https://atlas.mitre.org/>
12. Tabassi, E.: Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, National Institute of Standards and Technology (2023). <https://doi.org/10.6028/NIST.AI.100-1>
13. Government of India: The Digital Personal Data Protection Act, 2023. Act No. 22 of 2023. India Code. <https://www.indiacode.nic.in/indiacode/handle/123456789/22037>
14. Ministry of Electronics and Information Technology: Digital Personal Data Protection Rules, 2025. Government of India (2025). <https://www.meitv.gov.in/documents/act-and-policies/digital-personal-data-protection-rules-2025-gDOxUjMtOWa>
15. Indian Computer Emergency Response Team (CERT-In): Cyber Security Directions and related FAQs, 2022; reporting of specified incidents within six hours of noticing or being brought to notice. Government of India. <https://www.cert-in.org.in/>
16. NITI Aayog: National Strategy for Artificial Intelligence. Government of India (2018). <https://www.niti.gov.in/node/708>
17. NITI Aayog: Responsible AI / Principles for Responsible Management of AI Systems. Government of India (2021). <https://www.niti.gov.in/node/326>
18. Ministry of Electronics and Information Technology: IndiaAI Safety Institute initiative, including AI model evaluation, red-teaming, and risk assessment activities. Government of India (2026). <https://indiaai.gov.in/article/india-takes-the-lead-establishing-the-indiaai-safety-institute-for-responsible-ai-innovation>