JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

REVIEW ARTICLE
Vol. 3, No. 2 (2025), pp. 1-27
DOI: 10.63646/jaiaa.2025.030201

Artificial Intelligence for Cybersecurity Analytics: A Review of Deep Learning, Hybrid Architectures, and Operational Resilience

Emily Zhang[1], Michael Chen[2], Sophia Wang[3, *], Daniel Liu[4]

[1] College of Education and Human Development, Temple University, Philadelphia, PA 19122, USA

[2] Fox School of Business, Temple University, Philadelphia, PA 19122, USA

[3] College of Liberal Arts, Temple University, Philadelphia, PA 19122, USA

[4] College of Public Health, Temple University, Philadelphia, PA 19122, USA

* Corresponding E-mail: sophia.wang@temple.edu

ARTICLE INFO

Received
December 03, 2024
Revised
February 10, 2025
Accepted
March 18, 2025
Available Online
March 30, 2025
DOI
10.63646/jaiaa.2025.030201

Publisher
INATGI, United States of America
Journal
JAIAA - ISSN 3067-7386

ABSTRACT

This review examines how artificial intelligence is reshaping cybersecurity analytics across three tightly connected domains: threat detection, explanatory decision support, and operational resilience. Rather than treating AI as a stand-alone detection tool, the paper positions it as a layered analytical infrastructure that links telemetry collection, representation learning, anomaly scoring, malware and botnet recognition, threat correlation, analyst triage, and governance. The review synthesizes evidence from classical machine learning, deep learning, graph learning, transformers, explainable AI, federated intelligence, and early quantum-enhanced approaches. It argues that the main contribution of AI in cybersecurity lies not only in higher predictive accuracy but also in the ability to connect descriptive, predictive, and prescriptive analytics under fast-changing threat conditions. The paper further shows that operational value depends on data quality, benchmark realism, false-alarm control, model interpretability, lifecycle monitoring, and the fit between algorithmic outputs and security workflows. The strongest systems are not those that maximize laboratory metrics in isolation, but those that combine robust representations, calibrated uncertainty, explainability, and human oversight. Future directions are identified in multimodal cyber intelligence, continual and causal learning, graph-native detection, privacy-preserving collaboration, and adaptive architectures that treat resilience as an organizational capability rather than a model score.

Keywords: Artificial intelligence; Cybersecurity analytics; Intrusion detection; Deep learning; Explainable AI; Threat intelligence; Operational resilience

1. INTRODUCTION

Cybersecurity has become a decision problem as much as a technical one. Modern organizations now collect security logs, endpoint traces, authentication records, packet captures, DNS activity, vulnerability scans, malware indicators, and user behavior signals at a scale that would have overwhelmed most security teams only a few years ago. Yet a larger volume of telemetry has not automatically produced clearer judgment. Analysts still face alert floods, fragmented tooling, and threat

patterns that evolve faster than rule sets can be updated. In that environment, artificial intelligence has moved from the margins of cyber defense to the center of detection and triage. It is increasingly used not simply to classify packets or flag anomalies, but to connect raw evidence with operational action (Buczak & Guven, 2016; Garcia-Teodoro et al., 2009; Sperotto et al., 2010; Wu & Banzhaf, 2010; Xin et al., 2018; Berman et al., 2019; Macas et al., 2022; Yan et al., 2022).

The practical appeal of AI in cybersecurity comes from its ability to work across the full analytical stack. Classical machine learning can rank suspicious events, prioritize response queues, and identify nonlinear relationships that signature-based systems miss. Deep learning extends that capacity to high-dimensional, sequential, visual, and graph-structured data, enabling richer representations of malware binaries, network flows, command sequences, phishing content, and multi-stage attacks. More recent work on transformers, graph neural networks, explainable AI, and federated learning suggests that cyber analytics is shifting from isolated detection models toward broader decision architectures that combine pattern recognition, contextual reasoning, and governance (Xin et al., 2018; Berman et al., 2019; Macas et al., 2022; Yan et al., 2022; Hozouri et al., 2025; Sommer & Paxson, 2010).

Even so, the AI turn in cybersecurity should not be romanticized. The literature repeatedly shows that strong results in benchmark experiments do not necessarily translate into dependable performance in operational environments. Dataset artifacts, unrealistic class distributions, narrow task definitions, concept drift, weak labeling practices, and limited interpretability can all create a false sense of capability. A model that performs extremely well on NSL-KDD or a balanced slice of CICIDS2017 may still struggle in a live enterprise environment where benign traffic changes daily and attack signals are rare, noisy, and adversarially manipulated (Garcia-Teodoro et al., 2009; Macas et al., 2022; Tsai et al., 2009; Tavallaee et al., 2009; Moustafa & Slay, 2015; Moustafa & Slay, 2016; Sharafaldin et al., 2019; Ring et al., 2019).

This review therefore asks a broader question than whether AI improves threat detection. It examines how AI changes the structure of cybersecurity analytics itself. The article develops three arguments (Table 1).

Table 1. Representative Streams of AI-based Cybersecurity Analytics Literature.

Analytical stream	Typical methods	Primary task focus	Main practical value
Classical ML for IDS	RF, SVM, boosting, isolation forests	Flow-level attack detection	Strong baselines, low compute, easier inspection
Deep representation learning	CNN, RNN, autoencoders, GANs	High-dimensional and sequential telemetry	Richer feature extraction under weak structure
Relational and language-aware analytics	GNNs, transformers, LLM-assisted retrieval	Entity correlation, log mining, threat intelligence	Campaign-level reasoning and contextual enrichment
Governed hybrid defense	XAI, federated learning, robust optimization	Operational triage and cross-site learning	Actionable outputs under privacy and oversight constraints

First, the value of AI lies less in any single model family than in how models are embedded in data, workflow, and escalation architectures. Second, hybridization is now a defining feature of the field: high-performing systems increasingly combine multiple model types, multiple data sources, and multiple decision layers. Third, operational resilience depends not only on predictive power but also on explanation, calibration, monitoring, and human organizational capacity (Figure 1). Framed this way, the field is no longer just about intrusion detection accuracy; it is about building analytical systems that remain useful under uncertainty, drift, and attack adaptation (Berman et al., 2019; Macas et al., 2022; Yan et al., 2022; Hozouri et al., 2025; Sommer & Paxson, 2010; Mirsky et al., 2018; Meidan et al., 2018; Koroniotis et al., 2019; Alsaedi et al., 2020; Diro & Chilamkurti, 2018).

Figure 1. Layered AI-enabled Cybersecurity Analytics Workflow Linking Telemetry, Learning, Risk Scoring, and Action

The remainder of the paper is organized as follows. Section 2 clarifies cybersecurity analytics as a layered decision architecture rather than a single detection task. Section 3 reviews the major AI model families now used in cyber defense, from classical machine learning to deep, graph, transformer, federated, explainable, and quantum-enhanced approaches. Section 4 examines benchmark datasets and evaluation practices, with particular attention to what current performance numbers often hide. Section 5 discusses hybrid architecture and emerging design patterns for adaptive defense. Section 6 turns to deployment challenges and governance. Section 7 outlines future research directions, and Section 8 concludes by arguing for a more integrated and human-centered view of AI-enabled cybersecurity (Buczak & Guven, 2016; Xin et al., 2018; Macas et al., 2022; Yan et al., 2022; Hozouri et al., 2025; Sommer & Paxson, 2010).

There is also a historical reason why the AI debate in cybersecurity has become so intense. For years, security operations were dominated by signature logic, deterministic correlation rules, and manually tuned thresholds. Those approaches were not irrational; they were a rational response to environments where interpretability and control mattered more than broad generalization. What changed was the speed and heterogeneity of the threat landscape. Cloud-native infrastructure, software-defined perimeters, mobile work, API ecosystems, and machine-to-machine communication generated behaviors that were difficult to capture with static rules alone. AI did not replace the old logic so much as expose its scaling limits. That is why the most productive question today is not whether signatures or machine learning are better, but which combinations of symbolic knowledge and statistical inference are most effective under specific threat conditions (Buczak & Guven, 2016; Garcia-Teodoro et al., 2009; Sperotto et al., 2010; Wu & Banzhaf, 2010; Xin et al., 2018; Berman et al., 2019; Macas et al., 2022; Yan et al., 2022).

ISSN: 3067-7386 https://doi.org/10.63646/jaiaa.2025.030201

Another force behind the expansion of AI in cyber defense is the rise of cyber threat intelligence as a data problem. Intelligence programs now draw not only on indicators of compromise, but on narrative reports, social media signals, dark-web discussions, vulnerability disclosures, software supply chain advisories, and geopolitical context. Many of these sources are textual, uncertain, or only loosely structured. Natural language processing and retrieval-based models have therefore become increasingly relevant to cybersecurity analytics. They help link noisy threat reporting to operational evidence, support alert enrichment, and make historical knowledge more accessible to analysts. This development reinforces the central argument of the paper: AI in cybersecurity is not just a faster way to classify packets. It is becoming a broader infrastructure for turning scattered technical and contextual evidence into decisions (Xin et al., 2018; Berman et al., 2019; Macas et al., 2022; Yan et al., 2022; Hozouri et al., 2025; Sommer & Paxson, 2010; Yousefi-Azar et al., 2017; Wang et al., 2017; Lotfollahi et al., 2020; Apruzzese et al., 2018; Ribeiro et al., 2016; Guidotti et al., 2018; Adadi & Berrada, 2018; Doshi-Velez & Kim, 2017; Chen & Guestrin, 2016; Breiman, 2001; Cortes & Vapnik, 1995; Rumelhart et al., 1986).

2. CYBERSECURITY ANALYTICS AS A LAYERED DECISION ARCHITECTURE

Much of the early literature treated cyber defense as a classification problem given a record, a packet, a flow, or a file, decide whether it is benign or malicious. That framing remains useful, but it is incomplete. In practice, cybersecurity analytics unfold across several layers. Raw events must be collected and normalized; related events must be grouped into sessions or entities; statistical or learned representations must be constructed; suspicious patterns must be scored; and results must be placed in a queue, workflow, or playbook that a human analyst or automated control can actually use. The best way to understand AI in cybersecurity is therefore as a layered analytical architecture in which detection is only one stage in a larger sensemaking and response (Garcia-Teodoro et al., 2009; Sperotto et al., 2010; Xin et al., 2018; Sommer & Paxson, 2010; Yin et al., 2017; Tang et al., 2020).

At the first layer, organizations deal with heterogeneous and often messy telemetry. Some signals come from networks, some from hosts, some from identity systems, and some from emails, cloud APIs, or industrial controllers. These signals have different time granularities, schemas, and trust levels. They are also unevenly labeled. AI enters here as a tool for representation and fusion. Feature engineering, autoencoding, embedding, and sequence modeling all help convert raw operational traces into forms that support downstream inference. This is one reason why the strongest papers in the field pay careful attention to preprocessing, normalization, temporal windowing, feature selection, and data balancing rather than jumping directly to classifier design (Xin et al., 2018; Berman et al., 2019; Macas et al., 2022; Tavallaee et al., 2009; Moustafa & Slay, 2015; Moustafa & Slay, 2016; Sharafaldin et al., 2019; Ring et al., 2019; Mirsky et al., 2018; Meidan et al., 2018; Niyaz et al., 2016).

The second layer concerns detection and correlation (Figure 2). A binary classifier can be useful when the task is narrow, but most production settings require more nuanced outputs. Security teams need to know whether an event resembles credential abuse, malware staging, command-and-control traffic, lateral movement, or data exfiltration. They also need to understand how low-level signals combine across time. This is why recurrent, attention-based, and graph-based models have attracted growing interest. They are better aligned with the fact that many attacks are not atomic events but

ISSN: 3067-7386 https://doi.org/10.63646/jaiaa.2025.030201

sequences and relationships distributed across hosts, users, protocols, and time windows (Xin et al., 2018; Macas et al., 2022; Hozouri et al., 2025; Mirsky et al., 2018; Alom & Taha, 2017; Aminanto & Kim, 2017; Maimó et al., 2018; Yousefi-Azar et al., 2017; Wang et al., 2017; Lotfollahi et al., 2020; Apruzzese et al., 2018).

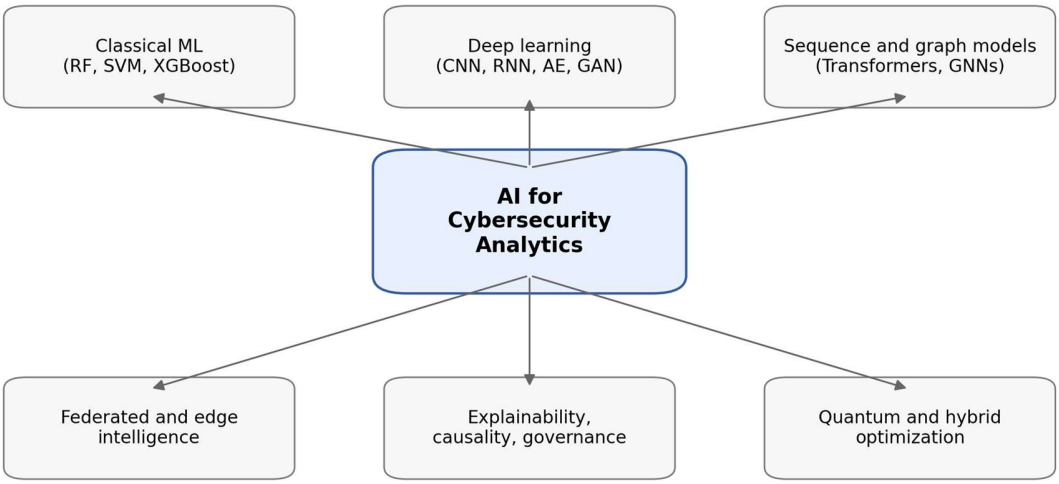

Figure 2. Taxonomy of Major AI Families Currently Used in Cybersecurity Analytics and Their Relationship to Hybrid System Design

The third layer is decision support. Here the challenge is not simply to identify suspicious activity but to make the output usable. Analysts care about ranking, uncertainty, similarity to previous incidents, likely blast radius, and whether the alert can be explained quickly enough to justify action. A model that raises slightly fewer false positives may still be inferior if it produces opaque decisions that force analysts to spend more time in manual verification. The literature on explainable machine learning and human-AI collaboration suggests that explanation is not a cosmetic add-on. In security operations, it changes triage speed, trust, escalation behavior, and the ability to translate an alert into a defensible response (Yan et al., 2022; Kloft & Laskov, 2010; Goodfellow et al., 2015; Papernot et al., 2018; Biggio & Roli, 2018; Carlini & Wagner, 2017; Tramèr et al., 1705; Shabtai et al., 2014).

Finally, there is an organizational layer that is often neglected in technical papers. Cybersecurity performance depends on staffing, policy, playbooks, asset visibility, business criticality, and legal constraints. AI outputs do not operate in a vacuum. They are interpreted and acted upon by people working under pressure. This organizational reality explains why resilience is a better endpoint than raw accuracy. A resilient cyber analytics system should help the organization anticipate, absorb, respond to, and learn from disruptive events. That requires calibrated models, robust data flows, version control, continuous evaluation, and explicit accountability for overrides and automated actions (Sperotto et al., 2010; Berman et al., 2019; Macas et al., 2022; Yan et al., 2022; Alsaedi et al., 2020;

Goodfellow et al., 2015; Tuor et al., 2017; Creech & Hu, 2014; Rigaki & Garcia, 2018; Vinayakumar et al., 2018).

A useful way to see this layered structure is to distinguish detection, correlation, attribution, and action. Detection asks whether something is unusual or matches a known malicious pattern. Correlation asks how multiple weak signals may belong to the same event chain. Attribution asks what actor, toolset, or campaign family the activity resembles, even if confidence remains partial. Action asks what the organization should do next given uncertainty, asset criticality, business impact, and resource constraints. AI contributes differently at each stage. Detection benefits from discriminative performance; correlation benefits from temporal and relational modeling; attribution benefits from richer knowledge representations; and action benefits from explanation, prioritization, and calibrated confidence. Treating all of these as one generic classification task obscures the real structure of the problem (Garcia-Teodoro et al., 2009; Sperotto et al., 2010; Xin et al., 2018; Macas et al., 2022; Hozouri et al., 2025; Alsaedi et al., 2020; Yin et al., 2017; Tang et al., 2020; Rosenblatt, 1958; McCulloch & Pitts, 1943; Kingma & Ba, 2015; Kingma & Welling, 2014; Goodfellow et al., 2014).

This layered perspective also clarifies why security teams often feel disappointed by technically impressive tools. A detector that is excellent at the first stage may add little value at the fourth if it cannot explain itself, cannot be tuned to business context, or cannot be mapped to playbooks. Conversely, a model with only moderate raw accuracy may still create value if it consistently surfaces the right cases early, reduces analyst search time, and integrates well with response workflows. In terms of operations, cybersecurity analytics is an orchestration problem. It depends on queues, context, thresholds, exception handling, and feedback loops. AI becomes strategically important when it supports that orchestration, not when it is judged only by leaderboard performance (Yan et al., 2022; Alsaedi et al., 2020; Kloft & Laskov, 2010; Goodfellow et al., 2015; Papernot et al., 2018; Biggio & Roli, 2018; Carlini & Wagner, 2017; Tramèr et al., 1705; Shabtai et al., 2014; Tuor et al., 2017; Creech & Hu, 2014; Rigaki & Garcia, 2018; Vinayakumar et al., 2018).

3. AI MODEL FAMILIES IN CONTEMPORARY CYBER DEFENSE

Classical machine learning remains highly relevant in cybersecurity, even in an era dominated by deep learning. Random forests, support vector machines, logistic regression, k-nearest neighbors, gradient boosting, and isolation-based methods continue to offer strong baselines, especially when data are tabular, labels are limited, and interpretability matters. Their importance is sometimes underestimated because the field tends to celebrate complexity. Yet classical methods often remain competitive on structured flow datasets and, crucially, are easier to inspect, recalibrate, and deploy under resource constraints. In settings where speed and accountability matter more than marginal gains on a benchmark, simpler models still have a meaningful place (Buczak & Guven, 2016; Wu & Banzhaf, 2010; Tsai et al., 2009; Tang et al., 2020; Kloft & Laskov, 2010; Millar et al., 2018; Kim et al., 2016; Vinayakumar et al., 2019; LeCun et al., 2015; Jordan & Mitchell, 2015; Hinton & Salakhutdinov, 2006).

Deep learning changed the field by making it feasible to learn higher-order representations directly from raw or minimally engineered inputs. Convolutional networks became attractive for packet-level or byte-level pattern learning, image-like traffic encodings, and malware analysis. Recurrent networks and

long short-term memory models became important for login streams, protocol dialogues, and other sequential behaviors. Autoencoders, especially stacked and variational variants, proved useful for anomaly detection when reliable malicious labels were scarce. Generative adversarial models later entered the field both as data augmentation tools and as a reminder that attackers can also use generative learning to evade detectors. The net effect was a shift from shallow feature dependence toward representation-rich cyber analytics (Xin et al., 2018; Berman et al., 2019; Macas et al., 2022; Mirsky et al., 2018; Niyaz et al., 2016; Aminanto & Kim, 2017; Maimó et al., 2018; Krizhevsky et al., 2017; He et al., 2016; Szegedy et al., 2015; Hochreiter & Schmidhuber, 1997; Cho et al., 2014; Vaswani et al., 2017; Devlin et al., 2019; Dosovitskiy et al., 2021; Kipf & Welling, 2017; Veličković et al., 2018).

This shift was visible across application areas. In intrusion detection, recurrent and autoencoder-based models improved the ability to capture temporal dependencies and subtle deviations in traffic. In malware analysis, deep architecture supported opcode-sequence modeling, image-based binary classification, and API-call interpretation. In phishing and malicious URL detection, character-level and sequence models helped detect patterns beyond hand-crafted lexical features. In IoT and industrial settings, distributed or edge-aware deep models improved the detection of attacks under limited computing and fragmented visibility. A common pattern emerges from these diverse cases: deep learning is most valuable when the raw signal is complex, weakly structured, or temporally dependent (Xin et al., 2018; Berman et al., 2019; Macas et al., 2022; Mirsky et al., 2018; Meidan et al., 2018; Niyaz et al., 2016; Alom & Taha, 2017; Aminanto & Kim, 2017; Maimó et al., 2018; Yousefi-Azar et al., 2017; Wang et al., 2017; Lotfollahi et al., 2020; Apruzzese et al., 2018; Cho et al., 2014; Vaswani et al., 2017; Devlin et al., 2019; Dosovitskiy et al., 2021; Kipf & Welling, 2017; Veličković et al., 2018; Hamilton et al., 2017; Lundberg & Lee, 2017).

Table 2. AI Model Families, Strengths, Limitations, and Deployment Fit in Cybersecurity

Model family	Strength in cyber settings	Typical weakness	Best-fit use cases
Classical ML	Fast on structured tabular features	May miss deep context or sequence structure	Flow IDS, risk scoring, interpretable baselines
CNN/RNN/AE	Captures local patterns, temporal dependencies, and anomaly signatures	Needs careful tuning and can be opaque	Traffic analytics, malware, endpoint and IoT monitoring
Transformers / GNNs	Handles long-range relations and entity context	Higher compute and data demands	Log mining, campaign correlation, knowledge-rich detection
Federated / hybrid / quantum-inspired	Supports distributed learning and specialized optimization	Operational complexity and immature tooling	Cross-site learning, privacy-sensitive defense, frontier experimentation

The recent arrival of transformers and self-attention architecture has expanded this trajectory (Table 2). Their appeal lies in their ability to model long-range dependencies and cross-feature interactions

without the strict sequential bottlenecks associated with classical recurrent networks. In cybersecurity, attention mechanisms are increasingly used for flow classification, log mining, malicious traffic analysis, insider-threat modeling, and cyber threat intelligence extraction from text (Figure 3). The rise of large language models has also revived interest in using pretrained architectures for report generation, alert summarization, retrieval over threat knowledge, and analyst copilots. However, this line of work remains uneven. Strong results exist, but the field is still working out how to preserve accuracy, reduce hallucination, and align language-generated explanations with security-grade evidence requirements (Yousefi-Azar et al., 2017; Wang et al., 2017; Lotfollahi et al., 2020; Apruzzese et al., 2018; Ribeiro et al., 2016; Guidotti et al., 2018; Adadi & Berrada, 2018; Doshi-Velez & Kim, 2017; Chen & Guestrin, 2016; Breiman, 2001; Cortes & Vapnik, 1995; Rumelhart et al., 1986).

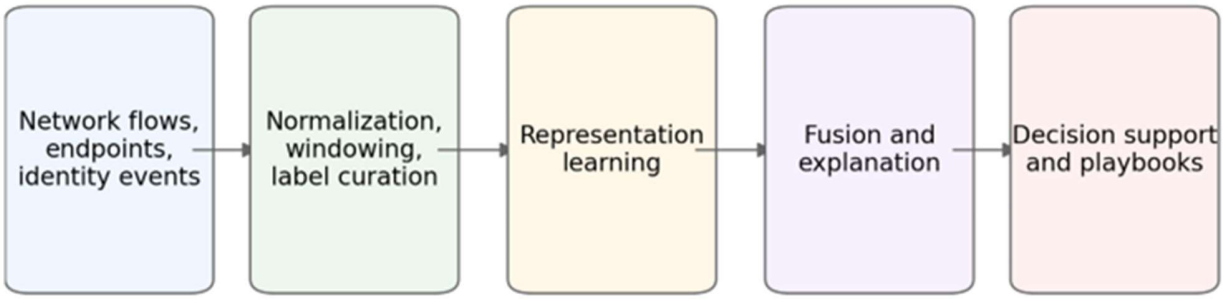

Figure 3. Representative Hybrid Cybersecurity Analytics Pipeline from Data Preparation to Decision Support

Graph neural networks respond to a different but equally important reality: cyber systems are relational. Devices connect to services, users log into hosts, processes spawn child processes, domains resolve to IPs, and alerts cluster around entities rather than isolated records. Graph-based learning has therefore become attractive for lateral movement detection, attack-path analysis, provenance reconstruction, botnet coordination, and knowledge-driven threat reasoning. By encoding nodes, edges, and neighborhood context, graph models often capture structural signals that tabular and sequential models miss. Their promise is especially strong in environments where the question is not simply whether an event is suspicious, but how activity propagates through a digital ecosystem (Wang et al., 2017; Lotfollahi et al., 2020; Rosenblatt, 1958; McCulloch & Pitts, 1943; Kingma & Ba, 2015; Kingma & Welling, 2014; Goodfellow et al., 2014).

Explainable AI sits across all these model families. In cybersecurity, explanation is valuable for at least four reasons. It can reduce analyst fatigue by showing why a case is unusual, improve model debugging when false alarms spike, support regulatory or policy-based accountability, and create a common language between data scientists and security operators. Techniques such as SHAP, LIME, attention visualization, feature attribution, and counterfactual reasoning are therefore increasingly discussed in cyber research. Yet explanation is not solved by adding a post hoc plot. Good explanations in security must be timely, context-aware, robust to adversarial manipulation, and aligned with how analysts actually reason about incidents (Yan et al., 2022; Kloft & Laskov, 2010; Goodfellow et al.,

2015; Papernot et al., 2018; Biggio & Roli, 2018; Carlini & Wagner, 2017; Tramèr et al., 1705; Shabtai et al., 2014; Vincent et al., 2008; Hinton et al., 2006; Baldi, 2012; Brownlee, 2016).

Federated and privacy-preserving learning has added another layer of complexity. Security data are often siloed across organizations, business units, clouds, and devices. Sharing raw telemetry may be impossible for legal, commercial, or privacy reasons. Federated learning provides a way to collaborate on model training while keeping data local. For cyber applications, this opens the possibility of cross-site anomaly detection, distributed IoT defense, and privacy-conscious threat modeling. At the same time, federated settings introduce their own risks, including poisoned updates, non-IID data, system heterogeneity, and difficulties in validation. The literature increasingly suggests that the future of cyber AI will involve not one centralized model but a family of coordinated models learning under constrained data-sharing conditions (Settles, 2009; He & Sun, 2015; Kraskov et al., 2004; Wang et al., 2017; McMahan et al., 2017; Kairouz et al., 2021; Li et al., 2020).

Quantum-enhanced learning remains exploratory, but it has already shaped how some researchers think about the next phase of cybersecurity analytics. Variational quantum circuits, quantum kernels, and hybrid quantum-classical workflows are being explored for optimization-intensive classification and anomaly detection problems. The current hardware limitations are substantial, and practical deployment remains distant. Still, this stream is analytically relevant because it reflects a broader move toward hybrid architecture: systems that combine conventional learning for representation with specialized computational modules for search, optimization, or uncertainty handling. In that sense, quantum work matters less because it is immediately deployable and more because it highlights how future cyber analytics may become increasingly modular and heterogeneous (Yang et al., 2019; Smith et al., 2017; Bonawitz et al., 2017; Shokri & Shmatikov, 2015; Carlini et al., 2019; Kasongo & Sun, 2023).

Self-supervised and representation-learning methods deserve special attention because they respond to one of the oldest problems in cybersecurity: the scarcity of trustworthy labels. Massive volumes of logs and flows exist, but only a tiny fraction is ever validated as malicious or benign with high confidence. Pretraining, contrastive learning, masked modeling, denoising objectives, and sequence prediction can exploit this unlabeled abundance to build stronger embeddings before downstream tuning. This direction is especially promising in environments where the signal of interest is rare or delayed, such as insider risk, cloud misuse, or low-and-slow lateral movement. It also fits well with transfer learning, which may reduce the dependence on narrow benchmark datasets and make it easier to adapt models across organizations (Krizhevsky et al., 2017; He et al., 2016; Szegedy et al., 2015; Hochreiter & Schmidhuber, 1997; Cho et al., 2014; Vaswani et al., 2017; Devlin et al., 2019; Dosovitskiy et al., 2021; Kipf & Welling, 2017; Veličković et al., 2018; Ribeiro et al., 2016; Guidotti et al., 2018; Adadi & Berrada, 2018; Doshi-Velez & Kim, 2017; Chen & Guestrin, 2016; Breiman, 2001; Cortes & Vapnik, 1995; Rumelhart et al., 1986; Settles, 2009; He & Sun, 2015; Kraskov et al., 2004; Wang et al., 2017; McMahan et al., 2017; Kairouz et al., 2021; Li et al., 2020).

Reinforcement learning has received less attention than supervised classification, but it is relevant wherever cyber defense involves sequential decision making. Examples include adaptive honeypot placement, dynamic thresholding, moving-target defense, response prioritization, and automated

ISSN: 3067-7386 https://doi.org/10.63646/jaiaa.2025.030201

containment under uncertainty. The attraction of reinforcement learning is obvious: cyber defense is full of trade-offs between action and delay, sensitivity and false positives, or containment and business disruption. Yet the literature also shows why reinforcement learning remains hard in this domain. Reward design is difficult, environments are partially observed, actions can have irreversible costs, and safe exploration is a major concern. For this reason, the most credible near-term uses of reinforcement learning are likely to be advisory or simulation-based rather than fully autonomous (Bonawitz et al., 2017; Shokri & Shmatikov, 2015; Carlini et al., 2019; Kasongo & Sun, 2023).

One further trend is the re-emergence of hybrid symbolic-statistical approaches. Security knowledge is often encoded in rules, attack graphs, ontologies, MITRE ATT&CK mappings, and expert heuristics. Purely neural systems may ignore that structure, while purely symbolic systems may be too brittle. Hybrid models seek to combine both forms of intelligence: learned representations for flexible pattern recognition and symbolic frameworks for constraints, causality, and explanation. In phishing detection, for example, lexical models can be enriched with domain registration rules and URL parsing logic. In endpoint defense, event embedding can relate to provenance rules and process hierarchies. This hybrid turn is likely to matter more over time because security operations still depend heavily on explicit semantics and auditable reasoning (Yan et al., 2022; Wang et al., 2017; Lotfollahi et al., 2020; Apruzzese et al., 2018; Kloft & Laskov, 2010; Goodfellow et al., 2015; Papernot et al., 2018; Biggio & Roli, 2018; Carlini & Wagner, 2017; Tramèr et al., 1705; Shabtai et al., 2014; Rosenblatt, 1958; McCulloch & Pitts, 1943; Kingma & Ba, 2015; Kingma & Welling, 2014; Goodfellow et al., 2014; Vincent et al., 2008; Hinton et al., 2006; Baldi, 2012; Brownlee, 2016).

4. DATASETS, EVALUATION PROTOCOLS, AND WHAT THE NUMBERS OFTEN HIDE

Benchmark datasets have played an indispensable role in the growth of AI-based cybersecurity research, but they have also shaped the field in problematic ways. KDD Cup 99, NSL-KDD, UNSW-NB15, CICIDS2017, Bot-IoT, TON_IoT, N-BaIoT, and Kitsune all made it easier to compare methods and reproduce experiments. At the same time, over-reliance on a small family of public benchmarks has encouraged narrow optimization around specific traffic profiles and label structures. Many papers report excellent accuracy without showing whether the underlying benchmark reflects realistic attack diversity, benign noise, temporal drift, cross-environment variability, or analyst-relevant costs such as false positive burden (Tavallaee et al., 2009; Moustafa & Slay, 2015; Moustafa & Slay, 2016; Sharafaldin et al., 2019; Ring et al., 2019; Mirsky et al., 2018; Meidan et al., 2018; Koroniotis et al., 2019).

Table 3. Common Benchmark Datasets and the Design Implications They Introduce

Dataset	Environment / emphasis	Key analytical value	Persistent caution
NSL-KDD	Legacy network intrusion benchmark	Historical comparability	Outdated traffic and attack assumptions
UNSW-NB15 / CICIDS2017	Enterprise-like modern traffic with richer features	Broader attack diversity and stronger tabular baselines	Possible environment-specific artifacts and leakage risks

ISSN: 3067-7386 https://doi.org/10.63646/jaiaa.2025.030201

Bot-IoT / TON_IoT / N-BaIoT	IoT and IIoT botnet and telemetry scenarios	Brings edge and device behavior into evaluation	Limited transfer to heterogeneous production IoT settings
Kitsune-style local datasets	Resource-aware online anomaly monitoring	Closer view of streaming local network behavior	Harder cross-study comparability

The problem begins with historical inertia. KDD Cup 99 and NSL-KDD remain deeply embedded in literature despite well-known weaknesses, including redundant records, outdated attacks, and unrealistic operating assumptions (Table 3). More recent datasets such as UNSW-NB15 and CICIDS2017 improved realism by expanding feature richness and attack diversity, while Bot-IoT, TON_IoT, and N-BaIoT helped bring IoT and industrial scenarios into view. Yet even these newer datasets are not neutral mirrors of production reality. They reflect the design choices of their creators: what attack tools were used, which traffic was simulated, how labels were assigned, what flows were retained, and what background noise was present (Figure 4). Models trained on them often learn not only malicious behavior but also dataset-specific artifacts (Moustafa & Slay, 2015; Moustafa & Slay, 2016; Sharafaldin et al., 2019; Ring et al., 2019; Mirsky et al., 2018; Meidan et al., 2018; Koroniotis et al., 2019; Hochreiter & Schmidhuber, 1997; Cho et al., 2014).

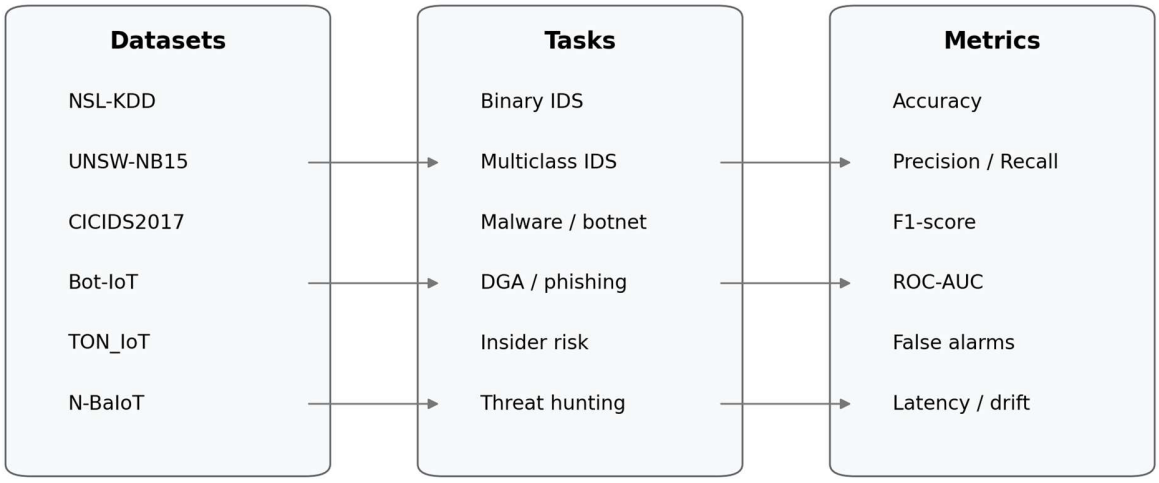

Figure 4. Relationship between Benchmark Datasets, Cybersecurity Tasks, and the Evaluation Metrics Most Often Reported

Class imbalance is another recurring issue. Real organizations generate huge volumes of benign activity and relatively few confirmed attacks, but public datasets handle this imbalance in different ways. Some balance classes for convenience, which can distort precision-recall tradeoffs. Others include rare attack types that are too underrepresented for stable learning or reliable comparison. This matters because accuracy can look impressive even when a model is operationally weak. In security

operations, the burden of false positives can be more consequential than a small drop in overall accuracy. Precision, recall, F1-score, ROC-AUC, PR-AUC, detection latency, calibration, and cost-sensitive metrics therefore need to be interpreted together rather than selectively reported (Garcia-Teodoro et al., 2009; Macas et al., 2022; Tavallaee et al., 2009; Moustafa & Slay, 2015; Moustafa & Slay, 2016; Sharafaldin et al., 2019; Ring et al., 2019; Diro & Chilamkurti, 2018).

Temporal design is also crucial. Many studies shuffle records randomly before train-test splitting, which makes the task easier but less realistic. In production, models confront future data that may differ from the past because of changing user behavior, infrastructure updates, new applications, seasonal patterns, and attacker adaptation. Random splits can leak environment-specific regularities into both training and testing. Time-aware evaluation, cross-dataset transfer, and out-of-domain testing offer a more demanding but more honest picture of model robustness. The literature on concept drift, continual learning, and domain adaptation suggests that future cyber benchmarking should focus much more explicitly on change over time (Macas et al., 2022; Alsaedi et al., 2020; Vaswani et al., 2017; Settles, 2009).

Another issue concerns labeling. Many cyber tasks involve weak supervision, partial labels, or after-the-fact incident reconstruction. What is called benign in a benchmark may include undetected malicious behavior; what is labeled malicious may represent a narrow subset of a broader campaign. This ambiguity is especially strong in insider risk, phishing campaigns, and advanced persistent threats, where intent and context matter. Weak labels do not make AI unusable, but they should temper how we interpret performance. They also strengthen the case for semi-supervised, self-supervised, contrastive, and anomaly-aware methods that do not depend entirely on clean binary labels (Mirsky et al., 2018; Alom & Taha, 2017; Maimó et al., 2018; Ribeiro et al., 2016; Rosenblatt, 1958).

In short, current metrics often hide the hardest part of the problem. High scores on a fixed dataset are informative, but they are not the same as dependable cyber defense. A mature evaluation culture would ask a broader set of questions: How stable is the model under drift? How many alerts does it generate per day at a given threshold? What kinds of false positives dominate? How quickly can an analyst understand the explanation? What happens when labels are delayed or partially wrong? Does the model transfer across networks or clouds? And how does performance change when the adversary adapts? Until these questions become standard, the field will continue to overestimate some capabilities and underestimate the social and operational costs of deployment (Yan et al., 2022; Alsaedi et al., 2020; Kloft & Laskov, 2010; Goodfellow et al., 2015; Papernot et al., 2018; Biggio & Roli, 2018; Carlini & Wagner, 2017; Tramèr et al., 1705; Shabtai et al., 2014).

Reproducibility deserves more attention than it typically receives. Many papers describe feature counts, training ratios, and headline metrics, but omit critical details such as feature leakage, exact preprocessing order, deduplication rules, imbalance treatment, threshold selection, and hyperparameter search strategy. These omissions matter because small design choices can generate large metric swings on widely used cyber datasets. The field has therefore reached a point where better reporting standards would create as much value as yet another incremental architecture paper. Transparent pipeline reporting, fixed evaluation splits, public code, and cross-dataset validation protocols would make comparisons more meaningful and reduce the risk of benchmark gaming (Macas et al., 2022;

Sharafaldin et al., 2019; Ring et al., 2019; Mirsky et al., 2018; Meidan et al., 2018; Koroniotis et al., 2019; Alsaedi et al., 2020; Diro & Chilamkurti, 2018).

There is also a deeper methodological issue. Most benchmark tasks are framed around retrospective detection: the model sees a curated slice of traffic and predicts a label already known to the researcher. Real security work is often prospective, open-ended, and resource constrained. Analysts do not inspect everything. They choose what to investigate. That means ranking quality, alert clustering, case narrative, and evidence sufficiency can matter more than raw classification on a balanced test fold. Some of the most useful future benchmarks may therefore look less like conventional machine learning datasets and more like analyst workbench simulations in which the model is judged by its contribution to triage quality and investigation efficiency (Yan et al., 2022; Alsaedi et al., 2020; Kloft & Laskov, 2010; Goodfellow et al., 2015; Papernot et al., 2018; Biggio & Roli, 2018; Carlini & Wagner, 2017; Tramèr et al., 1705; Shabtai et al., 2014; Tuor et al., 2017; Creech & Hu, 2014; Rigaki & Garcia, 2018; Vinayakumar et al., 2018).

The issue becomes even sharper in cloud, SaaS, and identity-centric environments, where the semantics of events can vary across vendors and change rapidly with software updates. A benchmark collected at one moment may lose realism surprisingly quickly. This is another reason to complement static dataset evaluation with synthetic stress tests, time-aware replay, and environment-sensitive validation. It may also explain why some of the strongest industrial systems emphasize continual monitoring, threshold management, and analyst feedback channels rather than relying on one high-performing model frozen in time (Alsaedi et al., 2020; Ribeiro et al., 2016; Settles, 2009; He & Sun, 2015; Kraskov et al., 2004; Wang et al., 2017; McMahan et al., 2017; Kairouz et al., 2021; Li et al., 2020).

5. HYBRID ARCHITECTURES AND EMERGING DESIGN PATTERNS

A major lesson from the literature is that no single model family is sufficient for the diversity of cybersecurity tasks. This is why hybrid architecture now dominates much of the serious work in the field. Hybridization can mean several things. It may refer to combining handcrafted features with learned embeddings, using CNNs to capture local patterns and RNNs or transformers to capture temporal dependencies, fusing graph structure with text or flow data, or linking predictive models with post hoc explanation modules. In some studies, hybridization also includes optimization layers, federated coordination, or quantum-enhanced kernels. The common idea is that cyber problems are heterogeneous and therefore benefit from analytical pipelines that are heterogeneous as well (Xin et al., 2018; Berman et al., 2019; Macas et al., 2022; Yan et al., 2022; Hozouri et al., 2025; Mirsky et al., 2018; Niyaz et al., 2016; Alom & Taha, 2017; Yousefi-Azar et al., 2017; Wang et al., 2017; Lotfollahi et al., 2020; Apruzzese et al., 2018; Yang et al., 2019; Smith et al., 2017; Bonawitz et al., 2017; Shokri & Shmatikov, 2015; Carlini et al., 2019; Kasongo & Sun, 2023).

A representative hybrid pipeline begins with data engineering. Raw flows, endpoint events, and identity traces are normalized into consistent time windows or entity-centric views. Feature selection and embedding then reduce sparsity while preserving operational semantics. A deep representation layer learns temporal, spatial, or relational patterns. A fusion layer integrates multiple sources, and a decision layer generates ranked outputs, confidence measures, or explanations. Finally, an analyst-

facing layer translates the result into a triage action, hunting lead, block recommendation, or escalation path. The advantage of this architecture is not merely better classification. It is the ability to connect model inference with downstream action, which is the step that many narrowly designed academic systems leave underdeveloped (Sperotto et al., 2010; Xin et al., 2018; Berman et al., 2019; Macas et al., 2022; Yan et al., 2022; Mirsky et al., 2018; Alsaedi et al., 2020; Alom & Taha, 2017; Kloft & Laskov, 2010).

Hybridization also appears in the growing relationship between supervised learning and anomaly detection. Supervised models work well when labels are abundant and attack categories are known. Anomaly detectors become more valuable when labels are scarce, attacks are novel, or the environment changes rapidly. The strongest systems increasingly combine these logics. Known threats can be handled with discriminative classifiers; unknown or weakly understood deviations can be surfaced through reconstruction error, density estimation, distance-based outlier detection, or one-class learning. This mixed strategy better matches operational reality, where defenders must recognize both what they have seen before and what does not fit known patterns (Garcia-Teodoro et al., 2009; Sperotto et al., 2010; Mirsky et al., 2018; Niyaz et al., 2016; He et al., 2016; Szegedy et al., 2015; Hochreiter & Schmidhuber, 1997; Cho et al., 2014).

Another important design pattern is multimodal fusion. Cyber incidents rarely reveal themselves in one source alone. A phishing campaign may combine email text, domain reputation, authentication anomalies, browser events, and endpoint behavior. Cloud intrusion may involve API abuse, privilege escalation, and lateral movement reflected in entirely different log types. Combining these signals is analytically hard because they have different schemas and time scales, but it is precisely where AI can offer value. Sequence models can align event streams, graph models can connect entities, and attention mechanisms can learn which modalities matter most in a given context. Multimodal fusion is therefore becoming central to the move from single-event alerting toward campaign-level reasoning (Yousefi-Azar et al., 2017; Wang et al., 2017; Lotfollahi et al., 2020; Apruzzese et al., 2018; Ribeiro et al., 2016; Guidotti et al., 2018; Adadi & Berrada, 2018; Doshi-Velez & Kim, 2017; Chen & Guestrin, 2016; Breiman, 2001; Cortes & Vapnik, 1995; Rumelhart et al., 1986; Rosenblatt, 1958; McCulloch & Pitts, 1943; Kingma & Ba, 2015; Kingma & Welling, 2014; Goodfellow et al., 2014).

Explainability is itself becoming hybrid. Post hoc explanation methods such as SHAP and LIME remain important, but many researchers now argue that explainability should be built into system design rather than appended afterward. Rule extraction, concept bottlenecks, prototype-based classification, attention mapping, and hierarchical evidence presentation all reflect this shift. In cybersecurity, this matters because the user of the explanation is often an analyst under time pressure who needs evidence that is actionable rather than philosophically satisfying. A useful explanation may consist of a handful of decisive features, a short behavioral narrative, and a comparison with similar incidents, rather than a full mathematical decomposition of the model (Yan et al., 2022; Kloft & Laskov, 2010; Goodfellow et al., 2015; Papernot et al., 2018; Biggio & Roli, 2018; Carlini & Wagner, 2017; Tramèr et al., 1705; Shabtai et al., 2014; Vincent et al., 2008; Hinton et al., 2006; Baldi, 2012; Brownlee, 2016).

Federated and distributed architectures are another branch of hybrid thinking. Large organizations and critical infrastructures rarely operate as one uniform environment. They contain business units, partner networks, operational technology segments, cloud regions, and edge devices with different data access rules and risk profiles. Federated learning and split-learning approaches offer a way to share model intelligence without centralizing raw data. This can strengthen collective defense, especially in IoT and cross-enterprise settings. However, hybrid federated systems must address non-IID data, communication cost, update poisoning, and the practical problem that each site often labels incidents differently. The most promising work therefore combines federated learning with trust scoring, robust aggregation, local explainability, and drift-aware adaptation (Settles, 2009; He & Sun, 2015; Kraskov et al., 2004; Wang et al., 2017; McMahan et al., 2017; Kairouz et al., 2021; Li et al., 2020).

Quantum-enhanced cybersecurity remains a frontier rather than a mature deployment domain, but it fits naturally into the hybrid design logic. Current proposals often use classical deep networks for feature extraction and quantum kernels or variational circuits for downstream optimization and classification. Whether these approaches will outperform strong classical baselines on a scale remains open. Still, they have already pushed the field to think more carefully about computational bottlenecks, search efficiency, and the possibility of modular analytic stacks that do not depend on one computational paradigm alone. Even if near-term practical impact is limited, the conceptual contribution is useful: cyber analytics is becoming an ecosystem of interoperable components rather than a single monolithic detector (Yang et al., 2019; Smith et al., 2017; Bonawitz et al., 2017; Shokri & Shmatikov, 2015; Carlini et al., 2019; Kasongo & Sun, 2023).

Security operations centers increasingly resemble multimodal decision rooms rather than pipeline-driven alert factories. A single incident may involve mail telemetry, identity anomalies, DNS requests, endpoint processes, ticket history, business-unit criticality, and prior threat reports. Hybrid architectures are well suited to this environment because they let designers allocate different representational tasks to different modules. Sequence models can summarize time dependence; graph modules can connect entities; language models can digest case notes or intelligence reports; and explanation layers can translate the result back into analyst-relevant evidence. In effect, hybridization is not just a model choice. It is a recognition that cyber incidents are composite objects that require composite analytics (Yousefi-Azar et al., 2017; Wang et al., 2017; Lotfollahi et al., 2020; Apruzzese et al., 2018; Ribeiro et al., 2016; Guidotti et al., 2018; Adadi & Berrada, 2018; Doshi-Velez & Kim, 2017; Chen & Guestrin, 2016; Breiman, 2001; Cortes & Vapnik, 1995; Rumelhart et al., 1986; Rosenblatt, 1958; McCulloch & Pitts, 1943; Kingma & Ba, 2015; Kingma & Welling, 2014; Goodfellow et al., 2014).

Operational technology and industrial control environments add another twist. These environments often involve strict latency constraints, limited tolerance for false shutdowns, and unusual protocol behavior that differs sharply from enterprise IT. Here, hybrid systems may combine lightweight edge inference with central contextual reasoning. Some signals must be analyzed close to the asset for speed, while others require broader cross-system context to become meaningful. This layered design reinforces the point that cybersecurity AI is increasingly an architectural discipline. Designers must think about where computation happens, where context is stored, how evidence is synchronized, and

when the system should escalate to humans rather than act autonomously (Meidan et al., 2018; Koroniotis et al., 2019; Alsaedi et al., 2020; Diro & Chilamkurti, 2018; Maimó et al., 2018; Tuor et al., 2017; Creech & Hu, 2014; Rigaki & Garcia, 2018; Vinayakumar et al., 2018; Settles, 2009; He & Sun, 2015; Kraskov et al., 2004; Wang et al., 2017; McMahan et al., 2017; Kairouz et al., 2021; Li et al., 2020).

Another important frontier is fusion between anomaly detection and external intelligence. Traditional anomaly detectors are strong at identifying deviation but weak at telling defenders why the deviation matters. Threat intelligence, by contrast, provides semantics but can be noisy, incomplete, and difficult to operate. Hybrid systems can potentially connect the two. For example, anomalous traffic toward infrastructure previously associated with a malware family carries a different meaning than the same anomaly directed toward an unknown destination. Likewise, a suspicious process lineage combined with intelligence about a newly abused toolchain may justify a faster response. This kind of fusion is still uneven in the literature, but it is one of the clearest paths from isolated analytics toward truly contextual cyber decision support (Yousefi-Azar et al., 2017; Wang et al., 2017; Lotfollahi et al., 2020; Apruzzese et al., 2018; Ribeiro et al., 2016; Guidotti et al., 2018; Adadi & Berrada, 2018; Doshi-Velez & Kim, 2017; Chen & Guestrin, 2016; Breiman, 2001; Cortes & Vapnik, 1995; Rumelhart et al., 1986; Rosenblatt, 1958; McCulloch & Pitts, 1943; Kingma & Ba, 2015; Kingma & Welling, 2014; Goodfellow et al., 2014).

6. DEPLOYMENT CHALLENGES, GOVERNANCE, AND THE LIMITS OF ALGORITHMIC CONTROL

The move from benchmark success to operational value remains the most difficult step in AI-based cybersecurity. One persistent challenge is data quality. Security logs are often incomplete, inconsistent, duplicated, or missing contextual fields needed for reliable interpretation. Asset inventories are inaccurate, identities are poorly resolved, and cloud environments generate telemetry with changing semantics. When model builders discuss feature importance without discussing the fragility of the underlying telemetry, they overstate what the model itself is doing. A technically sophisticated architecture cannot compensate indefinitely for weak instrumentation and unstable data pipelines (Sperotto et al., 2010; Berman et al., 2019; Macas et al., 2022; Alsaedi et al., 2020; Goodfellow et al., 2015; Tuor et al., 2017).

False positives remain another decisive barrier. Security operations centers can tolerate some level of imperfection, but they cannot sustainably absorb large volumes of low-quality alerts. This is where the literature on model calibration, threshold design, active learning, and cost-sensitive evaluation becomes practically important. A detector with slightly lower recall may create more operational value if it produces clearer, more stable, and more explainable cases. Conversely, a technically impressive model may fail because it shifts triage labor onto analysts without improving decision quality. Cyber AI should therefore be evaluated not only by classification metrics but also by analyst workload, time-to-triage, override frequency, and response outcomes (Yan et al., 2022; Alsaedi et al., 2020; Kloft & Laskov, 2010; Goodfellow et al., 2015; Papernot et al., 2018; Biggio & Roli, 2018; Carlini & Wagner, 2017; Tramèr et al., 1705; Shabtai et al., 2014).

Adversarial pressure makes cybersecurity different from many other AI application domains. Attackers adapt. They can poison data, probe model behavior, exploit blind spots, or mimic benign patterns (Table 4). They can also use AI themselves to generate phishing lures, craft malware variants, manipulate content, and accelerate reconnaissance. This means that cyber models should not be treated as static artifacts. They must be monitored for drift, stress-tested under adversarial conditions, and periodically revalidated against new tactics and environments. The literature on adversarial machine learning is therefore highly relevant to cybersecurity even when the two communities use different terminology (Tramèr et al., 1705; Lundberg & Lee, 2017).

Table 4. Operational Risks of AI-Enabled Cyber Defense and Corresponding Governance Responses

Operational risk	Why it matters	Governance response
False alarm overload	High alert volume erodes trust and wastes analyst time	Use calibrated thresholds, case ranking, and analyst feedback loops
Opacity and drift	Models become hard to justify and degrade under change	Maintain explanation layers, monitoring, retraining, and version control
Privacy and surveillance creep	Security analytics can exceed legitimate monitoring scope	Apply minimization, purpose limitation, and reviewable access governance
Vendor black-box dependence	Third-party tools may hide assumptions and blind spots	Require auditability, benchmark disclosure, and escalation transparency

Governance is equally important. As AI becomes more deeply embedded in cyber operations, organizations need clear policies for model ownership, versioning, approval, override, and audit. They need to know which models are advisory and which can trigger automated action. They need processes for documenting training data, monitoring drift, explaining major changes, and handling incidents in which the model itself becomes part of the problem. In regulated sectors, explainability and accountability are not optional extras; they shape whether AI can be used at all. The field is beginning to recognize this, but governance discussion is still far thinner than model innovation discussion (Yan et al., 2022; Alsaedi et al., 2020; Goodfellow et al., 2015; Tuor et al., 2017; Creech & Hu, 2014; Rigaki & Garcia, 2018; Vinayakumar et al., 2018; Vincent et al., 2008; Hinton et al., 2006; Baldi, 2012; Brownlee, 2016).

A final limitation concerns strategic overreach. AI can meaningfully improve certain parts of cyber defense, but it cannot eliminate structural vulnerability. Weak identity controls, poor patch discipline, unmanaged attack surface, brittle supply chains, or under-resourced security teams cannot be solved by a better classifier alone. In practice, the most valuable AI deployments are often those that augment disciplined operational programs rather than substitute for them. This is why resilience is such a useful framing. A resilient organization uses AI to improve visibility, prioritization, and learning, but it does not imagine that models alone will resolve the political, organizational, and infrastructural realities of

cybersecurity (Sperotto et al., 2010; Berman et al., 2019; Alsaedi et al., 2020; Goodfellow et al., 2015; Tuor et al., 2017; Creech & Hu, 2014; Vinayakumar et al., 2018).

The workforce dimension should also be taken seriously. AI changes what security analysts do. It can reduce repetitive screening, but it can also create new burdens in prompt engineering, threshold tuning, explanation interpretation, and model monitoring. Teams need not only data scientists and detection engineers, but people who can translate between operational pain points and analytical design. When that bridge is missing, models are often tuned to what is easy to measure rather than what analysts genuinely need. In this respect, cyber-AI is partly a capability-building problem. Success depends on whether organizations develop the human roles and collaboration routines required to make AI outputs actionable (Yan et al., 2022; Alsaedi et al., 2020; Kloft & Laskov, 2010; Goodfellow et al., 2015; Papernot et al., 2018; Biggio & Roli, 2018; Carlini & Wagner, 2017; Tramèr et al., 1705; Shabtai et al., 2014; Tuor et al., 2017; Creech & Hu, 2014; Rigaki & Garcia, 2018; Vinayakumar et al., 2018).

Ethics and privacy are equally central. Cybersecurity systems frequently process employee activity, communication metadata, browsing traces, and behavioral profiles. Even when the goal is legitimate defense, poorly designed analytics can create excessive surveillance, opaque risk scoring, and secondary uses of data that exceed original security purposes. Explainability and governance matter here not only because regulators may ask for them, but because internal legitimacy depends on them. Organizations need to distinguish narrowly tailored security monitoring from generalized behavioral control. As AI becomes more powerful, that distinction will be harder to maintain unless privacy, minimization, and purpose limitation are built into design and oversight (Yan et al., 2022; Kloft & Laskov, 2010; Goodfellow et al., 2015; Papernot et al., 2018; Biggio & Roli, 2018; Carlini & Wagner, 2017; Tramèr et al., 1705; Shabtai et al., 2014; Vincent et al., 2008; Hinton et al., 2006; Baldi, 2012; Brownlee, 2016; Settles, 2009; He & Sun, 2015; Kraskov et al., 2004; Wang et al., 2017; McMahan et al., 2017; Kairouz et al., 2021; Li et al., 2020).

A related concern is procurement and dependency risk. Many organizations adopt cyber-AI through vendors, managed services, or cloud-native security platforms rather than building systems internally. This creates efficiency, but it also introduces opacity. Buyers may receive a risk score without knowing how the model was trained, what telemetry it uses, how often it drifts, or how false positives are handled across customer environments. Vendor concentration can also create systemic exposure if widely deployed models share common blind spots. Future governance work therefore needs to address not only internal model control but also third-party assurance, benchmarking, and auditability for AI-enabled security products (Berman et al., 2019; Macas et al., 2022; Yan et al., 2022; Hozouri et al., 2025; Sommer & Paxson, 2010; Alsaedi et al., 2020; Goodfellow et al., 2015; Tuor et al., 2017; Creech & Hu, 2014; Rigaki & Garcia, 2018; Vinayakumar et al., 2018).

7. FUTURE RESEARCH DIRECTIONS

Future research is likely to move in several mutually reinforcing directions. The first is multimodal cyber intelligence. The field has already shown that logs, flows, binaries, text, and graph structures can each be modeled effectively. The next step is learning across them. This includes linking threat reports to telemetry, connecting identity context to endpoint anomalies, and combining graph structure with natural-language analyst notes. Progress here depends not only on larger models but also on better data

alignment, temporal synchronization, and entity resolution. The biggest gains may come less from bigger architectures and more from cleaner ways of connecting heterogeneous evidence (Yousefi-Azar et al., 2017; Wang et al., 2017; Lotfollahi et al., 2020; Apruzzese et al., 2018; Ribeiro et al., 2016; Guidotti et al., 2018; Adadi & Berrada, 2018; Doshi-Velez & Kim, 2017; Chen & Guestrin, 2016; Breiman, 2001; Cortes & Vapnik, 1995; Rumelhart et al., 1986; Rosenblatt, 1958; McCulloch & Pitts, 1943; Kingma & Ba, 2015; Kingma & Welling, 2014; Goodfellow et al., 2014).

A second direction concerns continual, transfer, and causal learning (Figure 5). Current cyber models often assume that the data-generating process is stationary enough for repeated retraining to be sufficient. Operational environments do not behave that way. Applications change, users change, attackers change, and infrastructure changes. Continual learning and drift-aware updating are therefore essential. Yet adaptation should not be purely reactive. Causal perspectives may help the field move beyond correlation-heavy detection toward reasoning about interventions: what would happen if access policies change, if a segment is isolated, or if a suspicious process is killed? Cybersecurity analytics will become more valuable when it can support not just detection but intervention design (Alsaedi et al., 2020; Ribeiro et al., 2016; Settles, 2009).

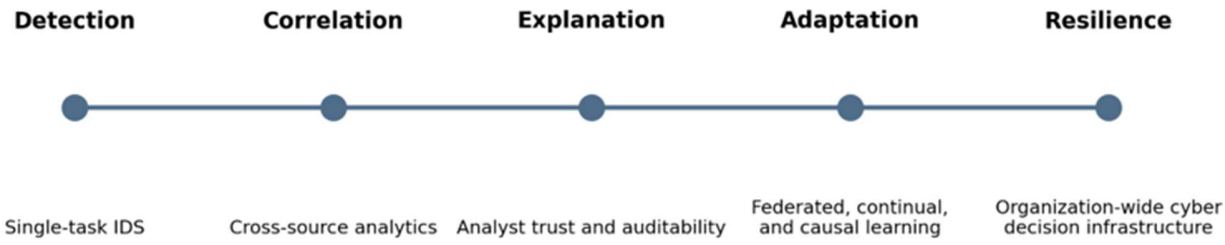

Figure 5. Research Roadmap from Narrow Detection Toward Resilient, Context-Rich, and Adaptive Cybersecurity Decision Infrastructure

A third direction is human-AI collaboration. Many papers still evaluate models as if decisions were taken directly from algorithmic outputs. In real operations, analysts investigate, override, escalate, and reinterpret alerts. Future work should therefore examine the interaction between explanation design, analyst trust, workload, and incident outcomes. Which explanation formats actually help? When does confidence information improve decisions, and when does it mislead? How should systems learn from analyst feedback without overfitting local habits or bias? These are not secondary questions. They shape whether AI becomes a force multiplier or merely another dashboard (Yan et al., 2022; Kloft & Laskov, 2010; Goodfellow et al., 2015; Papernot et al., 2018; Biggio & Roli, 2018; Carlini & Wagner, 2017; Tramèr et al., 1705; Shabtai et al., 2014; Vincent et al., 2008; Hinton et al., 2006; Baldi, 2012; Brownlee, 2016).

A fourth direction is privacy-preserving collaborative defense. As attacks increasingly cross organizational and sectoral boundaries, isolated learning becomes less attractive. Federated learning, secure aggregation, homomorphic techniques, split learning, and privacy-aware synthetic data generation may all contribute to a more cooperative defense ecosystem. But technical feasibility alone will not be enough. Incentives, liability, data governance, and trust between institutions matter. The

most interesting work in this area will likely sit at the intersection of machine learning, security engineering, and institutional design (Settles, 2009; He & Sun, 2015; Kraskov et al., 2004; Wang et al., 2017; McMahan et al., 2017; Kairouz et al., 2021; Li et al., 2020).

Finally, hybrid and specialized computing will remain an active frontier. This includes not only quantum-enhanced learning, but also graph-native inference engines, neuromorphic anomaly detection, and hardware-aware edge models for constrained devices. The future will probably not belong to one universal architecture. It will belong to interoperable systems that allocate different tasks to different analytical components depending on latency, data modality, privacy constraints, and evidentiary requirements. In that sense, the long-term trajectory of cybersecurity analytics is toward modular intelligence: architectures that are adaptive, explainable, and operationally grounded rather than merely accurate in isolation (Wang et al., 2017; Lotfollahi et al., 2020; Apruzzese et al., 2018; Rosenblatt, 1958; McCulloch & Pitts, 1943; Kingma & Ba, 2015; Kingma & Welling, 2014; Goodfellow et al., 2014; Yang et al., 2019; Smith et al., 2017; Bonawitz et al., 2017; Shokri & Shmatikov, 2015; Carlini et al., 2019; Kasongo & Sun, 2023).

Synthetic data generation is likely to play a larger role in future cyber research, but it should be approached carefully. Synthetic environments can help address privacy barriers, rare-event scarcity, and the need for stress testing. Generative models may also be useful for creating harder negative samples or simulating attacker adaptation. Yet synthetic data can easily reproduce the blind spots of the generator or create unrealistic regularities that make tasks artificially easy. The most valuable work in this space will probably combine synthetic augmentation with rigorous checks against real-world heterogeneity and analyst judgment, rather than treating synthetic generation as a substitute for difficult data collection (Hamilton et al., 2017; Lundberg & Lee, 2017; Cortes & Vapnik, 1995; Settles, 2009; He & Sun, 2015; Kraskov et al., 2004; Wang et al., 2017; McMahan et al., 2017; Kairouz et al., 2021; Li et al., 2020).

Large language models and agentic systems will also become more visible in cybersecurity, especially in alert summarization, threat intelligence extraction, case-note drafting, retrieval over knowledge bases, and natural-language interfaces for analysts. Their promise is clear: they can reduce cognitive overhead and make historical knowledge easier to use. Their risks are equally clear: hallucination, overconfidence, prompt injection, data leakage, and the possibility that fluent output may be mistaken for validated evidence. Research in this area should therefore focus less on generic enthusiasm and more on evidence grounding, source attribution, confidence signaling, and the design of human verification loops (Ribeiro et al., 2016; Guidotti et al., 2018; Adadi & Berrada, 2018; Doshi-Velez & Kim, 2017; Chen & Guestrin, 2016; Breiman, 2001; Cortes & Vapnik, 1995; Rumelhart et al., 1986; Vincent et al., 2008; Hinton et al., 2006; Baldi, 2012; Brownlee, 2016; Bonawitz et al., 2017; Shokri & Shmatikov, 2015; Carlini et al., 2019; Kasongo & Sun, 2023).

Public-private collaboration is another area likely to shape the next decade of cyber AI. Threats move across sectors, but data-sharing arrangements remain fragmented. More progress will depend on institutions that can support privacy-preserving joint analysis, common evaluation practices, interoperable schemas, and trusted mechanisms for distributing model updates or intelligence. Here again, the challenge is not purely technical. It involves governance, incentives, legal compatibility, and

long-term stewardship. If these institutional questions remain unresolved, even technically strong collaborative learning methods may struggle to achieve meaningful scale (Settles, 2009; He & Sun, 2015; Kraskov et al., 2004; Wang et al., 2017; McMahan et al., 2017; Kairouz et al., 2021; Li et al., 2020).

8. CONCLUSION

Artificial intelligence changes cybersecurity analytics from a largely retrospective and rule-bound activity into a more continuous, context-rich, and intervention-oriented decision process. Its strongest contribution lies not only in better pattern recognition but in the possibility of linking heterogeneous telemetry, layered inference, explanation, and action. The review has shown that deep learning, graph models, transformers, explainable AI, federated learning, and quantum-enhanced ideas each contribute pieces of this transformation. At the same time, the literature makes clear that predictive power alone is not enough. Data realism, evaluation discipline, explanation quality, drift management, governance, and organizational fit all shape whether AI creates practical security value (Buczak & Guven, 2016; Garcia-Teodoro et al., 2009; Sperotto et al., 2010; Wu & Banzhaf, 2010; Xin et al., 2018; Berman et al., 2019; Macas et al., 2022; Yan et al., 2022; Hozouri et al., 2025; Sommer & Paxson, 2010; Alsaedi et al., 2020; Kloft & Laskov, 2010; Goodfellow et al., 2015; Papernot et al., 2018; Biggio & Roli, 2018; Carlini & Wagner, 2017; Tramèr et al., 1705; Shabtai et al., 2014; Tuor et al., 2017; Creech & Hu, 2014; Rigaki & Garcia, 2018; Vinayakumar et al., 2018).

The field is therefore best understood not as a race to find a universally superior detector, but as a broader effort to design dependable cyber decision infrastructure. Models must be selected and combined in ways that reflect task structure, evidence quality, operational constraints, and human accountability. Benchmark culture must become more honest about what current numbers mean. And future systems must treat resilience, not just accuracy, as the central endpoint. Organizations that approach AI in this layered way are more likely to build cyber capabilities that remain useful when environments change, attackers adapt, and security teams need not only alerts, but defensible judgment (Sperotto et al., 2010; Berman et al., 2019; Macas et al., 2022; Yan et al., 2022; Hozouri et al., 2025; Sommer & Paxson, 2010; Alsaedi et al., 2020; Goodfellow et al., 2015; Tuor et al., 2017; Creech & Hu, 2014; Rigaki & Garcia, 2018; Vinayakumar et al., 2018).

The broad lesson of the current literature is therefore double-edged. AI has undeniably expanded the reach of cybersecurity analytics; it can surface patterns, correlations, and weak signals that manual methods or static rules often miss. But it has also forced the field to confront questions that are more organizational and epistemic than purely technical: what counts as reliable evidence, how much automation is defensible, where explanation becomes necessary, and how to preserve human judgment without surrendering speed. Those questions will shape the real trajectory of cyber AI at least as much as the next incremental gain in benchmark accuracy (Buczak & Guven, 2016; Garcia-Teodoro et al., 2009; Sperotto et al., 2010; Wu & Banzhaf, 2010; Xin et al., 2018; Berman et al., 2019; Macas et al., 2022; Yan et al., 2022; Hozouri et al., 2025; Sommer & Paxson, 2010; Alsaedi et al., 2020; Kloft & Laskov, 2010; Goodfellow et al., 2015; Papernot et al., 2018; Biggio & Roli, 2018; Carlini & Wagner, 2017; Tramèr et al., 1705; Shabtai et al., 2014; Tuor et al., 2017; Creech & Hu, 2014; Rigaki & Garcia, 2018; Vinayakumar et al., 2018).

Author Contributions

Emily Zhang conceptualized the study and led the manuscript drafting, Michael Chen developed the analytical framework and supported data interpretation, Sophia Wang supervised the project and coordinated the overall revision as the corresponding author, and Daniel Liu contributed to the literature synthesis, figure and table design, and final proofreading.

Declarations

Ethics approval was not applicable because this article does not report research involving human participants or animals. Consent to participate was not applicable. Consent for publication was not applicable. The author declares no competing financial or non-financial interests. This manuscript was prepared as a scholarly review and did not receive external funding. No proprietary dataset is distributed with this manuscript. The visual figures are interpretive syntheses designed for explanatory purposes.

REFERENCES

Adadi, A. & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alom, M. Z. & Taha, T. M. (2017). Network intrusion detection for cyber security using unsupervised deep learning approaches. 2017 IEEE National Aerospace and Electronics Conference, 63-69. https://doi.org/10.1109/NAECON.2017.8268767

Alsaedi, A., Moustafa, N., Tari, Z., Mahmood, A., & Anwar, A. (2020). TON_IoT Telemetry Dataset: A New Generation Dataset of IoT and IIoT for Data-Driven Intrusion Detection Systems. IEEE Access, 8, 165130-165150. https://doi.org/10.1109/ACCESS.2020.3022862

Aminanto, M. E. & Kim, K. (2017). Improving detection of Wi-Fi impersonation by fully unsupervised deep learning. in Information Security Applications, 212-223. https://doi.org/10.1007/978-3-319-56549-1_16

Apruzzese, G., Ferretti, L., Marchetti, M., Guido, A., & Colajanni, M. (2018). On the effectiveness of machine and deep learning for cyber security. 2018 10th International Conference on Cyber Conflict, 371-390. https://doi.org/10.23919/CYCON.2018.8405026

Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. Proceedings of ICML Workshop on Unsupervised and Transfer Learning, 37-49. https://doi.org/10.48550/arXiv.1206.5538

Berman, D. S., Buczak, A. L., Chavis, J. S., & Corbett, C. L. (2019). A Survey of Deep Learning Methods for Cyber Security. Information, 10, 4, 122. https://doi.org/10.3390/info10040122

Biggio, B. & Roli, F. (2018). Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. Pattern Recognition, 84, 317-331. https://doi.org/10.1016/j.patcog.2018.07.023

Bonawitz, K., Ivanov, V., Kreuter, B., et al. (2017). Practical Secure Aggregation for Privacy-Preserving Machine Learning. 2017 ACM SIGSAC Conference on Computer and Communications Security, 1175-1191. https://doi.org/10.1145/3133956.3133982

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2016). A gentle introduction to gradient boosting for machine learning. . https://doi.org/10.48550/arXiv.1603.02754

Buczak, A. L. & Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. IEEE Communications Surveys & Tutorials, 18, 2, 1153-1176. https://doi.org/10.1109/COMST.2015.2494502

Carlini, N. & Wagner, D. (2017). Towards Evaluating the Robustness of Neural Networks. 2017 IEEE Symposium on Security and Privacy, 39-57. https://doi.org/10.1109/SP.2017.49

Carlini, N., Liu, C., Erlingsson, U., Kos, J., & Song, D. (2019). The Secret Sharer: Measuring Unintended Neural Network Memorization & Extractability. 28th USENIX Security Symposium, 267-284. https://doi.org/10.48550/arXiv.1802.08232

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785

Cho, K., van Merrienboer, B., Gulcehre, C., et al. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. . https://doi.org/10.48550/arXiv.1406.1078

Cortes, C. & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273-297. https://doi.org/10.1007/BF00994018

Creech, G. & Hu, J. (2014). A semantic approach to host-based intrusion detection systems using contiguous and discontiguous system call patterns. IEEE Transactions on Computers, 63, 4, 807-819. https://doi.org/10.1109/TC.2012.249

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. . https://doi.org/10.48550/arXiv.1810.04805

Diro, A. A. & Chilamkurti, N. (2018). Distributed attack detection scheme using deep learning approach for Internet of Things. Future Generation Computer Systems, 82, 761-768. https://doi.org/10.1016/j.future.2017.08.043

Doshi-Velez, F. & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. . https://doi.org/10.48550/arXiv.1702.08608

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. . https://doi.org/10.48550/arXiv.2010.11929

Garcia-Teodoro, P., Diaz-Verdejo, J., Macia-Fernandez, G., & Vazquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. Computers & Security, 28, 1-2, 18-28. https://doi.org/10.1016/j.cose.2008.08.003

Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. . https://doi.org/10.48550/arXiv.1412.6572

Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1406.2661

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys, 51, 5, 93. https://doi.org/10.1145/3236009

Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1706.02216

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 770-778. https://doi.org/10.1109/CVPR.2016.90

He, K. & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1506.01497

Hinton, G. E. & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313, 5786, 504-507. https://doi.org/10.1126/science.1127647

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18, 7, 1527-1554. https://doi.org/10.1162/neco.2006.18.7.1527

Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9, 8, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hozouri, A., Mirzaei, A., & Effatparvar, M. (2025). A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges. Discover Artificial Intelligence, 5, 1. https://doi.org/10.1007/s44163-025-00578-1

Jordan, M. I. & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349, 6245, 255-260. https://doi.org/10.1126/science.aaa8415

Kairouz, P., McMahan, H. B., Avent, B., et al. (2021). Advances and Open Problems in Federated Learning. Foundations and Trends in Machine Learning, 14, 1-2, 1-210. https://doi.org/10.1561/2200000083

Kasongo, S. M. & Sun, Y. (2023). A deep learning technique for intrusion detection system using a Recurrent Neural Networks based framework. Computer Communications, 199, 113-125. https://doi.org/10.1016/j.comcom.2022.12.010

Kim, J., Kim, J., Thu, H., & Kim, H. (2016). Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection. 2016 International Conference on Platform Technology and Service, 1-5. https://doi.org/10.1109/PlatCon.2016.7456805

Kingma, D. P. & Ba, J. (2015). Adam: A Method for Stochastic Optimization. . https://doi.org/10.48550/arXiv.1412.6980

Kingma, D. P. & Welling, M. (2014). Auto-Encoding Variational Bayes. . https://doi.org/10.48550/arXiv.1312.6114

Kipf, T. N. & Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. . https://doi.org/10.48550/arXiv.1609.02907

Kloft, M. & Laskov, P. (2010). Online anomaly detection under adversarial impact. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 405-412. https://doi.org/10.48550/arXiv.1004.2514

Koroniotis, N., Moustafa, N., Sitnikova, E., & Turnbull, B. (2019). Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-IoT dataset. Future Generation Computer Systems, 100, 779-796. https://doi.org/10.1016/j.future.2019.05.041

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. Physical Review E, 69, 6, 066138. https://doi.org/10.1103/PhysRevE.69.066138

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM, 60, 6, 84-90. https://doi.org/10.1145/3065386

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436-444. https://doi.org/10.1038/nature14539

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated Optimization in Heterogeneous Networks. . https://doi.org/10.48550/arXiv.1812.06127

Lotfollahi, M., Zade, R. S. H., Siavoshani, M. J., & Saberian, M. (2020). Deep Packet: A Novel Approach For Encrypted Traffic Classification Using Deep Learning. Soft Computing, 24, 1999-2012. https://doi.org/10.1007/s00500-019-04030-2

Lundberg, S. M. & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1705.07874

Macas, M., Wu, C., & Fuertes, W. (2022). A survey on deep learning for cybersecurity: Progress, challenges, and opportunities. Computer Networks, 212, 109032. https://doi.org/10.1016/j.comnet.2022.109032

Maimó, L. F., Gómez, A. L., Clemente, F. J. G., & Pérez, M. G. (2018). A self-adaptive deep learning-based system for anomaly detection in 5G networks. IEEE Access, 6, 7700-7712. https://doi.org/10.1109/ACCESS.2018.2805768

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5, 115-133. https://doi.org/10.1007/BF02478259

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. . https://doi.org/10.48550/arXiv.1602.05629

Meidan, Y., Bohadana, M., Mathov, Y., Mirsky, Y., Breitenbacher, D., Shabtai, A., & Elovici, Y. (2018). N-BaIoT—Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders. IEEE Pervasive Computing, 17, 3, 12-22. https://doi.org/10.1109/MPRV.2018.03367731

Millar, S., McLaughlin, N., Rincon, J. Martinez-del, Miller, P., & Zhao, Z. (2018). DLAN: Deep Learning for Detection of Advanced Persistent Threats. 2018 International Joint Conference on Neural Networks, 1-8. https://doi.org/10.1109/IJCNN.2018.8489227

Mirsky, Y., Doitshman, T., Elovici, Y., & Shabtai, A. (2018). Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection. Network and Distributed System Security Symposium. https://doi.org/10.14722/ndss.2018.23211

Moustafa, N. & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). 2015 Military Communications and Information Systems Conference. https://doi.org/10.1109/MilCIS.2015.7348942

Moustafa, N. & Slay, J. (2016). The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 dataset and the comparison with the KDD99 dataset. Information Security Journal: A Global Perspective, 25, 1-3, 18-31. https://doi.org/10.1080/19393555.2015.1125974

Niyaz, Q., Sun, W., Javaid, A. Y., & Alam, M. (2016). A Deep Learning Approach for Network Intrusion Detection System. EAI Endorsed Transactions on Security and Safety, 3, 9, e2. https://doi.org/10.4108/eai.3-12-2015.2262516

Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. (2018). SoK: Security and Privacy in Machine Learning. 2018 IEEE European Symposium on Security and Privacy, 399-414. https://doi.org/10.1109/EuroSP.2018.00035

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. https://doi.org/10.1145/2939672.2939778

Rigaki, M. & Garcia, S. (2018). Bringing a GAN to a Knife-Fight: Adapting Malware Communication to Avoid Detection. 2018 IEEE Security and Privacy Workshops, 70-75. https://doi.org/10.1109/SPW.2018.00019

Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019). A survey of network-based intrusion detection data sets. Computers & Security, 86, 147-167. https://doi.org/10.1016/j.cose.2019.06.005

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65, 6, 386-408. https://doi.org/10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323, 533-536. https://doi.org/10.1038/323533a0

Settles, B. (2009). Active Learning Literature Survey. University of Wisconsin-Madison. https://doi.org/10.48550/arXiv.1409.2644

Shabtai, A., Elovici, Y., & Rokach, L. (2014). A survey of data leakage detection and prevention solutions. Springer Science Reviews, 2, 9-28. https://doi.org/10.1007/s40362-014-0013-8

Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2019). A Detailed Analysis of the CICIDS2017 Data Set. in Information Systems Security and Privacy, CCIS 977, 172-188. https://doi.org/10.1007/978-3-030-25109-3_9

Shokri, R. & Shmatikov, V. (2015). Privacy-Preserving Deep Learning. 22nd ACM SIGSAC Conference on Computer and Communications Security, 1310-1321. https://doi.org/10.1145/2810103.2813687

Smith, V., Chiang, C.-K., Sanjabi, M., & Talwalkar, A. (2017). Federated Multi-Task Learning. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1705.10467

Sommer, R. & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. 2010 IEEE Symposium on Security and Privacy, 305-316. https://doi.org/10.1109/SP.2010.25

Sperotto, A., Schaffrath, G., Sadre, R., Morariu, C., Pras, A., & Stiller, B. (2010). An overview of IP flow-based intrusion detection. IEEE Communications Surveys & Tutorials, 12, 3, 343-356. https://doi.org/10.1109/SURV.2010.032210.00054

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

Tang, T. A., Mhamdi, L., McLernon, D., Zaidi, S. A. R., Ghogho, M., & Moussa, F. El (2020). DeepIDS: Deep Learning Approach for Intrusion Detection in Software Defined Networking. Electronics, 9, 9, 1533. https://doi.org/10.3390/electronics9091533

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. https://doi.org/10.1109/CISDA.2009.5356528

Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., & McDaniel, P. (). Ensemble Adversarial Training: Attacks and Defenses. 2018 International Conference on Learning Representations.. https://doi.org/10.48550/arXiv.1705.07204

Tsai, C.-F., Hsu, Y.-F., Lin, C.-Y., & Lin, W.-Y. (2009). Intrusion detection by machine learning: A review. Expert Systems with Applications, 36, 10, 11994-12000. https://doi.org/10.1016/j.eswa.2009.03.012

Tuor, A., Kaplan, S., Hutchinson, B., Nichols, N., & Robinson, S. (2017). Deep learning for unsupervised insider threat detection in structured cybersecurity data streams. Workshops at the Thirty-First AAAI Conference on Artificial Intelligence. https://doi.org/10.48550/arXiv.1710.00811

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1706.03762

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph Attention Networks. . https://doi.org/10.48550/arXiv.1710.10903

Vinayakumar, R., Soman, K. P., Poornachandran, P., & Ganorkar, S. (2018). Evaluating Deep Learning Approaches to Characterize and Classify the DGAs at Scale. Journal of Intelligent & Fuzzy Systems, 34, 3, 1265-1276. https://doi.org/10.3233/JIFS-169424

Vinayakumar, R., Soman, K. P., Poornachandran, P., Alazab, V., & Jolfaei, A. (2019). A machine learning based cyber security analytics for intrusion detection systems. 2019 International Conference on Advances in Computing, Communications and Informatics, 1-6. https://doi.org/10.1109/ICACCI.2019.8823273

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. 25th International Conference on Machine Learning, 1096-1103. https://doi.org/10.1145/1390156.1390294

Wang, W., Zhu, M., Wang, J., Zeng, X., & Yang, Z. (2017). End-to-end encrypted traffic classification with one-dimensional convolution neural networks. 2017 IEEE International Conference on Intelligence and Security Informatics. https://doi.org/10.1109/ISI.2017.8004872

Wang, Q., Guo, W., Zhang, K., et al. (2017). A generic attack to ciphers. . https://doi.org/10.48550/arXiv.1709.08032

Wu, S. X. & Banzhaf, W. (2010). The use of computational intelligence in intrusion detection systems: A review. Applied Soft Computing, 10, 1, 1-35. https://doi.org/10.1016/j.asoc.2009.06.019

Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., Gao, M., Hou, H., & Wang, C. (2018). Machine Learning and Deep Learning Methods for Cybersecurity. IEEE Access, 6, 35365-35381. https://doi.org/10.1109/ACCESS.2018.2836950

Yan, F., Wen, S., Nepal, S., Paris, C., & Xiang, Y. (2022). Explainable machine learning in cybersecurity: A survey. International Journal of Intelligent Systems, 37, 12, 12305-12334. https://doi.org/10.1002/int.23088

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated Machine Learning: Concept and Applications. ACM Transactions on Intelligent Systems and Technology, 10, 2, 12. https://doi.org/10.1145/3298981

Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks. IEEE Access, 5, 21954-21961. https://doi.org/10.1109/ACCESS.2017.2762418

Yousefi-Azar, M., Varadharajan, V., Hamey, L., & Tupakula, U. (2017). Autoencoder-based feature learning for cyber security applications. 2017 International Joint Conference on Neural Networks, 3854-3861. https://doi.org/10.1109/IJCNN.2017.7966342