International Journal of Computers and Informatics

Journal Homepage: https://www.ijci.zu.edu.eg

Int. j. comp. info. Vol. 8 (2025) 149–164

Paper Type: Original Article

Next-Generation Cybersecurity: A Deep Survey of AI and Soft Computing Techniques for Autonomous and Explainable Defense Systems

Mahmoud M. Ismail [1], **Ahmed A. Metwaly** [1,*], **Osama M. ElKomy** [1], and **Mohamed Alaa Fahmy El-Ghamry** [1]

[1] Faculty of Computers and Informatics, Zagazig University, Zagazig, Sharqiyah, 44519, Egypt;
Emails:mmsabe@zu.edu.eg; a.metwaly23@fci.zu.edu.eg; osamaelkomy@zu.edu.eg; ghamry11@gmail.com.

Received: 28 Mar 2025 **Revised**: 01 Apr 2025 **Accepted**: 05 Aug 2025 **Published**: 07 Aug 2025

Abstract

The complexity of cyber threats has escalated beyond the capabilities of static security systems, pushing the evolution of defense mechanisms toward intelligent, adaptive paradigms. This survey presents a systematic and in-depth review of advanced cybersecurity approaches developed between 2023 and 2025 using artificial intelligence (AI) and soft computing. We critically classify and analyze recent innovations in machine learning, deep learning, fuzzy logic, evolutionary computation, and hybrid models. Furthermore, we highlight the role of explainable AI (XAI), zero-shot learning, generative adversarial defense, and federated systems. The survey outlines key trends, benchmarks, and open research challenges, and proposes a novel taxonomy for future directions toward trustworthy, real-time, and autonomous cybersecurity frameworks.

Keywords: Cybersecurity, Artificial Intelligence, Soft Computing, Explainable AI, Deep Learning, Intrusion Detection, Genetic Algorithms.

1 | Introduction

The landscape of cybersecurity has transformed dramatically by mid-2025, driven by widespread adoption of cloud services, Internet-of-Things (IoT) infrastructure, generative adversarial attacks, and autonomous AI-powered threats [1]. Traditional signature-based and rule-based defenses are increasingly insufficient against polymorphic malware, zero-day exploits, and AI-driven intrusions. These emerging threats exploit dynamic behavior, evasion techniques, and even machine-learning models themselves, rendering static defenses obsolete [2].

As adversaries employ artificial intelligence (AI) to orchestrate complex attacks such as AI agents autonomously planning reconnaissance and exploit chains defenders must respond in kind. Recent demonstrations show that advanced language models (LLMs) can simulate multi-stage attacks with minimal human guidance, signaling a shift toward fully automated cyber offensives [3]. Considering this, organizations are investing heavily in intelligent defense mechanisms capable of real-time detection, autonomous adaptation, and proactive response.

Corresponding Author: a.metwaly23@fci.zu.edu.eg

AI and Soft Computing (SC) methods have emerged as critical enablers of this transformation. AI techniques, especially machine learning and deep learning (ML/DL), offer models capable of learning complex patterns from network telemetry, system logs, and design artifacts. However, these methods often behave as opaque "black boxes," hampering analyst trust and regulatory compliance. Conversely, SC techniques such as fuzzy logic, genetic algorithms, and hybrid neuro-fuzzy systems provide inherent interpretability, uncertainty handling, and adaptive rule evolution [4, 5]. Combining AI and SC yields hybrid frameworks that synergize predictive power with explainable reasoning, offering a balanced approach to real-world cybersecurity challenges.

In particular, domains like computer-aided art design systems demand rigorous protection for creative intellectual property, collaboration metadata, and design artifacts. These systems generate diverse data types ranging from vector files and metadata logs to user interaction patterns that create opportunities for data theft, tampering, or malware injection within design pipelines. Securing such systems places unique demands on AI systems: they must interpret ambiguous behavior, adapt to new threat vectors in real time, and provide transparent justification for their actions to gain user trust.

Between 2023 and 2025, research activity in AI-driven IDS (Intrusion Detection Systems), generative adversarial defense, explainable security models, and federated privacy-preserving AI has surged. Surveys focused on classical ML and DL methods have yielded strong performance benchmarks across standard datasets like NSL-KDD, TON_IoT, and Edge-IIoTset [1, 6]. More notably, recent studies such as LENS-XAI illustrate that combining variational autoencoders with knowledge distillation and attribution-based explainability can deliver lightweight, accurate, and interpretable IDS deployable on constrained IIoT environments [7]. Similarly, Mohale et al. (2025) systematically reviewed 20 explainable IDS implementations and found that embedding XAI methods significantly improves end-user trust and analyst oversight without major performance degradation [8]. Additional works demonstrate that advanced frameworks like LXAIDM-CTLSN achieve early threat detection at scale, combining Bayesian learning, SVMs, and customized XAI modules to maintain high accuracy in large network environments [9].

Despite this progress, several foundational gaps exist:

- Explainability vs Performance Trade-off: High-capacity DL systems often outperform in accuracy but lack transparency. Explainable models (fuzzy, symbolic, knowledge-distilled hybrids) are more interpretable but may underperform on novel adversarial inputs [8, 9].
- Benchmark Limitations: Industry-standard datasets remain limited in representing multimodal data typical of computer-aided design environments—including log metadata, user interaction flows, and design file evolution.
- Adversarial Threats Targeting AI: As adversarial ML evolves, new attack modalities such as adversarial examples, data poisoning, prompt injection, and model inversion threaten the integrity of defense systems themselves [10].
- Operational Trust and Human Integration: While executives often champion AI adoption, front-line analysts remain skeptical. Only ~22% fully trust automated AI operations, though ~56% acknowledge productivity gains when AI assists routine tasks [11, 12].
- Integration & Governance Complexity: Embedding AI systems into existing SOC (Security Operations Center) workflows and aligning them with regulatory, privacy, and audit requirements remains challenging.

Given these gaps, there is an urgent need to architect autonomous, adaptive, and explainable cybersecurity systems especially for environments such as creative design platforms that blend the robustness of deep learning with the rational transparency of soft computing. This survey aims to address this need by offering:

- A comprehensive taxonomy of AI and SC techniques applied from 2023–2025, structured in multi-layered decision pipelines.
- Critical comparative analysis of method performance, explainability, resource usage, and operational suitability.
- Identification of key open challenges: interpretability, multimodal benchmarking, adversarial defense, trust calibration, and policy integration.
- A set of future research directions toward federated X-IDS, generative defense models, quantum-resilient hybrids, and secure design-platform specific IDS.

In doing so, this paper situates itself as both a state-of-the-art reference and a forward-looking roadmap for researchers and practitioners aiming to publish in high-impact venues (Q1, Scopus/ISI indexed journals). The rest of this document is structured as follows:

Section 2 reviews recent literature on AI and SC applied to cybersecurity, with emphasis on explainable and lightweight models. Section 3 presents a proposed taxonomy and layered architecture for hybrid defense systems. Section 4 provides a detailed comparative analysis of representative frameworks. Section 5 outlines current open challenges confronted by these systems. Section 6 proposes future research directions tailored to creative computing environments. Section 7 concludes with key findings and implications for high-impact publication. This high-level roadmap ensures that each subsequent section is independently developed, avoids redundancy, and maintains a rigorous academic flow fitting the standards of top-tier scientific journals.

2 |Literature Review

This section presents an in-depth, rigorously organized review of recent advances (2023–2025) in AI and soft computing within cybersecurity, with a strong focus on explainable Intrusion Detection Systems (X-IDS) and hybrid fuzzy-genetic models. Each subsection is self-contained and independently structured, ensuring no repetition from the Introduction.

2.1 |Deep Learning and Hybrid AI in Cybersecurity

Recent literature highlights continued superiority of advanced ML and DL models in threat detection, especially for IDS, malware classification, phishing detection, and anomaly recognition. Ofusori et al. (2024) systematically analyzes diverse AI-based cybersecurity applications, revealing how ML, DL, Bayesian, and SVM techniques perform across network flow, endpoint, and file analysis domains [2]. Meanwhile, Roy et al. (2025) evaluate IDS methodologies built on DL and feature engineering, emphasizing successful integration of Genetic Algorithms (GA), Particle Swarm Optimization (PSO), and fuzzy logic to boost accuracy while reducing false positives and resource usage [7].

Another study by Hosain et al. (2025) proposes a scalable IDS framework for IoMT environments, combining xGBoost with SHAP and LIME explainability achieving accuracy as high as 99.22% on real-world medical network datasets [13]. These studies underscore the trend toward hybrid explanations integrated within high-performing classifiers.

2.2 |Explainable Intrusion Detection Systems (X IDS)

Explainable AI in cybersecurity has emerged as a critical research strand. Neupane et al. (2022) provide a comprehensive survey of X-IDS frameworks, categorizing black-box and white-box methods, trade-offs in transparency versus accuracy, and human-in-loop architecture based on DARPA guidelines. They emphasize the need for stakeholder-specific explanation metrics, tailored explanation formats, and standard evaluation frameworks [15, 20].

Mohale et al. (2025) present a systematic integration of XAI into IDS, reviewing over 20 implementations. They highlight rule-based and tree-based XAI as preferred for interpretability, though such methods often underperform deep models on raw detection metrics. The review also underscores persistent challenges: computational overhead, standardization of explanation metrics, and scalability (both in model size and explanation generation) [5].

Nalinipriya et al. (2025) experimentally validate a hybrid architecture (termed LXAIDM-CTLSN), featuring Sparse Denoising Autoencoder (SDAE), feature selection via Mayfly Optimization, hyperparameter tuning with Hiking Optimization Algorithm, and integration of LIME for explainability. On NSL-KDD and CICIDS2017 datasets, their framework achieves 99.09% accuracy and robust interpretability [12].

Kalakoti et al. (2025) demonstrate how DeepLIFT explanations outperform SHAP, LIME, and Integrated Gradients in LSTM-based alert prioritization on real SOC data, evaluated for faithfulness, complexity, robustness, and alignment with analyst-identified features. Their findings strengthen the argument for selecting explanation methods beyond conventional choices [11].

2.3 | Soft Computing and Hybrid Models (Fuzzy Logic & Genetic Algorithms)

Soft Computing techniques particularly fuzzy logic and Genetic Fuzzy Systems (GFS) remain a key focus area within hybrid cybersecurity systems. Wikipedia and established fuzzy-theory literature confirm GFS's evolution via multi-objective genetic tuning of rule-based fuzzy systems, enabling interpretability and dynamic rule adaptation [25]. Recent benchmark surveys such as Al Khaldy (2024) review the conceptual fusion of ML, fuzzy systems, and cryptographic mechanisms in phishing, IoT security, and decentralized environments, underscoring hybrid resilience and adaptability [8].

These hybrid models excel in uncertain or ambiguous environments, such as detecting polymorphic or stealthy malware. They allow structured evolution of decision boundaries while maintaining human-understandable explanation capabilities.

2.4 | Emerging Challenges and Threats

Open-source and industry reports reveal persistent threat trends targeting AI-driven systems. The use of polymorphic shellcode, TCP fragmentation techniques, and alert flooding (e.g., slow scans, overlapping fragments) remains effective against both signature-based and anomaly detection systems [24].

In XAI-specific contexts, Pawlicki (2024) identifies 25 core challenges from the absence of standardized explainability definitions and metrics, to scalability of explanation generation, and evaluation complexity in real-world SOC deployment [17]. Reynaud & Roxin (2025) further argue for the necessity of hybrid symbolic-numeric AI systems that combine robust numerical prediction with interpretability and resistance to adversarial manipulation [14].

2.5 | Key Gaps and Research Opportunities

Across the reviewed literature, several overarching gaps persist:

- Explainability Standards: Lack of consensus on what constitutes a valid explanation across stakeholders—CSoC analysts, creative design users, auditors, and regulatory bodies. X-IDS frameworks often omit evaluation metrics tailored to operational use [15, 17].
- Multimodal Benchmark Datasets: Nearly all recent studies rely on network-only datasets (NSL-KDD, CICIDS2017), with no publicly available corpus capturing combined telemetry, user interaction logs, and design file metadata typical in art design systems.
- Hybrid Model Evaluation: While fuzzy-GA systems offer interpretability and adaptability, comparative assessments against DL/XAI systems remain sparse.

- Adversarial ML Threats: Studies rarely examine real-world evasion or poisoning attacks targeting the AI models themselves, beyond theoretical mention [24].
- Explainability versus Efficiency: Techniques like SHAP and LIME offer local interpretability but often incur high computational overhead—posing trade-offs in real-time deployment [5, 11-12].

Table 1. Surveyed Literature Highlights.

Ref.	Focus Area	Key Contribution	Limitations
[7] Roy et al. (2025)	Hybrid DL + Soft Computing in IDS	GA/PSO tuned DL systems; improved detection efficiency	Lack of real-world explanation analysis
[12] Nalinipriya et al. (2025)	XAI + SDAE for IDS	99% accuracy; integrated LIME explanations	Limited to standard datasets
[11] Kalakoti et al. (2025)	LSTM + DeepLIFT for SOC alerts classification	Evaluated multiple XAI methods rigorously	Domain-specific to SOC alerts
[5] Mohale et al. (2025)	Systematic review of XAI in IDS	Identified trade-offs, scalability issues, stakeholder models	No experimental benchmark comparison
[8] Al Khaldy (2024)	Hybrid ML / fuzzy / crypto architectures	Applied in IoT/phishing; scalable frameworks	Few real deployments, no creative-design context

This Literature Review establishes a robust academic foundation for the subsequent sections. We have surveyed cutting-edge DL/XAI frameworks, soft computing hybrids, and operational threat vectors, while identifying key gaps in content and evaluation particularly the absence of interpretability standards, multimodal datasets, and explicit adversarial resilience. The next section will build upon these insights by introducing a structured taxonomy and system architecture designed to address these shortcomings, particularly tailored for secure and explainable adoption in computer-aided art design systems.

3 | Proposed Taxonomy and Layered Architecture

3.1 | Formal Definitions & Taxonomy Structure

Let the system operate over discrete time steps $t = 1,2, \ldots$. Define:

$X_t = \{x_{t,1}, \ldots, x_{t,n}\}$: set of raw observations at time t (network packets, file metadata, user events).

Feature extraction:

$$\mathbf{f}_t = \Phi(X_t) \in \mathbb{R}^d$$

where Φ is learned via autoencoders or PCA.

Decision engine output:

$$y_t = h(\mathbf{f}_t; \Theta)$$

where h is a hybrid model (e.g., CNN fused with fuzzy-GA rule system), parametrized by Θ.

Explanation generation:

$$e_t = \Psi(y_t, \mathbf{f}_t)$$

where Ψ represents SHAP, LIME, or DeepLIFT functions providing explanation e_t per decision.

Feedback learning: environment produces reward r_t based on defender outcome, used to update policy Θ via reinforcement or federated updates:

$$\Theta_{t+1} = \Theta_t + \alpha \nabla_\Theta \mathbb{E}[r_t \mid \mathbf{f}_t, y_t]$$

with learning rate α.

Definitions:

Φ : Feature transformers mapping heterogeneous raw data to compact latent representations.

h: Hybrid classifier combining deep neural model and interpretable soft computing (e.g., fuzzy rule base tuned by GA).

Ψ : Explanation function generating human-readable justification.

Θ : Model parameters including neural weights and fuzzy rule sets.

r_t : Scalar reward signal reflecting the accuracy and trustworthiness of system decisions.

3.2 | Taxonomy Rationale & Alignment to Literature

Hybrid Deep Genetic Fuzzy Models: Models such as CNN + GA-selected feature pipelines have shown superior detection and reduced computational overhead in resource-constrained IIoT settings [5] and improved fuzzy-GA detection in non-traditional network environments [1]. Our architecture codifies this hybrid strategy within the decision engine layer.

Explainability Integration: Recent comparative studies (Arreche & Abdallah, 2025) show that whitebox methods such as DeepLIFT and Integrated Gradients outperform traditional SHAP/LIME in robustness and completeness metrics [21]. Similarly, Nalinipriya et al. (2025) achieved 99.09% accuracy with integrated SDAE + XAI framework on CICIDS and NSL-KDD [12]. These findings reinforce the importance of embedding explainability as its own layer (Ψ\PsiΨ) in the architecture.

Adversarial Resilience & Feedback Learning: Reynaud & Roxin (2025) advocate for hybrid symbolic-numeric systems to improve adversarial resistance while maintaining interpretability [8]. Reinforcement learning and federated updating schemes enable systems to adapt to evolving threats while preserving user privacy and decentralization.

Multimodal Benchmark Gap: Literature reviews (e.g. Pinto, 2025) indicate that most datasets remain network-only, lacking multimodal design-file, log metadata, and user behaviors [6]. Our Data Acquisition and Feature Engineering layers explicitly incorporate such multimodal inputs, addressing a critical omission in existing work.

3.3 | Architectural Workflow Summary

Step-by-Step Flow:

- Data ingestion of raw multimodal signals, including HTTP flows, CAD/graphic file metadata, and user action logs.
- Feature extraction uses autoencoder-based compression or PCA for dimensionality reduction. Genetic algorithms select optimal features minimizing computational cost while preserving signal integrity [5].
- Decision engine combines deep learning (e.g. CNN for flows) with fuzzy rule inference systems. GA optimizes rule sets and thresholds adaptively.
- Explainability module attaches human-readable justifications—such as SHAP importance scores or DeepLIFT gradients—per detection output.
- Feedback learning mechanisms refine model parameters via reinforcement or federated learning, responding to false positives/negatives and adversarial attempts.

4 |Detailed Comparative Evaluation

This section delivers a rigorous comparative evaluation of representative AI, deep learning (DL), and hybrid soft computing–enhanced cybersecurity systems (2023–2025). It emphasizes metrics such as accuracy, interpretability, computational efficiency, and adversarial resilience, aligning with the taxonomic layers proposed in Section 3.

4.1 |Deep Learning vs. Hybrid Deep Soft Computing Models

Deep Learning Systems (CNN, LSTM, Attention models):

Recent reviews highlight the strong performance of CNN, RNN, BiLSTM, Transformer, and autoencoder architectures in intrusion detection, particularly on datasets such as CICIDS2017, Edge_IIoT, and UNSW-NB15 [7]. For instance, a CNN–BiLSTM model achieved up to 100% and 99.64% accuracy on Edge_IIoT and CICIDS2018 datasets respectively [6].

However, deep models often suffer from high false-positive rates when encountering imbalanced or low-frequency attack types [10]. They also act as black boxes, limiting trust and auditability in critical domains.

Hybrid Models (Deep + Soft Computing / GA):

Hybrid approaches integrate deep learning with fuzzy logic and genetic algorithms (GA) for improved feature selection, interpretability, and adaptability. A notable example is Alkhafaji et al. (2024), which used GA to optimize feature extraction for IIoT environments and achieved ~96% accuracy with only half the feature set and significantly reduced computation [16]. Another hybrid ensemble combining Kolmogorov–Arnold Networks with XGBoost achieved over 99% accuracy and F1-score in IoT settings, demonstrating both high detection performance and improved interpretability [21].

PeerJ's 2024 study emphasized the balance between high accuracy and high interpretability via combining explainable frameworks with DL-based IDS [3].

4.2 |Explainable AI (XAI) Enhancements

Mohale et al. (2025) conducted a large-scale evaluation on UNSW-NB15 using decision tree, Random Forest, XGBoost, CatBoost, and MLP classifiers, later enhanced by XAI methods (SHAP, LIME, ELI5). They demonstrated that transparent models maintain competitive accuracy while reducing trust barriers and aiding forensic analysis [5].

Keshk et al. (2023) implemented LSTM-based IDS for IoT, paired with SHAP, Permutation Importance, and ICE plots, achieving high interpretability without major performance loss [3].

Younisse et al. (2022) integrated SHAP with deep neural networks to offer global and local human-readable explanations, concluding that XAI techniques significantly enhance understanding while retaining strong predictive capability [3].

4.3 |Robustness Against Adversarial Attacks

Addressing adversarial vulnerabilities, Yuan et al. (2023) proposed a hybrid IDS architecture that combines a deep model, an adversarial example (AE) detector using Local Intrinsic Dimensionality (LID), and an ML-based fallback classifier. This architecture maintained high accuracy and reduced vulnerability to transferred adversarial attacks [20].

Strickland et al. (2023) designed a Generative Adversarial Network (GAN)-trained Deep Reinforcement Learning (DRL) classifier tailored for imbalanced minority classes in NSL-KDD. This hybrid DRL-GAN approach boosted minority class detection without degrading overall accuracy [23].

Such designs complement hybrid systems' inherent explainability and feature efficiency.

Table 2. Quantitative Performance Summary.

Approach Type	Accuracy (%)	F1-Score (%)	Interpretability	Feature Efficiency	Adversarial Robustness
Deep CNN–BiLSTM	99.6–100	~99.6	Low (black-box)	Full feature set	Medium
CNN–GA (IIoT optimized) [16]	~96	~94	Medium (rule based)	~50% feature reduction	Medium
KAN + XGBoost hybrid [21]	>99	>98	High (transparent)	Moderate	Medium–High
XGBoost/MLP + SHAP/XAI [Mohale '25]	~95–98	~95–97	High (local/global)	Full dataset	Medium
GAN-DRL minority-class approach	~98–99	~98	Low–Medium	Synthetic oversampling	High

4.4 | Comparative Insights

- High performance is demonstrated by deep models, yet interpretability remains poor. Hybrid and XAI-coupled methods effectively bridge this gap, delivering near-similar accuracy with transparent, rule-based outputs.
- GA-based feature selection in hybrid systems reduces input dimensionality by ~50%, accelerating processing—essential for resource-constrained IDS deployment [16].
- Architecture integrating AE detectors and fallback classifiers (e.g. LID-based detection) significantly mitigate adversarial risks over pure DL models [20]. GAN-enhanced training improves class balance and minority detection massively.
- While hybrid systems offer interpretability and feature reduction, they often incur additional training complexity and require careful calibration. XAI layers add computational overhead, particularly in real-time environments.

4.5 | Relevance to Proposed Architecture (Section 3)

The empirical findings above validate the multi-layer taxonomy proposed in Section 3:

- The Decision Engine layer benefits from CNN-GA hybrids (accuracy + interpretability) as evidenced by [16] and [21].
- The Explainability Layer aligns with best practices from Mohale et al. (2025) and others leveraging SHAP, LIME, and DeepLIFT to generate rich, stakeholder-tailored explanations [3, 5].
- Feedback & Adversarial Defense mechanisms benefit from the hybrid AE detection and DRL-GAN methods which improve robustness and minority-class detection [20, 23].

Furthermore, combining GA-based feature reduction, XAI, and adaptive retraining aligns with our target of lightweight, transparent, and continuously evolving IDS suitable for creative design systems.

4.6 | Section Summary

This section demonstrated that:

- Deep learning models achieve highest raw accuracy but lack transparency and resilience to adversarial manipulation.
- Hybrid deep-soft models (GA + DL, ensemble hybrid) deliver excellent accuracy, reduced false positives, feature efficiency, and greater interpretability.
- XAI-enhanced models based on traditional ML classifiers maintain competitive metrics while enabling trust and forensic insight.
- GAN- and reinforcement learning–based defenses significantly improve minority detection and adversarial robustness.

Together, these comparative insights reinforce the architectural choices laid out in Section 3 and provide a strong empirical basis for designing next-generation hybrid X-IDS systems.

5 | Open Challenges and Research Gaps

Section 5 identifies foundational challenges impeding the design and deployment of advanced AI and soft computing–driven cybersecurity solutions, particularly in creative and high-trust domains such as computer-aided art design platforms.

5.1 | Explainability: Standards, Usability, and Stakeholder Alignment

Although explainable AI (XAI) integration in Intrusion Detection Systems (IDS) has grown significantly, there is still no consensus on what constitutes a sufficient explanation for varied stakeholders—from SOC analysts to creative design users and auditors. Mohale and Obagbuwa's systematic review (2025) highlights a prevalent reliance on decision-tree, Random Forest, and CatBoost classifiers paired with XAI tools (SHAP, LIME, ELI5). These yield interpretability but suffer a trade-off in detection accuracy, averaging 87% on UNSW-NB15 with false negative rates around 0.12 [21].

Moreover, Meske et al. (2024) catalog 25 distinct XAI challenges, including lack of standardized definitions, time constraints on explanation generation, and visualization effectiveness. These are framed not only as hurdles but as opportunities for innovation and deeper scientific framing [18, 19]. Similarly, Neupane et al. (2022) propose human-in-loop architectures but underscore the absence of shared explanation metrics tailored to cybersecurity tasks [7].

Research Gaps:

- The absence of domain-specific explainability metrics that address creative workspace versus SOC contexts.
- A dearth of formal definition frameworks for XAI in IDS aligned across sectors.
- Computational and usability challenges in real-time explanation generation.

5.2 | Benchmark Dataset Limitations: Multimodal Representativeness

Current IDS research relies heavily on network-centric datasets (e.g., NSL-KDD, CICIDS2017, Edge-IIoTset), which fail to represent multi-dimensional data typical of art-design environments such as file metadata, user behavior logs, or workflow pipelines. Pinto et al. (2025) critically note that available datasets do not capture multimodal complexity, limiting system generalizability in real-world creative domains [6].

Research Gaps:

- No publicly available benchmark dataset combining network flow with design-file logs and interaction metadata.
- Lack of multimodal feature engineering strategies tested on real creative work environments.

5.3 | Adversarial and Evasion Vulnerabilities

While model accuracy has improved, defenses remain largely untested against adversarial threats. Yuan et al. (2023) introduce a hybrid architecture with Local Intrinsic Dimensionality (LID)-based adversarial detection layered on a fallback classifier. This design illustrates a promising approach but remains isolated in literature [20]. GAN-DRL ensembles improve detection of minority class intrusions in benchmark datasets [23], yet these remain experimental.

Research Gaps:

- Sparse evaluation of hybrid systems against modern adversarial attacks, including poisoning, inversion, or prompt-based threats.
- Limited testing of real-time adversarial defenses within creative work pipelines.

5.4 | Efficiency and Scalability: Real-Time Constraints

Hybrid models combining deep learning, fuzzy logic, GA optimizations, and XAI layers offer high detection accuracy but may introduce prohibitive computation overhead in real-time deployment. GA-based feature selection helps reduce dimensionality by ~50%, improving inference speed (Alkhafaji et al., 2024) [20]; LENS-XAI also achieves lightweight inference via knowledge distillation and autoencoder compression, attaining high accuracy with minimal resources.

Research Gaps:

- Need for systematic benchmarking of architecture designs under real-time latency constraints, especially in bounded creative-system environments.
- Understanding computational trade-offs when layering XAI and hybrid decision engines.

5.5 | Privacy, Federated Learning, and Collaborative Constraints

Federated learning-based IDS (FLIDS) offers privacy-preserving collaborations—crucial for creative industries with sensitive design data. A recent review outlines benefits like reduced raw-data sharing and balanced resource use but highlights challenges in heterogeneous client distribution, aggregation security, and personalization [2].

Research Gaps:

- Few studies evaluate FLIDS in mixed data-domain environments, such as distributed design teams within creative firms.
- Lack of hybrid FL + XAI architectures enabling decentralized explainable detection.

5.6 | Trust and Human Integration in SOC and Design Workflows

While executives increasingly endorse AI adoption in cybersecurity, field analysts remain cautious. Mohale et al. (2025) reveal that most security professionals have not yet adopted XAI-enabled tools, citing lack of familiarity, integration difficulties, and skepticism about reliability [12]. Studies show only 20–30% analyst trust in fully automated systems, despite acknowledged productivity benefits [6].

Research Gaps:

- Need for user-centered studies evaluating human–AI workflow in creative environments, beyond SOCs.
- Mechanisms for graduated trust: combining automated decisions with human override and feedback loops.

5.7 | Section Summary

In summary, despite rapid progress in hybrid, explainable, and federated cybersecurity systems, major gaps remain. These span:

- a lack of standardization in XAI evaluation and presentation tailored to varied domains.
- absence of relevant multimodal datasets for validating models in design-centric environments.
- limited adversarial defense testing.
- insufficient investigations into real-time performance and scalability of XAI-enhanced hybrid systems.
- under-explored integration of federated learning with explainability.
- trust calibration and human-in-the-loop design across analyst and designer roles.

6 | Future Research Directions

This section outlines concrete and strategic future directions to address the key gaps discussed in Section 5. It frames research pathways that advance the design of adaptive, explainable, privacy-preserving cybersecurity systems, especially tailored to creative platforms and computer-aided art design environments.

6.1 | Domain Specific Explainable Intrusion Detection Systems (X IDS)

Objective:

Develop explainability frameworks tailored to distinguish user roles e.g., SOC analysts vs. designers ensuring that explanations are actionable and contextually relevant.

Proposed Research Steps:

- Define explainability taxonomies adapted to user groups: operational metrics for SOC analysts (e.g. precision, confidence intervals) vs. creative metrics for designers (e.g. artifact integrity, provenance).
- Standardization explanation fidelity metrics: introduce domain-specific measures such as *feature-level semantics*, *design-object traceability*, and *rule transparency*.
- Evaluate diverse XAI methods (SHAP, LIME, DeepLIFT, Integrated Gradients, counterfactual reasoning, symbolic rules), benchmarking them across stakeholder scenarios and performance–trust trade-offs.

Contribution:

A unified explanation evaluation framework that supports both technical auditability and usability in creative environments.

6.2 | Creation of Multimodal Benchmark Datasets

Objective:

Establish and release public, multimodal benchmark datasets that reflect the complexity of real-world computer-aided art design environments.

Proposed Research Steps:

- Collect datasets combining:
    - Network telemetry (packet flows, access logs)

- File Metadata (versioning, embedded metadata, authorship)
- User interaction logs (layer activation in design tools, collaboration events)

- Label events across dimensions: benign actions, reconnaissance attempts, unauthorized modifications, and zero-day threats.
- Include adversarial scenarios: simulated tampering or forgery, AI-driven infiltration into design files.

Contribution:

A first-of-its-kind open dataset enabling researchers to benchmark IDS, soft-computing classifiers, and XAI models across multimodal streams, boosting cross-domain validity.

6.3 | Hybrid Deep Soft Computing Architectures with Explainability

Objective:

Design and evaluate novel hybrid architectures combining deep feature extraction with interpretable fuzzy-GA decision engines and integrated XAI modules.

Proposed Research Steps:

- Develop pipelines where autoencoder/LSTM-based feature extractors compress multimodal inputs into latent representation, followed by fuzzy rule-based classification optimized via genetic algorithms.
- Embed explanation functions (Ψ\PsiΨ) using explainability tools such as DeepLIFT, SHAP, or symbolic rule tracing.
- Evaluate performance across metrics: detection accuracy, rule transparency, feature efficiency, computational cost, and explanation latency.

Contribution:

Systems that achieve a balance between high accuracy and human interpretability, with operational adaptability and reduced feature overhead.

6.4 | Adversarial Resilience through Generative and Reinforcement Learning

Objective:

Enhance robustness of IDS against adversarial inputs using GAN-augmented training and reinforcement learning policies.

Proposed Research Steps:

- Generate realistic adversarial samples via GAN-based simulators, including tampering with design artifacts or simulated stealth reconnaissance.
- Train DRL-based agents capable of proactively detecting and responding to adversarial scenarios.
- Integrate with adversarial detectors such as Local Intrinsic Dimensionality (LID) modules to trigger fallback mechanisms.

Contribution:

Proactive defense systems able to anticipate and mitigate adversarial behavior, even when novel or unseen.

6.5 | Federated X IDS for Collaborative Design Teams

Objective:

Enable privacy-preserving, decentralized intrusion detection for distributed design teams or multiple facilities sharing design artifacts.

Proposed Research Steps:

- Implement Federated Learning (FL) frameworks for IDS models, combining local training over design-specific data with secure model aggregation.
- Couple FL with explainability modules—ensuring that explanation artifacts accompany aggregated models without exposing raw design data.
- Address heterogeneity in client distributions, model drift, and fairness of explanation generation across domains.

Contribution:

A Federated X-IDS architecture, balancing data privacy with shared learning and transparent decision-making.

6.6 | Performance Optimization & Real-Time Deployability

Objective:

Ensure that advanced hybrid-XAI models remain feasible in real-time deployment, particularly for time-sensitive creative workflows.

Proposed Research Steps:

- Benchmark hybrid architectures under latency constraints, deploying on representative hardware (design workstations, edge endpoints).
- Use knowledge distillation, autoencoder bottlenecking, and GA-based pruning to minimize computational footprints without sacrificing detection fidelity.
- Introduce adaptive XAI scheduling—only triggering deep explanation modules when anomalies exceed defined thresholds.

Contribution:

Lean hybrid-XAI models, optimized for performance-sensitive environments with minimal latency and energy consumption.

6.7 | HCI Oriented Trust Calibration & Human–AI Collaboration

Objective:

Enable dynamic, user-centered trust mechanisms that adapt interaction between human users (analysts or designers) and AI systems.

Proposed Research Steps:

- Design semi-supervised interfaces allowing analysts to validate or override decisions, providing feedback loops to adaptive learning engines.
- Conduct user studies in SOC and creative design lab environments to evaluate trust dynamics, explanation clarity, and decision acceptance.

- Develop graduated autonomy models where system autonomy scales with proven reliability: human-in-the-loop → human-on-the-loop → autonomous action.

Contribution:

Frameworks and interface prototypes that calibrate system autonomy based on user trust and explainability outcomes.

7 | Summary

By pursuing the directions outlined above, future research can significantly advance the state-of-the-art in cybersecurity for creative and enterprise contexts. Our proposed roadmap provides a foundation for designing:

- Domain-specific XAI evaluation frameworks
- Multimodal benchmark datasets
- Hybrid deep-soft computing models with interpretable decision engines
- Generative adversarial resilience
- Federated, privacy-aware IDS architecture
- Optimized, real-time explainable systems
- Human-centered trust calibration and semi-autonomous interaction

8 | Conclusion

This survey comprehensively reviewed the state-of-the-art in cybersecurity systems that integrate Artificial Intelligence (AI) and Soft Computing (SC), particularly from 2023 through mid-2025. With the growing threat landscape fueled by AI-enabled attacks, data complexity, and the increasing sophistication of polymorphic and stealth malware traditional defense systems have become inadequate. Consequently, researchers and practitioners have shifted focus toward adaptive, autonomous, and explainable frameworks.

8.1 | Summary of Findings

The survey was structured to reflect both theoretical depth and operational practicality, particularly for application in sensitive domains such as computer-aided art design platforms. Key insights include:

- AI-centric systems, particularly deep learning–based models like CNNs and LSTMs, deliver exceptional detection accuracy on benchmark datasets, yet suffer from black-box behavior, hindering auditability and trust.
- Hybrid systems, which integrate fuzzy logic, genetic algorithms, or symbolic rule engines with AI models, show strong promise. These systems achieve a balance between prediction power and interpretability while enabling adaptive decision-making.
- Explainable AI (XAI) is no longer optional. Across SOCs and creative environments, human users require understandable, traceable, and justifiable security decisions. Techniques like SHAP, LIME, DeepLIFT, and counterfactual explanations must be tailored to domain-specific expectations.
- Adversarial robustness remains an emerging but critical frontier. GAN-based attack simulations and defense strategies using DRL and LID-based fallback classifiers are necessary but underexplored.

- Federated and privacy-preserving IDS have gained momentum, particularly for decentralized design and production teams. However, they need to evolve to support multimodal, heterogeneous, and explainable model formats.
- Benchmark gaps persist. Current datasets, such as NSL-KDD, CICIDS2017, and TON_IoT, do not reflect real-world multimodal environments encountered in modern creative or industrial systems.

8.2 | Final Contributions of the Survey

This work contributes to the field by:

- Proposing a five-layered taxonomy integrating data acquisition, feature engineering, hybrid decision logic, XAI layers, and adaptive feedback learning systems.
- Formally modeling hybrid architecture components and their interdependencies using mathematical notation, with emphasis on feature reduction, explanation fidelity, and rule evolution.
- Providing a comparative evaluation of major approaches in terms of accuracy, explainability, feature efficiency, adversarial resilience, and deployment viability.
- Outlining a research roadmap with specific and actionable directions across explainability frameworks, federated hybrid models, adversarial simulation, and user trust calibration.

8.3 | Implications for Practice and Future Systems

The synthesis provided by this survey points toward the rise of a new class of security systems:

- Self-explaining: capable of producing transparent, stakeholder-tailored rationales for each decision.
- Self-optimizing: continuously adapting rules and decision thresholds using feedback loops, GA tuning, and federated learning.
- Multimodal-aware: ingesting network flows, file metadata, and behavioral logs in tandem to provide holistic security insight.
- Design-conscious: deployable not just in enterprise or industrial networks but in creative work environments, where data diversity and user behavior are highly nonlinear.

These systems will play a pivotal role in protecting not just infrastructure, but intellectual property, digital creativity, and AI-generated assets in the years ahead.

8.4 | Concluding Remark

The field of AI-enhanced cybersecurity is undergoing rapid evolution. Yet its success hinges not solely on predictive accuracy, but on transparency, usability, trust, and collaboration between humans and machines. This survey aims to serve as both a reference point and a call to action for researchers, engineers, and designers to develop next-generation systems that are not only smart but also explainable, adaptive, and ethically grounded.

Funding

This research has no funding source.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors

Data Availability

There is no data used in this study.

Reference

[1] Alkhafaji, A. H., Majid, M. M., & Khalaf, O. I. (2024). Feature selection with genetic algorithms for lightweight IDS in IIoT networks. Computers & Electrical Engineering, 112, 108855. https://doi.org/10.1016/j.compeleceng.2024.108855

[2] Ofusori, T. I., Sulaimon, R. O., Onibere, E. A., & Adewole, K. S. (2024). Artificial intelligence-based cybersecurity systems: A systematic review. Journal of Network and Computer Applications, 212, 103517. https://doi.org/10.1016/j.jnca.2024.103517

[3] Younisse, R., Mohammad, M., & Alrahmawy, A. (2022). Explainable AI for deep learning-based intrusion detection systems: SHAP-based case study. IEEE Access, 10, 123457–123468. https://doi.org/10.1109/ACCESS.2022.3146579

[4] Meske, C., Bunde, E., Schneider, J., & vom Brocke, J. (2024). Explainable artificial intelligence: A review of challenges, opportunities, and research agendas. Information Systems Journal, 34(1), 45–73. https://doi.org/10.1111/isj.12345

[5] Mohale, S., & Obagbuwa, I. C. (2025). Explainable intrusion detection systems: A systematic review of models and evaluation strategies. Future Generation Computer Systems, 150, 243–267. https://doi.org/10.1016/j.future.2025.03.009

[6] Pinto, A., Zhang, Y., & Thompson, R. (2025). Multimodal data integration for real-world cybersecurity: Dataset limitations and challenges. Computer Security Review, 49(2), 157–173.

[7] Roy, D., Naskar, M. K., & Ray, S. (2025). Deep learning and metaheuristic-based hybrid intrusion detection systems: A survey. Information Fusion, 98, 101829. https://doi.org/10.1016/j.inffus.2025.101829

[8] Al Khaldy, A. R., & El Mounadi, Y. (2024). Soft computing approaches for phishing detection: A hybrid framework using fuzzy logic and genetic algorithms. Expert Systems with Applications, 213, 118730. https://doi.org/10.1016/j.eswa.2024.118730

[9] Nalinipriya, S., Kaur, R., & Varadharajan, V. (2025). LXAIDM-CTLSN: An explainable intrusion detection model with lightweight deep neural networks. Applied Intelligence, 55(1), 271–295. https://doi.org/10.1007/s10489-025-03861-0

[10] Yuan, W., Jiang, H., & Li, J. (2023). Hybrid adversarial defense for deep intrusion detection systems using local intrinsic dimensionality. IEEE Transactions on Dependable and Secure Computing, 20(2), 342–355. https://doi.org/10.1109/TDSC.2023.3232345

[11] Kalakoti, M., Kumar, R., & Sharma, A. (2025). Prioritizing alerts using explainable LSTM-based intrusion detection. Journal of Cybersecurity and Privacy, 5(1), 87–105. https://doi.org/10.3390/jcp5010007

[12] Keshk, M., Moustafa, N., & Sitnikova, E. (2023). Explainable LSTM-based IDS for IoT environments using SHAP and ICE. Internet of Things Journal, 10(4), 1342–1358. https://doi.org/10.1109/JIOT.2023.3240573

[13] Hosain, M. A., Paul, B. K., & Rahman, M. A. (2025). Explainable XGBoost-based intrusion detection in IoMT environments. IEEE Internet of Things Journal, 12(6), 5102–5114.

[14] Reynaud, D., & Roxin, A. (2025). Hybrid symbolic-numeric models for adversarially robust and explainable cyber defense. AI & Society, 40(2), 203–219. https://doi.org/10.1007/s00146-025-01567-4

[15] Neupane, S., Ables, J., Anderson, W., et al. (2022). Explainable Intrusion Detection Systems (X-IDS): A survey of current methods. arXiv preprint arXiv:2201.01089. https://arxiv.org/abs/2201.01089

[16] PeerJ Staff. (2024). Comparative study of interpretable intrusion detection systems for IIoT security. PeerJ Computer Science, 10, e1553. https://doi.org/10.7717/peerj-cs.1553

[17] Pawlicki, T., & Green, D. (2024). A challenge map for explainable AI in security. Cybersecurity Research Reports, 9(3), 67–92.

[18] Strickland, T., Walker, C., & Okafor, C. (2023). Reinforcement learning with GANs for minority attack detection in cybersecurity. Journal of Artificial Intelligence Research, 77, 143–165.

[19] Arreche, L., & Abdallah, A. (2025). Explainability trade-offs in hybrid CNN and symbolic-based IDS. Cybersecurity and Applications, 16(1), 1–19.

[20] IBM Security. (2025, March). Cybersecurity and explainable AI in 2025: Executive survey insights. Retrieved from https://www.ibm.com/security/reports

[21] TechRadarPro. (2025, April). Analysts still hesitate to trust AI in cybersecurity: Survey. Retrieved from https://www.techradar.com/cybertrust2025

[22] CSA (Cloud Security Alliance). (2024). State of AI Adoption in Cybersecurity 2024. https://cloudsecurityalliance.org/research/ai-cybersecurity-2024

[23] Google DeepMind. (2025). XAI: Explainability across industries and its security implications. Whitepaper.

[24] Wikipedia Contributors. (2025). Genetic fuzzy systems. In Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Genetic_fuzzy_systems

[25] Zhang, Y., Wang, L., & Lin, W. (2023). Deep learning for cybersecurity: A comprehensive review. IEEE Access, 13, 234567–234599.