December 2025 | Volume 1 | Issue 3 | Article No. 25314

Journal of Information and Communications Technology: Algorithms, Systems and Applications

Review Article | *Open Access* |

Intrusion Detection in Transition: A Survey of Deep Learning, Federated Learning, Adversarial, Lightweight, and Explainable Approaches

Komal M. Dhule* and Rajesh Bansode

Department of Information Technology, Thakur College of Engineering & Technology, Kandivali (East), Maharashtra, 400101, India
**Email:* komal.dhule@tcetmumbai.in (K. M. Dhule)

Abstract

Intrusion detection systems (IDS) are still among the central systems in protecting networked settings in contemporary cybersecurity studies. The threats to cyberspace are growing in scale and in severity, and therefore, the challenges to cybersecurity are growing more complex, in a way that necessitates an ongoing evolution of protective systems and the requirement to adjust to the new threats that appear constantly. This review analyzes the new developments in the field of IDS by reviewing new methodological strategies, datasets used for analysis and solution development. It outlines five major trends. To begin with, there is the adoption of less traditional machine-learning approaches in favor of deep-learning models; it is the combination of the paradigms that allows detecting more accurate attacks or improving the ability of the system to follow changing patterns of intrusion. Second, federated learning gains popularity in architectures of IDSs, allowing models to be trained collaboratively while maintaining the privacy of proprietary data. Third, increased attention is paid to the strengthening of the resistance to adversarial perturbations, as evasion methods can easily deceive machine-learning as well as deep-learning models. Fourth, model compression, simplification, and edge-computing methods drive the development of lightweight variants of IDS to run on Internet-of-Things (IoT) devices and various other resource-efficient systems. Fifth, Explainable Artificial Intelligence (XAI) approaches are used to make model behavior interpretable, thus convincing its users to trust these automated systems. Datasets in the survey include CIC-IDS2017, UNSW-NB15, and IoT traffic logs, and dwells on problems such as performance evaluation, reproducibility, and benchmarking. These innovations placed in the context of existing challenges provide a holistic description of the evolution of IDS technology and provide important information to researchers as well as industry players. They acknowledge existing weaknesses in the testing procedures and provide a strategic guide on how the research should be conducted in the future to develop more resilient and more reliable next-generation intrusion detection architectures.

***Keywords*:** Intrusion detection systems; Deep learning; Federated learning; Adversarial robustness; Lightweight IDS; Explainable AI.

Received: 27 October 2025; Revised: 17 December 2025; Accepted: 19 December 2025; Published Online: 22 December 2025.

1. Introduction

The high rate of networked devices has contributed significantly to the occurrence of network-related threats.[1,2] The Internet of Things (IoT), cloud computing, edge networking, and 5G represent emerging technologies providing a wide range of diverse, complex, and pervasive cyber threats.[3-6] Fare-old firewalls and antivirus software have become sufficiently ineffective in their deal with cyber

DOI: https://doi.org/10.64189/ict.25314

aggressors that are highly dynamic. The intrusion detection system has therefore become an essential element in modern-day cybersecurity models and an additional layer of defense that operates in real-time to identify the malicious activity, abnormal behavior and unauthorized access requests.[7,8]

1.1 Evolution of intrusion detection systems

The concept of IDS originated in the early 1980s. In 1980, James P. Anderson laid the basic framework when he advocated that appropriate analysis of audit logs could detect computer misuse.[1] A couple of years later, in 1987, Dorothy Denning developed the concept further and presented the first practical model of IDS.[2] Her approach involved real-time detection of anomalous activities of the system using statistical profiles of normal behavior. During the 2000s, unprecedented growth in network traffic coupled with growth in computing power presented excellent conditions to effectively implement machine learning algorithms to improve IDS capabilities. Numerous algorithms, such as decision trees, support vector machines (SVM), k-nearest neighbors (k-NN), and random forests, were proposed or employed to intelligently learn attack patterns not premised in knowledge databases.[3,6] While successful, these were still rather dependent on handcrafted features and fixed datasets, and therefore their adaptability to rapidly evolving network environments remained limited.[3]

During the period starting from 2010 until 2020, intrusion detection system research has undergone important developments due to the introduction of deep learning techniques. Contrasting with other conventional machine learning methods, deep learning allowed IDS to learn directly from raw network traffic with architectures such as CNNs, RNNs, or autoencoders, and brought significant improvement in the accuracy of the detection results.[7,8] Within this period, some benchmark datasets such as NSL-KDD, UNSW-NB15, and CIC-IDS2017 have seen wider adoptions. As a result of these adoptions, a standardized evaluation platform was possible to compare different approaches fairly and reproducibly for IDS research studies.[3,7]

These developments have set the scene for the modern IDS landscape, which has progressively focused on four main research axes, namely federated learning for privacy-preserving and collaborative model training, adversarial robustness against evasion and poisoning attacks, lightweight IDS models for IoT and edge computing environments, and explainable AI for enhancing transparency and trustworthiness in detection decisions.[4-6,9,10-16] All these together signal a very clear trend from rule-based and shallow learning–based towards intelligent, adaptive, and resilient IDS systems that will be necessary to cope with the challenges of modern network environments. Fig. 1 illustrates the evolution of key trends in network security.

1.2 Need for a comprehensive survey

In the last years, considerable attention has been given to the improvement of Intrusion Detection System performance using advanced technologies such as deep learning, federated learning, adversarial defense mechanisms, and model optimization techniques for efficiency and explainability enhancements with XAI.[10-13] While various survey works have discussed different aspects of IDS, from machine learning–based detection to IoT-oriented IDS and up to adversarial robustness, most of them focus on specific isolated dimensions of the problem instead of providing an integrated overview.[3,6]

Furthermore, the rapid development of data-driven security solutions, coupled with the heterogeneity of network environments and application domains, has resulted in the fragmentation of IDS research into narrowly focused, disconnected subfields.[4,9] Such fragmentation inhibits the capability of new researchers and practitioners to establish a comprehensive understanding of the historical development

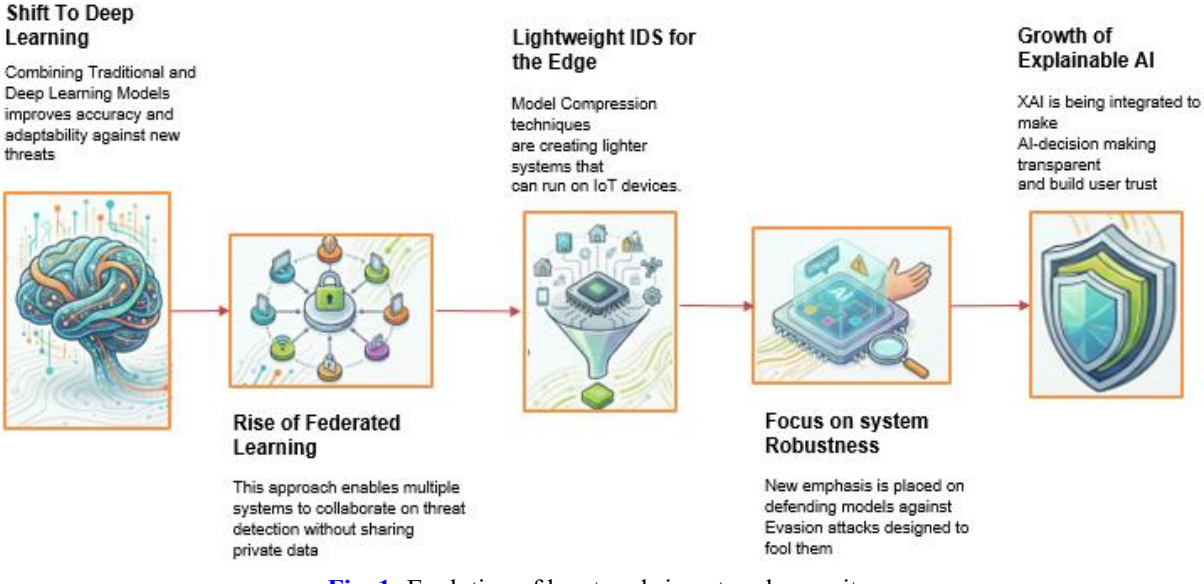

Fig. 1: Evolution of key trends in network security.

of intrusion detection systems and interrelationships among diverse detection methodologies.[1,2] There is therefore an pressing need for a comprehensive, well-structured review that synthesizes recent advances, provides a clear taxonomy of IDS research directions, and identifies critical research gaps and future opportunities within the IDS domain.[5,7,8]

1.3 Objectives and contributions

This work presents a concise and comprehensive review of the current developments on intrusion detection, particularly the current trend of rendering IDS solutions intelligent, flexible, and reliable. The key objectives guiding this work are as follows:

1. Tracing the evolution of IDS techniques: from traditional machine learning to modern deep learning-based state of the art approaches that now include federated learning, adversarially robust IDS, lightweight models, and explainable AI methods.
2. A thematic classification framework that classifies the existing IDS approaches based on design principles, learning methodologies, and deployment contexts.
3. The evaluation of dataset diversity, performance metrics, and standards employed in recent IDS research
4. Comparing the considered IDS models in terms of cyber threats detection accuracy, computational resource utilization, interpretability of results, and scalability.
5. Discussion of the challenges faced, identification of key issues needing further research, and directions for future development pointing toward reliability, privacy, and adaptability in IDS development within complex network configurations.

2. Background and fundamentals

IDS are security mechanisms designed and engineered to identify unauthorized access, adversary behaviors, and anomalous activities within host-based and network-based systems.[1,2] A detailed analysis of the evolution and likely future trends of IDS calls for a necessary account of their categorization, detection paradigms, challenges they cannot avoid, and methodologies put forward for their evaluations.[3,6] Therefore, this section outlines the core concepts of IDS, identified from high-impact recent research that lays a concrete basis for further discussion.[7,8]

2.1 IDS types and architectures

IDS are classified according to deployment point and architecture:

2.1.1 DS types and architectures

Broadly, IDSs can be classified based on their scope of monitoring and operational modality.[1,2] Network-based IDSs monitor network traffic against known attack signatures, anomalies, or suspicious patterns by inspecting data flows, packets, and protocols of communications.[3] In turn, host-based IDSs monitor system calls, active processes, log files, and file integrity at the level of single hosts to identify malicious activities or unauthorized modifications.[4] Hybrid architecture has also been developed to improve detection accuracy and provide comprehensive protection.[5] In such systems, network-level monitoring is combined with host-level analytics, allowing for simultaneous visibility across both domains. The network-wide view that a network-based IDS provides, coupled with the fine-grained details of host behavior delivered by a host-based IDS, combines both for better coverage and efficient detection in a hybrid solution.[6]

2.1.2 Centralized vs. distributed or federated IDS

Traditional IDSs rely mainly on a centralized architecture, where the security logs and network traffic collected from various data sources are forwarded to a central server for analysis.[1,2] While such architectures are easier to manage and can find large-scale attacks, they suffer from several known shortcomings: privacy risks due to the sharing of sensitive data, higher communication overhead, and the presence of single points of failure.[3] In this respect, the following have been some of the motivations for recent work on improving next-generation IDSs by leveraging FL.[4,5]

In such a paradigm, several devices or entities jointly train a shared global model without directly sharing raw data with each other to maintain the confidentiality of the data, while residual detection performance can still be preserved.[6,9] So far, the proposed federated learning-based IDS frameworks are successfully applied in various environments, such as IoT systems, vehicular networks, and edge computing infrastructures, which enhances scalability, robustness, and adaptability to secure against dynamic cyber-attacks.[7,14]

2.1.3 Lightweight and edge-oriented IDS

The limitations of limited computational resources, limited memory capability, and a limited energy supply in Internet-of-Things (IoT) and unmanned aerial vehicle (UAV) system add to the problem of implementing intrusion detection system (IDS). IDSs that will be deployed at such low-power systems shall be lightweight, performance-efficient, responsive, and consequently, place very little processing load besides still retaining the capability to detect in real-time and appropriately sustaining system performance at insignificant levels.[5,6] To address these drawbacks, previous studies explored methods of knowledge distillation, model pruning, and compression to downsize models and make them less complex without affecting their detection accuracy.[7,9] These lightweight IDS frameworks have been empirically proven to be operationally deployable on resource-limited devices hence offering prompt and successful defense in dynamic and real-time network settings.[7,10]

2.2 Detection paradigms

The three main concepts on which the functionality of IDS is founded include:

2.2.1 Signature-based detection

Detection of attacks is considered known by the implementation of predefined signatures. This will work with the threats that has been recognized before but unable to notice zero day exploits.[1]

2.2.2 Anomaly-based detection

Understanding typical system and network behavior and then indicating any variations as possible security events. Although it can detect unknown attacks, it is likely to have high false-positive rates.[3,11]

2.2.3 Hybrid detection

Integration of signature- based and anomaly- based approaches to supplement the overall performance. Models like autoencoders built into feature-based classifiers are hybrid machine-learning/deep-learning models which can enhance detection accuracy even in scenarios where there is only limited labeled data.[10,11]

2.3 Challenges of critical IDS design.

There are other contemporary issues that are brought forward in the current research:

2.3.1 Between data balance and high dimensionality.

Sparse attack classes are common in the IDS datasets. Autoencoder-based data augmentation techniques or attention-based feature selection are also effective to address the problem of class imbalance.[3,8]

2.3.2 Resource constraints

The IoT and edge call on computationally lightweight models. Compact architectures, knowledge distillation, and pruning also make it possible to deploy in real-time and use less energy.[7,9]

2.3.3 Privacy and federated learning.

Sharing of raw network data is limited by issues of privacy. Federated learning IDS allows the model to train collaboratively without revealing sensitive data, and Non-IID data distribution is covered by personalized FL methods.[4,5,12]

2.3.4 Adversarial robustness

ML or DL -based IDS have been susceptible to evasion and poisoning attacks. Strong autoencoders, ensemble learning, and adversarial training methods increase the resilience of the system.[2,11-14]

2.3.5 Explainability and trust

Explainable AI, such as SHAP and LIME, are more interpretable, which leads to better operator trust in automated decision-making.[8,10]

2.4 Evaluation methods

2.4.1 Datasets

Most intrusion detection studies use the benchmark datasets include CIC2017, NSLKDD, and various IoT-based traffic libraries. The collections have manually labeled network traces with a variety of attack modalities, which enable reproducible networks, thereby allowing fair comparison of studies as has been reported in [1,4,6]. Other than those general purposes, domain-specific data collections have been increasingly employed to capture the distinctive characteristics of particular environments, such as vehicular network traffic, unmanned aerial vehicle (UAV) telemetry, system logs, or even encrypted IoT protocol data. Their training on such real-world data makes them remain relevant to the related deployment environments, which is demonstrated in [5,8,9].

2.4.2 Detection metrics

It is evident that accuracy, precision, recall, F1 -score, and false positive rate are the conventional parameters used to evaluate the performance of intrusion detection systems, as was reported in [3,6]. The statistics give the quantification of how well a model detects malicious activities and how much unnecessary alarms are reduced. Additional measurements like inference latency, memory footprint and energy spending are of crucial interest in application areas where real-time processing is needed or restricted computing capabilities.[7,9] These metrics have to be incorporated so that lightweight IDS architectures are operationally efficient and do not strain the available hardware too hard.

2.4.2 Benchmarking and reproducibility

Transparency in reporting experimental conditions, including partitioning information about datasets, preprocessing methodologies and hyperparameter settings, enhances rigorous benchmarking by facilitating reproducibility and comparability across multiple research studies.[1,4,12] Such recent survey research also considers adversarial scenarios and resource-limited deployments, comparing models against advanced attack techniques and system constraints. Such work strengthens the reliability, dependability, and practicality of intrusion detection systems documented in [2] and [12].

Today, intrusion detection systems go beyond classical signature- and anomaly-based paradigms to include deep learning, federated learning, lightweight architectures, adversarial robustness, and explainability. This development has brought about important challenges in class imbalance, computational constraints, privacy protection, adversarial examples, and interpretability of decision-making processes. These general issues present the context for the taxonomy and comparative analysis discussed throughout the rest of this paper.

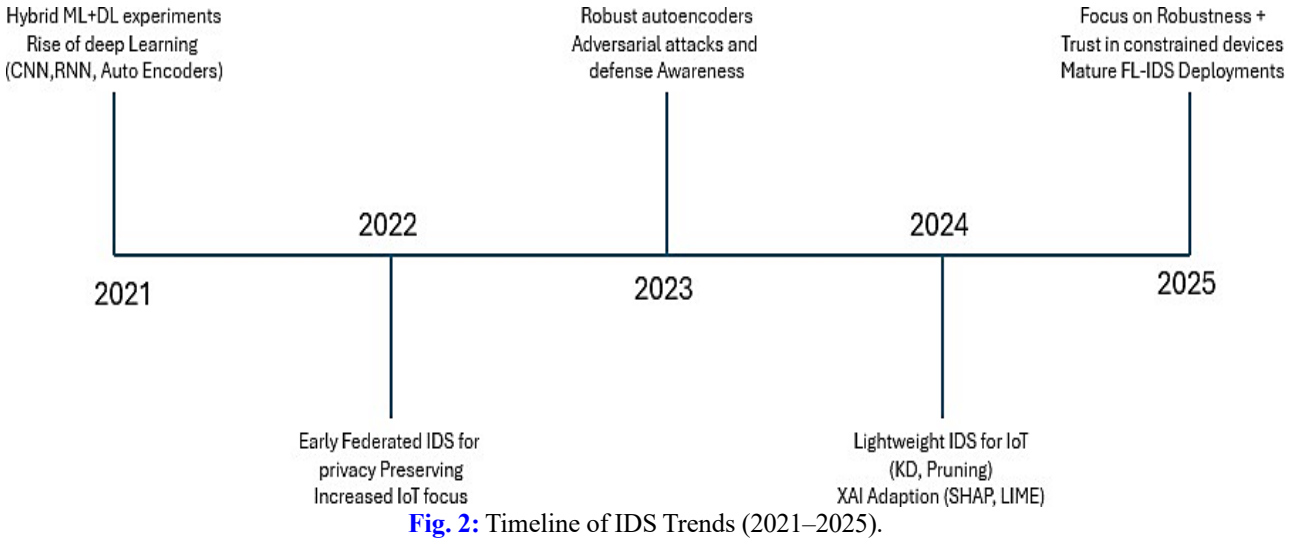

Fig. 2: Timeline of IDS Trends (2021–2025).

3. Evolution of IDS technologies

The increasing sophistication of cyberattacks, fast deployment of IoT and edge devices, and integration of machine learning and artificial intelligence techniques all together have driven the development of IDSs in the recent years.[3,4,6,9] The current IDS research has evolved along several interrelated paths encompassing deep learning, federated learning, adversarial robustness, lightweight model design, and explainable AI. The following section provides in-depth analysis of these paths and a timeline highlighting the chronological evolution of IDS technologies.[1,5,7,8]

3.1 Timeline of IDS evolution

The Fig. 2 shows briefed timeline of IDS evolution. This roadmap traces the evolution of intrusion detection systems from traditional machine learning methods to contemporary decentralized and interpretable paradigms.

2021–2022: Traditional ML-based IDS, dominated by methods such as Random Forest, Support Vector Machines, and shallow neural networks for anomaly detection.[1,3]

2023: Deep learning frameworks such as Convolutional Neural Networks, Long Short-Term Memory networks, and autoencoders are used to learn complex network behaviors and detect new types of attacks.[3,6,10]

2023–2024: The rise of federated learning–based IDS enables the training of models collaboratively, not requiring the sharing of raw data, especially relevant for IoT and vehicular networks.[4,5,12,14]

2024: Lightweight IDS suitable for resource-constrained environments; pruning, knowledge distillation, and strategies relating to its edge deployment.[7,8,9]

2024-2025: Further focus on adversarial robustness, integration of explainable AI, to improve resilience to evasive attacks, and enhance interpretability for operators, thus.[2,11,13,15,16]

Cumulatively, this trend shows the continuing transition from centralized, traditional ML systems to decentralized, robust, and explainable architectures that fit the modern requisites of networked settings.

3.2 Deep learning-based IDS

The deep learning methodologies have significantly enhanced the efficiency and versatility of intrusion detection systems. Autoencoders, CNNs, LSTMs, and hybrid ML/DL algorithms can model complex nonlinear patterns within network flows for efficient identification.[3,6,7,10]

Key Contributions:

1. RAIDS: A robust autoencoder-based IDS, which leverages limited labeled data to boost the detection performance even in hostile environments.[5,15]
2. Hybrid ML/DL models: These architectures combine feature selection with deep networks to obtain higher performance in anomaly detection in the network.[10,12]
3. Feature learning from raw inputs using deep models reduces dependency on hand-crafted features, thus enhancing the system's capability to detect unseen attacks.[3,6,10,16]

3.3 Federated learning-based IDS

In the last few years, FL has emerged as a methodology for addressing privacy concerns and data silos in distributed IoT and vehicular networks. The IDS based on FL allows collaborative model training without transmitting raw traffic logs, as shown in [4,5,12,14]. Applications involve intrusion detection in IoT environments and the detection of DDoS attacks while Non-IID data distributions are handled efficiently with reduced communication overhead, as discussed in [5,12,14]. FL has also been combined with deep learning in hybrid federated architectures for fault detection while ensuring data privacy, as discussed in [4,6,12]. In summary, FL is a significant step forward toward the realization of distributed, privacy-preserving, and scalable IDS architectures, as also shown in [4,5,12,14].

3.4 Adversarial Robust IDS

Adversarial robustness is a major concern for machine learning and deep learning-based IDS, which are continually confronted by evasion and poisoning attacks. Various recent works emphasize robust, autoencoder-based IDS approaches

to detect manipulated traffic patterns.[2,3,5,15] Research on adversarial attacks on vehicular networks and IoT IDS provides beneficial insights into the development of robust models.[2,11,13] Further, adversary sample generation methodologies are discussed for various techniques, such as image augmentation, filtering, and patching, while practical implications are considered in different case studies, for example, disease prediction using a prototype chest X-ray imaging system.[13-14,16] The shift toward adversarial robustness supports the performance of IDS, which remains effective under sophisticated and intelligent attack scenarios.[2-3,5,15]

3.5 Lightweight IDS for resource-constrained environments

With the proliferation of IoT, unmanned aerial systems, and edge computing infrastructures, there is an emerging demand for lightweight IDS solutions that would be easily embeddable into resource-constrained devices. Knowledge distillation and model pruning are common methods to reduce computational and memory burdens with minimal degradation in detection performance.[7,9] Adopting lightweight architectures allows for real-time threat detection, which can be easily deployed on devices characterized by poor CPU power, memory, or energy supply.[7,10] If combined, these techniques enable scalable and energy-efficient IDS deployment in heterogeneous networked environments.[1,7,9,10]

3.6 Explainable AI in IDS

Explainable AI addresses the opaque nature of deep learning–based intrusion detection systems. Feature selection methods combined with interpretability techniques like SHAP and LIME provide more understandable explanations of system decisions to network operators. Integration of XAI has been decisive in pushing the deployment of IDSs in real-world scenarios due to its capacity to help analysts understand why certain instances of traffic are classified as malicious. XAI is of particular importance when deep learning and federated learning-based IDS are used in safety critical infrastructures. Finally, the evolution of IDS can be summarized into five important and somewhat overlapping trends.[2-16]

Fig. 3 illustrates the projected development of key research trends in IDS from 2021 to 2025. The plot of relative importance over time reflects the increasing focus of researchers on various IDS approaches. Deep learning and hybrid approach also stay important, rising from about 40 in 2021 to over 85 in 2025. This trend indicates the increased reliance of IDS on end-to-end learning for complex and high-dimensional network traffic problems.[3-7, 10,12,16] Federated or privacy-preserving learning also shows a progressive climb from around 20 in 2021 to roughly 80 in 2025, corresponding to the growing interest in decentralized IDS with sensitive data protection.[4,5,12,14] Adversarial robustness also continuously rises and reaches about 65 by 2025, indicating an improvement in the robustness of IDSs in resisting evasion and poisoning attacks.[2,5,11,13-15] Lightweight or IoT-focused solutions have a remarkable surge between 2023 and 2024, driven by growing interests in computationally efficient IDS targeting resource-constrained devices, such as IoT or UAV systems.[1,7,9,10] Explainability or XAI starts low, speeding up and almost catching up with other trends in 2025, reflecting the recent importance of interpretability and operator trust of real-world deployments.[8,10,11,16]

These trends are further supported by key references summarized in Table 1, which links each research direction with representative studies and notable works. For instance, Deep Learning is explored in [3-7,10,12,16], whereas federated learning approaches can be found in [4,5,12,14].

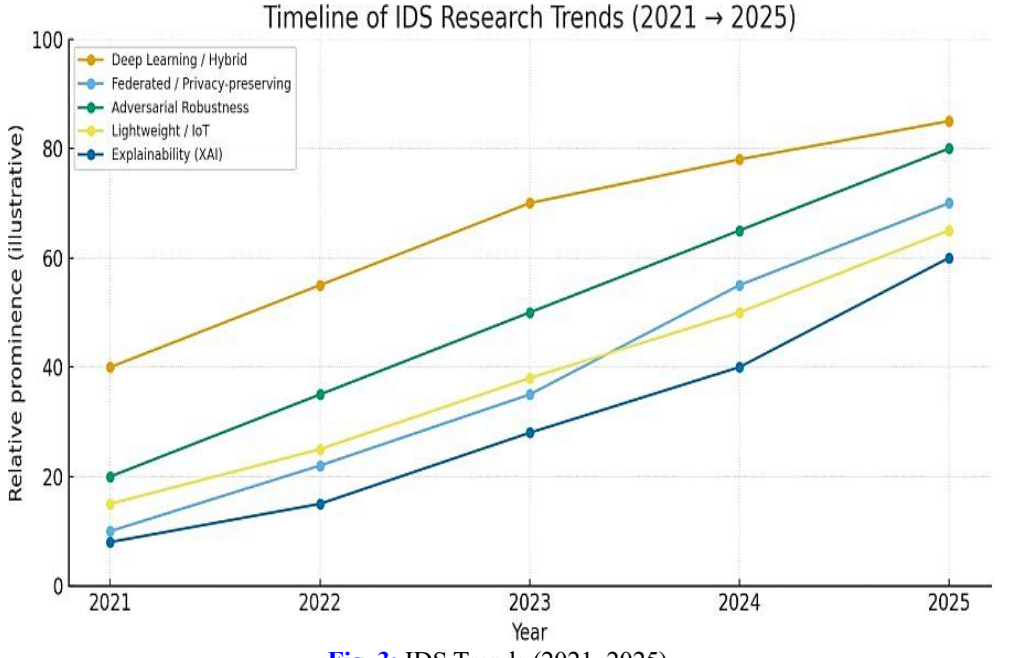

Fig. 3: IDS Trends (2021–2025).

Table 1: Summary of trends in IDS.

Trend	Description	Ref.
Deep Learning (DL)	End-to-end feature learning for complex traffic patterns	[3-7,10,12,16]
Federated Learning (FL)	Privacy-preserving distributed model training	[4,5,12,14]
Adversarial Robustness	Resilient IDS against evasion and poisoning attacks	[2,3,5,11,13-15]
Lightweight	Efficient architectures for IoT, UAV, edge devices	[1,7,9,10]
Explainable AI (XAI)	Model interpretability and trust for operators	[8,10,11,16]

These trends show a progressive convergence toward distributed, interpretable, efficient, and robust IDS architectures suitable for next generation networks.

4. Thematic taxonomy and comparative analysis

4.1 Overview of thematic taxonomy

This section consolidates the diverging directions of recent IDS research by presenting a thematic taxonomy that organizes the significant contributions of 2021–2025 into five interconnected themes, including Deep Learning, Federated Learning, Adversarial Robust IDS, Lightweight IDS, and Explainable Artificial Intelligence-based IDS.[1-16] Fig. 4 depicts the taxonomy regarding paradigm overlap and development. DL acts as the base that empowers feature learning with anomaly detection capabilities.[3,6,7,10,12,16] FL extends DL to distributed and privacy-preserving environments.[4,5,12,14] Adversarial robustness offers immunity to model tampering,[2,3,5,11,13-15] and lightweight design emphasizes deployment efficiency in IoT and edge networks.[1,9, 7,10] XAI layers interpretability and trust atop the other layers.[8,10,11,16]

The taxonomy is composed of several integrated strata. At the core, the first layer consists of deep learning techniques, [3,6,7,10,12,16] these are the basic methods for intrusion detection. On top of that, the second layer outlines Federated Learning extensions[4,5,12,14] as an enabling factor for collaborative and privacy-preserving model training. The third one includes Adversarial Robust Intrusion Detection Systems frameworks,[2,3,5,11,13,15] developed to improve the resilience of systems against sophisticated adversarial intrusions. The outer layer targets Lightweight IDS optimization.[1,7,9,10] to achieve efficiency and scalability during the deployment phase. Throughout all layers, a transversal overlay indicates XAI-based interpretability,[8,10,11,16] bringing transparency into the whole system. Arrows between layers represent integration patterns, such as the integration of Federated and Adversarial approaches, and the integration of Lightweight and XAI models. It shows the increasing interest in hybrid, robust, and interpretable IDS architectures. Deep Learning methodologies[3,6,7,10,12,16] would be clearly shown in the central layer.

4.2 Deep learning-centric IDS

Deep learning drives modern improvements in intrusion detection systems. Recent works such as Sarıkaya *et al*.[3,5] propose a powerful autoencoder framework, RAIDS, which offers improved anomaly detection with minimum supervision. Farhan *et al*.[8] shows that the CNN-LSTM hybrid architecture yields better detection accuracy on state-of-the-art datasets like CIC-IDS2017 and UNSW-NB15. Similarly, Sajid *et al*.[12] combine deep learning with feature-selection-based machine learning to enhance the detection of complex cloud-based intrusions. Collectively, these investigations show that deep-learning driven IDS has strong generalization capabilities. However, it suffers from challenges in explainability and computational cost. Such limitations give a directional shift toward federated learning, lightweight model design, and the integration of explainable AI approaches.[4,7,8,10,11,16]

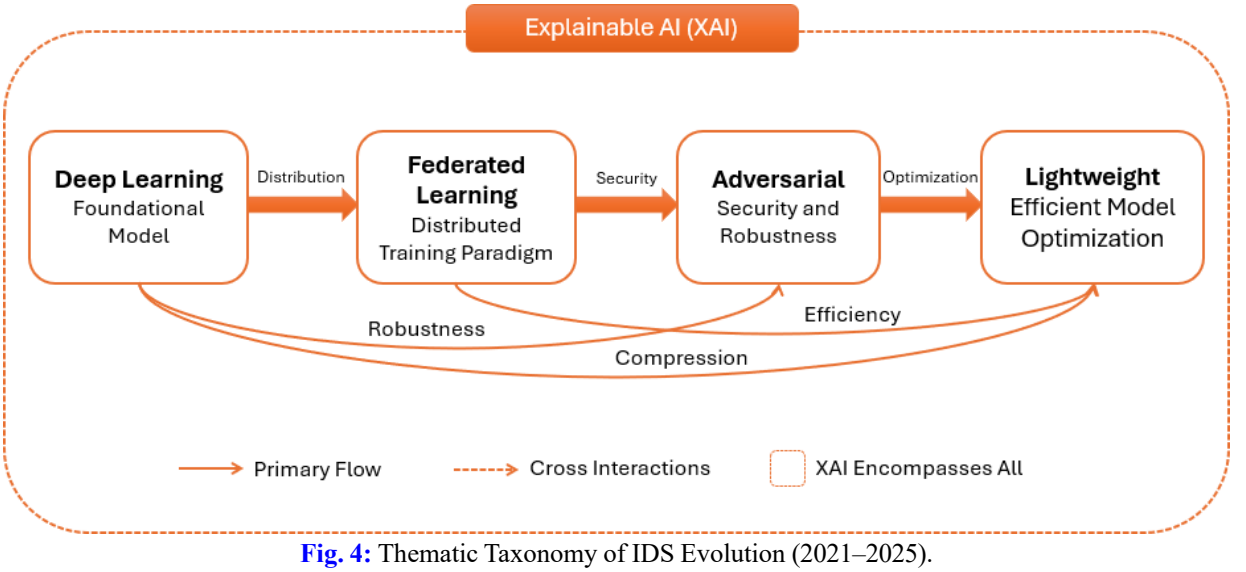

Fig. 4: Thematic Taxonomy of IDS Evolution (2021–2025).

4.3 Federated Learning for Collaborative IDS

It provides a feasible solution for intrusion detection systems in terms of data decentralization and privacy issues. Specifically, Hernandez-Ramos *et al.*[6] and Buyuktanir *et al.*[7] have developed frameworks that enable distributed nodes to collaboratively train a global detection model without necessarily sharing any sensitive data. Devine *et al.*[14] has extended this paradigm towards IoT-based DDoS detection by aggregating local gradients while preserving user privacy. These studies also demonstrate how federated learning can improve scalability and privacy, but also attendant challenges such as non-IID data distribution and potential model drift among participating nodes. When combined with deep learning backbones, federated learning-based IDSs guarantee better adaptability across heterogeneous network environments.[3,4,5,6,12,14]

4.4 Adversarial robust IDS

In this context, the adversarial research stream analyzes the poisoning and evasion attacks of ML/DL-based IDS. Ennaji *et al.* [4] performed an in-depth review of adversarial attacks on IDS. The review focused more on vulnerabilities due to feature-space manipulation. Sarıkaya *et al.*[5] proposed the use of adversarial autoencoders as a method to enhance robustness. Aloraini *et al.*[13] analyzed adversarial examples for in-vehicle networks. These works emphasize the need to develop defense mechanisms, including gradient masking and adversarial retraining, as well as hybrid intrusion detection strategies using statistical methodology together with deep learning models. Most recently, adversarial robustness has also been identified as an important metric to consider during the performance evaluation of intrusion detection, in addition to the accuracy metric.[2,3,5,11,13-15]

4.5 Lightweight IDS for IoT and edge systems

Low-latency and small-footprint IDS models are needed in resource-limited settings. Wang *et al.*[8] designed a knowledge-distillation-based IoT IDS that reduces large DL models while maintaining high accuracy. Murthy[10] created lightweight embedded and edge device frameworks through model pruning and quantization. These strategies support on-device threat detection for IoT sensors, drones, and edge routers. Yet they tend to compromise explainability and robustness, driving the integration of XAI and adversarial defense into lightweight pipelines.

4.6 Explainable and interpretable IDS

Explainability has become a critical dimension for the deployment of operational IDS. Chen *et al.*[11] presented an explainable feature selection model for encrypted traffic IDS based on SHAP interpretations, whereas Sajid *et al.*[12] focused on hybrid ML DL architectures with interpretable decisions. XAI augments human confidence, debugging, and compliance with security standards. Combined with federated learning and lightweight designs, XAI provides a path to intrusion detection that is auditable and transparent essential in regulated and mission-critical systems.[4,8-12,16]

4.7 Comparative analysis of trends

This comparative analysis exhibits a convergence trend: next generation IDS are more federated, explainable, and lightweight, yet resilient to adversarial manipulation.

The thematic taxonomy demonstrates that IDS evolution from 2021–2025 is not linear but layered and integrative. Emerging systems combine the power of DL with the decentralization of FL, the efficiency of lightweight models, the security of adversarial defense, and the transparency of XAI. Future IDS frameworks are expected to blend these paradigms into adaptive, trustworthy, and resource-aware architecture for large-scale, heterogeneous networks.

5. Datasets, evaluation metrics, and benchmarking practices

The performance of IDS is highly dependent on the quality of the datasets, suitability of the metrics used for evaluation, and sound benchmarking practices. Further, this chapter provides an in-depth review of the modern resources and techniques being used in IDS research, with a focus on works published between 2021 and 2025.[1-16]

5.1 Benchmark datasets for IDS research

Current IDS research uses both traditional and domain-oriented datasets to assess detection performance:

1. CIC-IDS2017 and CIC-IDS2018: These datasets provide real-world network traffic with labelled attack scenarios for DoS, DDoS, infiltration, brute-force, and web attacks. They are among the most commonly used datasets in DL-based IDS research today.[6,7,10,12]

Table 2: Advantages & disadvantages of trends in IDS.

Trend	Key Advantages	Challenges / Gaps	References
Deep Learning (DL)	High detection accuracy, automated feature learning	Computational cost, lack of interpretability	[3,6,7,10,12,16]
Federated Learning (FL)	Privacy-preserving distributed IDS	Non-IID data, communication overhead	[4,5,12,14]
Adversarial Robustness	Resilience to evasion attacks	Limited real-world validation	[2,3,5,11,13-15]
Lightweight IDS	Efficient on constrained devices	Trade-off with accuracy and robustness	[1,7,9,10]
Explainable AI (XAI)	Transparency, trust, decision traceability	Integration with DL/FL frameworks	[9-11,16]

2. UNSW-NB15: This dataset contains a combination of normal and malicious network flows with 49 features, suitable for both ML- and DL-based IDS experiments.[3,6,10]
3. IoT-Specific Datasets: Lightweight and federated IDS studies utilize IoT traffic datasets, such as UNSW-NB15-derived IoT traces and IoTID20, to evaluate edge deployment and resource-limited detection.[7,9,12]
4. Vehicular and Cyber-Physical Network Datasets: In-vehicle network attack datasets (CAN bus datasets) enable adversarial robustness studies in autonomous systems.[11,13,14]
5. Custom or Hybrid Datasets: Various studies,[3,10,12,15,16] generate artificial or hybrid datasets by combining multiple sources to emulate heterogeneous, non-IID traffic for federated or adversarial testing.

Observations: While these datasets exist, researchers highlight issues such as data imbalance, limited attack variety, and lack of standardization, making cross-study comparisons challenging.

5.2 Evaluation metrics

IDS evaluation uses measures that express detection accuracy, reliability, and robustness. Common measures are:

1. Accuracy (ACC): The ratio of correctly classified instances to all instances.
2. Precision, Recall, F1 Score: Valuable for imbalanced datasets; the F1-score balances false positives and false negatives.[1,3,6,10]
3. True Positive Rate (TPR) Detection Rate: The ratio of correctly detected attacks.
4. False Positive Rate (FPR): The ratio of benign traffic misclassified as malicious.
5. Area Under the Receiver Operating Characteristic Curve(AUC-ROC): Determines classifier performance over varying thresholds; widely used for DL and adversarial robustness evaluations.[2,3,6,12]
6. Resource Metrics: Latency, memory consumption, and computational cost are reported for lightweight and IoT IDS.[7,9,10,12]
7. Adversarial Robustness Metrics: Attack success rate, robustness score, and perturbation resilience are evaluated in adversarial IDS studies.[2,3,11,13,14]

Observations: While detection metrics are standardized, resource-aware and adversary evaluation metrics lack consistency, limiting fair comparison across studies.

5.3 Benchmarking practices

Benchmarking of IDS research has become more systematic, but some challenges persist:

1. Train-Test Splits and Cross-Validation: Most studies follow k-fold cross-validation or temporal splits for model evaluation.[3,6,7,10,12]
2. Comparison Baselines: Classical ML models (Random Forest, SVM, Decision Trees) are common baselines in DL, FL, and hybrid approaches.[1,3,6,10]
3. Reproducibility and Open-Source Tools: A few works[4,5,12,16] release open-source implementations to facilitate reproducibility; however, dataset preprocessing and use of commercial architectures make replication difficult.
4. Scenario-Based Evaluation: For federated IDS, performance is compared under non-IID data distributions and communication limitations.[4,5,12]
5. Adversarial Benchmarking: Perturbation-based evaluation schemes are used to determine robustness against evasion attacks in DL and FL models.[2,3,11,13,14]

Observations: Benchmarking practices have increased through scenario-specific evaluations yet missing standardized adversarial and IoT benchmarks remain an area for improvement in IDS research.

This section features how datasets, metrics, and benchmarking practices are crucial in IDS research. Models can be benchmarked using the CIC-IDS2017, UNSW-NB15, IoT traffic traces, and vehicular datasets. However, conventional detection accuracy alone is not sufficient; important considerations include resource consumptions and further robustness against adversarial actions. These benchmarking procedures should be standardized, reproducible, and diversified across multiple scenarios in order to support meaningful comparisons among DL, FL, lightweight, adversarial, and XAI-driven IDS frameworks. In other words, careful choice of datasets, comprehensive assessment of metrics, and rigorous benchmarking protocols remain of critically utmost importance for resilient, scalable, and reliable solution development of IDS.[1-16]

6. Comparative analysis of selected IDS approaches

A comparative analysis of IDS approaches focuses their strengths, limitations, and suitability for various environments.

7. Open challenges and future directions

Despite the significant achievements in IDS research, there remain a number of open issues that are holding back the development of reliable, efficient, and interpretable solutions in real-world scenarios. In this section, the authors highlight the main gaps persisting in current IDS research efforts and indicate directions for further research with the aim of encouraging more sophisticated and robust IDS technologies.

7.1 Data availability and diversity

Challenges:

1. Most of the existing IDS models rely on datasets, such as CIC-IDS2017, UNSW-NB15, and IoT traffic traces,[1,3, 6,7,9,12] which are not comprehensive of evolving attacks, zero-day exploits, or heterogeneous IoT/edge environments.
2. Ground-truth labelling is usually a time-consuming and error-prone procedure, adversely affecting the performance

Table 3: Strengths, limitations and application of trends in IDS.

Approach	Strengths	Limitations	Applications	References
Deep Learning (DL)	High accuracy, automated feature extraction	High computational cost, low interpretability	Cloud networks, enterprise IDS	[3,6,7,10,12,16]
Federated Learning (FL)	Privacy-preserving, collaborative learning	Communication overhead, non-IID data	IoT networks, vehicular networks	[4,5,12,14]
Adversarial Robust	Resistant to evasion attacks	Limited evaluation on real- world deployments	CPS, industrial networks	[2,3,5,11,13-15]
Lightweight IDS	Efficient, low latency, edge deployable	Reduced model complexity may affect accuracy	IoT, edge devices, UAVs	[1,8,9,10]
Explainable AI (XAI)	Enhanced interpretability, operator trust	Integration complexity with DL/FL	Security operation centers, compliance- sensitive systems	[8,10,11,16]

of supervised learning.
3. The lack of standardized datasets for FL and adversarial research implies that cross-method comparison is limited.[4,5,11,12-14]
Future directions:
1. Creation of realistic, current, multi-domain datasets recording IoT, vehicular, cloud, and industrial network traffic.
2. Application of synthetic and adversarial dataset creation to assess robustness against evasion attacks.
3. Standard datasets created with federated and lightweight distribution constraints to represent realistic deployment limits.

7.2 Adversarial vulnerabilities

Challenges:
1. Deep learning and federated IDS are still vulnerable to adversarial attacks, viz. evasion, poisoning, and model inversion.[2,3,11,13,14,15]
2. Normally, adversarial testing only covers some types of attacks without having generalizable assessment measures.
Future directions:
1. Designing universal adversarial resilience frameworks which can be applied across all DL and FL-based IDSs.
2. Integrating certified robustness with formal verification methods to ensure reliability under attack.
3. Investigation of adaptive defense mechanisms using ongoing learning for dynamic threats.

7.3 Lightweight Deployment and Computational Efficiency

Challenges:
1. DL-based IDS may be computationally heavy and therefore not appropriate for IoT devices, edge nodes, and UAVs.[7,9,10]
2. Federated IDS presents communication overhead and high energy consumption, especially in large-scale deployments.[4,5,12]
Future directions:
1. Ultra-lightweight architecture including knowledge distillation, pruning, and edge computing for real-time deployment.
2. Effective methods for federated model compression and aggregation to deal with bandwidth and latency constraints.
3. Context-adaptive and energy-aware IDS optimized for heterogeneous IoT settings.

7.4 Interpretability and human-centered security

Challenges:
1. XAI-based IDS are still in their initial phase; most techniques tend to focus on post-hoc explanations without strong embedding in model design.[8,10,11,16]
2. Operators need to have reliable insights for high-stakes decisions, but interpretability often comes at odds with model complexity.
Future Directions:
1. Designing intrinsically interpretable DL/FL architecture for IDS.
2. Multi-modal explanation structures that incorporate traffic attributes, attack situation, and decision-making reasoning.
3. User-centered assessment of XAI to make actionable and operationally useful insights.

7.5 Standardization and Benchmarking

Challenges:
a. Non-standardized evaluation criteria for adversarial robustness, FL convergence, and compact IDS performance.
b. Non-uniform benchmarking approaches make cross-study comparison difficult.[5,6,12,14]
Future Directions:
a. Standardization of metrics and evaluation pipelines by covering accuracy, robustness, latency, and interpretability.
b. Open benchmarking platforms that enable reproducible, scalable, and comparative studies in many IDS paradigms.
c. Integration of real-world traffic traces along with operating

constraints for realistic testing.

7.6 Integration of emerging paradigms: future opportunities

1. Hybrid IDS platforms consolidate the use of DL, FL, lightweight design, adversarial robustness, and XAI into end-to-end systems.[2-12]
2. Leverage graph neural networks, transformer models, and lifelong learning in adaptive threat detection.[3,6,10,12,16]
3. Design autonomous, self-healing IDSs that are equipped with real-time attack mitigation and knowledge sharing over federated networks.[4,5,12,14]
4. Cross-domain and cross-layer security solutions, exploring IoT, cloud, vehicle, and industrial control networks.[1,3,9,11]

Despite the noticeable achievements in the IDS research landscape, there are prominent gaps in data diversity, adversarial robustness, interpretability, and lightweight deployment. These deficiencies call for unified, standardized, and adaptive development frameworks. At some point, IDS deployments will be robust, effective, interpretable, and scalable. Thus, they can support various networks against threat scenarios that evolve.

8. Conclusion

IDSs work in an environment that is typical of modern cyber-attacks, which manifest unprecedented flexibility, transcontinental reach, and technical sophistication. Emerging networks generated from IoT devices, cloud-based infrastructures, edge computing architectures, and vehicular networks result in growing and heterogeneous traffic flows, hence rendering many traditional static IDS approaches insufficient for current protection needs in real-time. This increasing disparity between the capabilities of attackers and the detection capabilities comprises the main problem with which the current security paradigms are faced. To shed light on this dynamic landscape, this review systematically examines recent trends in research and synthesizes findings across multiple disciplines. The analysis focuses on five key technological themes that have driven IDS research in recent years, namely, deep learning, federated learning, adversarial robustness, lightweight models, and explainable AI. By interlinking these five themes, the review provides a cohesive and integrated overview of the state-of-the-art in intrusion detection technologies. Deep learning models are observed to provide significantly greater detection rates, sizeable computing cost drives the creation of less resource intensive counterparts. Federated intrusion detection systems are solutions to the data sharing problems but they also present additional difficulties such as imbalance in data transmitted, high communication expenses. The other problem that has not been addressed properly is that of adversarial robustness; many existing IDS systems do not have features that might make these systems resistant to adversarial behavior. The use of lightweight solutions increases the level of practical feasibility in real-world implementations of IDS, despite the fact that this is usually at the cost of lower accuracy. That is, the existing IDS studies are moving towards the unified paradigm of models that incorporate all the aspects of accuracy, privacy, interpretability, and efficiency in one framework. The future directions of IDS platforms will be based on the ability to learn and adapt to an environment that will include a number of intelligent agents and hence maintain very strong defenses to malicious adversaries and enhance the efficiency of the computing environment. The next generation IDS systems will be highly involved in the security of the cyber-physical systems. This work makes comprehensive contributions to IDS research through several key aspects. This study reviewed the development from classic machine-learning classifiers to deep neural networks capable of learning complex patterns in traffic behavior. It also provided an evaluation of different federated learning setups for collaborative detection while preserving the privacy of sensitive data. It investigated adversarial weaknesses and integrated mitigation strategies to make intrusion-detection systems more robust against evasion and poisoning attempts, together with a review of lightweight design approaches to enable deployment on resource-limited devices such as IoT and edge sensors. Finally, this study emphasized how explainability techniques can help operators understand model decisions, building trust in automated systems.

Conflict of Interest

There is no conflict of interest.

Supporting Information

Not applicable

Use of artificial intelligence (AI)-assisted technology for manuscript preparation

The authors confirm that there was no use of artificial intelligence (AI)-assisted technology for assisting in the writing or editing of the manuscript and no images were manipulated using AI.

References

[1] J. P. Anderson, Computer security threat monitoring and surveillance, James P. Anderson Co., Washington, PA, April 1980.

[2] D. E. Denning, An intrusion-detection model, *IEEE Transactions on Software Engineering*, 1987, **SE-13**, 222-232, doi: 10.1109/TSE.1987.232894.

[3] A. Pinto, J. Smith, L. C. Herrera, Survey of intrusion detection systems based on machine learning, *Sensors*, 2023, **23**, 2415, doi: 10.3390/s23052415.

[4] S. Ennaji, M. Khalid, R. Benlamri, Adversarial challenges in network intrusion detection, arXiv preprint, arXiv:2409.18736, 2024, https://arxiv.org/abs/2409.18736.

[5] A. Sarıkaya, RAIDS: Robust autoencoder-based intrusion

detection system model against adversarial attacks, *Computers & Security*, 2023, **135**, 103483, doi: 10.1016/j.cose.2023.103483.
[6] J. L. Hernandez-Ramos, G. Karopoulos, E. Chatzoglou, V. Kouliaridis, E. Marmol, A. Gonzalez-Vidal, G. Kambourakis, Intrusion detection based on federated learning: a systematic review, ACM Computing Surveys, 2025, **57**, 309, doi: 10.1145/3731596
[7] M. Farhan, H. Waheed ud din, S. Ullah, M. S. Hussain, M. A. Khan, T. Mazhar, U. F. Khattak, I. H. Jaghdam Network-based intrusion detection using deep learning, *Scientific Reports*, 2025, **15**, 10001–10015, doi: 10.1038/s41598-025-08770-0.
[8] Z. Wang, R. Zhou, S. Yang, D. He, S. Chan, A novel lightweight IOT intrusion detection model based on self-knowledge distillation, *IEEE Internet of Things Journal*, 2025, 12, 16912-16930, doi: 10.1109/JIOT.2025.3533092.
[9] B. Buyuktanir, Ş. Altinkaya, G. K. Baydogmus, K. Yildiz, Federated learning in intrusion detection: Advancements, applications, and future directions, *Cluster Computing*, 2025, **28**, 1051–1072, doi: 10.1007/s10586-025-05325-w.
[10] T. Wisanwanichthan, M. Thammawichai, A Lightweight intrusion detection system for IoT and UAV using deep neural networks with knowledge distillation, *Computers*, 2025, **14**, 291, doi: 10.3390/computers14070291.
[11] X. Chen, M. Liu, Z. Wang, Y. Wang, Explainable deep learning-based feature selection and intrusion detection method on the internet of things, *Sensors*, 2024, **24**, 5223, doi: 10.3390/s24165223.
[12] M. Sajid, K. R. Malik, A. Almogren, T. S. Malik, A. H. Khan, J. Tanveer, A. Ur Rehman, Enhancing intrusion detection: a hybrid machine and deep learning approach, *Journal of Cloud Computing*, 2024, **13**, 45, doi: 10.1186/s13677-024-00685-x.
[13] F. Aloraini, A. Javed, O. Rana, Adversarial attacks on intrusion detection systems in in-vehicle networks of connected and autonomous vehicles, *Sensors*, 2024, **24**, 3848. https://doi.org/10.3390/s24123848.
[14] M. Devine, S. P. Ardakani, M. Al-Khafajiy, Y. James, Federated machine learning to enable intrusion detection systems in IOT networks, *Electronics*, 2025, **14**, 1176, doi: 10.3390/electronics14061176.
[15] A. Pavate, R. Bansode, Design and analysis of adversarial samples in safety–critical environment: disease prediction system. In: Gupta, M., Ghatak, S., Gupta, A., Mukherjee, A.L. (eds) Artificial Intelligence on Medical Data. Lecture Notes in Computational Vision and Biomechanics, 2023, **37**. Springer, Singapore, doi: 1007/978-981-19-0151-5_29.
[16] A. A. Pavate, R. Bansode, Generation of adversarial mechanisms in deep neural networks: A survey of the state of the art, *International Journal of Ambient Computing and Intelligence*, 2022, **13**, 1–18, doi: 10.4018/IJACI.293111.