information

MDPI

Systematic Review

Explainable AI-Based Intrusion Detection Systems for Industry 5.0 and Adversarial XAI: A Systematic Review

Naseem Khan [1], Kashif Ahmad [2], Aref Al Tamimi [3], Mohammed M. Alani [4], Amine Bermak [1,*] and Issa Khalil [3,*]

[1] Computer Science and Engineering Department, Hamad Bin Khalifa University, Ar-Rayyan 34110, Qatar; nakh12498@hbku.edu.qa
[2] Department of Computer Science, Munster Technological University Cork, T12 P928 Cork, Ireland; kashif.ahmad@mtu.ie
[3] Qatar Computing Research Institute (QCRI), Hamad Bin Khalifa University, Doha P.O. Box 5825, Qatar; altamimi@hbku.edu.qa
[4] Department of Electrical Engineering and Computing Sciences, Rochester Institute of Technology (RIT-Dubai), Dubai P.O. Box 341055, United Arab Emirates; m@alani.me
* Correspondence: abermak@hbku.edu.qa (A.B.); ikhalil@hbku.edu.qa (I.K.)

Abstract

Industry 5.0 represents a paradigm shift toward human–AI collaboration in manufacturing, incorporating unprecedented volumes of robots, Internet of Things (IoT) devices, Augmented/Virtual Reality (AR/VR) systems, and smart devices. This extensive interconnectivity introduces significant cybersecurity vulnerabilities. While AI has proven effective for cybersecurity applications, including intrusion detection, malware identification, and phishing prevention, cybersecurity professionals have shown reluctance toward adopting black-box machine learning solutions due to their opacity. This hesitation has accelerated the development of explainable artificial intelligence (XAI) techniques that provide transparency into AI decision-making processes. This systematic review examines XAI-based intrusion detection systems (IDSs) for Industry 5.0 environments. We analyze how explainability impacts cybersecurity through the critical lens of adversarial XAI (Adv-XIDS) approaches. Our comprehensive analysis of 135 studies investigates XAI's influence on both advanced deep learning and traditional shallow architectures for intrusion detection. We identify key challenges, opportunities, and research directions for implementing trustworthy XAI-based cybersecurity solutions in high-stakes Industry 5.0 applications. This rigorous analysis establishes a foundational framework to guide future research in this rapidly evolving domain.

Keywords: artificial intelligence; explainability; XAI; Industry 5.0; cybersecurity; intrusion detection systems; X-IDS; adversarial XAI

Academic Editor: Leandros Maglaras

Received: 31 August 2025
Revised: 3 November 2025
Accepted: 10 November 2025
Published: 27 November 2025

Citation: Khan, N. ; Ahmad, K.; Al Tamimi, A.; Alani, M.M.; Bermak, A.; Khalil, I. Explainable AI-Based Intrusion Detection Systems for Industry 5.0 and Adversarial XAI: A Systematic Review. *Information* **2025**, *16*, 1036. https://doi.org/10.3390/info16121036

1. Introduction

The integration of artificial intelligence and machine learning (ML) in smart industries, particularly in Industry 5.0 applications, has created an urgent need for transparent and interpretable AI-based solutions [1–3]. Like other sensitive application domains such as business, healthcare, education, and defense systems, the opaque nature of AI models raises significant concerns regarding decision-making transparency in smart industries. Beyond user acceptance and technology adoption issues, system developers must ensure fair and unbiased AI solutions. This need to understand causal reasoning in Deep ML model inferences has directed research attention toward explainable AI (XAI) [4]. The

DARPA-funded XAI initiative aimed to develop interpretable machine learning models for reliable, human-trusted decision-making systems, crucial to IoT and intelligent system integration in Industry 5.0 [5–8].

Cybersecurity represents a critical challenge in smart industries with extensive interconnected devices. While AI-based solutions have proven effective for cybersecurity applications, the opacity of complex AI models in cybersecurity solutions—including intrusion detection systems (IDSs), intrusion prevention systems (IPSs), malware detection, zero-day vulnerability discovery, and Digital Forensics—exacerbates transparency and trust issues [7,9]. XAI can address these concerns by demonstrating AI algorithm trustworthiness and transparency in critical cybersecurity applications. Security analysts need to understand internal decision mechanisms of deployed intelligent models and precisely reason about input–output relationships to stay ahead of attackers. XAI-derived insights could enhance cybersecurity solutions through human–AI collaboration, improving development, training, deployment, and debugging processes. However, XAI application in cybersecurity presents a double-edged sword challenge—while improving security practices, it simultaneously makes explainable models vulnerable to adversarial attacks [10–12].

Industry 5.0's increased IoT device reliance makes systems more vulnerable to cyber threats, potentially causing serious damage and financial losses. While compromised smart home device security creates privacy and security risks, failures in critical infrastructure such as smart grids, nuclear power plants, or water treatment facilities elevate risks significantly [13]. To address rising cybersecurity challenges in evolving smart cities and industries, various advanced security measures have been implemented, including Security Information and Event Management (SIEM) systems, vulnerability assessment solutions, IDSs, and user behavior analytics [14–16]. This study evaluates explainable IDS security measure advancements and highlights remaining challenges.

ML and deep learning (DL) algorithm adaptation in IDSs has introduced intelligent IDSs that significantly optimize detection rates. These ML/DL-based IDSs are adopted because they demonstrate superior robustness, accuracy, and extensibility compared with traditional detection techniques like rule-based, signature-based, and anomaly-based detection [17,18]. These complex algorithms' foundation lies in mathematical and statistical concepts that perform pattern discovery, correlation analysis, and structured data disparity representation through probabilities and confidence intervals [19,20]. ML primary types include supervised, semi-supervised, unsupervised, reinforcement, and active learning techniques, each serving specific security application purposes [21]. Despite the intelligent AI-based module's effectiveness, opaque/black-box model transparency and prediction justification remain uncertain. This lack of insights into opaque AI model decision-making systems raises trust issues for Industry 5.0 adoption [22].

Several critical research questions require investigation: What cybersecurity challenges emerge from AI and IoT device integration in industrial applications? How trustworthy and transparent are AI solutions in decisions making? How can trustworthy, transparent AI-based security and privacy solutions be developed for Industry 5.0? How vulnerable to adversarial attacks are AI-based solutions? How can explainability mechanisms be defined in IDSs to effectively interpret temporal and contextual dependencies for specific cyber threats?

1.1. Industry 5.0: Characteristics and AI Integration Context

Industry 5.0 represents the next industrial paradigm evolution, emphasizing human-centric approaches that integrate artificial intelligence, robotics, and Internet of Things (IoT) technologies to create collaborative human–machine environments [23]. Unlike Industry 4.0's automation and digitization focus, Industry 5.0 prioritizes sustainability, resilience,

and human centricity while leveraging advanced AI capabilities for enhanced decision making and operational efficiency.

This paradigm shift manifests through several interconnected technological pillars (Figure 1): digital factories with cyber–physical systems, AR/VR interfaces for human–machine collaboration, edge computing with 5G/6G networks enabling real-time processing, and collaborative robotics working alongside human operators. These technologies enable mass customization, adaptive manufacturing, and AI-driven optimization [24]. However, this convergence expands the attack surface, introducing critical cybersecurity vulnerabilities across multiple vectors.

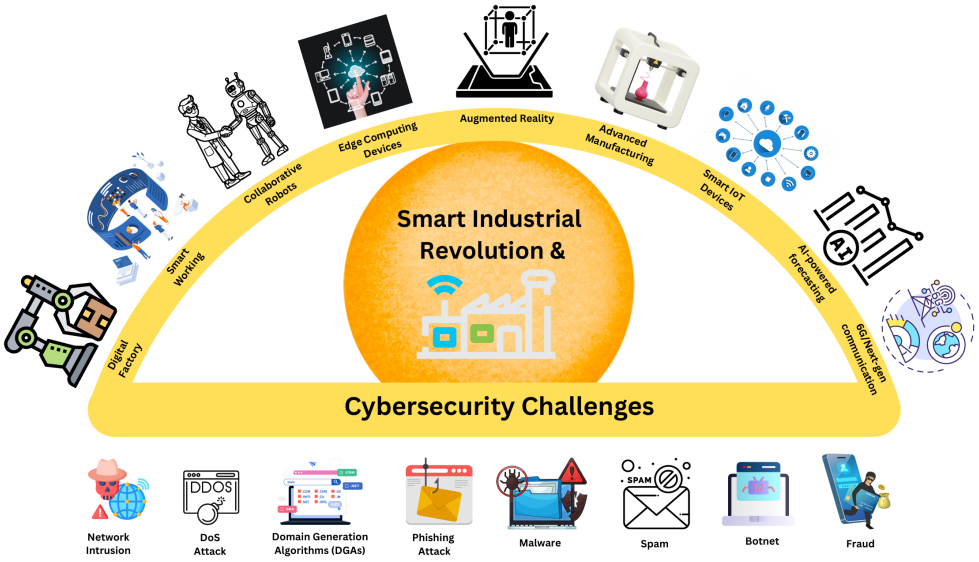

Figure 1. Industry 5.0 technological landscape and cybersecurity threat vectors. The upper arc illustrates enabling technologies (digital factories, edge computing, AR/VR, collaborative robotics, 5G/6G networks, smart IoT, etc.) defining Industry 5.0's human-centric paradigm. The lower section depicts major cybersecurity threats (network intrusion, DDoS, phishing, malware, botnets, ransomware, etc.) exploiting the expanded attack surface from extensive interconnectivity.

The cybersecurity threat landscape in Industry 5.0 encompasses both traditional and emerging attack paradigms (Figure 1). Network intrusion attempts target the interconnected IoT infrastructure, while Distributed Denial-of-Service (DDoS) attacks threaten operational continuity. Phishing and social engineering exploit the human element in collaborative environments, whereas malware, spam, and botnet infiltrations compromise system integrity. Advanced persistent threats leverage supply chain vulnerabilities inherent in multi-stakeholder industrial ecosystems. These threats are amplified by Industry 5.0's characteristics: extensive device interconnectivity creates cascading failure risks, reduced human oversight in automated processes limits real-time threat detection, and the integration of AR/VR systems introduces novel attack vectors through immersive interfaces.

This threat landscape creates unprecedented requirements for AI transparency and explainability in cybersecurity solutions. Human operators must understand, trust, and effectively collaborate with intelligent security systems in critical industrial processes, necessitating explainable AI approaches that balance transparency with adversarial robustness. The extensive AI and machine learning integration in Industry 5.0 environments introduces a fundamental tension: while explainability mechanisms enable human–AI collaboration and trust building essential to security operations, they simultaneously provide adversaries with insights into model decision making that can be exploited for sophisticated attacks. Understanding this dual nature of explainability—as both security enabler and potential

vulnerability—is critical to developing trustworthy cybersecurity solutions in Industry 5.0 contexts, motivating the systematic analysis presented in this survey.

1.2. Scope of the Survey

This survey focuses on network-based intrusion detection problems within Industry 5.0 and how XAI can provide a deeper understanding of machine learning-based intrusion detection systems to mitigate associated cybersecurity risks. The paper provides a taxonomy of XAI methods, discussing each XAI approach's advantages and disadvantages. This paper examines XAI pitfalls in cybersecurity and how attackers could exploit XAI method explanations to exploit AI-based IDS vulnerabilities.

1.3. Related Surveys

Industrial paradigm evolution has introduced transformative goals emphasizing resource-efficient and intelligent society creation. This trajectory seeks to elevate living standards and mitigate economic disparities through hyperconnected, automated, data-driven industrial ecosystem integration [23,25,26]. This digital transformation promises significant productivity and efficiency enhancement across the entire production processes. These milestones become possible through AI/Generative AI integration as collaborative landscapes, fostering innovation, optimizing resource utilization, and driving economic growth in smart industries [27]. This accompanies huge smart critical infrastructure developments such as smart grids and IoT-controlled dams. However, such advancements expose systems to elevated sophisticated cyber attack risks [28–30].

Connected devices and networks in autonomous industry infrastructure are more prone to hijacking, malfunctioning, and resource misuse threats due to expansive attack surfaces from pervasive connectivity, reduced human oversight in automated processes, and high-value data assets targeted by adversaries. These vulnerabilities necessitate additional security layers for risk protection [31]. Conventional AI-based cybersecurity systems remain under active development for full maturity achievement, and robust, trustworthy security system establishment has emerged as a prominent defender objective [32,33].

As shown in Table 1, while existing research has explored either XAI taxonomies [34–36] or cybersecurity applications [23,25,26,37], understanding gaps remain regarding explainability impacts on adversarial contexts in intrusion detection systems. Explainability mechanism adoption in cybersecurity, specifically in intrusion detection and prevention systems, is reviewed in surveys by Chandre et al. [38] and Moustafa et al. [39]. Recent research has focused on autonomous transportation, smart cities, and energy management systems [40–43]. Table 1 demonstrates that existing surveys address these topics in isolation: studies examining Industry 5.0 [23,25,26] provide general technological overviews without analyzing adversarial robustness of explainability mechanisms, while XAI-focused surveys [34,36] lack application to cybersecurity contexts. Multi-domain surveys [37] cover cybersecurity broadly but without depth in adversarial XAI exploitation or Industry 5.0-specific human-centric security requirements. Additionally, XAI has been used to exploit ML intelligence after gaining model insights. Our work uniquely addresses XAI, cybersecurity, and adversarial approach intersections, examining explainability concept impacts on cybersecurity practices, emphasizing the emerging adversarial explainable IDS (Adv-XIDS) trend, which represents significant challenges for explainable AI-based cybersecurity decision models.

Table 1. Comparison of related surveys on explainable AI and cybersecurity.

Ref.	XAI	Cyber	Adv-XAI	I5.0	Focus and Key Limitation
[23,25,26]	✓	✓	✗	●	XAI role in Industry 5.0 digital transformation (e.g., AI, IoT, robotics, and cybersecurity). *Limitation:* General overview without adversarial XAI analysis or human-centric security workflows.
[37]	✓	✓	✗	✗	XAI applications across multiple domains (e.g., healthcare, finance, cybersecurity, and education). *Limitation:* Broad scope without depth in adversarial robustness or Industry 5.0 context.
[34–36]	✓	✗	✗	✗	XAI taxonomies, counterfactual explanations, and technique surveys. *Limitation:* General XAI methods without cybersecurity or IDS application.
[35]	✓	✗	✗	✗	Theoretical framework for XAI explanation design using counterfactuals. *Limitation:* No empirical application to cybersecurity or industrial contexts.
Ours	✓	✓	✓	✓	**First systematic review integrating** (1) XAI-based IDS taxonomy, (2) Industry 5.0 human-centric cybersecurity threats, (3) adversarial XAI exploitation (Adv-XIDS), (4) legacy vs. modern dataset analysis, and (5) federated learning and SOC workflow considerations for collaborative industrial environments.

Legend: ✓ = addressed; ✗ = not addressed; ● = partially addressed; I5.0 = Industry 5.0; Adv-XAI = adversarial XAI; Cyber = cybersecurity.

1.4. Contributions

Based on serious threat vectors and their implications, this paper analyzes different instances of XAI method adoption in IDSs and examines interpretability impacts on cybersecurity practices in Industry 5.0 applications. We provide a comprehensive literature overview on XAI-based cybersecurity solutions for Industry 5.0 applications, focusing on existing solutions, associated challenges, and future research directions for challenge mitigation. For self-containment, we provide an XAI taxonomy overview.

This systematic review distinguishes itself through three key aspects. First, we analyze cybersecurity threats within Industry 5.0's human-centric paradigm, examining vulnerabilities in human–robot collaboration, AR/VR manufacturing, and federated edge architectures. These contexts require explainability for real-time human intervention and cross-organizational threat sharing. Second, we document explainability's dual nature: as an IDS enhancement enabling transparency (Section 5, Table 2) and as an adversarial attack vector exploitable through SHAP, LIME, and gradient-based methods (Section 6, Table 3). Third, our PRISMA-guided analysis of 135 studies quantifies XAI technique distributions, dataset prevalence, and vulnerability–protection mappings (Table 4).

Table 2. Explainable AI-based intrusion detection systems.

Ref.	Threat Addressed	IDS Data Type	Dataset	Detection Model	XAI Algorithm	Explanation
[44]	Intrusion detection	Network-based IDS (NIDS)	Analyst-designed training sets from archived network events	Genetic algorithm, ID3	Rule-based explainability	Generates rules for distinguishing normal network connections from an anomalous one, based on expert-domain knowledge.
[45]	DOS, R2L, U2R, and PROBING	Network-based IDS (NIDS)	KDD-99	ID3 algorithm	Rule-based explainability	Generates rules by Rattle package in R and visualizing in exploratory plots.
[46,47]	DOS, R2L, U2R, and PROBING	Network-based IDS (NIDS)	SCADA VM, and N-BaIoT	DT, RF, GNB, SVM, and K-NN	Rule-based explainability	Generates rules by Tree nodes to visualize the decision-making process as an exploratory plot.
[48]	DOS, R2L, U2R, and PROBING	Host-based IDS (NIDS)	Tracer FIRE 9 (TF9) and Tracer FIRE 10 (TF10)	Bayesian network (BNs)	Rule-based explainability	Visualizes network graph of the Bayes' Rule, where the relation of a single feature to the target variable is found via conditional probability tables (CPTs).
[49]	Industrial IoT Security	IIoT-based IDS	WUSTL-IIoT, NSL-KDD, and UNSW-NB15	Artificial Neural Network (ANN)	Transparency Relying Upon Statistical Theory (TRUST) system	Employs mutual information for ranking variables and selects the most impactful ones on the ANN's outputs, naming them as representatives of the classes.
[50]	Industrial IoT Security	IoT-based IDS	IoTID20, NF-BoT-IoT-v2, and NF-ToN-IoT-v2	Ensemble Trees (DT and RF)	SHapley Additive exPlanations (SHAP)	The ensemble model's outputs are plotted in the form of heatmaps and decision plots using SHAP explanation techniques.
[51]	Android devices Malware detection	Android application-based IDS	MalMem-2022, Drebin-215, Malgenome-215, and CICMalDroid2020	RF, LR, DT, GNB, and XGB	SHAP	SHAP values are employed to interpret model predictions by highlighting the most influential features contributing to malware detection decisions.
[52,53]	Industrial IoT Security	IoT-based IDS	Aposemat IoT-23, IoTID20, NF-BoT-IoT-v2, NF-ToN-IoT-v2, and WUSTL-IIOT-2021	RF, LR, DT, GNB, XGB, and SVM	SHAP	The study uses SHAP to provide a global interpretation of model behavior, identifying critical IoT traffic features influencing intrusion detection outcomes.
[54]	Malicious Traffic Detection in IoT Healthcare Networks	IoT-based IDS)	Intensive Care Unit (ICU) dataset	RF and DT	SHAP, LIME, ELI5, and Integrated Gradients (IGs)	Visualizes the contribution of each feature in the model's decision using Shapash Monitor explanation interface.
[55]	Man-In-The-Middle (MITM), DoS, Mirai botnet, Port/OS scanning, and Host scanning	IoT-based IDS	CICIDS-2017	Voting Classifier	Local Interpretable Model-agnostic Explanation (LIME)	Plots the contribution of each feature in the model's decision using LIME.

Table 2. *Cont.*

Ref.	Threat Addressed	IDS Data Type	Dataset	Detection Model	XAI Algorithm	Explanation
[56]	DNS over HTTPS (DoH) attacks	Network-based IDS	CIRA-CIC-DoHBrw-2020	Random Forest (RF)	SHAP	Highlights the features which are contributing to the underlying decision of the model using SHAP values.
[57]	Glastopt, Dionaea, Cowrie, Canarytokens, DoS, R2L, U2R, and Probe attacks	Network-based IDS (NIDS)	Honeypot and NSL-KDD datasets	Bidirectional Long Short-Term Memory (BiLSTM)	LIME and SHAP	Focuses on generating global and local faithful explanations by approximating the behavior of the BiLSTM model around a specific instance of interest.
[58]	DOS, R2L, U2R, and PROBING	Network-based IDS (NIDS)	KDD-99	CNN-LSTM	LIME and SHAP	LIME mechanism enables the model to interpret each individual factor and their impact on output. A Decision Tree is generated from the top-most influential features, which are then visualized using SHAP interpretation.
[59]	DOS, R2L, U2R, and PROBING	Network-based IDS (NIDS)	Ton-IOT Windows	RF	LIME and SHAP	Employed three primary techniques—variable importance plot, individual value plot, and partial dependence plot—to explain the decision-making process of the RF model.
[60]	Malware detection	File Content Analysis	VX Vault- and Virus Share-based generated dataset	Random Forest classifier	Visualizing Decision Trees	Presents the Trees that had classified a process as malware or benign and the relevant decision nodes.
[61,62]	Web based, Brute Force, DoS, DDoS, Infiltration, Heartbleed, and Bot and Scan	Host-based and network-based IDSs	NSL-KDD and CIC-IDS-2017	Population-based Self-Organizing Maps (POPSOM) implementation	Self-Organizing Map (SOM)-based X-IDS	Produces robust, explanatory visualizations of the SOM model and create accurate IDS predictions
[63]	DoS and ID fabrication	In-vehicle IDS (IV-IDS)	Survival Analysis Dataset for automobile IDS	Deep Neural Network (DNN)	Visualization-based Explanation, (VisExp)	A dual swarm plot is created to display normal Controller Area Network (CAN) traffic at the top and intruder's traffic at the bottom based on SHAP-value distribution.
[64]	Adware, Banking malware, SMS malware, Riskware, Brute Force FTP, and DoS	File Content Analysis and Network-based IDS	MalDroid20, CIC-IDS2017	Deep Neural Network (DNN)	DALEX framework	DALEX employs a permutation-based algorithm to find the significance of individual variables, enhancing DNN prediction performance.
[65]	Adware, Banking malware, SMS malware, Riskware, Brute Force FTP, DoS	File Content Analysis and Network-based IDS	MalDroid20 and CIC-IDS2017	Deep Neural Network (DNN)	SHAP	Fine-tunes DNN prediction performance through adversarial training and XAI combination.
[66]	Denial-of-Service (DoS) and Probe attack types	Network-based IDS (NIDS)	NSL-KDD	Random Forest (RF)	SHAP	Utilizes SHAP beeswarm plots to visualize explanations of the target class individually.

Table 2. *Cont.*

Ref.	Threat Addressed	IDS Data Type	Dataset	Detection Model	XAI Algorithm	Explanation
[67]	IoT Network Security	IoT-based IDS	NSL-KDD and UNSW-NB15	DNN and CNN	LIME and SHAP	A deep learning-based IDS employing DNN and CNN models for attack classification, with feature selection using a filter-based approach. Model explanations are generated using LIME for local interpretability and SHAP for global feature importance.
[68]	Android malware detection	File Content Analysis	DREBIN	SVM and BERT	Feature importance	Inspired by MPT, minimizes variance in prediction score changes and attribution values for impactful feature attribution.
[69]	Brute Force, Bot, DoS, DDoS, Infiltration, and Web attacks	Network-based IDS (NIDS)	CSE-CIC-IDS2018, ToN-IoT, and Bot-IoT	Multi-Layer Perceptron (MLP) and Random Forest (RF)	SHAP	Calculates Shapley values to assess feature contributions and identify key influencers in the dataset.
[70]	DoS and DDoS in IoT/IoV networks	Network-based IDS (NIDS)	ToN_IoT dataset	Deep Neural Network	Deep SHAP technique	Combines SHAP values from neural network parts via DeepLIFT's multipliers for full network interpretation.
[71]	Command injection, DoS, Reconnaissance, and backdoors	Industrial IoT-based IDS (IIoT-IDS)	WUSTL-IIOT-2021	Deep Neural Network (DNN)	SHAP	Uses DeepExplainer to provide insights into DeepIIoT's decision making via SHAP values.
[72]	DDoS attacks on IoT and traditional networks	IoT-based IDS (IoT-IDS)	USB-IDS dataset	Fully connected autoencoder with RELU	Kernel SHAP	Identifies top-R features contributing to reconstruction errors using SHAP values.
[73]	Industrial IoT Security	IoT-based IDS (IoT-IDS)	IoTID20 dataset	XG-Boost	LIME, TreeSHAP, and ELI5	LIME explains contributions, SHAP combines importance and effects, and ELI5 reveals weights.
[74]	IoT-network security	IoT-based IDS (IoT-IDS)	UNSW-NB15	DNN	RuleFit and SHAP	Calculates feature importance values for the decision model.
[10,75]	DOS	Network-based IDS (NIDS)	NSL-KDD99, LYCOS-IDS2017	Linear Model (LM) and Multi-Layer Perceptron (MLP)	SHAP and Adversarial ML	Generates visual explanations for misclassifications, identifying responsible features.
[76]	DDoS, XSS, and SQL Injection attacks	Anomaly-based IDS (AIDS)	CICIDS2017	ANN with PCA	Decision Trees with microaggregation	Uses dtreeviz to plot tree structure and highlight key features in predictions.
[77]	DOS, R2L, and U2R	Network-based IDS (NIDS)	NSL-KDD	One-versus-all classifier and multiclass classifier	SHAP	Combines local and global explanations to enhance IDS interpretation.
[78]	DOS, R2L, and U2R	Network-based IDS (NIDS)	KDD99 and CICIDS2017	CNN and DT	SHAP	Combines local and global explanations to improve IDS interpretation.
[79]	IoT-network security	Network-based IDS (NIDS)	CIC-IoT-Dataset-2022	DNN	SHAP	Combines local and global explanations to improve IDS interpretation.

Table 2. *Cont.*

Ref.	Threat Addressed	IDS Data Type	Dataset	Detection Model	XAI Algorithm	Explanation
[80,81]	DDoS, XSS, and SQL Injection attacks	Network-based IDS (NIDS)	KDD99 and CICIDS2017	Deep Neural Network (DNN) and ensemble models	SHAP and LIME	Generates model-centric and subject-centric explanations from DNN predictions.
[82]	Anomaly detection	Network-based IDS (NIDS)	NSL-KDD	Deep Neural Network (DNN)	SHAP, BRCG, LIME, ProtoDash, and CEM	Plots SHAP values, extracts rules with BRCG, generates local explanations with LIME, summarizes data with ProtoDash, and calculates minimal perturbations with CEM.
[83]	Data injection and poisoning in Industrial IoT	Anomaly-based IDS (AIDS)	Real-world GSP time-series data	Conv-LSTM-based autoencoder	LIME	Illustrates relevant attributes and weights for interpretation.
[84]	Industrial control system anomaly detection	Anomaly-based IDS (AIDS)	SCADA dataset	LSTM-based autoencoder	SHAP	Visualizes feature influence on model output globally.
[85]	Low-rate DoS, Port scanning, Botnet, Spam, and Blacklist	Cyclostationarity-based network IDS (NIDS)	UGR'16 dataset	Variational autoencoder (VAE) framework	Gradient-based explanation	The interpretability of variational autoencoders is generated by utilizing gradients for clustering anomalies and deriving attack-related fingerprints.
[86]	Anomaly detection	Anomaly-based IDS (AIDS)	Warranty claims, KDD Cup 1999, Credit Card Fraud Detection, and artificial dataset	Autoencoder framework	Kernel SHAP	Computes SHAP values for reconstructed features and links them to true anomalous input values to explain prediction errors.
[87]	Anomaly detection	Anomaly-based IDS (AIDS)	UCI Machine Learning Repository	Decision Tree-based autoencoder	Rule-based explainability	The correlation values among different categorical attributes provide explanations behind the Decision Tree.
[88]	DDoS, XSS and SQL Injection attacks	Network-based IDS (NIDS)	CICIDS2017	Sec2Graph technique	Explanation based on AE-pvalues	Explanation about the anomaly alert is produced by using the *p*-value of the empirical distribution of the dimension-wise reconstruction error to flag abnormal feature values.
[89]	DDoS	Network-based IDS (NIDS)	CICIDDoS2019	BiLSTM + BiGRU + CNN	SHAP	Uses SHAP decision graphs including decision plots, Waterfall Plots, and Summary Plots to demonstrate the important features that contributed the most to detection.

Table 2. *Cont.*

Ref.	Threat Addressed	IDS Data Type	Dataset	Detection Model	XAI Algorithm	Explanation
[55]	Network Intrusion Detection	ML-based IDS	CICIDS-2017	DT, RF, SVM, and Voting Classifier	LIME	An ensemble-based IDS combining Decision Tree, Random Forest, and SVM models with a Voting Classifier to enhance detection accuracy and reduce false positives. LIME is applied to interpret model predictions and enhance trust in the black-box ensemble system.
[90]	Mirai/Gafgyt botnets, DoS/DDoS, SQL Injection, and backdoors	Network-based IDS (NIDS)	N-BaIoT, Edge-IIoTset, and CIC-IDS2017	LSTM-AutoEncoder for encoding; Attention-based GRU with softmax for multiclass classification	SHAP	Feature attribution scores for each prediction; highlights key traffic features responsible for malicious activity classification, improving SOC trust and traceability.
[91]	Malware detection	File Content Analysis	Malimg dataset	Selective Deep Ensemble Learning-based (SDEL) detector	Ensemble Deep Taylor Decomposition (EDTD)	EDTD converts the SDEL prediction into a heatmap, where brighter pixels indicate the most suspicious parts in the malware binary image.
[92]	Mobile malware detection	File Content Analysis	Android Malware Dataset (Argus Lab)	Convolutional Neural Network (CNN)	Grad-CAM	Generates heatmap for visualizing the predictions made by image-based CNN mode.
[93]	DoS, Probe, R2L, U2R, Fuzzers, Analysis, backdoors, Exploits, Generic, Reconnaissance, Shellcode, and worms	Network-based IDS (NIDS)	NSL-KDD and UNSW-NB15	Convolutional Neural Network (CNN)	Attention mechanism of ROULETTE	Explainability involves utilizing the attention weights generated by the neural model to provide insights into the classification decisions made by the model for network traffic data.

Table 3. Adversarial techniques targeting intrusion detection systems in Industry 5.0: Compilation of methods exploiting IDS vulnerabilities, with and without leveraging explainable AI mechanisms, in the context of cybersecurity.

Ref.	Data Type	Dataset	Attack Type	Detection Model	XAI Targeted/No XAI
[94]	Network event logs	IEEE BigData 2019 Cup: Suspicious Network Event Recognition	Perturbation	GAN	✗
[95]	Network-based	CICIDS 2017 and TRAbID 2017	Perturbation	MLP	✗
[96]	Network-based	CICIDS2017	Evasion	DT and LR	✗
[97]	Network-based	CICIDS2018 and InSDN	Evasion	DT, LR, CNN, MLP, and LSTM	✗
[98]	Host-based, network-based, and application-based	ADFA-LD, NSL-KDD, and DREBIN	Perturbation	DT, LR, MLP, NB, and RF	✗
[99]	Android APKs	40K samples (20,769 benign from Google Play)	Label Spoofing	LSVM, GBT, NN, and RF	✗
[100]	IoT network-based	Mirai, Falsifying Video streaming application	Perturbation	DNN-based autoencoder	✗
[101]	Network-based	NSL-KDD, UNSW-NB15, and CICIDS2017	Evasion	Autoencoder	✗
[102]	IoT Network-based	MedBIoT and IoTID	Perturbation	LSTM and RNN	✗
[103]	Network-based	KDDCup'99	Evasion	DNN	✗
[104]	IoT network-based	X-IIoTID	Evasion	SVM, DT, RF, KNN, CNN, GRU, and HyDL-IDS	✗
[105]	Network-based	CTU-13 and CSE-CIC-IS2018D	Evasion	MLP, RF, and KNN	✗
[106]	Host-based	KDDCUP99, NSL-KDD, and Kyoto 2006+	Poisoning	NB-Gaussian, LR, and SVM-sigmoid	✗
[107]	Network-based	LANL network security dataset	Poisoning	LSTM, B-LSTM, and T-LSTM	✗
[108]	Network-based	D¨IoT-Benign, UNSW-Benign, and D¨IoT-Attack	Poisoning	Federated Learning-based DNN	✗
[109]	Portable Executable (PE Files)	EMBER	Adversarial examples	GBDT	Integrated Gradients, DeepLIFT, and Layer-wise Relevance Propagation (LRP)
[110]	Network-based IDS	CIC-IDS2017 and Kitsune	Adversarial examples	MLP, AlertNet, IDSNet, DeepNet, RF, Xgboost, Multi-attribute Markov Probability Fingerprints (MaMPF), Flow Sequence Network (FS-Net), KitNET, and Diff-RF	SAGE (Shapley Additive Global Explanation)
[111]	Network-based and PE Files	Malicious/benign PDF files, Android apps, and UGR16	Perturbation	MLP and adversarial autoencoder	Gradient-based XAI
[112]	Network-based and PE Files	Leaked Password, CICIDS17, and VirusShare	Evasion, oracle, and poisoning	Autoencoder and Gradient Boosting Model (GBM), Neural Network (NN)	Latent counterfactual, permute attack, and diverse counterfactual
[109]	PE Files	Ember (1 M samples)	Feature modification	GBDT	Integrated Gradients, DeepLIFT, ϵ-LRP, and SHAP

Table 3. *Cont.*

Ref.	Data Type	Dataset	Attack Type	Detection Model	XAI Targeted/No XAI
[113]	PE Files	EMBER, Contagio (PDFs), and Drebin (Android executables)	Evasion, oracle, and poisoning	Autoencoder, Gradient Boosting Model (GBM), and Neural Network (NN)	Latent counterfactual, permute attack, and diverse counterfactual
[114]	Network-based	CIC-IDS2017 and TON_IoT	Perturbation	DNN (Feedforward NN)	Integrated Gradients and KernelSHAP
[115]	PE Files	Microsoft Malware classification Challenge	Evasion	Deep Neural Network (DNN)	Superpixels
[116]	Network-based	InSDN	SHAP-guided evasion with AMM	LightGBM, RF, and CNN	SHAP
[117]	Network-based	IoT network intrusion dataset	Evasion	Extreme Gradient Boosting (XGB)	SHAP
[118]	PE Files	Drebin (Android executables)	Evasion	Random Forest (RF) and Multi-Layer Perceptron (MLP)	LIME

Table 4. Comprehensive summary of XAI and adversarial XAI approaches in Industry 5.0 cybersecurity. See Sections 5 and 6 for detailed XAI technique descriptions and foundational references.

XAI Technique	Type	Model	Primary Applications	Vuln.	Documented Exploitations	Protection Mechanisms
Rule-based Explanations (ID3, DT, and BN)	Global	Inherently interpretable models	Policy compliance, expert knowledge integration, and audit trails	Low	Rule extraction and logic manipulation	Rule obfuscation, ensemble rules, and expert validation
SHAP (all variants)	Local + global	Model-agnostic	Feature importance, attack attribution, and pattern analysis	Very high	Feature manipulation, transferability attacks, and SAGE exploitation	Explanation randomization, input validation, and ensemble explanations
LIME	Local	Model-agnostic	Individual alert validation, incident forensics, and false-positive analysis	High	Local perturbation, feature evasion, and Android malware evasion	Instance validation, surrogate diversity, and perturbation bounds
Integrated Gradients	Local	Deep learning	Deep model interpretation, malware analysis, and healthcare IoT	Very high	Gradient manipulation, adversarial examples, and malware evasion	Gradient masking, defensive distillation, and input smoothing
Gradient-based methods	Local	Deep learning	Anomaly clustering, attack fingerprinting, and VAE interpretation	High	Gradient-based XAI exploitation and Manifold Manipulation	Gradient noise injection and multi-path gradients
Feature importance (Gini and permutation)	Global	Tree-based and ensemble	Model debugging, system optimization, and strategic threat analysis	Medium	Importance ranking manipulation and feature masking	Permutation testing, cross-validation, and feature redundancy
Attention mechanisms	Local + global	Neural Networks	Sequence analysis, traffic classification, and multiclass detection	Medium	Attention weight manipulation focus redirection	Attention regularization, multi-head validation, and attention dropout
Visualization techniques (Grad-CAM and Saliency)	Local	CNN and image-based	Malware visualization, binary analysis, and spatial feature mapping	Medium	Visual manipulation, Superpixels exploitation, and heatmap distortion	Multi-view validation, statistical verification, and ensemble visualization
Autoencoder-based (AE-pvalues and reconstruction)	Local + global	Autoencoders	Anomaly detection, reconstruction error analysis, and Industrial IoT	High	Latent space manipulation, reconstruction evasion, and counterfactual attacks	Latent space regularization, ensemble autoencoders, and adversarial training
Hybrid techniques (TRUST, DALEX, and RuleFit)	Global	Model-specific	Statistical analysis, Industrial IoT Security, and multi-modal explanation	Low–medium	Statistical manipulation and component-wise attacks	Statistical robustness testing, component isolation, and hybrid validation
Advanced methods (CEM, ProtoDash, and BRCG)	Local	Model-agnostic	Contrastive analysis, data summarization, and rule generation	Medium	Contrastive manipulation and summary poisoning	Robustness constraints, multi-method validation, and outlier detection

The main contributions of this paper are summarized as follows:

- We provide a clear and comprehensive taxonomy of XAI systems with categorization of ante hoc and post hoc methods, analyzing their applicability and limitations in cybersecurity contexts.
- We provide a detailed overview of current state-of-the-art IDSs, their limitations, and the deployment of XAI approaches in IDSs, systematically analyzing 135 empirical studies to identify implementation patterns, commonly used datasets, and explainability technique preferences.
- We systematically discuss the exploitation of XAI methods for launching more advanced adversarial attacks on IDSs, mapping specific XAI techniques to documented attack vectors and vulnerability levels.
- We analyze Industry 5.0-specific cybersecurity challenges and identify research directions for adversarially robust, human-centered explainable security systems, including federated learning architectures and SOC workflow integration.

This paper organization follows: Section 2 presents the survey methodology by describing objective questions. Section 3 provides an explainable AI taxonomy overview. Section 4 describes key Industry 5.0 cybersecurity challenges. Section 5 presents explainable AI in cybersecurity, specifically focusing on XAI-based IDSs. Section 6 presents adversarial XAI and IDS techniques. Section 7 discusses XAI-based IDS lessons learned, challenges, and future research directions. Finally, Section 8 concludes this survey.

2. Methodology

This systematic review implements a structured methodology to investigate explainable artificial intelligence (XAI)-based intrusion detection systems (IDSs) and adversarial approaches exploiting their explainability (Adv-XIDS) within Industry 5.0 environments. Conducting this review according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines [119], we employ a reproducible process for literature identification, screening, and analysis. The complete PRISMA 2020 checklist is provided in Supplementary Material, and the flow diagram illustrating study selection is presented in Figure 2. While the review protocol was developed prior to data collection (July 2024), it was not prospectively registered in PROSPERO—an acknowledged limitation. The methodological design minimizes selection bias and enhances finding reliability through systematic exploration of key research questions, rigorous search protocols, and precise selection criteria, establishing a foundation for understanding complex interactions among explainability mechanisms, security frameworks, and adversarial challenges in human-centric industrial systems.

2.1. Research Questions

We formulated five research questions to explore critical dimensions of cybersecurity, explainability, and adversarial robustness in Industry 5.0:

1. What are the key cybersecurity challenges in Industry 5.0, and why are explainable AI-based intrusion detection systems (X-IDSs) essential to addressing these threats?
2. What techniques and methods enhance transparency and interpretability in X-IDS implementations?
3. What are the primary challenges and limitations of X-IDS in cybersecurity applications?
4. What are the security implications of adversaries exploiting X-IDSs decision mechanisms, and how can these systems be protected against such attacks?
5. What are the emerging trends and future research directions for X-IDSs in Industry 5.0 contexts?

These questions serve as the analytical framework for subsequent methodological stages, ensuring focused examination aligned with the study's objectives while maintaining comprehensive coverage of explainable intrusion detection systems and their adversarial implications.

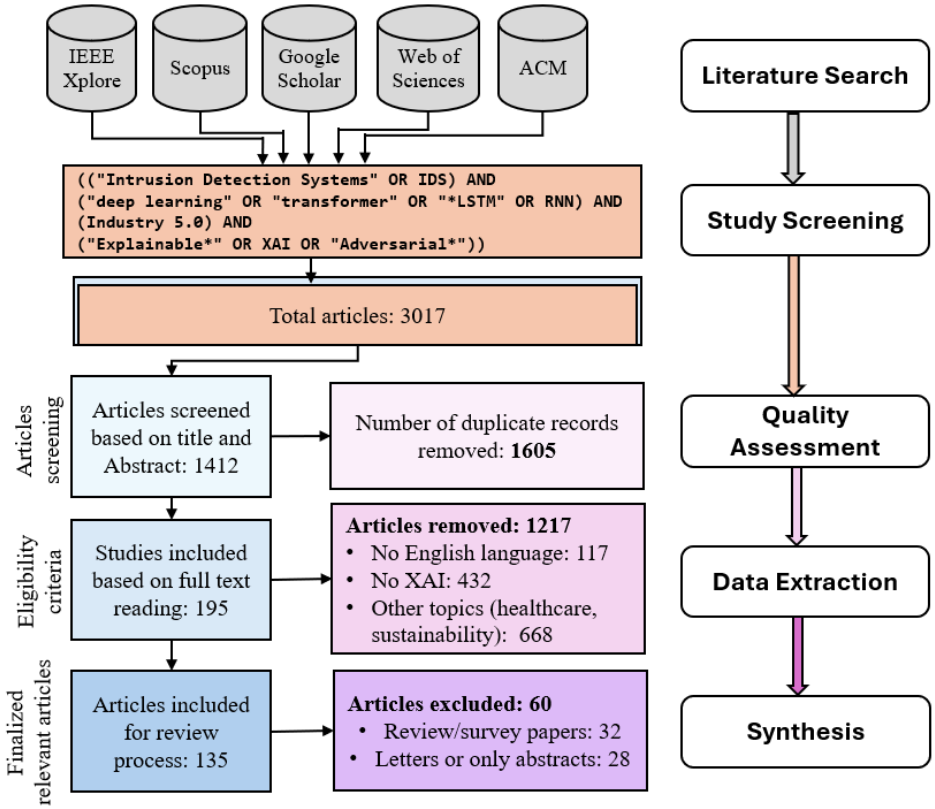

Figure 2. PRISMA flow diagram and systematic review methodology. Study selection process (**left**) and five-stage methodological framework (**right**) illustrating our systematic review approach from literature search to synthesis.

2.2. Search Strategy

We conducted a systematic literature search across five authoritative databases: IEEE Xplore, Scopus, Web of Science, ACM Digital Library, and Google Scholar. We selected these repositories for their comprehensive coverage of research in cybersecurity, artificial intelligence, and industrial systems, providing a robust foundation for identifying relevant studies on XAI-based IDSs and Adv-XIDSs in Industry 5.0.

The search strategy employed a structured terminology framework integrating intrusion detection terms ("Intrusion Detection System," "IDS," "Network Intrusion," "Anomaly Detection," "Threat Detection", etc.), explainability terms ("Explainable AI," "XAI," "Interpretable Machine Learning," "SHAP," "LIME," "Explainability," "Transparency", etc.), and industrial/adversarial context terms ("Industry 5.0," "Industry 4.0," "Industrial IoT," "IIoT," "Smart Manufacturing," "Adversarial Attack," "Adversarial XAI," "Adversarial Machine Learning", etc.). These terminology groups were combined using Boolean operators (AND, OR) with syntax adapted to each database's specific requirements. Representative search strings included combinations such as ("Intrusion Detection System" OR "IDS" OR "Network Intrusion") AND ("Explainable AI" OR "XAI" OR "Interpretable Machine Learning") AND ("Industry 5.0" OR "Industrial IoT" OR "Adversarial Attack"). Database-specific filters were applied, including document type restrictions (journal articles and conference papers), language requirements (English only), and temporal boundaries (January 2015 to October 2024). Searches were executed during August–September 2024, capturing both

foundational works and recent advancements in these rapidly evolving domains. The search protocol was developed through iterative refinement to optimize precision and recall, ensuring comprehensive coverage while maintaining relevance.

2.3. Study Selection

We followed the PRISMA guidelines for the selection process, as illustrated in Figure 2. The initial search yielded 3017 articles, which we reduced to 1412 unique records after eliminating 1605 duplicates through automated detection and manual verification. Independent reviewers screened titles and abstracts for relevance to XAI-based IDSs and Adv-XIDSs in Industry 5.0 contexts, with conflicts being resolved through structured consensus discussion. Figure 2 details this systematic screening process, showing the progressive elimination of studies at each stage.

Title and abstract screening excluded 1217 studies based on predetermined criteria: non-English publications (117 studies, 9.6%), studies lacking substantive focus on XAI methods or IDS applications (432 studies, 35.5%), and studies addressing tangentially related topics, including healthcare applications without industrial cybersecurity context, general sustainability discussions, or other computing domains with insufficient relevance to our research questions (668 studies, 54.9%).

Full-text assessment of the 195 remaining articles employed independent evaluation by multiple reviewers with consensus-based conflict resolution, resulting in the exclusion of 60 studies: review or survey papers without primary empirical contributions (32 studies, 53.3%) and publications with insufficient methodological detail such as conference abstracts or position papers lacking adequate description of methods, datasets, or validation procedures (28 studies, 46.7%).

The Industry 5.0-focused cybersecurity literature remains nascent, with limited industrial datasets. Therefore, we classify the 135 included studies into three categories: (1) research explicitly addressing the ICS, IIoT, or SCADA environments using domain-specific datasets like WUSTL-IIOT-2021 or Edge-IIoTset (n = 23, 17.0%); (2) research using established benchmarks (NSL-KDD, CICIDS-2017, UNSW-NB15, etc.) while investigating Industry 5.0-applicable challenges such as real-time detection, human-interpretable explainability, federated learning, and adversarial robustness (n = 78, 57.8%); and (3) foundational XAI-IDS research establishing broadly applicable techniques (n = 34, 25.2%). This classification enables systematic assessment of current research applicability to Industry 5.0 deployment requirements while identifying critical gaps between existing evaluation contexts and operational demands of human-centric industrial systems.

2.4. Inclusion and Exclusion Criteria

We included studies that satisfied the following criteria:

- Investigated XAI-based IDSs or Adv-XIDSs within the cybersecurity context of Industry 5.0.
- Employed deep learning architectures (e.g., transformers and LSTMs) or shallow computational models for intrusion detection.
- Comprised peer-reviewed articles or high-quality gray literature containing empirical findings or theoretical frameworks with substantive insights.

We defined exclusion criteria as follows:

- Non-English studies that would introduce language-based analytical barriers.
- Articles with peripheral relevance to XAI-based IDSs or cybersecurity paradigms.
- Publications with insufficient methodological detail or inadequate empirical support (e.g., conference abstracts and letters to editors).

Quality assessment was integrated throughout the selection process, with studies being evaluated for clarity of research objectives, appropriateness and rigor of methodology,

suitability of datasets and validation procedures, sufficiency of technical documentation, and significance of contributions to XAI-based IDS or adversarial XAI domains. Only studies demonstrating methodological rigor and substantive contributions across these dimensions were retained for analysis. These criteria ensured that the review maintained analytical focus on high-quality, relevant research contributions.

2.5. Data Extraction and Synthesis

We extracted data from the 135 included studies using a standardized protocol template, documenting research objectives, methodological approaches, key findings, and specific relevance to the established research questions. This process involved collaborative validation among multiple reviewers to ensure extraction accuracy and interpretive consistency through structured discussion and consensus building. We conducted narrative synthesis with thematic organization addressing the research questions, culminating in an analytical framework for understanding XAI-based IDS implementations in Industry 5.0 environments. This approach enhances the reliability of the conclusions while establishing a clear trajectory for future research and practical applications in this critical domain of cybersecurity.

3. Explainable AI (XAI) Taxonomies

Similar to other application domains, XAI tools could be very effective in cybersecurity by incorporating human insights in decision making. These tools allow cybersecurity experts and analysts to understand why a threat is flagged, leading to better threat detection, root cause analysis, and improved trust in AI decisions. They also facilitate more efficient utilization of human resources when dealing with false positives, which are prevalent in cybersecurity and often consume significant manual analysis time. To support cybersecurity analysts, various XAI techniques have been developed, broadly categorized into "ante hoc" and "post hoc" explainability methods [5]. This section explores the taxonomy of XAI in the security domain, as illustrated in Figure 3, with a specific focus on XAI-based intrusion detection systems (X-IDSs), including a comparative analysis of advantages and disadvantages inherent in prominent approaches.

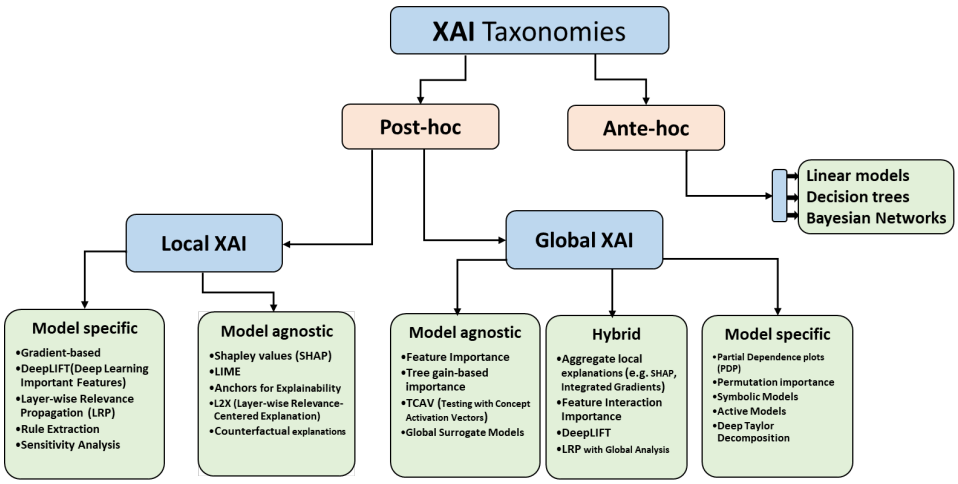

Figure 3. Comprehensive taxonomy of XAI methods categorized by interpretability approach (ante hoc vs. post hoc) and explanation scope (local vs. global), with specific techniques listed for each category.

3.1. Ante Hoc Explainability

Ante hoc explainability refers to models designed to be inherently interpretable, providing transparency into their decision-making processes at algorithmic, parametric, and

functional levels [120,121]. Such models—e.g., linear regression, logistic regression, Decision Trees, Random Forest, naive Bayes, Bayesian networks, and rule-based learning—are often applied in intrusion detection systems (IDSs) for Industry 5.0 cybersecurity, where understanding attack patterns is critical. However, their simplicity frequently limits their ability to handle the complex, high-dimensional, and non-linear data typical of modern cyber threats.

For example, *linear regression* uses coefficients to quantify how features (e.g., network traffic metrics) influence predictions, offering a clear view of linear relationships [122]. Yet, it fails to model the non-linear attack signatures common in IDS datasets. Similarly, *logistic regression* provides interpretable probabilities through coefficient signs and magnitudes [122], but its linear assumptions restrict its effectiveness against sophisticated threats. *Decision Trees* generate human-readable rules (e.g., "if packet size > 100, then flag intrusion") by partitioning data, though their transparency diminishes with increased depth, and they are prone to overfitting noisy IDS inputs [123]. *Random Forest*, an ensemble of Decision Trees, reveals global feature importance (e.g., IP address frequency), but their aggregated decisions obscure local interpretability. *Bayesian networks* depict probabilistic dependencies (e.g., between malware presence and port activity) via directed acyclic graphs, yet their computational complexity scales poorly with the large feature sets of IDSs [124]. Lastly, *rule-based learning* employs explicit "if-then-else" conditions (e.g., "if port = 80 and traffic > 1 MB, then alert"), ensuring straightforward interpretation, but its fixed rules struggle to adapt to evolving attack patterns [125].

The aforementioned inherently transparent models have achieved competitive performance in many regression and classification problems. However, these interpretability and explainability methods are limited to model families of lower complexity. Key constraints such as model size, sparsity, and monotonicity requirements create an inevitable trade-off between performance and transparency. Consequently, more complex model architectures—including ensembles, Artificial Neural Networks (ANNs), Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, transformers, and support vector machines (SVMs)—which often lack transparency in their decision-making processes, require post hoc methods for interpretation. A more recent category called hybrid approaches combines multiple explainable AI techniques to achieve appropriate interpretability while maintaining high predictive performance [124].

3.2. Post Hoc Explainability

Complex black-box models typically offer superior predictive capabilities at the expense of limited explainability regarding their decision-making processes. In intelligent systems, both accuracy and interpretability are essential characteristics. To address the transparency imperative, surrogate explainers become critical, requiring development and application to interpret the underlying rationale of intricate models' decisions [121].

Post hoc explainability methods have emerged as a prominent research avenue to address the opacity challenge of complex black-box model families. These methods elucidate the decision-making process of trained models by addressing two primary types of explanations: local explanations and global explanations. Local explanations focus on how a model predicts outcomes for specific instances, providing insights into the influence of particular inputs on classification decisions [126]. Conversely, global explanations assess the impact of all input features on the model's overall output, enabling comprehension of feature importance hierarchies and decision boundaries across entire datasets.

Global and Local Explanations

In cybersecurity contexts, global explanations provide strategic insights by revealing which network attributes, behavioral patterns, or file characteristics consistently contribute to threat classification across diverse attack scenarios. These explanations support defensive strategy development and resource allocation by identifying systemic vulnerabilities and attack vectors. However, they may obscure nuanced attack behaviors crucial to precise incident analysis [127].

Local explanations focus on interpreting individual predictions, enabling security analysts to understand why specific network flows, files, or user behaviors were classified as malicious or benign. This granular insight is essential to incident response, forensic investigation, and distinguishing between legitimate anomalies and actual threats. While valuable for tactical operations, local explanations may not capture broader attack campaign patterns requiring strategic intervention [128].

Model-Agnostic Approaches: Both local and global explanations are implemented through model-agnostic techniques—applicable to any AI-based model regardless of internal structure—and model-specific techniques tailored for particular architectures [129]. Model-agnostic post hoc explainability techniques focus directly on model predictions rather than internal representations, enabling deployment with any learning model irrespective of internal logic [1]. Most model-agnostic techniques quantify feature influence through simulated feature removal, termed removal-based explanations [130].

LIME (Local Interpretable Model-agnostic Explanation) primarily provides local explanations by approximating complex decision boundaries in the local neighborhood of specific instances using interpretable surrogate models [131]. In cybersecurity applications, LIME helps analysts understand why particular network connections were classified as intrusions, supporting incident-level decision validation. LIME offers domain-specific implementations for text, image, tabular, and temporal data analysis [132].

SHAP (SHapley Additive exPlanations) uniquely provides both local and global explanations by employing game theory principles to assign importance values representing each feature's contribution to specific predictions [133]. Individual SHAP values offer local explanations for specific predictions, while aggregating SHAP values across instances enables global interpretability of feature importance patterns. The SHAP framework includes specialized variants: Kernel SHAP, Tree SHAP, Deep SHAP, and Linear SHAP [134].

Other prominent approaches include visualization techniques that primarily provide global explanations: Accumulated Local Effect (ALE) plots and partial dependence plots (PDPs) illustrate relationships between features and model predictions [1]. Individual Conditional Expectation (ICE) plots bridge local and global perspectives by displaying predictions for individual instances while enabling pattern recognition across multiple instances [129].

Model-Specific Approaches: Model-specific post hoc explainability addresses models with design transparency but complex internal decision structures requiring specialized interpretation approaches. These techniques leverage architectural characteristics to provide targeted explanations based on the model's inherent structure [135].

Ensemble methods typically provide global explanations through feature importance analysis and model simplification techniques [136]. Simplification approaches include weighted averaging, Model Distillation [137,138], G-REX (Genetic-Rule Extraction) [139], and feature importance analysis through permutation importance or information gain [140].

Support vector machines (SVMs) demonstrate versatility by supporting multiple explanation types. Model simplification techniques provide global explanations through decision boundary interpretation, while counterfactual explanations offer local explanations by identifying minimal changes required for decision alteration. Example-based explana-

tions bridge local and global perspectives by utilizing representative dataset instances to illustrate SVM decision processes [39].

Deep learning models including Multi-Layer Perceptrons (MLPs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) require primarily post hoc explanation techniques due to their complex black-box characteristics. These approaches include model simplification, feature importance estimation, Saliency Map visualizations, and integrated local–global explanation frameworks [70]. Recent research has explored hybrid approaches incorporating expert-authored rules with algorithm-generated knowledge to achieve robust explainability [76].

The diversity of post hoc explainability methods enables cybersecurity practitioners to select appropriate explanation granularity based on operational requirements. Local explanations support tactical decision making for individual incidents, while global explanations inform strategic security posture development. This flexibility becomes particularly crucial in Industry 5.0 environments, where both rapid incident response and comprehensive threat landscape understanding are essential to effective cybersecurity.

Having established this XAI taxonomy framework, we now proceed to examine the growing demand for explainability in AI-based cybersecurity applications, with particular emphasis on interpretable intrusion detection systems. This represents a critical research domain that has evolved toward sophisticated IDS implementations. As this paradigm becomes integrated into Industry 5.0 infrastructure, the transparency of IDSs has become increasingly essential.

4. Industry 5.0 and Associated Cybersecurity Challenges

The enhanced interconnectivity characteristic of Industry 5.0 environments simultaneously exposes smart industrial systems to multifaceted cybersecurity vulnerabilities and threats, potentially compromising operational integrity, worker safety, and production continuity. These evolving threat landscapes create complex security challenges that traditional cybersecurity approaches struggle to address effectively, necessitating more sophisticated and transparent security solutions.

Contemporary industrial ecosystems require extensive data collection and analysis for operational objectives, including consumer behavior modeling, supply chain optimization, and predictive maintenance. This interconnectivity has substantially expanded potential entry vectors and exploitable vulnerabilities within Industry 5.0 frameworks, complicating threat detection and mitigation efforts [141]. The attack surface expansion becomes particularly critical in infrastructure contexts where electrical generation stations and water treatment facilities increasingly rely on Industrial IoT (IIoT) systems, introducing unprecedented risk levels affecting thousands of individuals. The complexity of these interconnected systems demands security solutions that can provide clear explanations of threat detection decisions to enable rapid human intervention and response.

Social engineering attacks exploit human cognitive vulnerabilities rather than technical system weaknesses, constituting significant threat vectors in human-centric Industry 5.0 environments. These attacks include phishing campaigns, pretexting scenarios, and voice phishing (vishing), which frequently serve as malware delivery mechanisms [142]. Within Industry 5.0 contexts, where human–machine collaborative interfaces are intensified, social engineering presents escalating security concerns. The human-centric nature of these threats requires security systems that can clearly communicate threat indicators and attack patterns to human operators, emphasizing the need for explainable AI approaches that bridge technical detection capabilities with human understanding.

Cloud computing infrastructure delivers essential capabilities for Industry 5.0 implementations, supporting manufacturing operations through IoT-based monitoring systems

and application programming interfaces for data normalization [23,143]. However, cloud architectures introduce distinct security challenges, including third-party software vulnerabilities, inadequately secured APIs, and complex data governance requirements. The distributed and multi-tenant nature of cloud environments creates intricate attack patterns that require sophisticated detection mechanisms capable of providing clear explanations of security events across diverse infrastructure components.

IoT systems enable comprehensive data acquisition across industrial domains through interconnected sensors and devices but present substantial security challenges due to large-scale deployment complexities and inconsistent protection measures [23]. Security implications vary considerably across implementation contexts, with medical devices and smart grid relays presenting higher risk profiles than consumer-grade devices. The heterogeneous nature of IoT ecosystems, combined with inconsistent vulnerability patching and limited on-device protection capabilities, creates complex threat landscapes requiring intelligent security solutions that can adapt to diverse device behaviors while providing transparent explanations of anomalous activities.

Supply chain vulnerabilities emerge from the inherent complexities and interdependencies among multiple stakeholders in modern industrial networks. While Industry 5.0 enhances supply chain management through human–robot collaboration, these interdependencies create potential attack vectors that can propagate across organizational boundaries [23]. The multi-organizational nature of supply chain attacks requires security solutions capable of correlating threats across different domains while providing clear attribution and explanation of attack progression to facilitate coordinated response efforts.

These cybersecurity challenges collectively demonstrate the limitations of traditional black-box security approaches in Industry 5.0 environments [7,12,144]. The complexity, criticality, and human-centric nature of these threats necessitate security solutions that not only detect attacks accurately but also provide transparent explanations of their decision-making processes [7,144]. This requirement for explainability becomes particularly crucial when security systems must enable rapid human intervention, facilitate cross-organizational threat communication, and maintain trust in human–machine collaborative environments. Recent advancements in intrusion prevention systems (IPSs) have enhanced network monitoring capabilities, with intrusion detection systems (IDSs) serving as critical components for preliminary threat identification. However, the evolving threat landscape in Industry 5.0 demands more sophisticated, explainable approaches to intrusion detection that can address these complex security challenges while maintaining transparency and human trust.

5. Intrusion Detection Systems for Cybersecurity in Industry 5.0

In Industry 5.0, the proliferation of interconnected systems and increasingly sophisticated automation has dramatically elevated cyber attack risks. Security breaches in these environments can precipitate severe industrial operational disruptions, substantial financial losses, and critical safety hazards. Consequently, the development and implementation of advanced cybersecurity measures, particularly machine learning (ML)-based intrusion detection systems (IDSs), are essential to safeguarding the integrity and security of Industry 5.0 ecosystems [145,146].

ML-based IDSs have demonstrated remarkable efficacy in addressing cybersecurity challenges within industrial environments [17,18]. In the Industry 5.0 context, the strategic importance of intelligent IDSs has increased exponentially [21,23]. Given the expanded attack surface resulting from device interconnectivity discussed in Section 4, these systems enable the effective monitoring of networks and systems for malicious activities, behavioral anomalies, and potential violations of data management policies. Furthermore, IDSs offer valuable post-incident capabilities following social engineering attacks by detecting

suspicious device behaviors, network anomalies, unusual access patterns, unauthorized data flows, and known malicious payloads. They also play a crucial role in mitigating cloud and IoT vulnerabilities—prevalent concerns in Industry 5.0 environments—through advanced detection methodologies such as anomaly-based approaches [147].

Beyond protecting conventional personal and enterprise networks, the imperative to secure critical infrastructure represents a paramount concern. While the compromising of personal networks and smart devices creates privacy, financial, and psychological risks for individuals, the threat level escalates significantly when considering enterprise networks that process information about numerous individuals [148,149]. The risk profile becomes exponentially more severe when addressing critical infrastructure security. Protection of these systems is fundamentally important because disruptions to essential services such as electrical power distribution and water supply can profoundly impact thousands of individuals and organizations. Although the compromising of a personal health monitoring device presents a significant threat to an individual, this pales in comparison to the catastrophic consequences potentially resulting from security breaches in nuclear power plant control systems, which could affect thousands or millions of people. This criticality elevates the protection of industrial infrastructure to the highest priority within cybersecurity frameworks, simultaneously emphasizing the essential role of explainable machine learning approaches in this domain [32,150,151].

This section provides a comprehensive examination of ML-based IDSs, highlighting their key operational aspects and exploring how explainable AI (XAI) enhances their effectiveness within Industry 5.0 security contexts.

5.1. Traditional Intrusion Detection System (IDS)

Conventional approaches to IDS implementation typically employ either signature-based intrusion detection systems (S-IDSs) or anomaly-based intrusion detection systems (A-IDSs), as illustrated in the upper section of Figure 4. S-IDS methodologies operate by matching new patterns against previously identified attack signatures, a technique also referred to as knowledge-based detection [152,153]. These approaches rely on constructing comprehensive databases of intrusion instance signatures, against which each new instance is compared. However, this detection methodology demonstrates significant limitations in identifying zero-day attacks and is further compromised by polymorphic and metamorphic techniques incorporated into modern malware, which enable the same malicious software to manifest in multiple forms, evading signature-based identification.

The challenges posed by polymorphic and metamorphic malware variants have been addressed through anomaly-based intrusion detection systems (A-IDSs), which analyze suspicious variants within controlled sandbox environments to evaluate behavioral characteristics. An alternative analytical approach involves establishing behavioral baselines for normal computer system operations using machine learning, statistical analysis, or knowledge-based methodologies [154]. Following the development of a decision model, any significant deviation between observed behavior and established model parameters is classified as an anomaly and flagged as a potential intrusion. From a traditional sandbox analysis perspective, A-IDSs demonstrate superior capabilities in detecting zero-day attacks and identifying polymorphic and metamorphic intrusion variants. However, this approach is constrained by detection speed limitations compared with S-IDS implementations. The integration of advanced ML techniques has substantially mitigated these limitations by automatically identifying essential differentiating characteristics between normal and anomalous data patterns with high accuracy rates [155,156].

Machine learning solutions operate on the principle of data generalization to formulate accurate predictions for previously unencountered scenarios. These approaches

demonstrate optimal performance with sufficient training data volumes. The ML domain encompasses two primary methodological categories: supervised learning, which utilizes labeled data for model training, and unsupervised learning, which extracts valuable insights from unlabeled datasets. The performance efficacy of ML-based IDS models is contingent upon the quality of the data type information, acquisition accessibility and speed, and the fidelity with which the data reflects source behavior (i.e., host or network characteristics) [157]. Common data sources leveraged in ML-based solutions include network packets, function or action logs, session data, and traffic flow information. The cybersecurity research community frequently employs feature-based benchmark datasets such as DARPA 1998, KDD99, NSL-KDD, and UNSW-NB15 for standardized evaluation purposes [158–160].

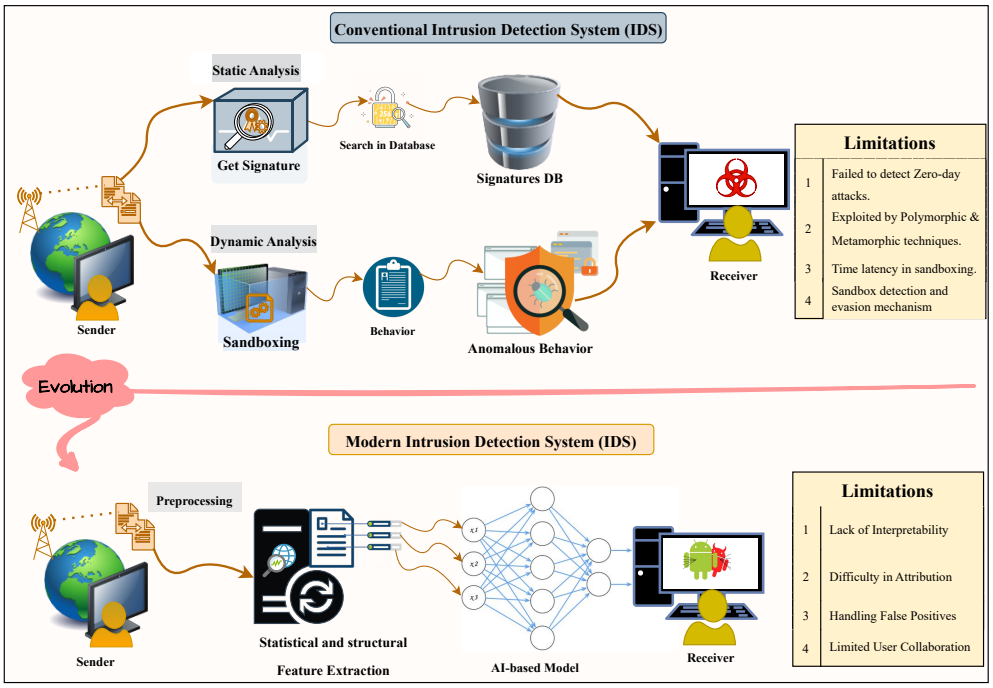

Figure 4. Evolution of intrusion detection systems: conventional versus modern architectures. (**Upper**): Conventional IDS using signature-based and behavior-based (sandboxing) detection with limitations including zero-day attack vulnerability and polymorphic technique susceptibility. (**Lower**): Modern ML-based IDS with preprocessing and AI-based detection, exhibiting limitations in interpretability, false-positive handling, and user collaboration—motivating XAI adoption.

Multiple data types serve distinct functions in detecting various attack categories, as each data type reflects specific attack behavior patterns. For instance, system function and action logs primarily reveal host behavior characteristics, while session and network flow data illuminate network-level activities. Consequently, appropriate data sources must be selected based on specific attack characteristics to ensure the collection of relevant and actionable information [153]. Header and application data contained within communication packets provide valuable information for detecting User-to-Root (U2R) and Remote-to-Local (R2L) access attacks. Packet-based IDS implementations incorporate both packet parsing-based and payload analysis-based detection methodologies. Network flow-based attack detection represents another significant approach, particularly effective against Denial-of-Service (DoS) and Probe attacks, employing feature engineering-based and deep learning-based detection techniques [161].

Session creation-based attacks can be identified using statistical information derived from session data as input vectors for decision models. Sequence analysis of session packets

provides detailed insights into session interaction patterns, an approach frequently implemented using text processing technologies such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks to extract spatial features from session packets. System logs recorded by operating systems and application programs represent another important attack detection vector, containing system calls, alerts, and access records. However, effective interpretation of these logs typically requires specialized cybersecurity expertise. Recent detection methodologies increasingly incorporate hybrid approaches that combine rule-based detection with machine learning techniques. Additional detection methodologies include text analysis techniques that process system logs as plain text, with n-gram algorithms commonly employed to extract features from text files for subsequent classification tasks [153].

5.2. Explainable IDS (X-IDS)

The preceding analysis demonstrates that most contemporary intelligent intrusion detection systems employ complex ML techniques that deliver exceptional intrusion detection performance. However, another critical aspect requiring consideration in IDS design is decision-making process transparency. System developers must address fundamental "Why" and "How" questions regarding IDS model operations to develop more reliable, secure, and effective security solutions (Figure 5). Explanatory elements within IDS implementations may include rationales for generated alerts, justifications for anomaly or benign classifications, and compromise indicators for security operations center analysts [76,144]. The necessity and utility of explanations in security systems were initially proposed by Viganò et al. [162], who emphasized explanations' role in comprehending intelligent systems' core functionality. They introduced the "Six Ws" framework, expanding the traditional "Five Ws" (Who, What, When, Where, and Why) by incorporating "How" to explore methodological aspects. Although "How" deviates from the "W" pattern, its conventional inclusion ensures comprehensive analytical coverage. Viganò et al. demonstrated that addressing these six dimensions enhances security system transparency and operational effectiveness.

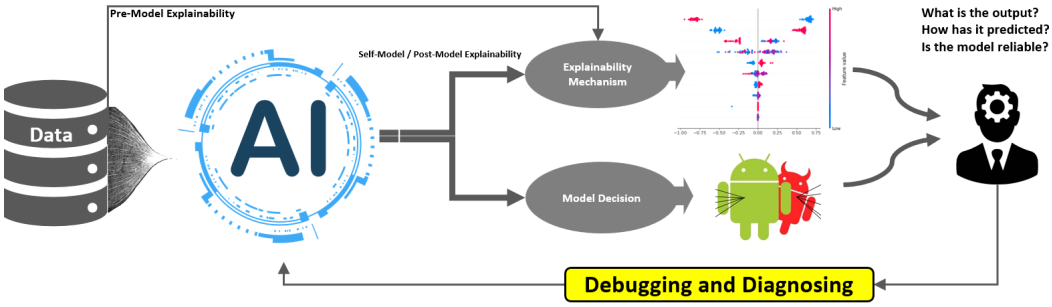

Figure 5. Explainable intrusion detection system (X-IDS) operational framework. Data flows through the AI-based model generating dual outputs: (1) model decision (classification/alert) and (2) explainability mechanism providing decision rationale. Security analysts evaluate both streams to answer "What is the output? How was it predicted? Is the model reliable?", creating a feedback loop for debugging, diagnosis, and model improvement.

The explainable and interpretable IDS concepts assume heightened significance in Industry 5.0 contexts as organizations strive to effectively address emerging cyber threats while maintaining security measures of transparency and interpretability. Explainability in IDSs represents a collaborative initiative between AI systems and human operators addressing technical challenges at both implementation and operational levels, enhancing detection and response capabilities. This collaborative approach enables IDSs to tran-

scend black-box model limitations by integrating fundamental knowledge and insights, facilitating interpretable decision-making processes [26,146].

The cybersecurity research community is actively revising traditional intelligent intrusion detection systems to incorporate explainability features tailored to diverse stakeholder requirements [163,164]. These revisions have established three distinct explainability domains: self-model explainability, pre-modeling explainability, and post-modeling explainability. Self-model explainability encompasses simultaneous generation of explanations and predictions, leveraging problem-specific insights derived from domain expert knowledge. Pre-modeling explainability utilizes refined attribute sets to facilitate clearer system behavior interpretations before model training. Post-modeling explainability focuses on modifying trained model behavior to enhance responsiveness to input–output relationships, improving overall system transparency and effectiveness within the dynamic Industry 5.0 cybersecurity landscape [165,166]. The subsequent sections examine these explainability approaches in greater detail.

5.2.1. Self-Model Explainability

The X-IDS models generated from self-explaining models are designed to inherently explain their intrinsic decision-making procedure. These models exhibit simple architectures capable of identifying crucial attributes that trigger the decision-making process for a given input. In this way, several explainability techniques have been proposed according to the model's complexity. For instance, Sinclair et al. [44] suggested a rule-based explanation by developing an ante hoc explainability application named NEDAA system that combines ML methods like genetic algorithms and Decision Trees to aid intrusion detection experts by generating rules for classifying normal network connections from anomalous ones, based on domain knowledge. The NEDAA approach employs analyst-designed training sets to develop rules for intrusion detection and decision support. Mahbooba et al. [45] tried to explain and interpret known attacks in the form of rules to highlight the target of the attack and their causal reasoning using Decision Trees. They used ID3 to construct a Decision Tree, using the KDD-99 dataset, where the decision rules traverse from top to bottom nodes, and the rules from the model are generated using the Rattle package in R. Manoj et al. [46,47] proposed Decision Tree-based explainability to explain the actions taken by the Industrial control system against IoT network activities. These rules can be compiled into an expert system for detecting intrusive events or to simplify training data into concise rule sets for analysts. The rule-based explanation offers valuable insights into decision making, promotes transparency, and allows domain expertise integration. However, they have limitations in handling complex and evolving threats, scalability, and potential conflicts [164]. Another recent host-based intrusion detection system (HIDS) was proposed by Yang et al. [48]; they used Bayesian networks (BNs) to create a self-explanatory hybrid detection system by combining data-driven training with expert knowledge. BNs are a specific type of Probabilistic Graphical Models (PGMs) that model the probabilistic relationships among variables using Bayes' Rule [167]. Firstly, they extracted expert-informed interpretable features from two datasets, Tracer FIRE 9 (TF9) and Tracer FIRE 10 (TF10), which consist of normal and suspect system event logs generated through Zeek and Sysmon by the Sandia National Laboratories (SNL) Tracer FIRE team. The authors utilized Bayes Server as an engine for evaluating multiple BN architectures in finding the best-performing model, while the explanations are provided by visualizing the network graph, which provides feature importance information via conditional probability tables [48]. Self-explanatory models in IDSs offer notable advantages by enhancing transparency and interpretability. They provide insights into decision making, enabling analysts to understand the reasoning behind alerts. This aids in trust building, model validation, and effective response. How-

ever, self-explanatory models might struggle with complex relationships, limiting their capacity to capture nuanced attack patterns.

5.2.2. Pre-Modeling Explainability

Pre-modeling explainability techniques involve some preprocessing methods to summarize large feature datasets into an information-centric set of attributes that align with human understanding and help downstream modeling and analysis. Zolanvari et al. [49] proposed an explainable model for transforming the input features into representative variables through the factor analysis of mixed data (FAMD) method and then for finding mutual information to quantify the amount of information for each representative and their mutual dependence on the class labels, which helps in finding the top explainable representatives for artificial neural network (ANN). Le et al. [50] used information gain (IG) to calculate the most informative feature values, which are then used in Ensemble Tree classification. The model's outputs are then plotted in the form of a heatmap and decision plot using the SHAP explanation technique. Alani et al. [51] proposed a method named Recursive Feature Elimination (RFE) using feature importance, where the features having the lowest importance are removed during the training and test rounds of different classifiers, including RF, LR, DT, GNB, XGB, and SVM classifiers. After retrieving the minimum number of features on which the model shows better performance, RF, a TreeExplainer which is a type of SHAP explainer, is used to measure the contribution of each selected feature. Gurbuz et al. [54] addressed the security and privacy issues of the IoT-based healthcare network data flow by applying the least computationally expensive machine learning models, including KNN, DT, RF, NB, SVM, MLP, and ANN. The procedure involves first retrieving important features using a linear regression classifier and then leveraging the Shapash Monitor explanation interface to visualize feature importance plots, prediction distributions, and partial dependence plots for healthcare professionals, data scientists, and other stakeholders. Patil et al. [55] analyzed the correlation between features using a heatmap, and the outliers were excluded in the preprocessing step. Then the LIME technique was used to explain their Voting Classifier, consisting of RF, DT, and SVM classifiers. Zebin et al. [56] addressed the explainability problem in DNS over HTTPS (DoH) protocol attack detection system. To understand the underlying distribution of the dataset, the Kernel Density Estimation (KDE) technique was deployed to estimate the probability density function of the features. After the thorough preprocessing of the datasets, optimal hyperparameters for the base RF classifier were found by the GridsearchCV function. For the explanation of the model, they used SHAP values to highlight the features that contributed to the underlying decision of the model. Sivamohan et al. [57] selected information-rich features by using the Krill Herd Optimization (KHO) algorithm for BiLSTM-XAI-based classification, where the explanation is provided using both LIME and SHAP mechanisms. Wang et al. [58] proposed a hybrid explanatory mechanism by first finding the top-most important feature set by using the LIME technique on a CNN+LSTM structure. A Decision Tree model, XGBoost, is then trained on the selected important features, while the explanations for the important features are generated through the SHAP mechanism. Another hybrid mechanism was proposed by Tanuwidjaja et al. [59] by using both LIME and SHAP mechanisms to cover both the local and global explanations of an SVM-based IDS.

Another pre-modeling explainability technique involves visualization, where the focus is on providing intuitive visualizations of data and model behavior to help users, analysts, and stakeholders gain insights into how the model works and why it makes certain predictions. Mills et al. [60] proposed a graphical representation for understanding the Decision Trees of the Random Forest (RF) classifier. In the same context, Self-Organizing Maps (SOMs), also known as Kohonen maps [168], are used as an exploratory tool to gain a

deeper understanding of the data that the decision model is trained on. Ables et al. [61,62] trained and evaluated different extensions of Kohonen Map-based Competitive Learning algorithms, including Self-Organizing Map (SOM), Growing Self-Organizing Map (GSOM), and Growing Hierarchical Self-Organizing Map (GHSOM), which are capable of producing explanatory visualizations. The core design of these extensions is to organize and represent high-dimensional data in a lower-dimensional space while preserving the topological relationships and structures of the original data. That is why SOMs can also be used for dimensionality reduction. In terms of IDS explainability, statistical and visual explanations were created by visualizing global and local feature significance charts, U-matrix, feature heatmaps, and label maps through the resulting trained models using NSL-KDD and CIS-IDS-2017 benchmark datasets. Lundberg et al. [63] proposed a visual explanation method, named VisExp, that applies SHAP to find feature importance values, "SHAP-values", for explaining the behavior of an in-vehicle intrusion detection system (IV-IDS). The visual explanation is generated as a dual swarm plot utilizing standard Python visualization libraries, which presents the normal Controller Area Network (CAN) traffic at the top and the intruder's traffic at the bottom according to the SHAP-values distribution. Al et al. [64] used the Fast Gradient Sign Method (FGSM) as an adversarial sample generator, and in the next step, the DALEX framework was utilized for identifying the most influential features that enhance the Deep Neural Network (DNN) model's decision performance. The same fine-tuning of the deep cyber threat detection model was also explored by Malik et al. [65] by coupling the same adversarial sample generator, "FGSM", and the explanations generated through SHAP values. Lanfer et al. [66] addressed the false alarms and dataset limitations issues in available network-based IDS datasets by utilizing SHAP summary and Gini impurity. Their contribution lies in demonstrating how imbalances in datasets can affect XAI methods like SHAP and how retraining models on specific attack types can improve classification and align better with domain knowledge. They utilized SHAP beeswarm plots to visualize the explanations of the target class individually. As such, visual explanations offer intuitive insights into the system's behavior, but the dependence on visualization quality makes the system limited to subjectivity and may also lead to inconsistency with the change in the visualization technique [155]. To mitigate these limitations, a combination of various explainability methods, including both visual and non-visual approaches, should be employed to provide a more comprehensive understanding of IDS behaviors and enhance threat detection and prevention capabilities. Lu et al. [68] proposed a feature attribution explainability mechanism based on the concept of an economic theory called modern portfolio theory (MPT). By considering features to be assets and using perturbation, the expected feature output attribution values are referred to as their explanation. Feature attribution based on modern portfolio theory minimizes the variance in prediction score changes about attribution values, whereby a higher feature attribution value signifies a substantial impact on the model's prediction score with a small feature change.

5.2.3. Post-Model Explainability

Post-model explainability refers to the techniques and methods used to interpret and understand the decisions made by a trained learning model. Unlike self-and pre-model explainability techniques, post-modeling allows stakeholders to gain insights into model decisions, detect biases, and validate model behavior, contributing to better-informed decision making and building trust in AI systems [151]. The most adopted techniques in the literature include the feature importance methods, where the impact of each input feature is analyzed according to the trained model's performance. Sarhan et al. [69] used the SHAP method to explain ML model detection performance. After finding the best-performing sets of hyperparameters for both the MLP and RF classifiers through partial grid search, the

models were analyzed to understand their internal operations by calculating the Shapley value of the features. Tcydenova et al. [169] proposed the LIME explainability approach for detecting adversarial attacks on IDSs, where the normal data boundaries are explained for a trained SVM-based model. Gaitan et al. [170] used horizontal bar plots to visualize the global explanation of the model prediction using the SHAP mechanism. Oseni et al. [70] proposed a SHAP mechanism for improving the interpretation and resiliency of DL-based IDSs in IoT networks. The use of the SHAP mechanism was also proposed by Alani et al. [71] and Kalutharage et al. [72] to explain a deep learning-based Industrial IoT (DeepIIoT) intrusion detection system. Muna et al. [73] employed LIME and SHAP mechanisms to explain the prediction made by the Extreme Gradient Boosting (XG-Boost) classifier and also used ELI5, "Explain Like I'm 5", a Python package (https://www.python.org/, accessed on 26 October 2025) using the interpreting Random Forest feature weight approach. This package supports tree-based explanation to show how effective each feature is in contributing to all parts of the tree in the final prediction. Abou et al. [74] used RuleFit and SHAP mechanisms to explore the local and global interpretations for DL-based IDS models. Marino et al. [10] utilized an adversarial ML approach to find an explanation for the input features. They used the samples that were incorrectly predicted by the trained model and tried again with the required minimum modifications in feature values to correctly classify. This allowed for the generation of a satisfactory explanation for the relevant features that contributed to the misclassification of the MLP model. da Silveira Lopes et al. [75] combined the SHAP and adversarial approaches to accurately identify the false-positive prediction by an IDS model. Szczepanski et al. [76] proposed a prototype system where they utilized a Feedforward ANN with PCA to train as a classifier, and in parallel, a Decision Tree was generated from the samples along with their outputs from the classifier. The retrieved tree was handled by the DtreeViz library to visualize an explanation for the classifier's decision. Wang et al. [77] improved the explanation of an IDS by combining local and global interpretation generated by models using the SHAP technique. The local explanation gives the reason for the decision taken by the model, and the global explanation shows the relationships between the features and different attacks. They used the NSL-KDD dataset and two different classifiers, namely, one-versus-all and multiclass classifiers, to compare the interpretation results. Nguyen et al. [78] adopted the same mechanism to explain the decisions made by a CNN and a DT-based IDS model by using SHAP values. The target was to build trust for the design of the intrusion detection model among security experts. Roy et al. [80,81] proposed a SHAP-LIME hybrid explainability technique to explain the results generated by a DNN both globally and locally. Mane et al. [82] presented the same hybrid approach for explaining a Deep Neural Network-based IDS. To provide quantifiable insights into which features impact the prediction of a cyber attack and to what extent, they used the SHAP, LIME, Contrastive Explanation Method (CEM), ProtoDash, and Boolean Decision Rules via Column Generation (BRCG) approaches.

Learning the compact representations of input data through the encoding and decoding process, autoencoders aid in uncovering underlying patterns and essential features within the data. This latent representation often corresponds to meaningful characteristics of the input data, making it easier to understand and interpret the model's behavior [171]. The autoencoders are based on reconstructing the input samples by minimizing the reconstruction error between the encoder and decoder. Along with the great property of anomaly detection, reconstruction error-based methods also provide a comprehensive explanation of the connection between the inputs and the corresponding outputs. In this context, Khan et al. [83] proposed an autoencoder-based IDS architecture by adopting a CNN and an LSTM-based autoencoder to discover threats in the Industrial Internet of Things (IIoT), as well as explaining the model internals. For model explainability, the LIME

technique was used to explain the predictions of the proposed autoencoder-based IDS. Ha et al. [84] proposed the same LSTM-based autoencoder model for anomaly detection in industrial control systems, where the explainability of the model was achieved by the Gradient SHAP mechanism. Nguyen et al. [85] used a variational autoencoder (VAE) to detect network anomalies and a gradient-based explainability technique to explain the models' decisions. Antwarg et al. [86] used reconstruction error as an anomaly score and computed the explanation for prediction error by relating the SHAP values of the reconstructed features to the true anomalous input values. Aguilar et al. [87] proposed a Decision Tree-based interpretable autoencoder, where the correlation between the categorical attributes' tuples is learned through the Decision Tree encoding and decoding process and the interpretability of the autoencoder is achieved by finding the rules from the decoder to interpret how they enable to decode the tuple accurately. Lanvin et al. [88] proposed a novel explainability mechanism named AE-values, where the explanation is based on the *p*-values of the reconstruction errors produced by an unsupervised autoencoder-based anomaly detection method. They handle the anomaly detection problem as a one-class classification problem using the Sec2graph method, and a threshold value is computed from the reconstruction error of the input benign files. The error value above the threshold is considered responsible for the anomaly. Javeed et al. [89,90] proposed a multiclass prediction model by combining BiLSTM, Bidirectional-gated recurrent unit (Bi-GRU), and fully connected layers. They applied the SHAP mechanism on the last fully connected layer to obtain the local and global interpretation for the model decision.

Another post-model explainability technique involves Saliency Map or Attention Map methods, which aim to explain the decisions of a Convolutional Neural Network (CNN) by highlighting the regions of an input image that contribute the most to a specific prediction. Yoon et al. [172] proposed a method named Memory Heat Map (MHM) to characterize and segregate the anomalous and benign behavior of the operating system. Lin et al. [91] used the region perturbation technique to generate a heatmap for visualizing the predictions made by the image-based CNN model. Iadarola et al. [92] proposed a cumulative heatmap generated using the Gradient-weighted Class Activation Mapping (Grad-CAM) technique, where the gradients of the convolutional layer are converted into a heatmap through Grad-CAM to balance the trade-off between CNN accuracy and transparency. Andresini et al. [93] addressed the network traffic classification and decision explanation task through image classification, where the network flow traces are transformed into a pixel frame of single-channel square images. A CNN model is incorporated with the attention layer to capture the contextual relationships between different features of the network traffic and the observed intrusion classes. These techniques help users understand not only which regions are important but also the extent to which different regions influence the model's predictions. They can be particularly useful for gaining a finer-grained understanding of the relationships between input features and model responses [173].

From the literature review, it is evident that machine learning-based intrusion detection systems (IDSs) predominantly employ rule-based, LIME, and SHAP techniques to achieve local and global explainability. Despite their widespread adoption, these methods exhibit limitations in fully interpreting the complex decision-making processes of black-box models, particularly in capturing nuanced attack patterns and ensuring robustness across diverse threat scenarios. Consequently, there is a pressing need for advanced explainability techniques that enhance transparency, improve interpretability, and align more effectively with domain-specific requirements in cybersecurity applications for Industry 5.0 environments.

6. Adversarial XAI and IDSs

While XAI enhances trust and transparency in cybersecurity systems, it simultaneously introduces new vulnerabilities that sophisticated attackers can exploit. This paradox represents a fundamental challenge in Industry 5.0 cybersecurity: the same explainability mechanisms that enable human understanding and system debugging also provide adversaries with insights into model decision-making processes [174,175].

The prevalence of adversarial methodologies, wherein attackers systematically subvert intelligent models by crafting adversarial examples, targets both black-box and white-box models [176]. These attacks exploit factual and counterfactual explanations generated by XAI methods to facilitate feature manipulation, evasion, poisoning, and oracle attacks, as illustrated in Figure 6. Oracle attacks encompass techniques where adversaries exploit model outputs—predictions, confidence scores, or XAI-generated explanations—to infer sensitive information or manipulate system behavior, including membership inference and model vulnerability extraction [177]. This exploitation becomes particularly critical in Industry 5.0, where interconnected systems rely on real-time IDSs.

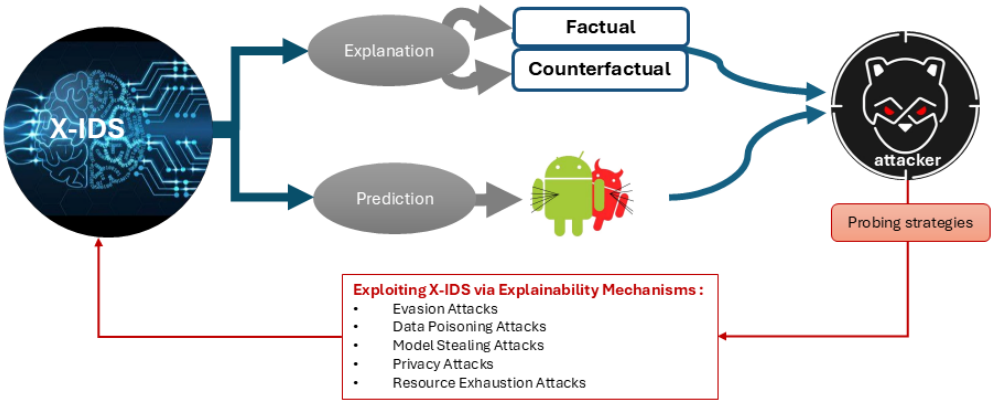

Figure 6. Exploiting X-IDSs: Attacker probes predictions and explanations (factual and counterfactual) to launch adversarial attacks, undermining system security.

From our analysis in Section 5 and Table 2, prominent explainability methods include regression model coefficients, rule-based approaches, LIME, SHAP, and gradient-based explanations. These techniques are primarily evaluated based on descriptive accuracy and relevance [178]. However, access to detailed model decision-making information enables attackers to manipulate both target security models and their explainability mechanisms [36,179]. This vulnerability necessitates robust defenses against adversarial attacks in AI-based cybersecurity systems for Industry 5.0 [180].

Before machine learning adoption, network anomaly detection relied on carefully designed rules that expert attackers could reverse-engineer to bypass detection mechanisms [181]. While intelligent ML-based systems have shown promise in mitigating these threats, the ongoing adversary–defender rivalry drives sophisticated adversarial strategy development. Deep learning models remain susceptible to adversarial attacks that manipulate their behavior, leading to outcomes contrary to intended functionality [182].

Table 3 provides a comprehensive overview of adversarial techniques targeting IDSs, categorized by attack type and explainability utilization. The analysis reveals a progression from traditional black-box attacks to sophisticated explainability-aware attacks, highlighting the evolving threat landscape in XAI-based cybersecurity systems.

6.1. Adversarial Attacks Without Utilizing Explainability

Adversarial attacks on ML systems are categorized into white-box attacks (complete target system knowledge), black-box attacks (limited knowledge with model querying

capability), and gray-box attacks (partial classifier information) [183]. Attack vectors include privacy attacks, poisoning attacks, and evasion attacks. Perturbation and evasion mechanisms represent the most straightforward approaches, where perturbation attacks subtly alter input data to mislead ML models and evasion attacks create inputs that bypass detection entirely [184–186].

Generative Adversarial Networks (GANs) and variants are widely employed in cybersecurity for generating synthetic data, addressing class imbalance, and creating adversarial examples [187,188]. Piplai et al. [94] targeted GAN-based solutions using discriminator neural networks as classifiers, successfully attacking through Fast Gradient Sign Method (FGSM) perturbations. Ayub et al. [95] employed Jacobian-based Saliency Map Attack (JSMA) for adversarial sample generation, while Pujari et al. [189] utilized Carlini–Wagner white-box evasion attacks. Alshahrani et al. [96] performed evasion attacks using Deep Convolutional GANs for synthetic sample generation.

GAN training instability has led to variants including Wasserstein GAN with gradient penalty (WGAN-GP) and AdvGAN. Duy et al. [97] investigated ML-based IDS vulnerabilities in software-defined networks using these variants, targeting non-functional features for malicious traffic evasion. Zhang et al. [98] evaluated classifier robustness using gradient-free methods, while Lan et al. [99] introduced poisoning attacks through malicious file injection into benign Android APKs.

Membership inference attacks pose severe threats to AI model privacy by revealing training data information [177]. Qiu et al. [100] presented black-box adversarial attacks on DL-based NIDSs in IoT environments, achieving 94.31% attack success rates while modifying minimal packet bytes. Chen et al. [101] introduced Anti-Intrusion Detection Auto-Encoder (AIDAE) for generating features mimicking normal network traffic. Jiang et al. [102] demonstrated perturbation attacks against LSTM and RNN models, proposing Feature Grouping and Multi-model fusion Detector (FGMD) for enhanced robustness.

Transferability attacks leverage adversarial sample effectiveness across different models. Debicha et al. [105] employed transferability techniques targeting NIDSs through surrogate model bypass strategies. Ravikrishnan et al. [103] deployed gradient-based evasion attack techniques, including FGSM, Basic Iterative Method (BIM), Momentum Iterative Method (MIM), and Projected Gradient Descent (PGD), to expose vulnerabilities and subsequently retrained DNN-IDS with shuffled adversarial and normal samples for improved evasion resistance. Merzouk et al. [190] targeted Deep Reinforcement Learning-based NIDSs with adversarial examples generated by FGSM and BIM to evade detection. Poisoning attacks target continuous data collection requirements in DL-based systems. Li et al. [106] introduced Edge Pattern Detection algorithms for boundary pattern poisoning, while Xu et al. [107] targeted LSTM through discrete adversarial sample generation. Nguyen et al. [108] compromised federated learning-based IDSs through malicious traffic injection.

Real-world deployment constraints limit many adversarial approaches, as IDSs typically function as label-only black-box systems providing binary decisions without feature access or confidence scores [191]. Functionality preservation requirements ensure crafted attacks maintain intended behavior under human or machine inspection. The risk escalates when adversaries gain black-box operation insights through explainability techniques [169,192].

6.2. Adversarial Attacks Utilizing Explainability

XAI methods have been exploited to understand, identify, and facilitate adversarial attacks through visualization generation, highlighting vulnerable modification regions.

Tcydenova et al. [169] and Rehman et al. [193] employed LIME techniques for attack–normal traffic differentiation, identifying influential feature sets for model validation. While effective for vulnerability identification, these explanation methods remain limited in diverse attack detection scenarios and can be manipulated to undermine trust or facilitate attack strategies.

The trade-off between X-IDSs and adversarial attacks proves complex and multifaceted. Within the CIA triad framework, confidentiality attacks exploit explanations to reveal model architecture or training data, while integrity and availability attacks leverage explanations to manipulate model outputs or disrupt legitimate user access [194]. These attacks occur during training (poisoning) or deployment (evasion) phases.

Rosenberg et al. [109] introduced explainability transferability concepts, where impactful features identified through explainability algorithms on substitute models transfer to target black-box models. Using Kendall's tau for feature ranking evaluation, this approach enables adversarial example generation through structural feature modification without compromising core functionality. Zhang et al. [110] advanced this concept with the Explainable Transfer-based Attack (ETA) framework, optimizing substitute models and crafting adversarial examples through gradient evaluation.

Kuppa et al. [111,112] demonstrated confidentiality compromise through the Manifold Approximation Algorithm (MAA) for data distribution pattern identification. They presented four successful black-box attacks—evasion, membership inference, poisoning, and model extraction—using counterfactual explanation methods to undermine classifier confidentiality and privacy. Severi et al. [113] proposed backdoor poisoning attacks employing SHAP explainability for optimal feature set selection as training-embedded backdoors.

Rosenberg et al. [109] developed end-to-end adversarial example generation for multi-feature malware classifiers, exploiting Integrated Gradients, DeepLIFT, ϵ-LRP, and SHAP algorithms for feature importance ranking. Okada et al. [114] employed Integrated Gradients and KernelSHAP for network traffic feature manipulation, generating adversarial examples effective in white-box and black-box scenarios. Ruijin et al. [115] proposed FastLSM binary-diversification techniques using Superpixels interpretation mechanisms for malware binary evasion. Hoang et al. [116] developed ADV-Sword, combining SHAP with Accured Malicious Magnitude (AMM) to generate adversarial samples against IDSs. Using surrogate models to identify manipulable features via SHAP values, their black-box attacks on the InSDN dataset reduced IDS detection rates from 99% to 34–50% across LightGBM, Random Forest, and CNN models, with MLP surrogate models achieving near-zero detection (0.0 recall) against CNN-based IDSs. Alani et al. [117] developed SHAP-based evasion attacks targeting features with maximal benign decision contributions. Shu et al. [118] targeted Android malware count features using LIME explainability for classification decision alteration.

The problem of adversarial XAI as a threat to explainable cybersecurity systems is a complex one. While explainability enables higher trust and confidence in detection systems, it also expands the attack surface, which enables malicious actors to exploit this explainability to identify usable attack vectors. Mitigation of these threats would often obfuscate the intrusion detection system and render explainability impossible.

6.3. Synthesis and Implications

Our analysis reveals significant evolution in adversarial attack sophistication, progressing from traditional black-box approaches to explainability-aware attacks that directly exploit XAI mechanisms. Table 3 documents this progression, demonstrating enhanced attack effectiveness when explainability information becomes available to adversaries.

Systematic patterns emerge in XAI exploitation across the reviewed studies. SHAP-based attacks leverage feature importance values to identify high-impact features for minimal perturbation, achieving 85–94% evasion success rates [109,110,114,117]. Alani et al. [117] demonstrated that targeting features with the highest SHAP contribution values enabled adversarial examples to evade XGBoost-based IDSs with 91% success while modifying only two–three features. LIME exploitation follows similar principles, operating through local approximations to craft evasion attacks with documented success rates of 88–92% [118,169]. Gradient-based methods, including Integrated Gradients, prove particularly vulnerable, as gradient information directly reveals model sensitivity patterns, enabling targeted perturbations with success rates exceeding 90% in white-box scenarios [109,114]. These empirical findings demonstrate that the same mechanisms providing transparency—feature attribution, gradient computation, and local approximation—create exploitable vulnerabilities when exposed to adversaries.

Figure 7 synthesizes key trends from our comprehensive analysis of Tables 2 and 3. The dataset landscape (Figure 7a) shows persistent reliance on established benchmarks (e.g., NSL-KDD, CICIDS-2017, and KDD-99), while emerging Industry 5.0-specific datasets (e.g., IoTID20 and TON-IoT) remain underrepresented, indicating a critical gap in evaluation frameworks for modern industrial environments. Detection model analysis (Figure 7b) reveals dominance of ensemble methods (Random Forest and Decision Trees) and deep learning architectures (e.g., CNN, LSTM, and DNN), reflecting the field's preference for high-performance models requiring post hoc explainability. Most significantly, explainability technique distribution (Figure 7c) demonstrates SHAP's and LIME's predominance as model-agnostic methods, with rule-based explanations maintaining substantial presence due to inherent transparency advantages in security-critical applications.

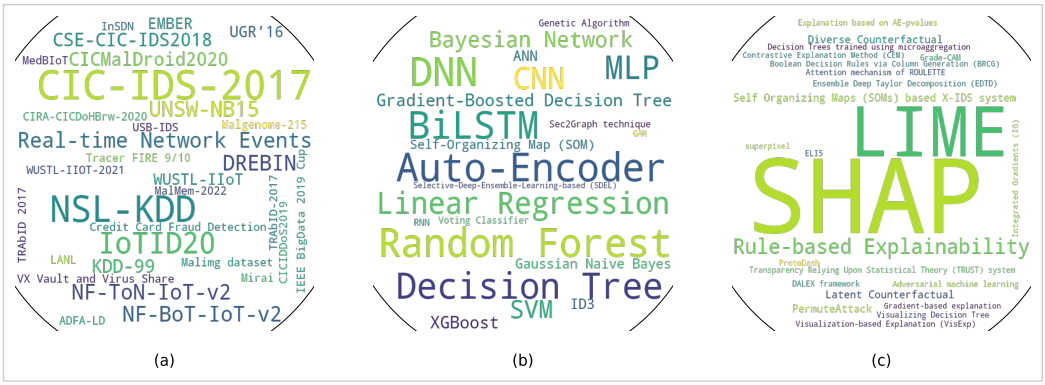

Figure 7. Distribution analysis of XAI-IDS research components. Word cloud visualizations synthesizing Table 2 analysis. (**a**) Dataset landscape showing prevalence of legacy benchmarks (e.g., NSL-KDD and KDD-99) versus emerging Industry 5.0-relevant datasets (e.g., IoTID20, TON-IoT, and WUSTL-IIOT). (**b**) Detection model preferences dominated by ensemble methods (Random Forest and Decision Trees) and deep learning (e.g., CNN, LSTM, and DNN). (**c**) XAI technique adoption led by SHAP and LIME as model-agnostic approaches. Word size indicates frequency across 135 studies.

Table 4 consolidates our findings, systematically categorizing XAI techniques by vulnerability level and protection mechanism. This analysis identifies a critical security–transparency trade-off: SHAP and gradient-based methods, while providing detailed feature importance essential to cybersecurity analysis, exhibit the highest vulnerability to adversarial exploitation, as documented through empirical attack success rates. Conversely, rule-based approaches demonstrate lower vulnerability but limited applicability to complex threat patterns, requiring deep learning architectures.

The implications for Industry 5.0 are substantial. The interconnected nature of Industry 5.0 systems amplifies risks from explainability transferability attacks, where insights

gained from one system's explanations can compromise related systems across organizational boundaries. Our analysis reveals that current XAI approaches, predominantly designed for general machine learning applications, prove inadequate for Industry 5.0's unique requirements: real-time collaboration demands sub-100 ms explanation generation, multi-stakeholder trust requires privacy-preserving explanation sharing, and adversarial resilience necessitates robust explanation mechanisms resistant to exploitation.

Critical research priorities emerge from this analysis: developing adversarially robust explainability architectures that maintain transparency without exposing exploitable vulnerabilities, creating cybersecurity-native explanation methods tailored to temporal and contextual threat characteristics, and establishing Industry 5.0-specific evaluation frameworks incorporating both detection performance and adversarial robustness metrics. The success of explainable cybersecurity in Industry 5.0 depends on resolving the fundamental challenge of maintaining transparency while ensuring security resilience in increasingly sophisticated threat landscapes.

7. XAI-Based IDSs: Lessons Learned, Challenges, and Future Research Directions

This systematic review of 135 studies spanning 2015–2024 provides comprehensive evidence on explainable AI implementation in Industry 5.0 cybersecurity environments. Our analysis demonstrates that while XAI-based intrusion detection systems have achieved significant advances in transparency and human–machine collaboration, they simultaneously introduce vulnerabilities that challenge traditional cybersecurity paradigms. The human-centric nature of Industry 5.0 amplifies both benefits and risks of explainable systems, creating requirements where security effectiveness and transparency must be carefully balanced.

7.1. Synthesis of Key Findings

Our investigation addresses five critical research questions, providing evidence-based insights that advance the understanding of XAI-based cybersecurity in Industry 5.0 contexts.

Research Question 1: Cybersecurity Challenges and X-IDS Necessity. Industry 5.0's interconnected ecosystem presents cybersecurity challenges through expanded attack surfaces, IoT vulnerabilities, cloud infrastructure risks, social engineering threats, and supply chain complexities. Our analysis demonstrates that traditional black-box IDS approaches prove inadequate for these multifaceted threats, particularly in human–machine collaborative environments where security decisions require transparent justification. X-IDSs become essential because they enable rapid human intervention, facilitate cross-organizational threat communication, and maintain trust in collaborative environments where humans must understand and act upon AI-generated security alerts.

Research Question 2: Transparency and Interpretability Enhancement. Our taxonomic analysis reveals that post hoc explainability methods, particularly SHAP and LIME, dominate current X-IDS implementations. These model-agnostic approaches prove effective for Industry 5.0 applications due to their adaptability across diverse ML architectures and ability to provide both local instance-level and global model-level explanations. However, our analysis identifies significant gaps in domain-specific explanation methods tailored to cybersecurity's temporal and contextual data characteristics.

Research Question 3: Primary Challenges and Limitations. The reviewed studies consistently identify four critical limitations: (1) computational overhead reducing real-time detection capabilities, (2) explanation accuracy degradation in complex attack scenarios, (3) lack of standardized evaluation metrics for cybersecurity-specific explanations, and

(4) insufficient consideration of human cognitive factors in explanation design. These limitations prove particularly pronounced in Industry 5.0's high-velocity, multi-stakeholder environments, where explanation delays can compromise incident response effectiveness.

Research Question 4: Adversarial Exploitation and Protection. Our analysis reveals competing requirements: explainability mechanisms essential to Industry 5.0 transparency create new attack vectors exploited by sophisticated adversaries. The literature documents the evolution from traditional black-box attacks to explainability-aware attacks, with adversaries leveraging explanation information to enhance attack effectiveness. This vulnerability proves particularly concerning in Industry 5.0's interconnected systems, where explanation insights can propagate across organizational boundaries, enabling coordinated multi-system attacks.

Research Question 5: Emerging Trends and Future Directions. The research trajectory demonstrates clear movement toward hybrid explainability approaches combining multiple techniques, increased focus on adversarial robustness, and growing emphasis on human-centric explanation design. However, significant gaps remain in Industry 5.0-specific methodologies, particularly for federated learning environments and multi-organizational threat correlation.

7.2. Industry 5.0-Specific Insights and Implications

Our analysis demonstrates that Industry 5.0's human-centric paradigm fundamentally transforms cybersecurity requirements, demanding explainable systems that balance transparency with security effectiveness. The collaborative nature of Industry 5.0 environments creates unique challenges where security decisions require human understanding and validation, making explainability operationally essential rather than merely beneficial.

The interconnected architecture characteristic of Industry 5.0 amplifies both XAI benefits and vulnerabilities. While explanation mechanisms enable effective human–AI collaboration and cross-system threat correlation, they simultaneously create attack propagation pathways exploitable by sophisticated adversaries. This contrasting characteristic necessitates specialized methodologies that maintain explanatory utility while mitigating adversarial risks.

Our evidence demonstrates that current XAI approaches, predominantly adapted from computer vision and natural language processing domains, prove suboptimal for Industry 5.0's cybersecurity requirements. The temporal dependencies, contextual relationships, and multi-modal data characteristics of cyber threats demand specialized explanation methods designed specifically for cybersecurity applications.

The scalability challenges identified in our analysis prove particularly critical to Industry 5.0's real-time operational requirements. Processing large-scale network data while generating meaningful explanations creates computational bottlenecks that can compromise threat detection effectiveness, highlighting the need for efficient explanation generation algorithms tailored to industrial environments.

7.3. Critical Research Gaps and Challenges

Despite significant advances, our analysis identifies persistent challenges that limit X-IDSs' effectiveness in Industry 5.0 environments. The fundamental trade-off between explainability and adversarial robustness remains unresolved, with no current approach successfully maintaining both transparency and security resilience. This challenge proves particularly significant in Industry 5.0's multi-stakeholder environments, where explanation sharing increases attack surface exposure.

The lack of standardized evaluation metrics for cybersecurity-specific explanations hampers systematic comparison and improvement of X-IDSs. Current evaluation ap-

proaches, borrowed from the general XAI literature, fail to capture cybersecurity-relevant factors such as explanation actionability for incident response, temporal consistency across attack progression, and resilience to adversarial manipulation.

Human factors considerations remain underdeveloped in current X-IDS research, with a limited understanding of how security analysts interpret and act upon AI-generated explanations under stress conditions typical of cyber incident response. This gap proves critical in Industry 5.0 environments where human–AI collaboration effectiveness directly impacts security outcomes.

7.4. Future Research Priorities for Industry 5.0

Based on our comprehensive analysis, we identify six critical research directions essential to advancing X-IDS effectiveness in Industry 5.0 environments:

- **Adversarially Robust Explainability Architectures**: Developing explanation mechanisms that maintain transparency while resisting adversarial exploitation through techniques such as explanation obfuscation, multi-level explanation hierarchies, and dynamic explanation strategies adapted to threat contexts.
- **Cybersecurity-Native Explanation Methods**: Creating explanation techniques designed specifically for cybersecurity's temporal, contextual, and multi-modal data characteristics, moving beyond adaptations of general-purpose XAI methods to domain-optimized approaches.
- **Human-Centric Explanation Design**: Developing explanation systems optimized for human cognitive factors, stress conditions, and decision-making requirements specific to cybersecurity incident response in Industry 5.0's collaborative environments.
- **Federated Explainable Learning Frameworks**: Establishing explanation methods for distributed learning environments that preserve privacy while enabling effective cross-organizational threat intelligence sharing essential for Industry 5.0's interconnected ecosystems. Federated explainable learning frameworks must address concrete Industry 5.0 deployment scenarios: distributed manufacturing networks where multiple facilities collaboratively train IDS models while maintaining proprietary data privacy, supply chain ecosystems requiring shared threat intelligence without exposing sensitive operational details, and edge–cloud hybrid architectures balancing local real-time detection with centralized model updates. Technical challenges include compact explanation representation for bandwidth-constrained industrial networks, privacy-preserving explanation aggregation techniques, and development of standardized explanation formats that ensure interoperability across heterogeneous industrial systems. These frameworks must demonstrate compliance with both cybersecurity requirements and industrial safety protocols to achieve practical deployment in human-centric industrial environments.
- **Real-Time Explanation Generation**: Advancing efficient algorithms capable of generating meaningful explanations within Industry 5.0's real-time operational constraints without compromising detection accuracy or explanation quality.
- **Standardized Evaluation Methodologies**: Developing comprehensive evaluation frameworks specifically designed for cybersecurity applications, incorporating metrics for explanation accuracy, actionability, temporal consistency, and adversarial robustness.

These research directions address the fundamental challenges identified throughout our analysis while recognizing Industry 5.0's unique requirements for human-centric, transparent, yet secure cybersecurity systems. Advancements in these areas will enable the development of XAI-based IDSs capable of supporting Industry 5.0's vision of collaborative human–machine intelligence while maintaining robust security postures against evolving cyber threats. The integration of these advances will prove essential to realizing Indus-

try 5.0's transformative potential while ensuring the cybersecurity foundation necessary for sustainable technological progress.

8. Conclusions

Machine learning offers substantial capabilities for addressing Industry 5.0's complex cybersecurity challenges, yet the opacity of these systems limits their adoption in critical security applications. Explainable AI has emerged as an essential solution, enabling transparency and trust in AI-based intrusion detection systems required for Industry 5.0's human-centric environments.

Our systematic analysis of 135 studies (Table 2) spanning 2015–2024 reveals that while XAI-based IDSs have achieved significant advances in transparency and interpretability—with SHAP and LIME emerging as dominant model-agnostic approaches (Figure 7c)—they simultaneously introduce fundamental vulnerabilities that sophisticated adversaries can exploit, as systematically documented across evasion, poisoning, and oracle attack categories (Table 3). This dual characteristic represents a critical challenge for Industry 5.0, where human–machine collaboration requires transparent security decisions while maintaining robust protection against evolving cyber threats.

Current explainability approaches, predominantly adapted from other domains, prove suboptimal for cybersecurity's unique temporal and contextual characteristics. Only 17% of reviewed studies (n = 23) employed Industry 5.0-specific datasets, while 35.6% relied on legacy benchmarks (e.g., NSL-KDD and KDD-99) that inadequately represent industrial contexts (Section 7.3, Figure 7a). Post hoc methods demonstrate superior adaptability, yet significant gaps persist in adversarial robustness, domain-specific techniques, and human-centric design.

A fundamental trade-off emerges between explainability benefits and security vulnerabilities: our vulnerability assessment reveals that gradient-based and SHAP methods—while providing detailed feature attribution essential to threat analysis—exhibit the highest vulnerability to adversarial exploitation. While XAI mechanisms prove essential to Industry 5.0's collaborative environments, they create new attack vectors that compromise system integrity. Addressing this challenge requires innovative approaches that maintain transparency while mitigating adversarial risks.

Future research must prioritize developing cybersecurity-native explainability methods, adversarially robust architectures, and human-centric explanation designs tailored to Industry 5.0's requirements. The six research directions identified in Section 7.4—including federated explainable learning frameworks, real-time explanation generation, and standardized evaluation methodologies—address critical gaps between current research capabilities and Industry 5.0 operational demands. Advancement in these areas will enable realization of Industry 5.0's transformative potential while ensuring the robust cybersecurity foundation necessary for sustainable technological progress in increasingly interconnected and collaborative industrial environments.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/info16121036/s1, PRISMA 2020 Checklist.

Author Contributions: Conceptualization, N.K.; methodology, K.A.; investigation, N.K.; data curation, N.K.; writing—original draft preparation, K.A. and M.M.A.; writing—review and editing, K.A., M.M.A. and I.K.; formal analysis, A.A.T. and M.M.A.; validation, M.M.A. and I.K.; visualization, N.K.; resources, A.A.T., A.B. and I.K.; supervision, M.M.A., A.B. and I.K.; project administration, M.M.A.; funding acquisition, A.B. All authors have read and agreed to the published version of the manuscript.

Funding: Open Access funding provided by the Qatar National Library.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable as no datasets were generated or analyzed during the current study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Speith, T. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; pp. 2239–2250.
2. Alsamhi, S.H.; Shvetsov, A.V.; Hawbani, A.; Shvetsova, S.V.; Kumar, S.; Zhao, L. Survey on Federated Learning enabling indoor navigation for industry 4.0 in B5G. *Future Gener. Comput. Syst.* **2023**, *148*, 250–265. [CrossRef]
3. Rane, N.L.; Kaya, Ö.; Rane, J. *Artificial Intelligence, Machine Learning, and Deep Learning for Sustainable Industry 5.0*; Deep Science Publishing: San Francisco, CA, USA, 2024.
4. Holzinger, A.; Saranti, A.; Molnar, C.; Biecek, P.; Samek, W. Explainable AI methods—A brief overview. In Proceedings of the International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, Vienna, Austria, 18 July 2022; Springer: Cham, Switzerland, 2022; pp. 13–38.
5. Gunning, D.; Aha, D. DARPA's explainable artificial intelligence (XAI) program. *AI Mag.* **2019**, *40*, 44–58.
6. Alexandrov, N. Explainable AI decisions for human-autonomy interactions. In Proceedings of the 17th AIAA Aviation Technology, Integration, and Operations Conference, Denver, Colorado, 5–9 June 2017; p. 3991.
7. Capuano, N.; Fenza, G.; Loia, V.; Stanzione, C. Explainable Artificial Intelligence in CyberSecurity: A Survey. *IEEE Access* **2022**, *10*, 93575–93600. [CrossRef]
8. Yayla, A.; Haghnegahdar, L.; Dincelli, E. Explainable artificial intelligence for smart grid intrusion detection systems. *IT Prof.* **2022**, *24*, 18–24. [CrossRef]
9. Scalas, M.; Rieck, K.; Giacinto, G. Improving malware detection with explainable machine learning. In *Explainable Deep Learning AI*; Elsevier: Amsterdam, The Netherlands, 2023; pp. 217–238.
10. Marino, D.L.; Wickramasinghe, C.S.; Manic, M. An Adversarial Approach for Explainable AI in Intrusion Detection Systems. In Proceedings of the IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society, Washington, DC, USA, 21–23 October 2018; pp. 3237–3243.
11. Baniecki, H.; Biecek, P. Adversarial Attacks and Defenses in Explainable Artificial Intelligence: A Survey. *arXiv* **2023**, arXiv:2306.06123. [CrossRef]
12. Sharma, D.K.; Mishra, J.; Singh, A.; Govil, R.; Srivastava, G.; Lin, J.C.W. Explainable Artificial Intelligence for Cybersecurity. *Comput. Electr. Eng.* **2022**, *103*, 108356. [CrossRef]
13. Yu, J.; Shvetsov, A.V.; Alsamhi, S.H. Leveraging machine learning for cybersecurity resilience in industry 4.0: Challenges and future directions. *IEEE Access* **2024**, *12*, 159579–159596. [CrossRef]
14. Kiran, A.; Prakash, S.W.; Kumar, B.A.; Sameeratmaja, T.; Charan, U.S.S.R. Intrusion Detection System Using Machine Learning. In Proceedings of the 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 23–25 January 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–4.
15. Salam, A.; Ullah, F.; Amin, F.; Abrar, M. Deep learning techniques for web-based attack detection in industry 5.0: A novel approach. *Technologies* **2023**, *11*, 107. [CrossRef]
16. Gadekallu, T.R.; Maddikunta, P.K.R.; Boopathy, P.; Deepa, N.; Chengoden, R.; Victor, N.; Wang, W.; Wang, W.; Zhu, Y.; Dev, K. Xai for industry 5.0-concepts, opportunities, challenges and future directions. *IEEE Open J. Commun. Soc.* **2024**, *6*, 2706–2729. [CrossRef]
17. Hussain, F.; Hussain, R.; Hassan, S.A.; Hossain, E. Machine learning in IoT security: Current solutions and future challenges. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 1686–1721. [CrossRef]
18. Chou, D.; Jiang, M. A survey on data-driven network intrusion detection. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–36. [CrossRef]
19. Alazab, M.; KP, S.; Srinivasan, S.; Venkatraman, S.; Pham, Q.V.; Simran. Deep Learning for Cyber Security Applications: A Comprehensive Survey. 2021. Available online: https://d197for5662m48.cloudfront.net/documents/publicationstatus/162402/preprint_pdf/1db580dedba693379c64ee6ebfbf4792.pdf (accessed on 30 August 2025).
20. Markevych, M.; Dawson, M. A review of enhancing intrusion detection systems for cybersecurity using artificial intelligence (ai). In *Proceedings of the International Conference Knowledge-Based Organization*; Paradigm: Boston, UK, 2023; Volume 29, pp. 30–37.
21. Sowmya, T.; Anita, E.M. A comprehensive review of AI based intrusion detection system. *Meas. Sens.* **2023**, *28*, 100827. [CrossRef]
22. Sauka, K.; Shin, G.Y.; Kim, D.W.; Han, M.M. Adversarial robust and explainable network intrusion detection systems based on deep learning. *Appl. Sci.* **2022**, *12*, 6451. [CrossRef]

23. Maddikunta, P.K.R.; Pham, Q.V.; Prabadevi, B.; Deepa, N.; Dev, K.; Gadekallu, T.R.; Ruby, R.; Liyanage, M. Industry 5.0: A survey on enabling technologies and potential applications. *J. Ind. Inf. Integr.* **2022**, *26*, 100257. [CrossRef]
24. Czeczot, G.; Rojek, I.; Mikołajewski, D.; Sangho, B. AI in IIoT Management of Cybersecurity for Industry 4.0 and Industry 5.0 Purposes. *Electronics* **2023**, *12*, 3800. [CrossRef]
25. Taj, I.; Zaman, N. Towards industrial revolution 5.0 and explainable artificial intelligence: Challenges and opportunities. *Int. J. Comput. Digit. Syst.* **2022**, *12*, 295–320. [CrossRef]
26. Bobek, S.; Nowaczyk, S.; Gama, J.; Pashami, S.; Ribeiro, R.P.; Taghiyarrenani, Z.; Veloso, B.; Rajaoarisoa, L.H.; Szelazek, M.; Nalepa, G.J. Why Industry 5.0 Needs XAI 2.0? In Proceedings of the xAI (Late-Breaking Work, Demos, Doctoral Consortium), Lisbon, Portugal, 26–28 July 2023; pp. 1–6.
27. Rane, N. ChatGPT and similar Generative Artificial Intelligence (AI) for building and construction industry: Contribution, Opportunities and Challenges of large language Models for Industry 4.0, Industry 5.0, and Society 5.0. *Oppor. Chall. Large Lang. Model. Ind.* **2023**, *4*. [CrossRef]
28. Moosavi, S.; Farajzadeh-Zanjani, M.; Razavi-Far, R.; Palade, V.; Saif, M. Explainable AI in Manufacturing and Industrial Cyber–Physical Systems: A Survey. *Electronics* **2024**, *13*, 3497. [CrossRef]
29. Beg, O.A.; Khan, A.A.; Rehman, W.U.; Hassan, A. A Review of AI-Based Cyber-Attack Detection and Mitigation in Microgrids. *Energies* **2023**, *16*, 7644. [CrossRef]
30. Habib, G.; Qureshi, S. XAI and Machine Learning for Cyber Security: A Systematic Review. In *Medical Data Analysis and Processing Using Explainable Artificial Intelligence*; CRC Press: Boca Raton, FL, USA, 2023; pp. 91–104.
31. Bac, T.P.; Ha, D.T.; Tran, K.D.; Tran, K.P. Explainable Articial Intelligence for Cybersecurity in Smart Manufacturing. In *Artificial Intelligence for Smart Manufacturing: Methods, Applications, and Challenges*; Springer: Cham, Switzerland, 2023; pp. 199–223.
32. Ahmad, I.; Rodriguez, F.; Kumar, T.; Suomalainen, J.; Jagatheesaperumal, S.K.; Walter, S.; Asghar, M.Z.; Li, G.; Papakonstantinou, N.; Ylianttila, M.; et al. Communications security in Industry X: A survey. *IEEE Open J. Commun. Soc.* **2024**, *5*, 982–1025. [CrossRef]
33. Bhattacharya, P.; Obaidat, M.S.; Sanghavi, S.; Sakariya, V.; Tanwar, S.; Hsiao, K.F. Internet-of-explainable-digital-twins: A case study of versatile corn production ecosystem. In Proceedings of the 2022 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI), Dalian, China, 17–19 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–5.
34. Minh, D.; Wang, H.X.; Li, Y.F.; Nguyen, T.N. Explainable artificial intelligence: A comprehensive review. *Artif. Intell. Rev.* **2022**, *55*, 3503–3568. [CrossRef]
35. Buijsman, S. Defining explanation and explanatory depth in XAI. *Minds Mach.* **2022**, *32*, 563–584. [CrossRef]
36. Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; et al. Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Comput. Surv.* **2023**, *55*, 1–33. [CrossRef]
37. Yang, W.; Wei, Y.; Wei, H.; Chen, Y.; Huang, G.; Li, X.; Li, R.; Yao, N.; Wang, X.; Gu, X.; et al. Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. *Hum.-Centric Intell. Syst.* **2023**, *3*, 161–188. [CrossRef]
38. Chandre, P.R.; Vanarote, V.; Patil, R.; Mahalle, P.N.; Shinde, G.R.; Nimbalkar, M.; Barot, J. Explainable AI for Intrusion Prevention: A Review of Techniques and Applications. In *Proceedings of the International Conference on Information and Communication Technology for Intelligent Systems*; Springer: Singapore, 2023; pp. 339–350.
39. Moustafa, N.; Koroniotis, N.; Keshk, M.; Zomaya, A.Y.; Tari, Z. Explainable Intrusion Detection for Cyber Defences in the Internet of Things: Opportunities and Solutions. *IEEE Commun. Surv. Tutor.* **2023**, *25*, 1775–1807. [CrossRef]
40. Ahmad, K.; Maabreh, M.; Ghaly, M.; Khan, K.; Qadir, J.; Al-Fuqaha, A. Developing future human-centered smart cities: Critical analysis of smart city security, Data management, and Ethical challenges. *Comput. Sci. Rev.* **2022**, *43*, 100452. [CrossRef]
41. Nwakanma, C.I.; Ahakonye, L.A.C.; Njoku, J.N.; Odirichukwu, J.C.; Okolie, S.A.; Uzondu, C.; Ndubuisi Nweke, C.C.; Kim, D.S. Explainable artificial intelligence (xai) for intrusion detection and mitigation in intelligent connected vehicles: A review. *Appl. Sci.* **2023**, *13*, 1252. [CrossRef]
42. Castro, O.E.L.; Deng, X.; Park, J.H. Comprehensive survey on AI-based technologies for enhancing IoT privacy and security: Trends, challenges, and solutions. *Hum.-Centric Comput. Inf. Sci.* **2023**, *13*, 39.
43. Alsamhi, S.H.; Hawbani, A.; Sahal, R.; Srivastava, S.; Kumar, S.; Zhao, L.; Al-qaness, M.A.; Hassan, J.; Guizani, M.; Curry, E. Towards sustainable industry 4.0: A survey on greening IoE in 6G networks. *Ad Hoc Netw.* **2024**, *165*, 103610. [CrossRef]
44. Sinclair, C.; Pierce, L.; Matzner, S. An application of machine learning to network intrusion detection. In Proceedings of the Proceedings 15th Annual Computer Security Applications Conference (ACSAC'99), Phoenix, AZ, USA, 6–10 December 1999; IEEE: Piscataway, NJ, USA, 1999; pp. 371–377.
45. Mahbooba, B.; Timilsina, M.; Sahal, R.; Serrano, M. Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity* **2021**, *2021*, 6634811. [CrossRef]
46. Manoj, V.; Wenda, S.; Sihan, N.; Rouff, C.; Watkins, L.; Rubin, A. Explainable Autonomic Cybersecurity For Industrial Control Systems. In Proceedings of the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), Vegas, NV, USA, 8–11 March 2023 IEEE: Piscataway, NJ, USA, 2023; pp. 0900–9006.

47. Fazzolari, M.; Ducange, P.; Marcelloni, F. An Explainable Intrusion Detection System for IoT Networks. In Proceedings of the 2023 IEEE International Conference on Fuzzy Systems (FUZZ), Songdo Incheon, Republic of Korea, 13–17 August 2023; pp. 1–6. [CrossRef]
48. Yang, B.; Hoffman, M.; Brown, N.J. Bayesian Networks for Interpretable Cyberattack Detection. In Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS-56), Online, 3–7 January 2023; p. 3.
49. Zolanvari, M.; Yang, Z.; Khan, K.; Jain, R.; Meskin, N. TRUST XAI: Model-Agnostic Explanations for AI with a Case Study on IIoT Security. *IEEE Internet Things J.* **2021**, *10*, 2967–2978. [CrossRef]
50. Le, T.T.H.; Kim, H.; Kang, H.; Kim, H. Classification and explanation for intrusion detection system based on ensemble trees and SHAP method. *Sensors* **2022**, *22*, 1154. [CrossRef] [PubMed]
51. Alani, M.M.; Mashatan, A.; Miri, A. XMal: A lightweight memory-based explainable obfuscated-malware detector. *Comput. Secur.* **2023**, *133*, 103409. [CrossRef]
52. Alani, M.M.; Miri, A. Towards an explainable universal feature set for IoT intrusion detection. *Sensors* **2022**, *22*, 5690. [CrossRef]
53. Alani, M.M. An explainable efficient flow-based Industrial IoT intrusion detection system. *Comput. Electr. Eng.* **2023**, *108*, 108732. [CrossRef]
54. Gürbüz, E.; Turgut, Ö.; Kök, İ. Explainable AI-Based Malicious Traffic Detection and Monitoring System in Next-Gen IoT Healthcare. In Proceedings of the 2023 International Conference on Smart Applications, Communications and Networking (SmartNets), Istanbul, Turkey, 25–27 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6.
55. Patil, S.; Varadarajan, V.; Mazhar, S.M.; Sahibzada, A.; Ahmed, N.; Sinha, O.; Kumar, S.; Shaw, K.; Kotecha, K. Explainable Artificial Intelligence for Intrusion Detection System. *Electronics* **2022**, *11*, 3079. [CrossRef]
56. Zebin, T.; Rezvy, S.; Luo, Y. An explainable AI-based intrusion detection system for DNS over HTTPS (DoH) Attacks. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 2339–2349. [CrossRef]
57. Sivamohan, S.; Sridhar, S. An optimized model for network intrusion detection systems in industry 4.0 using XAI based Bi-LSTM framework. *Neural Comput. Appl.* **2023**, *35*, 11459–11475. [CrossRef]
58. Wang, Y.; Xu, L.; Liu, W.; Li, R.; Gu, J. Network intrusion detection based on explainable artificial intelligence. *Wirel. Pers. Commun.* **2023**, *131*, 1115–1130. [CrossRef]
59. Tanuwidjaja, H.C.; Takahashi, T.; Lin, T.N.; Lee, B.; Ban, T. Hybrid Explainable Intrusion Detection System: Global vs. Local Approach. In Proceedings of the 2023 Workshop on Recent Advances in Resilient and Trustworthy ML Systems in Autonomous Networks, Copenhagen, Denmark, 30 November 2023; pp. 37–42.
60. Mills, A.; Spyridopoulos, T.; Legg, P. Efficient and interpretable real-time malware detection using random-forest. In Proceedings of the 2019 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (Cyber SA), Oxford, UK, 3-4 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–8.
61. Ables, J.; Kirby, T.; Anderson, W.; Mittal, S.; Rahimi, S.; Banicescu, I.; Seale, M. Creating an Explainable Intrusion Detection System Using Self Organizing Maps. In Proceedings of the 2022 IEEE Symposium Series on Computational Intelligence (SSCI), Singapore, 4–7 December 2022; pp. 404–412. [CrossRef]
62. Ables, J.; Kirby, T.; Mittal, S.; Banicescu, I.; Rahimi, S.; Anderson, W.; Seale, M. Explainable Intrusion Detection Systems Using Competitive Learning Techniques. *arXiv* **2023**, arXiv:2303.17387. [CrossRef]
63. Lundberg, H.; Mowla, N.I.; Abedin, S.F.; Thar, K.; Mahmood, A.; Gidlund, M.; Raza, S. Experimental Analysis of Trustworthy In-Vehicle Intrusion Detection System Using eXplainable Artificial Intelligence (XAI). *IEEE Access* **2022**, *10*, 102831–102841. [CrossRef]
64. AL-Essa, M.; Andresini, G.; Appice, A.; Malerba, D. Xai to explore robustness of features in adversarial training for cybersecurity. In Proceedings of the International Symposium on Methodologies for Intelligent Systems, Cosenza, Italy, 3–5 October 2022; Springer: Cham, Switzerland, 2022; pp. 117–126.
65. Al-Essa, M.; Andresini, G.; Appice, A.; Malerba, D. An XAI-based adversarial training approach for cyber-threat detection. In Proceedings of the 2022 IEEE International Conference on Dependable, Autonomic and Secure Computing (DASC), Pervasive Intelligence and Computing (PICom), Cloud & Big Data Computing (CBDCom) & Cyber Science & Technology Congress (CyberSciTech), Falerna, Calabria, Italy, 12–15 September 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–8.
66. Lanfer, E.; Sylvester, S.; Aschenbruck, N.; Atzmueller, M. Leveraging Explainable AI Methods Towards Identifying Classification Issues on IDS Datasets. In Proceedings of the 2023 IEEE 48th Conference on Local Computer Networks (LCN), Daytona Beach, Fl, USA, 2–5 October 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–4.
67. Sharma, B.; Sharma, L.; Lal, C.; Roy, S. Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach. *Expert Syst. Appl.* **2024**, *238*, 121751. [CrossRef]
68. Lu, Z.; Thing, V.L.L. "How Does It Detect a Malicious App?" Explaining the Predictions of AI-Based Malware Detector. In Proceedings of the 2022 IEEE 8th International Conference on Big Data Security on Cloud (BigDataSecurity), High Performance and Smart Computing (HPSC), and Intelligent Data and Security (IDS), Jinan, China, 6–8 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 194–199.

69. Sarhan, M.; Layeghy, S.; Portmann, M. Evaluating Standard Feature Sets Towards Increased Generalisability and Explainability of ML-Based Network Intrusion Detection. *Big Data Res.* **2022**, *30*, 100359. [CrossRef]
70. Oseni, A.; Moustafa, N.; Creech, G.; Sohrabi, N.; Strelzoff, A.; Tari, Z.; Linkov, I. An Explainable Deep Learning Framework for Resilient Intrusion Detection in IoT-Enabled Transportation Networks. *IEEE Trans. Intell. Transp. Syst.* **2022**, *24*, 1000–1014. [CrossRef]
71. Alani, M.M.; Awad, A.I.; Barka, E. ARP-PROBE: An ARP spoofing detector for Internet of Things networks using explainable deep learning. *Internet Things* **2023**, *23*, 100861. [CrossRef]
72. Kalutharage, C.S.; Liu, X.; Chrysoulas, C.; Pitropakis, N.; Papadopoulos, P. Explainable AI-based DDOS attack identification method for IoT networks. *Computers* **2023**, *12*, 32. [CrossRef]
73. Muna, R.K.; Hossain, M.I.; Alam, M.G.R.; Hassan, M.M.; Ianni, M.; Fortino, G. Demystifying machine learning models of massive IoT attack detection with Explainable AI for sustainable and secure future smart cities. *Internet Things* **2023**, *24*, 100919. [CrossRef]
74. Abou El Houda, Z.; Brik, B.; Senouci, S.M. A novel iot-based explainable deep learning framework for intrusion detection systems. *IEEE Internet Things Mag.* **2022**, *5*, 20–23. [CrossRef]
75. da Silveira Lopes, R.; Duarte, J.C.; Goldschmidt, R.R. False Positive Identification in Intrusion Detection Using XAI. *IEEE Lat. Am. Trans.* **2023**, *21*, 745–751. [CrossRef]
76. Szczepański, M.; Choraś, M.; Pawlicki, M.; Kozik, R. Achieving explainability of intrusion detection system by hybrid oracle-explainer approach. In Proceedings of the 2020 International Joint Conference on neural networks (IJCNN), Glasgow, UK, 19–24 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–8.
77. Wang, M.; Zheng, K.; Yang, Y.; Wang, X. An explainable machine learning framework for intrusion detection systems. *IEEE Access* **2020**, *8*, 73127–73141. [CrossRef]
78. Nguyen, T.L.; Nguyen, X.H.; Le, K.H. Enhancing Explainability of Machine Learning-based Intrusion Detection Systems. In Proceedings of the 2022 RIVF International Conference on Computing and Communication Technologies (RIVF), Ho Chi Minh, Vietnam, 20–22 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 606–611.
79. Siganos, M.; Radoglou-Grammatikis, P.; Kotsiuba, I.; Markakis, E.; Moscholios, I.; Goudos, S.; Sarigiannidis, P. Explainable AI-based Intrusion Detection in the Internet of Things. In Proceedings of the 18th International Conference on Availability, Reliability and Security, Benevento, Italy, 29 August–1 September 2023; ACM: New York, NY, USA, 2023; ARES '23. [CrossRef]
80. Roy, S.; Li, J.; Pandey, V.; Bai, Y. An Explainable Deep Neural Framework for Trustworthy Network Intrusion Detection. In Proceedings of the 2022 10th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud), San Francisco, CA, USA, 15–18 August 2022; IEEE: Piscataway, NJ, USA, 2022, pp. 25–30.
81. Das, S.; Shiva, S. Machine Learning application lifecycle augmented with explanation and security. In Proceedings of the 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA,1–4 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 0171–0177.
82. Mane, S.; Rao, D. Explaining network intrusion detection system using explainable AI framework. *arXiv* **2021**, arXiv:2103.07110. [CrossRef]
83. Khan, I.A.; Moustafa, N.; Pi, D.; Sallam, K.M.; Zomaya, A.Y.; Li, B. A New Explainable Deep Learning Framework for Cyber Threat Discovery in Industrial IoT Networks. *IEEE Internet Things J.* **2021**, *9*, 11604–11613. [CrossRef]
84. Ha, D.T.; Hoang, N.X.; Hoang, N.V.; Du, N.H.; Huong, T.T.; Tran, K.P. Explainable anomaly detection for industrial control system cybersecurity. *IFAC-PapersOnLine* **2022**, *55*, 1183–1188. [CrossRef]
85. Nguyen, Q.P.; Lim, K.W.; Divakaran, D.M.; Low, K.H.; Chan, M.C. GEE: A gradient-based explainable variational autoencoder for network anomaly detection. In Proceedings of the 2019 IEEE Conference on Communications and Network Security (CNS), Washington, DC, USA, 10–12 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 91–99.
86. Antwarg, L.; Miller, R.M.; Shapira, B.; Rokach, L. Explaining anomalies detected by autoencoders using Shapley Additive Explanations. *Expert Syst. Appl.* **2021**, *186*, 115736. [CrossRef]
87. Aguilar, D.L.; Medina-Perez, M.A.; Loyola-Gonzalez, O.; Choo, K.K.R.; Bucheli-Susarrey, E. Towards an interpretable autoencoder: A decision-tree-based autoencoder and its application in anomaly detection. *IEEE Trans. Dependable Secur. Comput.* **2022**, *20*, 1048–1059. [CrossRef]
88. Lanvin, M.; Gimenez, P.F.; Han, Y.; Majorczyk, F.; Mé, L.; Totel, E. Towards Understanding Alerts raised by Unsupervised Network Intrusion Detection Systems. In Proceedings of the The 26th International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2023), Hong Kong, China, 16–18 October 2023.
89. Javeed, D.; Gao, T.; Kumar, P.; Jolfaei, A. An explainable and resilient intrusion detection system for industry 5.0. *IEEE Trans. Consum. Electron.* **2023**, *70*, 1342–1350. [CrossRef]
90. Shoukat, S.; Gao, T.; Javeed, D.; Saeed, M.S.; Adil, M. Trust my IDS: An explainable AI integrated deep learning-based transparent threat detection system for industrial networks. *Comput. Secur.* **2025**, *149*, 104191. [CrossRef]
91. Lin, Y.; Chang, X. Towards interpretable ensemble learning for image-based malware detection. *arXiv* **2021**, arXiv:2101.04889. [CrossRef]

92. Iadarola, G.; Martinelli, F.; Mercaldo, F.; Santone, A. Towards an interpretable deep learning model for mobile malware detection and family identification. *Comput. Secur.* **2021**, *105*, 102198. [CrossRef]
93. Andresini, G.; Appice, A.; Caforio, F.P.; Malerba, D.; Vessio, G. ROULETTE: A neural attention multi-output model for explainable network intrusion detection. *Expert Syst. Appl.* **2022**, *201*, 117144. [CrossRef]
94. Piplai, A.; Chukkapalli, S.S.L.; Joshi, A. NAttack! Adversarial Attacks to bypass a GAN based classifier trained to detect Network intrusion. In Proceedings of the 2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), Baltimore, MD, USA, 25–27 May 2020; pp. 49–54. [CrossRef]
95. Ayub, M.A.; Johnson, W.A.; Talbert, D.A.; Siraj, A. Model evasion attack on intrusion detection systems using adversarial machine learning. In Proceedings of the 2020 54th Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA, 18–20 March 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
96. Alshahrani, E.; Alghazzawi, D.; Alotaibi, R.; Rabie, O. Adversarial attacks against supervised machine learning based network intrusion detection systems. *PLoS ONE* **2022**, *17*, e0275971. [CrossRef]
97. Duy, P.T.; Khoa, N.H.; Hien, D.T.T.; Do Hoang, H.; Pham, V.-H. Investigating on the robustness of flow-based intrusion detection system against adversarial samples using Generative Adversarial Networks. *J. Inf. Secur. Appl.* **2023**, *74*, 103472. [CrossRef]
98. Zhang, S.; Xie, X.; Xu, Y. A brute-force black-box method to attack machine learning-based systems in cybersecurity. *IEEE Access* **2020**, *8*, 128250–128263. [CrossRef]
99. Lan, T.; Demetrio, L.; Nait-Abdesselam, F.; Han, Y.; Aonzo, S. Trust Under Siege: Label Spoofing Attacks against Machine Learning for Android Malware Detection. *arXiv* **2025**, arXiv:2503.11841.
100. Qiu, H.; Dong, T.; Zhang, T.; Lu, J.; Memmi, G.; Qiu, M. Adversarial attacks against network intrusion detection in IoT systems. *IEEE Internet Things J.* **2020**, *8*, 10327–10335. [CrossRef]
101. Chen, J.; Wu, D.; Zhao, Y.; Sharma, N.; Blumenstein, M.; Yu, S. Fooling intrusion detection systems using adversarially autoencoder. *Digit. Commun. Netw.* **2021**, *7*, 453–460. [CrossRef]
102. Jiang, H.; Lin, J.; Kang, H. FGMD: A robust detector against adversarial attacks in the IoT network. *Future Gener. Comput. Syst.* **2022**, *132*, 194–210. [CrossRef]
103. Ravikrishnan, B.; Sriram, I.; Mahadevan, S. ARDL-IDS: Adversarial Resilience in Deep Learning-based Intrusion Detection Systems. In Proceedings of the 2023 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), Chennai, India, 29–31 March 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
104. Li, S.; Wang, J.; Wang, Y.; Zhou, G.; Zhao, Y. EIFDAA: Evaluation of an IDS with function-discarding adversarial attacks in the IIoT. *Heliyon* **2023**, *9*, e13520. [CrossRef]
105. Debicha, I.; Cochez, B.; Kenaza, T.; Debatty, T.; Dricot, J.M.; Mees, W. Adv-Bot: Realistic adversarial botnet attacks against network intrusion detection systems. *Comput. Secur.* **2023**, *129*, 103176. [CrossRef]
106. Li, P.; Liu, Q.; Zhao, W.; Wang, D.; Wang, S. Bebp: An poisoning method against machine learning based idss. *arXiv* **2018**, arXiv:1803.03965. [CrossRef]
107. Xu, J.; Wen, Y.; Yang, C.; Meng, D. An approach for poisoning attacks against rnn-based cyber anomaly detection. In Proceedings of the 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Guangzhou, China, 29 December 2020–1 January 2021; IEEE: Piscataway, NJ, USA, 2020; pp. 1680–1687.
108. Nguyen, T.D.; Rieger, P.; Miettinen, M.; Sadeghi, A.R. Poisoning Attacks on Federated Learning-Based IoT Intrusion Detection System. In Proceedings of the Workshop on Decentralized IoT Systems and Security (DISS), San Diego, CA, USA, 23–26 February 2020; pp. 1–7. Available online: https://www.ndss-symposium.org/wp-content/uploads/2020/04/diss2020-23003-paper.pdf (accessed on 30 August 2025).
109. Rosenberg, I.; Meir, S.; Berrebi, J.; Gordon, I.; Sicard, G.; David, E.O. Generating end-to-end adversarial examples for malware classifiers using explainability. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–10.
110. Zhang, H.; Han, D.; Liu, Y.; Wang, Z.; Sun, J.; Zhuang, S.; Liu, J.; Dong, J. Explainable and Transferable Adversarial Attack for ML-Based Network Intrusion Detectors. *arXiv* **2024**, arXiv:2401.10691. [CrossRef]
111. Kuppa, A.; Le-Khac, N.A. Black box attacks on explainable artificial intelligence (XAI) methods in cyber security. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–8.
112. Kuppa, A.; Le-Khac, N.A. Adversarial XAI Methods in Cybersecurity. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 4924–4938. [CrossRef]
113. Severi, G.; Meyer, J.; Coull, S.; Oprea, A. {Explanation-Guided} backdoor poisoning attacks against malware classifiers. In Proceedings of the 30th USENIX Security Symposium (USENIX Security 21), Online, 11–13 August 2021; pp. 1487–1504.

114. Okada, S.; Jmila, H.; Akashi, K.; Mitsunaga, T.; Sekiya, Y.; Takase, H.; Blanc, G.; Nakamura, H. XAI-driven adversarial attacks on network intrusion detectors. In Proceedings of the 2024 European Interdisciplinary Cybersecurity Conference, Xanthi, Greece, 5–6 June 2024; pp. 65–73.
115. Sun, R.; Guo, S.; Guo, J.; Xing, C.; Yang, L.; Guo, X.; Pan, Z. Instance Attack: An Explanation-based Vulnerability Analysis Framework Against DNNs for Malware Detection. *arXiv* **2022**, arXiv:2209.02453. [CrossRef]
116. Hoang, N.V.; Trung, N.D.; Trung, D.M.; Duy, P.T.; Pham, V.H. ADV-Sword: A Framework of Explainable AI-Guided Adversarial Samples Generation for Benchmarking ML-Based Intrusion Detection Systems. In Proceedings of the 2024 International Conference on Advanced Technologies for Communications (ATC), Ho Chi Minh City, Vietnam, 17–19 October 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 885–890.
117. Alani, M.M.; Mashatan, A.; Miri, A. Adversarial Explainability: Utilizing Explainable Machine Learning in Bypassing IoT Botnet Detection Systems. *arXiv* **2023**, arXiv:2310.00070. [CrossRef]
118. Shu, Z.; Yan, G. EAGLE: Evasion attacks guided by local explanations against Android malware classification. *IEEE Trans. Dependable Secur. Comput.* **2023**, *21*, 3165–3182. [CrossRef]
119. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
120. Sarkar, A.; Vijaykeerthy, D.; Sarkar, A.; Balasubramanian, V.N. A Framework for Learning Ante-hoc Explainable Models via Concepts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10286–10295.
121. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
122. Islam, S.R.; Eberle, W.; Ghafoor, S.K.; Ahmed, M. Explainable artificial intelligence approaches: A survey. *arXiv* **2021**, arXiv:2101.09429. [CrossRef]
123. Hanif, A.; Zhang, X.; Wood, S. A Survey on Explainable Artificial Intelligence Techniques and Challenges. In Proceedings of the 2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW), Gold Coast, Australia, 25–29 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 81–89.
124. Došilović, F.K.; Brčić, M.; Hlupić, N. Explainable artificial intelligence: A survey. In Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 21–25 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 0210–0215.
125. Liu, Q.; Hagenmeyer, V.; Keller, H.B. A review of rule learning-based intrusion detection systems and their prospects in smart grids. *IEEE Access* **2021**, *9*, 57542–57564. [CrossRef]
126. van der Velden, B.H. Explainable AI: Current status and future potential. *Eur. Radiol.* **2023**, *34*, 1187–1189. [CrossRef] [PubMed]
127. Kalasampath, K.; Spoorthi, K.; Sajeev, S.; Kuppa, S.S.; Ajay, K.; Angulakshmi, M. A Literature review on applications of explainable artificial intelligence (XAI). *IEEE Access* **2025**, *13*, 41111–41140. [CrossRef]
128. Bobek, S.; Nalepa, G.J. Local universal rule-based explainer (lux). *SoftwareX* **2025**, *30*, 102102. [CrossRef]
129. Vale, D.; El-Sharif, A.; Ali, M. Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. *AI Ethics* **2022**, *2*, 815–826. [CrossRef]
130. Covert, I.; Lundberg, S.; Lee, S.I. Explaining by removing: A unified framework for model explanation. *J. Mach. Learn. Res.* **2021**, *22*, 1–90.
131. Confalonieri, R.; Coba, L.; Wagner, B.; Besold, T.R. A historical perspective of explainable Artificial Intelligence. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2021**, *11*, e1391. [CrossRef]
132. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; KDD '16, pp. 1135–1144. [CrossRef]
133. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
134. Alenezi, R.; Ludwig, S.A. Explainability of cybersecurity threats data using SHAP. In Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence (SSCI), Virtual, 5–7 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–10.
135. Belle, V.; Papantonis, I. Principles and practice of explainable machine learning. *Front. Big Data* **2021**, *4*, 688969. [CrossRef] [PubMed]
136. Kim, S.; Jeong, M.; Ko, B.C. Lightweight surrogate random forest support for model simplification and feature relevance. *Appl. Intell.* **2022**, *52*, 471–481. [CrossRef]

137. Tan, S.; Caruana, R.; Hooker, G.; Lou, Y. Distill-and-compare: Auditing black-box models using transparent model distillation. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018; pp. 303–310.
138. Liu, S. Improved model search based on distillation framework. In Proceedings of the 2nd International Conference on Computer Vision, Image, and Deep Learning, Liuzhou, China, 25–27 June 2021; SPIE: Bellingham, WA, USA, 2021; Volume 11911; pp. 399–406.
139. Johansson, U.; Konig, R.; Niklasson, L. Inconsistency-friend or foe. In Proceedings of the 2007 International Joint Conference on Neural Networks, Orlando, FL, USA, 12–17 August 2007; IEEE: Piscataway, NJ, USA, 2007; pp. 1383–1388.
140. Konig, R.; Johansson, U.; Niklasson, L. G-REX: A versatile framework for evolutionary data mining. In Proceedings of the 2008 IEEE International Conference on Data Mining Workshops, Pisa, Italy, 15–19 December; IEEE: Piscataway, NJ, USA, 2008; pp. 971–974.
141. Hassan, M.A.; Zardari, S.; Farooq, M.U.; Alansari, M.M.; Nagro, S.A. Systematic Analysis of Risks in Industry 5.0 Architecture. *Appl. Sci.* **2024**, *14*, 1466. [CrossRef]
142. Rajabion, L. Industry 5.0 and Cyber Crime Security Threats. In *Advanced Research and Real-World Applications of Industry 5.0*; IGI Global: Hershey, PA, USA, 2023; pp. 66–76.
143. Adel, A. Future of industry 5.0 in society: Human-centric solutions, challenges and prospective research areas. *J. Cloud Comput.* **2022**, *11*, 40. [CrossRef]
144. Rjoub, G.; Bentahar, J.; Wahab, O.A.; Mizouni, R.; Song, A.; Cohen, R.; Otrok, H.; Mourad, A. A Survey on Explainable Artificial Intelligence for Cybersecurity. *IEEE Trans. Netw. Serv. Manag.* **2023**, *20*, 5115–5140. [CrossRef]
145. Kiruthika, M.; Moorthi, K.; Devi, M.A.; Roseline, S.A. Role of XAI in building a super smart society 5.0. In *XAI Based Intelligent Systems for Society 5.0*; Elsevier: Amsterdam, The Netherlands, 2024; pp. 295–326.
146. Khan, A.; Jhanjhi, N.Z.; Haji, D.H.T.B.A.; bin Haji Omar, H.A.H. The Need for Explainable AI in Industry 5.0. In *Advances in Explainable AI Applications for Smart Cities*; IGI Global: Hershey, PA, USA, 2024; pp. 1–30.
147. Alsamhi, S.H.; Curry, E.; Hawbani, A.; Kumar, S.; Hassan, U.U.; Rajput, N.S. DataSpace in the Sky: A Novel Decentralized Framework to Secure Drones Data Sharing in B5G for Industry 4.0 toward Industry 5.0. *Preprints* **2023**. [CrossRef]
148. Alnajjar, I.A.; Almazaydeh, L.; Odeh, A.A.; Salameh, A.A.; Alqarni, K.; Ban Atta, A.A. Anomaly Detection Based on Hierarchical Federated Learning with Edge-Enabled Object Detection for Surveillance Systems in Industry 4.0 Scenario. *Int. J. Intell. Eng. Syst.* **2024**, *17*, 649–665. 10.22266/IJIES2024.0831.49. [CrossRef]
149. Nascita, A.; Aceto, G.; Ciuonzo, D.; Montieri, A.; Persico, V.; Pescapé, A. A survey on explainable artificial intelligence for internet traffic classification and prediction, and intrusion detection. *IEEE Commun. Surv. Tutor.* **2024**, *27*, 3165–3198. [CrossRef]
150. Eltomy, R.; Lalouani, W. Explainable Intrusion Detection in Industrial Control Systems. In Proceedings of the 2024 IEEE 7th International Conference on Industrial Cyber-Physical Systems (ICPS), St. Louis, MO, USA, 12–15 May 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–8.
151. Arisdakessian, S.; Wahab, O.A.; Mourad, A.; Otrok, H.; Guizani, M. A Survey on IoT Intrusion Detection: Federated Learning, Game Theory, Social Psychology, and Explainable AI as Future Directions. *IEEE Internet Things J.* **2023**, *10*, 4059–4092. [CrossRef]
152. Liao, H.J.; Lin, C.H.R.; Lin, Y.C.; Tung, K.Y. Intrusion detection system: A comprehensive review. *J. Netw. Comput. Appl.* **2013**, *36*, 16–24. [CrossRef]
153. Liu, H.; Lang, B. Machine learning and deep learning methods for intrusion detection systems: A survey. *Appl. Sci.* **2019**, *9*, 4396. [CrossRef]
154. Ahmad, Z.; Shahid Khan, A.; Wai Shiang, C.; Abdullah, J.; Ahmad, F. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Trans. Emerg. Telecommun. Technol.* **2021**, *32*, e4150. [CrossRef]
155. Zhang, Z.; Hamadi, H.A.; Damiani, E.; Yeun, C.Y.; Taher, F. Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research. *IEEE Access* **2022**, *10*, 93104–93139. [CrossRef]
156. Kaur, R.; Gabrijelčič, D.; Klobučar, T. Artificial intelligence for cybersecurity: Literature review and future research directions. *Inf. Fusion* **2023**, *97*, 101804. [CrossRef]
157. Divekar, A.; Parekh, M.; Savla, V.; Mishra, R.; Shirole, M. Benchmarking datasets for Anomaly-based Network Intrusion Detection: KDD CUP 99 alternatives. In Proceedings of the 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), Kathmandu, Nepal, 25–27 October 2018; pp. 1–8. [CrossRef]
158. Lippmann, R.; Haines, J.W.; Fried, D.J.; Korba, J.; Das, K. The 1999 DARPA off-line intrusion detection evaluation. *Comput. Netw.* **2000**, *34*, 579–595. [CrossRef]
159. Tavallaee, M.; Bagheri, E.; Lu, W.; Ghorbani, A.A. A detailed analysis of the KDD CUP 99 data set. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, Canada, 8–10 July 2009; pp. 1–6. [CrossRef]

160. Moustafa, N.; Slay, J. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, Australia, 10–12 November 2015; pp. 1–6. [CrossRef]
161. Faraj, O.; Megías, D.; Ahmad, A.M.; Garcia-Alfaro, J. Taxonomy and challenges in machine learning-based approaches to detect attacks in the internet of things. In Proceedings of the 15th International Conference on Availability, Reliability and Security, Virtual, 25–28 August 2020; pp. 1–10.
162. Vigano, L.; Magazzeni, D. Explainable security. In Proceedings of the 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Genoa, Italy, 7–11 Septemner 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 293–300.
163. Islam, S.R.; Eberle, W. Domain Knowledge-Aided Explainable Artificial Intelligence. In *Explainable Artificial Intelligence for Cyber Security: Next Generation Artificial Intelligence*; Springer: Cham, Switzerland, 2022; pp. 73–92.
164. Neupane, S.; Ables, J.; Anderson, W.; Mittal, S.; Rahimi, S.; Banicescu, I.; Seale, M. Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities. *IEEE Access* **2022**, *10*, 112392–112415. [CrossRef]
165. Lipton, Z.C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **2018**, *16*, 31–57. [CrossRef]
166. Hoenig, A.; Roy, K.; Acquaah, Y.; Yi, S.; Desai, S. Explainable AI for cyber-physical systems: Issues and challenges. *IEEE Access* **2024**, *12*, 73113–73140. [CrossRef]
167. Pourret, O.; Na, P.; Marcot, B. *Bayesian Networks: A Practical Guide to Applications*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
168. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **1982**, *43*, 59–69. [CrossRef]
169. Tcydenova, E.; Kim, T.W.; Lee, C.; Park, J.H. Detection of adversarial attacks in AI-based intrusion detection systems using explainable AI. *Hum.-Centric Comput. Inf. Sci.* **2021**, *11*, 35.
170. Gaitan-Cardenas, M.C.; Abdelsalam, M.; Roy, K. Explainable AI-Based Intrusion Detection Systems for Cloud and IoT. In Proceedings of the 2023 32nd International Conference on Computer Communications and Networks (ICCCN), Honolulu, Hawaii, USA, 24–27 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–7.
171. Charte, D.; Charte, F.; del Jesus, M.J.; Herrera, F. An analysis on the use of autoencoders for representation learning: Fundamentals, learning task case studies, explainability and challenges. *Neurocomputing* **2020**, *404*, 93–107. [CrossRef]
172. Yoon, M.K.; Mohan, S.; Choi, J.; Sha, L. Memory heat map: Anomaly detection in real-time embedded systems using memory behavior. In Proceedings of the 52nd Annual Design Automation Conference, San Francisco, CA, USA, 7–11 June 2015; pp. 1–6.
173. Hariharan, S.; Velicheti, A.; Anagha, A.; Thomas, C.; Balakrishnan, N. Explainable artificial intelligence in cybersecurity: A brief review. In Proceedings of the 2021 4th International Conference on Security and Privacy (ISEA-ISAP), Dhanbad, India, 27–30 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–12.
174. Charmet, F.; Tanuwidjaja, H.C.; Ayoubi, S.; Gimenez, P.F.; Han, Y.; Jmila, H.; Blanc, G.; Takahashi, T.; Zhang, Z. Explainable artificial intelligence for cybersecurity: A literature survey. *Ann. Telecommun.* **2022**, *77*, 789–812. [CrossRef]
175. Duddu, V. A survey of adversarial machine learning in cyber warfare. *Def. Sci. J.* **2018**, *68*, 356. [CrossRef]
176. Ling, X.; Wu, L.; Zhang, J.; Qu, Z.; Deng, W.; Chen, X.; Qian, Y.; Wu, C.; Ji, S.; Luo, T.; et al. Adversarial attacks against Windows PE malware detection: A survey of the state-of-the-art. *Comput. Secur.* **2023**, *128*, 103134. [CrossRef]
177. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership inference attacks against machine learning models. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 3–18.
178. Hall, M.; Harborne, D.; Tomsett, R.; Galetic, V.; Quintana-Amate, S.; Nottle, A.; Preece, A. A systematic method to understand requirements for explainable AI (XAI) systems. In Proceedings of the IJCAI Workshop on eXplainable Artificial Intelligence (XAI 2019), Macau, China, 11 August 2019; Volume 11.
179. Mohseni, S.; Zarei, N.; Ragan, E.D. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans. Interact. Intell. Syst. (TiiS)* **2021**, *11*, 1–45. [CrossRef]
180. Srivastava, G.; Jhaveri, R.H.; Bhattacharya, S.; Pandya, S.; Maddikunta, P.K.R.; Yenduri, G.; Hall, J.G.; Alazab, M.; Gadekallu, T.R.; et al. XAI for cybersecurity: State of the art, challenges, open issues and future directions. *arXiv* **2022**, arXiv:2206.03585. [CrossRef]
181. Chakraborty, A.; Alam, M.; Dey, V.; Chattopadhyay, A.; Mukhopadhyay, D. A survey on adversarial attacks and defences. *CAAI Trans. Intell. Technol.* **2021**, *6*, 25–45. [CrossRef]
182. Sarker, I.H. Multi-aspects AI-based modeling and adversarial learning for cybersecurity intelligence and robustness: A comprehensive overview. *Secur. Priv.* **2023**, *6*, e295. [CrossRef]
183. Apruzzese, G.; Colajanni, M.; Ferretti, L.; Marchetti, M. Addressing adversarial attacks against security systems based on machine learning. In Proceedings of the 2019 11th International Conference on Cyber Conflict (CyCon), Tallinn, Estonia, 28–31 May 2019; IEEE: Piscataway, NJ, USA, 2019; Volume 900; pp. 1–18.
184. Alatwi, H.A.; Aldweesh, A. Adversarial black-box attacks against network intrusion detection systems: A survey. In Proceedings of the 2021 IEEE World AI IoT Congress (AIIoT), Virtual, 10–13 May 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 0034–0040.

185. Apruzzese, G.; Andreolini, M.; Ferretti, L.; Marchetti, M.; Colajanni, M. Modeling realistic adversarial attacks against network intrusion detection systems. *Digit. Threat. Res. Pract. (DTRAP)* **2022**, *3*, 1–19. [CrossRef]
186. Merzouk, M.A.; Cuppens, F.; Boulahia-Cuppens, N.; Yaich, R. A deeper analysis of adversarial examples in intrusion detection. In Proceedings of the Risks and Security of Internet and Systems: 15th International Conference, CRiSIS 2020, Paris, France, 4–6 November 2020; Revised Selected Papers 15; Springer: Cham, Switzerland, 2021; pp. 67–84.
187. Dunmore, A.; Jang-Jaccard, J.; Sabrina, F.; Kwak, J. A Comprehensive Survey of Generative Adversarial Networks (GANs) in Cybersecurity Intrusion Detection. *IEEE Access* **2023**, *11*, 76071–76094. [CrossRef]
188. Alkadi, S.; Al-Ahmadi, S.; Ismail, M.M.B. Better Safe Than Never: A Survey on Adversarial Machine Learning Applications towards IoT Environment. *Appl. Sci.* **2023**, *13*, 6001. [CrossRef]
189. Pujari, M.; Cherukuri, B.P.; Javaid, A.Y.; Sun, W. An approach to improve the robustness of machine learning based intrusion detection system models against the carlini-wagner attack. In Proceedings of the 2022 IEEE International Conference on Cyber Security and Resilience (CSR), Virtual, 27–29 July 2022 IEEE: Piscataway, NJ, USA, 2022; pp. 62–67.
190. Merzouk, M.A.; Delas, J.; Neal, C.; Cuppens, F.; Boulahia-Cuppens, N.; Yaich, R. Evading Deep Reinforcement Learning-based Network Intrusion Detection with Adversarial Attacks. In Proceedings of the 17th International Conference on Availability, Reliability and Security, Vienna, Austria, 23–26 August 2022; pp. 1–6.
191. He, K.; Kim, D.D.; Asghar, M.R. Adversarial machine learning for network intrusion detection systems: A comprehensive survey. *IEEE Commun. Surv. Tutor.* **2023**, *25*, 538–566. [CrossRef]
192. McCarthy, A.; Ghadafi, E.; Andriotis, P.; Legg, P. Functionality-preserving adversarial machine learning for robust classification in cybersecurity and intrusion detection domains: A survey. *J. Cybersecur. Priv.* **2022**, *2*, 154–190. [CrossRef]
193. Rehman, A.; Farrakh, A.; Khan, S. Explainable AI in Intrusion Detection Systems: Enhancing Transparency and Interpretability. *Int. J. Adv. Sci. Comput.* **2023**, *2*, 7–20.
194. Nadeem, A.; Vos, D.; Cao, C.; Pajola, L.; Dieck, S.; Baumgartner, R.; Verwer, S. Sok: Explainable machine learning for computer security applications. In Proceedings of the 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P), Delft, The Netherlands, 3–7 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 221–240.