This document contains the synthesis of how the literature addresses the "black box" limitations and the vulnerability of models to adaptive attackers. The authors utilized **NotebookLM** (processing 2–5 papers per prompt) to identify specific explainability techniques (e.g., SHAP, LIME, Attention weights). The AI agent was also asked to identify adversarial defense mechanisms such as poisoning defense and generative hardening. All extracted findings regarding model interpretability and security were cross-checked manually by the authors to ensure that the specific details of the robustness evaluations and the types of adversarial attacks (e.g., evasion vs. poisoning) were captured accurately by the AI agent.

Prompt

Task: You are a research assistant specializing in Trustworthy AI and Cybersecurity. Analyze the uploaded papers and fill the **XAI and Adversarial Robustness Table** based on the following criteria.

Instructions for Columns:

- **Paper ID:** The identifier for this study (e.g., author (year)).
- **Algorithm / Model:** List the primary machine learning or deep learning model used in this study.
- **XAI Technique:** Identify any interpretability or explainability tools used. Look for: *SHAP, LIME, Attention Mechanisms, Layer-wise Relevance Propagation (LRP), Feature Importance Maps, or Rule-based explanations.*
- **Scope (Global / Local):** Specify the level of explanation.
    - *Global:* Explaining the entire model's logic.
    - *Local:* Explaining why a specific user action was flagged as a threat.
- **Adversarial Defense / Testing:** Identify if the authors tested the model against malicious manipulation. Look for: *Evasion attacks, Poisoning attacks, Adversarial Training, or Robustness testing.*
- **Impact on Trust / Robustness:** Briefly describe how these techniques improved the system. Did it help security analysts understand the alerts (Trust), or did it make the model harder to hack (Robustness)?

Output Format: Provide the results in a single-row table with these columns: | Authors (year) | Algorithm / Model | XAI Technique | Scope (Global-Local) | Adversarial Defense / Testing | Impact on Trust / Robustness |

Constraint: Use ONLY information from this paper. If the paper does not mention XAI or Adversarial testing (which is common), write "None mentioned" or "Not specified."

Paper ID	**Algorithm / Model**	**XAI Technique**	**Scope** **Global- Local**	**Adversarial Defense / Testing**	**Impact on Trust / Robustness**
Adun (2023)	SVM, ANFIS	None	N/A	None	Implicit via high prediction accuracy/precision.
Alabdulkareem (2022)	MWF-IDLDC (DL Fusion)	None	N/A	None	Focus on performance for robust leakage detection.
Ahmadi (2025)	RF, GB, K-means	Context-aware Risk Scores	Local	Generic threat adaptation; no specific ML testing.	Dynamic access governance and false positive reduction.
Ahmed 2025	Random Cut Forest (RCF)	None mentioned	None mentioned	None mentioned	Improved detection accuracy and reduced false positives, enhancing trust in alerts.
Alhammadi 2021	Adaptive Boosting, Random Forest, 2D CNN, 1D CNN, MLP with CNN, KNN	Explainable AI (general), permutation importance	Global	None mentioned	Improved understanding of model predictions, identifying key EEG features, enhancing trust in identifying insider risks.
Ali et al 2025	BERT, BERTopic, Ensemble Model (BERT + BERTopic + multi-class logistic regression)	None mentioned	None mentioned	None mentioned	Optimized human analyst efforts and improved insider threat detection through human-machine partnership, enhancing trust.
ALmihqani (2021)	AD-DNN (ADASYN + Deep Neural Network)	None mentioned	Not specified	None mentioned	Not specified
almusawi (2024)	Hybrid (ML algorithms + Expert Rule-based Policies)	Rule-based explanations (from expert policies)	Local	None mentioned	Improved trust by reducing false positives and enhancing detection accuracy through understandable expert rules.
AL-Mihqani (2022)	Multilayer framework (Random Forest, K-Nearest Neighbors, VIKOR)	None mentioned	Not specified	None mentioned	Not specified
Alshehari (2023) IF	Isolation Forest (IF)	Inherent interpretability of isolation (path length in trees)	Local (identifying specific anomalous instances)	None mentioned	Improves detection performance and effectiveness for imbalanced datasets, contributing to model reliability.
Al-shehari et al. (2023)	Decision Tree, KNN, Random Forest, XGBoost (with resampling)	None mentioned	Not applicable	None mentioned	
Al-shehari et al. (2024)	CNN (with SMOTE, Borderline-SMOTE, ADASYN)	None mentioned	Not applicable	None mentioned	Significantly improves detection accuracy and discriminative ability, offering a promising direction for cybersecurity applications.
AL-SHEHARI (2024)	DBLOF (Density-Based Local Outlier Factor)	Inherent interpretability of local density deviation for outlier identification	Local (identifying specific outliers based on local density)	None mentioned	Achieves high F-score, demonstrating effectiveness in accurately detecting insider threats without data-level modifications, contributing to robust systems.
Alshehri (2022)	Relational RNN (Rel-RNN) with LSTM units and Direct Graphs	None mentioned	Not specified	None mentioned	Not specified

Paper ID	**Algorithm / Model**	**XAI Technique**	**Scope** **Global- Local**	**Adversarial Defense / Testing**	**Impact on Trust / Robustness**
Amiri-Zarandi (2023)	Deep Autoencoder (within a Federated Learning framework)	**SHAP** (Shapley Additive Explanation)	**Global and Local** (Summarizes overall feature impact	None mentioned	**Improved Trust:** Helps security experts identify "causing factors" of threats and provides visibility into how the model makes predictions to justify investigations.
Al-Shehari (2021)	ML Classifiers (LR, DT, RF, NB, KNN, KSVM) with SMOTE	None mentioned	Not specified	None mentioned	Not specified
Amuda (2022)	**Hybrid CNN-GRU** (Convolutional Neural Network and Gated Recurrent Unit)	**None mentioned**	**Not specified**	**None mentioned**	**Not specified**
Anju (2024)	**Stacked CNN-Attentional BiGRU**	**Attention Mechanisms**	**Local** (Captures long-range/temporal aspects of specific activity sequences	**None mentioned** (Notes that some related works lack adversarial techniques)	**Improved Performance:** Allows the model to focus on complex features and temporal representations without increasing network parameters.
Anakath (2022)	**Deep Belief Neural Network (DBN)** using Restricted Boltzmann Machines (RBM)	**None mentioned**	**Not specified**	**None mentioned**	**Not specified**
Asha S (2023)	Double-layer architecture (Sampling + Anomaly Detection models like **One-Class SVM**),	None mentioned,	Not specified	None mentioned (related works on adversarial properties are cited but not tested),	Not specified
Cai X (2024)	**LAN** (Learning Adaptive Neighbors) using LSTM and Graph Neural Networks,	**Attention Mechanisms** (Multi-head attentive pooling),	**Local** (Models temporal dependencies and relationships specific to the detected activity),	Discusses strategies for **Adaptive Attacks** (suggests frequent incremental updates and ensembling)	**Robustness:** Strategies like ensembling multiple approaches are used to enhance overall robustness against adversarial samples
Dong J (2025)	**DDCC** (Denoising Diffusion Probabilistic Models + Curriculum Learning)	None mentioned,	Not specified	None mentioned (discusses handling "noise" via diffusion but not malicious manipulation),	**Robustness:** The curriculum strategy facilitates controlled optimization to handle behavioral heterogeneity
Eshmawi (2026)	SVM, RF, KNN, DNN, and NB	**General Model Interpretability** (inherent to classical ML models)	Not specified	None mentioned (Adversarial resilience is discussed only in related work)	**Trust:** Classical models were specifically selected over deep sequential models to establish an interpretable baseline.

Paper ID	Algorithm / Model	XAI Technique	Scope Global- Local	Adversarial Defense / Testing	Impact on Trust / Robustness
Feng (2025)	MG-UABD (**Random Forest**)	**Feature Importance** (identification of significant features)	Global	None mentioned	**Trust:** The model's ability to evaluate feature importance helps analysts identify which behaviors have the most significant impact on anomaly representation.
Ferraro (2025)	GABM (**LLaMA-3.1 8B** hierarchical agents)	**Chain-of-Thought (CoT) reasoning**	Local (Explaining specific activity flags)	None mentioned	**Trust:** Enhances interpretability by producing natural language reports with "assessments, justifications, and potential red flags" for specific log entries.
Gayathri (2025)	**SNN-MLP, SNN-1DCNN, and TabNet**	**Shapley Additive Explanations (SHAP)** and **Tree SHAP**	**Global and Local**; SHAP provides local feature attribution while Tree SHAP can determine global importance.	**Adversarial Training** tested against **Poison-label Backdoor attacks**, FGSM, DeepFool, Carlini & Wagner (CW), and JSMA.	**Robustness:** Adversarial training (specifically using CWGAN-GP) significantly reduced backdoor effects and improved model stability and resistance to perturbations.
Gayathri, B. (2025)	**Jordan Neural Network (JNN)**	**None mentioned** (Relies on manual domain expertise for extraction)	**Not specified**	**None mentioned** (Focuses on missing value imputation and cloud threat mitigation)	**Trust:** High detection accuracy (96%) and specialized recurrent architecture (JNN) improved system reliability in identifying authorized users acting as intrusive parties.
Gayathri (2024)	**Hybrid Bayesian Neural Network (BNN)** with MLP and 1DCNN	**t-SNE (t-distributed Stochastic Neighbor Embedding)**	**Global**; used for visualizing data distributions to increase interpretability in lower dimensions.	**Adversarial Training** using **SPCAGAN**; tested against **Fast Gradient Sign Method (FGSM)** and **DeepFool** white-box attacks.	**Robustness:** The hybrid BNN model combined with SPCAGAN-augmented data yielded better resistance to one-step attacks and reduced predictive uncertainty.
Gonzales (2025)	**C3P** and **C3P-SMART** (integrating pattern mining and ABC framework)	**None mentioned** (Utilizes internal "Local Context Weighting" (LCW) to score behavior)	**Local**: Focuses on specific three-action segments (trigrams) rather than entire sequences	**None mentioned** (Discusses vulnerabilities to insiders aware of detection but does not perform tests)	**Trust**: Improved scalability and efficiency by automating detection and removing the need for human validation or expert intervention.
Gupta (2024)	**FedMUP** (AFed, CFed, and DFed architectures)	**None mentioned** (Utilizes "User Behavior Evaluation" (UBE) to compute security parameters)	**Global**: Aggregates various local models into a central updated global version	**None mentioned** (Identified as a future work goal to develop adaptive privacy-preserving approaches)	**Robustness**: Enhanced data security and privacy by utilizing a "model-to-data" approach that avoids transferring raw sensitive data.
Hafizu Rhman (2022)	**NARX** (Nonlinear Autoregressive Exogenous) neural network	**None mentioned** (Conceptualizes dimensions like "Access" and "Motivation" for modeling)	**Not specified**	**None mentioned**	**None mentioned**

Paper ID	**Algorithm / Model**	**XAI Technique**	**Scope** **Global- Local**	**Adversarial Defense / Testing**	**Impact on Trust / Robustness**
Haq (2022)	**Hybrid LSTM** (Word2vec/GloVe) and **ML Ensemble** (XGBoost, AdaBoost, RF, KNN, LR)	None mentioned	Not specified	None mentioned	**Not specified**; study focuses on bridging gaps in accuracy and real data availability.
He (2022)	**L-XGB** (LSTM/Bi-LSTM + XGBoost) and **BiLA-ITD** (Bi-LSTM + Attention)	**Attention Mechanism**	**Local**: Captures "key information" in long-range user behavior sequences.	None mentioned	**Trust**: Supports "after-the-fact investigation" and aids analysts in reviewing specific insiders.
He (2024)	**L-XGB** (LSTM/Bi-LSTM + XGBoost) and **BiLA-ITD** (Bi-LSTM + Attention)	**Attention Mechanism**	**Local**: Specifically used to "capture the attack words" within behavioral sequences.	None mentioned; identifies a weakness against phishing links in images/QR codes.	**Trust**: Provides early warning and facilitates the "timely review" of insiders by explaining threat indicators
Huang (2025)	**DenseAttDNN** (Dense Connection + Attention Mechanism)	**Attention Mechanism** (Scaled Dot-Product)	**Local**: Enables the model to focus on crucial feature information and dynamic behavioral anomalies.	**None mentioned** (Identifies lack of dynamic recalibration as a limitation).	**Trust**: Aids security analysts in "analyzing and explaining detected threats" by highlighting relevant behavioral features.
Jaiswal (2024)	**TSITD Models** (Combinations of sampling and classifiers like XGB, ANN, RVFL)	**None mentioned**	**Not specified**	**None mentioned**	**Not specified**; focus is on mitigating dataset imbalance.
Janjua (2021)	**Semi-supervised framework** (K-Means + Decision Tree)	**None mentioned**	**Not specified**	**None mentioned**	**Not specified**; focus is on classification accuracy for unlabeled email content.
Kamatchi (2025)	**Federated Bi-LSTM** (Bidirectional Long Short-Term Memory)	**None mentioned** (abstractly refers to "detailed logs explaining the specific behaviors" triggering alerts)	**Local**: Logs explain specific behaviors causing an alert.	**Robustness Testing**: Proposes **FAWO HMAC-SHA** to defend against **model poisoning** and tampering with local updates,.	**Robustness**: Reduces the risk of malicious insiders or compromised devices manipulating local updates to corrupt the global model,.
Kong (2025)	**DPI-ITDD (FMLP)** and **DPI-ITDM (XGBoost)**	**None mentioned**	**Not specified**	**None mentioned**	**Not specified**
Kotb (2025)	**Binary Deep Learning** (Fully connected hidden blocks)	**None mentioned** (identifies implementing XAI as future work)	**Not specified**	**Robustness Testing**: Specifically tests the model's ability to distinguish between real users and **AI-generated synthetic threats** (CTGAN, TVAE),.	**Robustness**: Enhances the model's ability to identify sophisticated "fake user profiles" created by generative AI to mimic legitimate behavior,.

Paper ID	Algorithm / Model	XAI Technique	Scope Global- Local	Adversarial Defense / Testing	Impact on Trust / Robustness
Lavanya et al. (2024)	**EBiGAN + DNN-PI** (Enhanced Bidirectional GAN + Deep Neural Network)	**None mentioned** (Paper explicitly states it "lacks interpretability in real-time environment")	**Not specified**	**None mentioned** (Uses adversarial generation for data balancing, but does not test against adversarial attacks)	**Robustness**: Improved PCA and outlier estimators are used to ensure "standard stability and robustness" to the detection model.
Lavanya et al. (2025)	**HOGPNN-ITD** (ABWGAN-EHI + L2-SP regularized Pretrained AGCN)	**Attention Mechanisms** (Assigns "attention scores" to neighbor nodes to identify relevant behavioral context)	**Local**: Allocates importance scores to neighbor nodes to update feature rules for specific behavior identification.	**Robustness Testing**: Evaluated for stability and robustness against class imbalance using an optimized generative model to ensure high-quality adversarial samples.	**Trust/Robustness**: Attention helps identify specific user patterns; Adabelief and Hinge loss optimization improve stability and convergence.
Le & Zincir-Heywood (2021)	**Unsupervised Ensembles** (AE, IF, LODA, and LOF)	**Feature Identification** (LODA identifies specific features of an outlier to explain the causes of events)	**Local**: Explaining the causes of specific anomalous events.	**Poisoning Attacks**: Extensively tested the model's response to training data poisoning where malicious data is injected to corrupt the model.	**Robustness**: Ensembles (especially IF, LODA, and VOTE2) were found to be "very robust to the data poisoning attack" compared to individual models like AE
Li et al. (2024)	**GMFITD** (Graph modularized-based Meta-learning Framework) incorporating GAE and MAML,	**Attention Mechanism** and **Edge Importance Estimation Mechanism**,,	**Local**: Assigns importance weights to node-level local topology and prunes suspicious edges,,	**Robustness Testing** against targeted (**Nettack**) and non-targeted (**Mettack**) structural attacks,	**Robustness**: Pruning perturbed edges and weighting critical ones restored performance to levels comparable to non-attack scenarios,.
Li et al. (2023)	**DD-GCN** (Dual-Domain Graph Convolutional Network),	**Attention Mechanism** to learn adaptive importance weights of features,,	**Local**: Identifies specific features (e.g., web browsing vs. device usage) that contribute most to a specific detection,	**None mentioned** (Focuses on adaptive detection and domain constraints),	**Trust**: Helps analysts understand the "why" by showing higher attention values for features related to malicious scenarios.
Liu et al. (2025)	**Ripple2Detect** (BERT-based semantic similarity + Attack Knowledge Graph)	**Path-tracing** on Knowledge Graph; **Attention mechanisms** (inherent to BERT)	**Local**: Decomposes specific attack sequences into "motive" and "behavioral" evidence to explain a user's threat assessment.	**None mentioned** (Explicitly listed as a "Future research direction").	**Trust**: Enhances forensic accuracy and legal accountability by connecting sensitive semantics to clear attack mechanisms.
Mehmood et al. (2023)	**Ensemble Learning** (LightGBM, XGBoost, AdaBoost, Random Forest)	**Feature Importance Maps** (Heatmap of specified features)	**Global**: Identifies which specific features (e.g., file size, attachments) are most critical for the models' overall classification logic.	**None mentioned** (Notes ensemble models are generally "robust," but no adversarial tests are performed).	**Trust**: Helps security analysts understand the most important attributes in the dataset for better classification results.

Paper ID	Algorithm / Model	XAI Technique	Scope Global- Local	Adversarial Defense / Testing	Impact on Trust / Robustness
Medvedev et al. (2025)	**Siamese Neural Network (SNN)** with CNN branches and triplet loss	**t-SNE Dimensionality Reduction** (Visualization of embeddings)	**Local**: Visualizes specific authentication attempts (genuine vs. impostor) relative to user clusters to explain classification decisions.	**None mentioned** (Discusses general robustness to behavioral variability, but specific adversarial attacks are listed as background threat info).	**Trust**: Visualization of embeddings provides insight into learned representations and helps identify patterns/anomalies.
Mehnaz (2021)	**Finite State Automata (FSA)** and **Unsupervised Clustering**,,	**Anomaly Flags (AF1-AF7)** mapped to a defined **Taxonomy of Anomalies**	**Local**: Explains specific alerts based on deviation in access size, frequency, or segment,	**None mentioned** (Adversarial concept drift and collusion attacks are listed as future work),	**Trust**: Mapping flags to a taxonomy helps security staff understand the nature of alerts to minimize the disruption of normal activities,.
Mladenovic (2024)	**XGBoost** and **AdaBoost** optimized by **HARFO** (Hybrid Adaptive Red Fox Optimization),	**SHAP (Shapley Additive Explanations)**,,	**Global and Local**: Identifies key features across the dataset and factor-level drivers for individual samples,,	**None mentioned** (Focuses on hyperparameter optimization and sentiment classification),	**Trust**: Elucidates the reasoning behind decisions to ensure mathematical intricacies do not obscure feature-outcome relationships,.
Nasir (2021)	**LSTM-Autoencoder**,	**None mentioned**,	**Not specified**,	**None mentioned**,	**Not specified** (Focuses on achieving high accuracy and low false positive rates),.
Nikiforova et al. (2024)	**K-means clustering** refined with the **Elbow method** and **Markov chains**,.	**Behavioral Graphs** and **Suspiciousness Coefficients** used to visualize action sequences,,.	**Local**: Graphs explain specific suspicious user sessions (e.g., user "Ava") by highlighting atypical action transitions,.	**None mentioned** (identifies "polluting" models as a future research area).	**Trust**: Visualizing atypical behavior helps security analysts verify alerts and reduces manual investigation time,,.
Pal et al. (2023)	**Ensemble** of **stacked-LSTM** and **stacked-GRU** models,.	**Attention Mechanisms**,.	**Local**: Assigns weights to prioritize "critical sections" of a user's single-day activity sequence to explain the threat label,,.	**None mentioned**.	**Trust**: Helps analysts identify which specific activities within a working day most heavily influenced the model's decision,.
Patel & Iyer (2025)	**SiaDNN** (Siamese Convolutional Neural Network + Deep Neural Network),.	**None mentioned** (identifies an explainable framework as future work).	**Not specified**.	**None mentioned**.	**Not specified**.
Peccatiello et al. (2023)	**Isolation Forest (ISOF)**, **Elliptic Envelope**, and **Local Outlier Factor**,.	**None mentioned**,.	**Not specified**.	**None mentioned** (mentions zero-day vulnerabilities but no formal adversarial testing).	**Not specified**.

Paper ID	Algorithm / Model	XAI Technique	Scope Global- Local	Adversarial Defense / Testing	Impact on Trust / Robustness
Pennada et al. (2024)	**Stacking and Voting classifiers** (Ensemble of Random Forest, Adaboost, and DT).	**Interpretability** of DT and **Feature Significance** analysis in RF.	**Global**: Uses feature significance to provide insights into behavior.	**None mentioned**.	**Trust**: The interpretable nature of the base models helps analysts understand why certain features lead to a threat classification.
Pennada et al. (2025)	**Hybrid Model** combining Generative AI (**DAEs**, **VAEs**) with ML (**Random Forest**, **XGBoost**).	**None mentioned** (identifies interpretable fault detection as an area of related research but not used in the study).	**Not specified**.	**None mentioned**.	**Not specified**.
Perez-Miguel et al. (2025)	**Not Specified** (Focuses on design/generation of the **SPEDIA dataset**).	**None mentioned**.	**Not specified**.	**None mentioned**	**Not specified**.
Qawasmeh & AlQahtani (2025)	**XGBoost** (primary), Random Forest, SVM, and **Logistic Regression**.	**Model Transparency** and interpretability of Logistic Regression; **Feature Importance** for Random Forest.	**Global**: Highlights feature importance and transparency for application-specific requirements.	**Continuous learning** and **Dynamic profiling** intended to prevent adversaries from "tricking" the system by gradually modifying behavior.	**Trust/Robustness**: Interpretability aids transparency; continuous learning makes the model harder to evade through behavioral shifts.
Randive et al. (2023)	**Wavelet CNN (WCNN)** based on VGG-19.	**Visual Pattern Recognition**; represents behavioral logs as grayscale images to reveal spectral/spatial patterns.	**Local**: Allows detection of specific malicious patterns within a single day's activity image.	**None mentioned**.	**Trust**: Reveals "visual rhythms" in user behavior that help analysts manually verify detected malicious patterns.
Rauf et al. (2021)	Bio-inspired auto-resilient policy regulation framework using machine learning (DBSCAN, Random Forest, SVM) and Z3 SMT solver.	**Formal Verification** and Property Checking; the system uses a property-based diagnosis to identify error-prone configurations.	**Global/Local**: Property (2) verifies the safety of all configurations (Global), while Property (1) allows specific error-prone configurations to be diagnosed (Local).	**Formal Verification** of the system's correctness and safety; tested against a real-life threat test dataset.	**Trust**: Helps security analysts diagnose human errors in configurations. **Robustness**: The auto-resilient framework allows the system to autonomously change its state to mitigate threats in real-time.
Roy & Chen (2024)	**GraphCH**: A Cyber-Human Graph Neural Network (GNN) framework utilizing CH-GLM (BiLSTM and RWR).	**TF-IDF feature weighting** and **Cyber-human mapping**; correlates cyber actions like logon failures with psychological traits (impulsiveness/risk-taking).	**Global/Local**: Mapping defines general model behavior across user types (Global), while ranking suspicious users explains specific flags in clusters (Local).	None mentioned.	**Trust**: Makes detection results explainable via psychological scores (e.g., BART) and facilitates further forensics efforts.

Paper ID	Algorithm / Model	XAI Technique	Scope Global- Local	Adversarial Defense / Testing	Impact on Trust / Robustness
Bin Sarhan & Altwaijry (2023)	SVM (Classification), Anomaly Detection (OCSVM, iForest), NN, AdaBoost, and Random Forest.	**Deep Feature Synthesis (DFS)** (features described in natural language) and **PCA** (identifies most important uncorrelated features).	**Global/Local**: PCA identifies critical features across the dataset (Global), while DFS creates features characterizing specific employee usage behavior (Local).	None mentioned; however, the paper notes that formal verification is a general approach that should be conducted for ML systems.	**Trust**: DFS generates features that are easy to understand in natural language, helping security experts and data scientists interpret the model inputs.
Senevirathna et al. (2025)	**CNN and Random Forest** (Behavioral); **MobileNetV2 and LSTM** (Physical Security),.	**None currently implemented**; identifies its hybrid CNN-RF as "more interpretable" than complex alternatives. Plans to implement **XAI methodologies** in the future.	Not specified	**None mentioned** (Adversarial training and GANs are discussed only in the Literature Review for other researchers' work),,.	**Trust**: Future implementation of XAI is intended to increase "decision-making transparency" so analysts can better understand and trust AI-driven insights.
Song et al. (2024)	**BRITD** (Stacked Bidirectional LSTM and Feedforward Neural Network),.	**None mentioned**; discusses using a **Self-attention** layer to capture long-distance dependence but found it led to overfitting.	Not specified	**None mentioned** (Mentions "Adversarial Recurrent Autoencoders" only in the context of external related works).	Not specified
Tabassum et al. (2024)	**Isolation Forest (IForest) and SVM** (also uses Decision Trees and Random Forests).	**Rule-based interpretability**; specifically notes that **Decision Trees** offer "simplicity and interpretability" by providing "clear decision paths".	Not specified	**None mentioned**; notes a general need for "security and resilience culture" to address evolving threats but performs no specific adversarial testing.	**Trust**: Highlights that the interpretation effectiveness of decision paths is valuable for analysts to understand the anomaly detection process.
Tian T et al. (2025)	**Transformer Encoder** utilizing a **multi-head attention mechanism**,.	**Multi-head attention mechanism**; **TF-IDF** for semantic text topic analysis.	**Global/Local**: Captures long-distance dependencies and local information to provide depth feature representation,.	None mentioned.	**Robustness**: Improves the model's ability to discern subtle differences in patterns and distinguishes normal from abnormal behaviors with higher accuracy.
Tian Z et al. (2024)	**Attention-LSTM** (LSTM-RNN combined with **multi-head attention**) and	**Dempster-Shafer theory** (fusion engine); **Multi-head attention mechanism**.	**Local**: The D-S fusion engine aids analysts in interpreting specific system decisions.	None mentioned.	**Trust**: The D-S fusion engine helps human analysts interpret why data was flagged as a threat, facilitating their review process.

Paper ID	Algorithm / Model	XAI Technique	Scope Global- Local	Adversarial Defense / Testing	Impact on Trust / Robustness
Villarreal-Vasquez et al. (2023)	**LADOHD** (Long Short-Term Memory - LSTM),.	None mentioned.	Not specified.	None mentioned (Evasion/mimicry attacks are discussed as a motivation for the model but not specifically tested as a defense).	**Robustness**: Tackles the **Order-Aware Recognition (OAR)** problem, allowing the system to identify attack sequences mixed with benign actions over long periods.
Wall & Agrafiotis (2021)	**Bayesian Network** utilizing the **Bayes-Ball algorithm**,.	**Probabilistic Graphical Modeling** (inherent interpretability) and **prior knowledge encoding**,.	**Global/Local**: Learning the structure of "normal behavior" (Global) and performing specific queries for outcomes (Local),.	None mentioned.	**Trust**: Allows analysts to encode prior situational and behavioral knowledge into the model, providing justified and informed decisions for security operations
Wang & El Saddik (2023)	**Transformer** (DistilledTrans), BERT, and RoBERTa.	**Self-attention mechanisms**; Feature importance metrics; Digital Twin visualization.	Monitors organizational risk profiles (**Global**) and allows analysts to find root causes of specific alerts (Local).	None mentioned.	**Trust**: Provides transparency and interpretability to the detection process, helping analysts investigate and revert damages.
Wang Zhi et al. (2024) FedITD	**Federated LLMs** (BERT, RoBERTa, XLNet, DistilBERT) with PETuning.	None mentioned (Explicitly notes that Shapley values are unsuitable for their temporal sequence data).	Not specified.	**Privacy/Evasion**: Tested against **embedding inversion (data reconstruction) attacks**; mentions defense against **data poisoning**.	**Robustness**: PETuning and encryption effectively protect against data reconstruction, significantly reducing the accuracy of recovered text.
Wang Jiarong et al. (2023) Deep Cluster	**Encoder-Decoder** (RNN/GRU) with **Deep Clustering**.	None mentioned.	Not specified.	None mentioned.	**Robustness**: Improves detection accuracy by automatically adjusting feature representations to better suit the clustering task.
Wei Yichen et al. (2021)	**Cascaded Autoencoders (CAEs)**, BiLSTM, and Hypergraph.	None mentioned.	Not specified.	**Robustness:** Evaluated the ability of the CAE purification module to filter anomalies at varying ratios (5% - 40%).	**Robustness**: Unsupervised data purification enables the system to handle noisy training sets without relying on human-labeled "clean" data.
Wei Zhiyuan et al. (2024)	**Hybrid**: Information Gain (Entropy-based) and Local Outlier Factor (LOF).	**Information Theory** (Information Gain) for rule-based transparency.	**Local**: Focuses on the unique behavioral profile of each individual employee.	**Robustness testing**: Rigorously tested against **benign noise** (up to 10%) added to the dataset.	**Trust/Robustness**: High resilience to noise (95% accuracy at 10% noise); allows analysts to fine-tune statistical thresholds for personalized monitoring.
Wen et al. (2023)	SVD and Eigenvector Centrality	**Visualization** of abnormal character patterns and communication behavior patterns	**Local**: Suggests potential members colluding with key staff and reconstructs individual behavior patterns	None mentioned	**Trust**: Helps analysts identify key staff and visualize the synergistic relationships between them

Paper ID	**Algorithm / Model**	**XAI Technique**	**Scope Global- Local**	**Adversarial Defense / Testing**	**Impact on Trust / Robustness**
Xiao J. et al. (2023)	MEWRGNN (R-GCN, GCN, CAN-GAT)	**Feature Ranking**: Ranks the contribution of different edge-representation features	**Local**: Provides security analysts with insights for investigating specific *detected* threats	None mentioned	**Trust**: Provides understandable insights for security analysts to investigate the root causes of alerts
Xiao F. et al. (2025)	SENTINEL (ST-GNN: GCN, GRU, EGAT)	**Attention Mechanisms**: Edge Graph Attention (EGAT) and Hierarchical Contextual Attention (HCA)	**Local**: Provides fine-grained threat intelligence for specific abnormal behaviors	None mentioned	**Trust**: Enhances the representative ability of user behavior to provide fine-grained intelligence
Xiao H. et al. (2024)	CATE (Convolutional Attention & Transformer)	**Self-Attention Mechanism**: Integrated into both convolutional and transformer modules	Not specified (Authors note more local action-tracing is needed as future work)	None mentioned	**Trust**: empowers the model to focus on essential features, though authors acknowledge the need for clearer explanations in future work
Ye et al. (2025)	SqueezeNet (CNN) in Federated Learning	**DeepInsight**: Groups similar features together on a 2D plane for visual recognition	Not specified	**Robustness Testing**: Explicitly notes the framework *currently lacks* defenses against adversarial attacks	**Trust**: Adjacent grouping of similar features makes the data representation more valuable than individual handling
Yildirim & Anarim (2022)	Ensemble Learning (XGBoost/GBM)	**Feature Masking**: Handles correlated features and identifies weights of high-performance combinations	Not specified	None mentioned	**Robustness**: Feature masking allows the model to safely combine features from different domains (time and frequency) without performance drops
Zhu et al. (2024)	TL-AAE (TCN + LSTM Adversarial Autoencoder)	**Reconstruction Error Comparison**: Comparing errors of time features vs. event features	**Local**: Pinpoints which specific sessions have high probabilities of threatening behavior	**Adversarial Training**: introducing a discriminator to align latent features with a Gaussian prior	**Robustness**: Adversarial learning reduces reconstruction uncertainty, making the detection of stealthy threats more reliable

The following table is what the authors included in the SLR after manually cross-checking on the accuracy of the studies reported to have used XAI and adversarial training .

TABLE 9. Explainability And Adversarial Robustness

XAI / Robustness Approach	Description	Count (N)	Representative Studies
1. Feature Importance & SHAP	Technique: Post-hoc explanation methods like SHAP, LIME, and Feature Maps.	11	Focus: Explaining which inputs (e.g., "File Size" or " Logon Time") drove the decision. SHAP: [16], [40], [66] Feature Importance/ Feature Maps: [23], [30], [52], [54], [60], [73], [91], [93]
2. Attention Mechanisms	Technique: Interpretability via Attention Weights, Heatmaps, and Multi-head attention.	15	Focus: highlighting when in the sequence the threat occurred (Temporal localization). [25], [33], [34], [46], [57], [59], [67], [71], [72], [73], [76], [77], [83], [89], [92]
3. Rule & Logic-Based	Technique: Expert rules, Decision Tree paths, Knowledge Graphs, and Formal Verification.	10	Focus: White-box transparency where the logic is human-readable and follows defined taxonomies (graphs) or policies (rules). Rules/Graphs: [22], [58], [64], [82], [88], [90] Model Logic: [13], [28] [39], [56]
4. Visual & Natural Language	Technique: Converting logs to Images (CNN visualization), Text (LLM reasoning) and t-SNE.	8	Focus: Intuitive reporting and pattern visualization for non-technical analysts. LLM/ Natural Language : [8],[32] Visual: [50], [68], [69], [94], [95], [96]

Defense Against Adversarial Attacks

Technique Category	**Description**	**Count (N)**	**Representative Studies**
1. Poisoning & Noise Resistance	Hardening models against corrupted training data, local update tampering in federated settings, Unsupervised data purification and environmental noise reduction.	5	[22], [24], [47], [74], [87]
2. GAN & Adversarial Training	Defending against evasion attacks (FGSM, DeepFool), structural graph attacks, AI-generated samples, and behavioral shifts.	7	[15], [16], [31],[54], [69], [77], [89]