-

# CISSAN

## Collective intelligence supported by security aware nodes

## D1.2 Detection and analysis of weak signals - Update

**Editors:** Kurt Tutschku and Jianguo Ding (contact: kurt.tutschku@bth.se), BTH

**Abstract**

The deliverable D1.2 provides an update to the state-of-the-art (SoA) for selected technology areas of the CISSAN project, as presented in its predecessor D1.1. Its objective is to improve SoA descriptions and to highlight new developments, particularly with a focus on the notion of collective intelligence of security-aware nodes.

**Project** **CISSAN**

**Public Report**

**March 2026**

Participants in project CISSAN are (in alphabetical order with project coordinator first):

- University of Jyväskylä (coordinator)
- Affärsverken Karlskrona AB
- Arctos Labs
- Bittium Biosignals Ltd
- Bittium Wireless Ltd
- Blekinge Tekniska Högskolan
- Blue Science Park
- Clavister AB
- Councilbox Ltd
- Geodata ZT GmbH
- Mattersoft
- Mint Security Ltd
- Netox Ltd
- Nodeon Ltd
- Savantic AB
- Scopesensor Ltd
- Techinova AB
- Wirepas Ltd

CISSAN-Collective intelligence supported by security aware nodes

D1.2 Detection and analysis of weak signals – Update

Editors: Kurt Tutschku and Jianguo Ding, Blekinge Tekniska Högskola

Project coordinator: Ilgin Safak, University of Jyväskylä

CELTIC published project result

# Executive Summary

In the context of the rapidly changing cybersecurity landscape, driven by the evolution of artificial intelligence (AI), distributed systems, and the emergence of new technologies such as quantum computing, it is important to have a clear view of the State-of-The-Art (SoTA) in the field. The purpose of this report is to provide such an analysis, as it covers the SoTA in the main areas of research of interest to the CISSAN project, thus positioning the results of the project in the current scientific/technological context, including opportunities as well as risks, which have a significant impact on the long-term sustainability of the results.

The report is an update to previously identified SoTA areas for CISSAN in deliverable D1.1. This document, D1.2, provides both deepened SoTA in some of the areas outlined in D1.1 and new SoTA in new ones. It addresses the following areas: collective intelligence (CI) mechanisms in the context of cybersecurity (deepened and new), the use of Generative Adversarial Networks (GANs) in the context of adversarial defence (deepened and new), the use of Large Language Models (LLMs) in the context of threat hunting and intrusion detection (new), as well as the use of quantum computing as a potential threat to the current cryptographic foundations (new), including an analysis of the impact of the latter, along with the associated risk management and mitigation, including the use of post-quantum cryptography (PQC).

As a result of the SoTA analysis, the CISSAN project's strategy has been reviewed and improved to address the identified key technological risks and opportunities. In particular, the findings regarding emerging threats, including quantum computing and the emerging role of AI in offensive and defensive cyber operations, have been used to adjust the project's technical priorities. This includes the potential adoption of PQC in the CISSAN platform to mitigate the risks posed by the emergence of quantum computing attacks, along with appropriate risk management and mitigation strategies. This ensures that the CISSAN solutions are not only robust but also future-proof, able to address potential changes in the cybersecurity landscape.

For stakeholders, the report provides a well-structured, up-to-date overview of the technologies likely to influence CISSAN's future capabilities, investment strategies, and risk management approaches. It can be used for decision-making, as it explains where the field is going, which paths are becoming feasible, and which risks need to be addressed early.

In conclusion, the SoTA analysis is a strategic reference for guiding the project's technical direction and improving the robustness and futureproofing of its outcomes. As a result of the SoTA analysis, the project strategy has been reviewed and adjusted to address identified risks and opportunities, ensuring that the project's solutions remain relevant to the dynamic cybersecurity landscape and continue to provide value to stakeholders and the European ecosystem.

# List of Authors

(in alphabetical order by partner name)


- Jianguo Ding, BTH, Sweden
- Kurt Tutschku, BTH, Sweden
- Rodrigo Martinez, Councilbox
- Klaus Chmelina, Geodata, Austria
- Veikko Markkanen, University of Jyväskylä, Finland
- Sara-Päivi Paukkeri, University of Jyväskylä, Finland
- Ilgin Safak, University of Jyväskylä, Finland

# Table of Contents

# Abbreviations

| | |
|---|---|
| AI | Artificial Intelligence |
| AIQUSEC | AI-based QUantum-safe cyberSECurity automation and orchestration for edge intelligence in future networks |
| API | Application Programming Interface |
| ASR | Automatic Speech Recognition |
| AUC | Area Under the Curve |
| BFT | Byzantine Fault Tolerant |
| BTH | Blekinge Tekniska Högskolan |
| CCI | Collaborative Cybersecurity Intelligence |
| CGAN | Conditional Generative Adversarial Network |
| CI | Collective Intelligence |
| CLI | Command-Line Interface |
| CIA | Confidentiality, Availability, and Integrity |
| CISSAN | Collective Intelligence Supported by Security Aware Nodes |
| DCGAN | Deep Convolutional Generative Adversarial Network |
| DDoS | Distributed Denial of Service |
| DP | Differential Privacy |
| ECC | Elliptic Curve Cryptography |
| ECDSA | Elliptic Curve Digital Signature Algorithm |
| ELBO | Evidence Lower Bound |
| EU | European Union |
| FID | Fréchet Inception Distance |
| FIPS | Federal Information Processing Standards |
| FN | False Negative |
| GAN | Generative Adversarial Network |
| GNSS | Global Navigation Satellite System |
| GPU | Graphics Processing Unit |
| GPS | Global Positioning System |
| HMI | Human Machine Interface |
| ICT | Information and Communications Technology |
| IDS | Intrusion Detection System |
| IIoT | Industrial Internet of Things |
| IoT | Internet of Things |
| IS | Inception Score |

| | |
|---|---|
| JYU | University of Jyväskylä |
| LLM | Large Language Model |
| LSTM | Long Short-Term Memory |
| MES | Manufacturing Execution System |
| MIA | Membership Inference Attack |
| ML | Machine Learning |
| ML-DSA | Module-Lattice-Based Digital Signature Standard |
| ML-KEM | Module-Lattice-Based Key Encapsulation Mechanism |
| MMD | Maximum Mean Discrepancy |
| MQTT | Message Queuing Telemetry Transport |
| NIDS | Network Intrusion Detection System |
| NIS | Network and Information Systems |
| NIST | National Institute of Standards and Technology |
| OT | Operational Technology |
| PoC | Proof-of-Concept |
| PQC | Post-Quantum Cryptography |
| REST | Representational State Transfer |
| RL | Reinforcement Learning |
| ROC | Receiver Operating Characteristic |
| RSA | Rivest-Shamir-Adleman |
| SCADA | Supervisory Control and Data Acquisition |
| SH-DSS | Stateless Hash-Based Digital Signature Standard |
| SMOTE | Synthetic Minority Over-sampling Technique |
| SoTA | State-of-The-Art |
| VAE-GAN | Variational Autoencoder Generative Adversarial Network |
| WGAN-GP | Wasserstein Generative Adversarial Network with Gradient Penalty |
| WP | Work Package |
| XAI | eXplainable AI |

# 1 Introduction

CISSAN Work Package (WP) 1 examines the current state of relevant technologies, organizing and clarifying those needed for the project. Since local signals alone may be insufficient to detect complex attacks, the work explores intelligent, secure, and collaborative approaches—such as artificial intelligence (AI) (Deep Learning, Collaborative AI, and Generative AI), advanced log analysis, and blockchain-based information sharing among Internet of Things (IoT) and Operational Technology (OT) devices without relying on vulnerable central systems. In addition, modern cryptographic concepts need to be applied to secure the system's features in the Confidentiality, Availability, and Integrity (CIA) context.

## 1.1 Objective of this Document

The report is an update to previously identified the State-of-The-Art (SoTA) areas for CISSAN in D1.1 [1]. It deepens the SoTA and covers new SoTA in the following areas: collective intelligence mechanisms in the context of cybersecurity (deepening and new SoTA), the use of Generative Adversarial Networks (GANs) in the context of adversarial defence (new SoTA), the use of Large Language Models (LLMs) in the context of threat hunting and intrusion detection (new SoTA), as well as the use of quantum computing ((new SoTA),) as a potential threat to the current cryptographic foundations, including an analysis of the impact of the latter, along with the associated risk management and mitigation, including the use of post-quantum cryptography (PQC).

## 1.2 The Structure of this Document

The deliverable is structured as follows: Section 2 provides an overview of the CISSAN project and its architecture to prepare for discussions on the SoTA areas. Section 3 outlines and discusses three main SoTA areas: postquantum cryptography, GAN-based adversarial defence in cybersecurity, and collective Intelligence for Cybersecurity. The latter topic comprises deeper discussions of subareas of collective intelligence relevant to CISSAN, including data quality verification methods, anomaly detection mechanisms, blockchain-based trust management, LLMs for cybersecurity intelligence and collaborative cybersecurity intelligence, and threat hunting for cybersecurity. Finally, Section 4 provides a short summary, risk analysis, and outlook, ensuring that the project's solutions remain relevant to the dynamic cybersecurity landscape and continue to provide value to stakeholders and the European ecosystem.

# 2 Overview of the CISSAN Project

The CISSAN project aims to enhance the cybersecurity, cyber resilience, and automation of Internet of Things (IoT) and Operational Technology (OT) ecosystems that utilize device, edge, and cloud computing capabilities (thus, including IT elements). CISSAN proposes and implements algorithms, develops use case solutions and a CISSAN platform for countering IoT security and operational threats. It focuses on the collective intelligence (CI) of IoT network nodes, techniques for distributed security and operational monitoring, event tracking, attack detection, and response in IoT networks. Its objective is to apply collaborative and smart mechanisms to enhance local security awareness in case where local information and indications for an attack are not sufficient. Hence, the collaboration of nodes, i.e., collective intelligence, for gathering and analysing locally available security information is needed. As a results, CISSAN expects that systems based on its architectural concepts will be more reliable and more secure. The expected outcomes of the project include a collectively intelligent security platform consisting of a set of innovative algorithms, technologies, and solutions interconnected and integrated to the project use cases and experimental environments, evidence of value in and beyond the project use cases, plans and models for production use and commercial products and services.

There are five use cases in the CISSAN project, including transportation, smart grids, tunnel construction, manufacturing execution system (MES) and a joint use case. The **transportation** use case is represented by Mattersoft and Nodeon. Its focus is performing anomaly detection of Global Positioning System (GPS) data to detect jamming and spoofing attacks, malfunction of devices, other anomalies in buses and Global Navigation Satellite System (GNSS) satellite sensors. The **smart grids** use case is represented by Affärsverken and Techinova. This includes smart grid monitoring and control. Its focus in CISSAN is local and hybrid artificial intelligence (AI)-based anomaly detection in network traffic and physical sensor data to identify attacks and faults. The **tunnel construction** use case is represented by Geodata. This includes IoT systems for underground construction monitoring that securely signs, transmits, anchors and analyses IoT sensor data. The current focus in CISSAN is performing sensor data quality analysis by calculating believability scores to identify potential attacks, faults and abuse. Data integrity protection is achieved by securely logging data on the bitcoin blockchain and the use of security chips for sensor data signing. Bittium represents the **manufacturing execution system** use case, which is virtual manufacturing execution environment that controls the manufacturing of Bittium products and devices throughout the life cycle until the maintenance phase of the products. Its focus in CISSAN is performing local and system level anomaly detection for securing the virtual environment. Arctos Labs and JYU represents the **joint use case**, which is related to the mitigation of multi-domain critical infrastructure attacks using security task distribution, blockchain-based device and trust management and automated disaster recovery. The aim is to optimally distribute security tasks across the network and to leverage collective intelligence of each use case for the risk assessment and trust management of the network including using global data quality verification, anomaly detection and trust scoring. This is used in identifying, blacklisting and isolating malicious or anomalous devices. Distributed disaster recovery of devices that are unresponsive will be performed by optimally re-distributing their security tasks and activating failover nodes. This will enable the business continuity of these critical infrastructures. Details of the CISSAN platform and its use cases are available in the CISSAN D6.1 report.

## 2.1 The CISSAN Platform Architecture

The updated CISSAN architecture and its implementation is documented in the CISSAN deliverable D2.3 report. The layered architecture of the CISSAN platform, as shown in Figure 1, reflects a converged IoT/OT perspective and highlights its core components and their interactions within the overall system. It provides a clear modular integration, experimentation, and collaboration architecture for

IoT/OT domains, thus reducing the risks of development and the likelihood of duplication of efforts among the CISSAN partners.

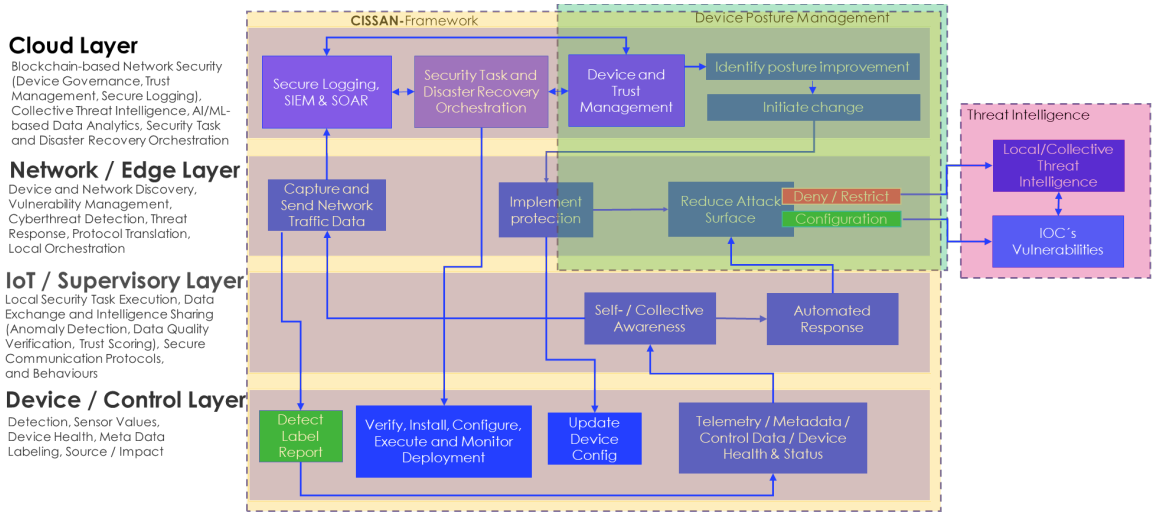


**Figure 1.** CISSAN platform layered architecture

The CISSAN platform is primarily deployed on-premise at JYU and consists of four layers, including the cloud layer, network / edge layer, IoT / supervisory layer and the device / control layer. These layers perform the following functionalities in the CISSAN platform:

- **Cloud layer:**
  - Blockchain-based network security, which includes device governance, trust management, secure logging, and data aggregation.
  - Aggregates and analyses collective threat intelligence, performs AI/Machine Learning (ML)-based data analytics, and security task and disaster recovery orchestration.
  - Devices' posture including health statuses, device behaviour, security posture, configuration posture, identity and trust postures, are continuously monitored by the cloud.
  - Appropriate actions are performed to enhance the device posture. This entails segregating the compromised or low-trust devices from the rest of the network, updating the security state of the devices in the device ledger, initiating remediation actions through the orchestrator, and enforcing access control rules at the network level, including triggering device isolation at the switch level.

- **Network / edge layer:**
  - Device and network discovery
  - Vulnerability management
  - Incidence detection
  - Threat response
  - Protocol translation
  - Local data and task orchestration

- **IoT / supervisory layer:**
  - Local security task execution
  - Data exchange and intelligence sharing (anomaly detection, data quality verification, trust scoring)
  - Secure communication protocols, and behaviours

- **Device / control layer:**

- Detection, sensor values, device health, meta data labeling, source / impact. Implementation of device security features include certificate-based device authentication, access control.
- In the device layer of IoT networks, devices sense and act on the physical world while running embedded logic, and send telemetry, metadata, and health/status information to the IoT layer, while receiving control commands and configuration in return.
- In OT networks, the control layer executes control of the physical process using sensor inputs and control logic, and sends process variables, states, alarms, and events to the supervisory layer for monitoring and higher-level control.

## 2.2 Related Research Areas and Technological Innovations

The CISSAN project strives to provide collective intelligence and collaborative defence solutions with the CISSAN platform and its use case solutions for enhancing the overall network security posture of IoT and OT networks with security aware nodes.

Recently AI-related threats have been one of the prominent areas of threat vectors, considering that the adversary is leveraging generative models and automation for carrying out attacks. For instance, generative models, which can encompass generative adversarial networks (GANs) and LLMs, can create personalized phishing content, deep fakes, telemetry data, and logs that can trick both human operators and machine-based detection systems. If the data is fed into collective intelligence and shared threat models, the information can create incorrect perceptions of the environment, corrupt trust models, and trick machine-based response models. Automated and adaptive attack orchestration is also becoming more prevalent, wherein reinforcement learning and planning models have been leveraged by the adversary for carrying out attacks.

Concurrently, the advent of quantum computing is also a near- to medium-term concern, wherein while quantum cryptanalysis has yet to be developed, the advancement of quantum computer hardware makes the possibility of ‘harvest now, decrypt later’ attacks on encrypted coordination and telemetry communication a near-term concern. This also highlights the requirement to ensure the inclusion of post-quantum cryptographic (PQC) techniques in the mechanisms of secure communication and trust exchange.

Conventional ML-based mechanisms used in the detection of anomalies and determination of trust also face the challenges of ‘data poisoning’ and ‘adversarial examples,’ wherein the misclassification of threats or normal behaviour by the collective intelligence mechanisms may be caused by the inclusion of malicious or ‘perturbed’ data. The last of the threats is the systemic risk of ‘trust and misinformation manipulation’ in the mechanisms of collective intelligence, wherein the ability of adversaries to inject false information or manipulate the identities of contributors to the collective intelligence mechanisms may be used to corrupt the integrity of the mechanisms of distributed decision-making.

These evolving threats highlight the requirement to ensure the development of collective intelligence mechanisms with high ‘adversarial resilience.’

# 3 Updates on CISSAN-Related Research and Challenge Areas

## 3.1 Postquantum Cryptography against Future Quantum Computing Attacks

This rapid development in the field of quantum computing represents a potential threat to the security of cryptographic primitives that are currently being used to protect the security and integrity of the IoT and OT networks. The cryptographic algorithms such as Rivest-Shamir-Adleman (RSA) and elliptic curve cryptography (ECC), which are being used to provide the necessary security and integrity to the communication protocols, identity management systems, and software updates, are likely to be compromised once the necessary quantum computers are available. The long lifespan of the industrial and OT systems, along with the “harvest now, decrypt later” threat model, makes the adoption and evaluation of post-quantum cryptography (PQC) techniques an immediate need to provide the necessary security and integrity to the systems [2].

The adoption and use of PQC techniques in the IoT and OT systems, specifically in critical infrastructures, are likely to pose specific challenges to the systems due to the nature and requirements of the systems [3]. The PQC techniques are likely to have longer key sizes and signatures compared to the classical cryptography techniques, which could be a problem given the nature and requirements of the systems. The longer key sizes and signatures could affect the performance and availability of the systems, which are critical and have long deployment cycles [4].

The SoTA analysis, therefore, takes into consideration the level of maturity of candidate PQC schemes, including those that are being standardized under the National Institute of Standards and Technology (NIST) PQC process, as well as their viability for use within heterogeneous IoT and OT infrastructures. This includes their potential impact on aspects such as device authentication, secure boot mechanisms, firmware updates, secure communication, and trust mechanisms. In addition, migration paths, use cases for classical and PQC-based schemes, and deployment paths are relevant for consideration.

The NIST PQC standardization process has paved the way of the development, standardization and adoption of new cryptographic techniques. NIST PQC Federal Information Processing Standards (FIPS) include Module-Lattice-Based Key-Encapsulation Mechanism Standard (ML-KEM) [5] (formerly known as CRYSTALS-Kyber) for key establishment and Module-Lattice-Based Digital Signature Standard (ML-DSA) [6] (formerly CRYSTALS-Dilithium) and Stateless Hash-Based Digital Signature Standard (SH-DSS) [7] (formerly Sphincs+) for digital signatures, focusing on cryptographic agility, long-term security, and preparation for quantum-based cyber threats in critical infrastructure systems. The European Network And Information Systems (NIS) PQC roadmap [8] is intended to offer strategic recommendations to European Member States and operators of critical infrastructures on how to address the question of preparing for and transitioning to quantum-resistant cryptography in a coordinated and risk-based manner. Its purpose is to address the long-term confidentiality, integrity, and authenticity of digital systems and communications in the face of the quantum computing threat, without pre-selecting certain technologies or algorithms, but rather focusing on the inventory of existing cryptographic systems, risk assessment related to quantum computing, and aligning national and sectoral strategies with internationally recognized standards, such as those promoted by NIST, to enable the selection of the best solution based on risk, constraints, and regulations.

As described in the D7.2 report, as part of WP1 efforts, upon the identification of quantum computing as a potential threat to the project results in the future, a collaboration between the CISSAN and AI-based Quantum-Safe Cybersecurity Automation and Orchestration for Edge Intelligence In Future Networks (AIQUSEC) [9] projects was initiated to ensure the sustainability of the project results long

term. As part of the cooperation, the consortium analysed the most economically viable and technically feasible approach for incorporating PQC in the CISSAN platform, where it was identified that securing external communications between the CISSAN platform would need to be prioritized. PQC schemes that could potentially be suitable for future-proofing secure communications between the CISSAN management server and external partner systems were selected as a result of a technical analysis. From the analysis, it was determined that the ML-KEM and ML-DSA algorithms are suitable for supporting a crypto-agile approach and enabling the use of post-quantum secure key exchange and authentication for the Application Programming Interface (API) and Message Queuing Telemetry Transport (MQTT)-based communications within the CISSAN platform. In particular, the considered schemes are ML-DSA as a signature scheme, ML-KEM as a key encapsulation mechanism with 32-byte encryption, while also allowing for parameter tuning based on specific security level requirements, such as Dilithium2, Dilithium3, and Dilithium5 and Kyber512, Kyber786, and Kyber1024.

A PQC command-line interface (CLI) tool developed by JYU outside of CISSAN was studied to determine the feasibility and potential compliance of the CISSAN platform with NIST standards. CISSAN partners, including Arctos Labs, Clavister, Councilbox, and Mattersoft, intend on integrating such a PQC tool post-project for securing their interfaces connecting to the CISSAN platform, which would include encrypting and digitally signing data prior to sending it via the interfaces defined in the D6.2 annexes.

As a result of examining the PQC CLI tool, the ML-KEM algorithm is gathered to be suitable for key establishment with low computational cost, even at higher security levels [4], and thus can be used for enhancing communications security. The ML-DSA algorithm, although computationally expensive, especially when generating signatures, is suitable for the proposed authentication and integrity checks during the setup of the connections, handling of certificates, and verification of messages. The main issue with MQTT-based communications and other constrained system components such as sensors, is anticipated to be related to the increased size of the keys, signatures, and messages used during the key exchange and setup, rather than the execution speed of the algorithms. For future PQC deployments, the lower and medium security parameter sets, such as Kyber512/Kyber768 and Dilithium2/Dilithium3, are identified to be suitable for most of the CISSAN use case system components with less stringent latency and bandwidth requirements, while the proposed higher parameter sets, Kyber1024 and Dilithium5, can be used for system components that do not have such stringent requirements. In all CISSAN use cases, high assurance security is critical. The most appropriate components to utilize for deploying the strongest post-quantum security parameter sets would be those residing in the management, control, and orchestration layers, which include the CISSAN management server, including the trust and governance components and external interfaces such as the metadata blockchain API and optimization APIs, and cross-domain integration interfaces therein, as well as the CISSAN orchestration server. These components reside in server-class environments, utilize API-based and MQTT communication protocols, and do not have to contend with hard real-time latency constraints. Therefore, they would be best positioned to support the more computationally expensive yet stronger post-quantum security mechanisms. Conversely, real-time control loops and highly constrained edge and sensor devices in the CISSAN use cases, will have significant latency, bandwidth, and resource constraints. Therefore, these components would best be serviced by terminating the post-quantum security mechanisms at gateways and during session establishment, while using symmetric keys for lightweight cryptography at edge devices and sensors.

In addition to increased computational complexity, increased size of the keys and messages, and increased latency of the operations, it is anticipated that in the integration of the PQC tool, an additional complexity may also arise due to the management of the cryptographic assets, the management of multiple parameter sets, and the consistent deployment of the CLI tool. The interoperability of the system with the existing APIs and communication protocols, such as Representational State Transfer (REST) APIs and MQTT, may also create technical difficulties. These issues can be addressed by employing a crypto-agile integration strategy, whereby PQC is first enabled in the non-

latency-critical interfaces (e.g., management server APIs) and then introduced in the MQTT links after profiling and validation in the target environments. Additionally, backward compatibility can be ensured through the use of hybrid configurations and parameter sets, which can allow performance-security trade-offs to be made per interface or class of devices. Operational security issues can also be addressed by packaging the CLI as a managed service or container with strict versioning, incorporating automated key and certificate management (generation, rotation, revocation), and incorporating monitoring and audit logging to detect failures or misconfiguration in the system. Lastly, end-to-end testing, benchmarking, and validation, as part of the hardening process, can help ensure consistency in behaviour across heterogeneous clients and servers from different partners.

## 3.2 GAN-Based Adversarial Defence in Cybersecurity

This section presents a technical review of Generative Adversarial Network (GAN) architectures for adversarial defense in cybersecurity. We synthesize findings from 185 peer-reviewed studies (2021–2025), introducing a four-dimensional taxonomy covering defensive functions, GAN architectures, cybersecurity domains, and threat models. Key results show that WGAN-GP achieves 14–30% accuracy improvements in intrusion detection, while CGANs enable targeted synthesis with F1-score gains of 10–20%. Additionally, hybrid models improve phishing detection recall by 18%. We identify persistent challenges, including training instability, computational overhead, and standardization gaps, then outline future research directions for scalable, trustworthy GAN-powered defenses.

### 3.2.1 GAN Foundations for Cybersecurity

GANs formulate generative modeling as a minimax game between generator G and discriminator D [10]. The generator maps latent vectors z ~ pz(z) to synthetic samples, while D distinguishes real from generated data:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))] \quad (1)$$

#### 3.2.1.1 Advanced GAN Variants

| Variant | Key Innovation | Cybersecurity Application |
|---|---|---|
| Deep Convolutional GAN (DCGAN) [11] | Convolutional architectures, batch normalization | Stable feature generation for Intrusion Detection System (IDS), biometric security |
| Wasserstein GAN with Gradient Penalty (WGAN-GP) [12] | Wasserstein distance, gradient penalty | Overcomes mode collapse; 14-30% IDS accuracy gain |
| Conditional GAN (CGAN) [13] | Conditional generation on labels y | Targeted synthesis: Distributed Denial of Service (DDoS) traffic, malware families |
| GAN-Reinforcement Learning (RL) [14] | Policy gradient for goal-directed generation | Adaptive adversarial crafting, real-time defence |
| Variational Autoencoder (VAE)-GAN [15] | Evidence Lower Bound (ELBO) + adversarial loss combination | Latent-structured synthesis, privacy-preserving |

WGAN Objective: Replaces JS divergence with Wasserstein distance:

$$W(P,Q) = \inf_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}\left[\|x-y\|\right] \quad (2)$$

Evaluation Metrics: Classification (Accuracy, F1-Score, Area Under the Curve (AUC)-Receiver Operating Characteristic (ROC)), Robustness (Automatic Speech Recognition (ASR) Robust Accuracy), Sample Quality (Fréchet Inception Distance (FID) and Inception Score (IS)), Statistical (Maximum Mean Discrepancy (MMD)).

### 3.2.2 Four-Dimensional Taxonomy of GAN-Based Defences

#### 3.2.2.1 Dimension 1: Defensive Function

a) **Data Augmentation:** CGANs generate minority-class samples. Results: F1-Score ↑ 12-20% vs Synthetic Minority Over-sampling Technique (SMOTE) on CIC-IDS2017 [16], [17]. See (1).

b) **Adversarial Training:** GANs generate perturbations δ to harden classifiers. Results: 25% IDS resilience improvement [18].

$$\min_{\theta} \mathbb{E}_{(x,y)}\left[\max_{\delta\in\Delta} \mathcal{L}(f_\theta(x+\delta),y)\right], \quad \delta \sim G(z,x) \quad (3)$$

c) **Privacy-Preserving Generation:**

$$\mathcal{L}_{DP} = \mathcal{L}_{GAN} + \lambda \cdot \text{PrivacyLoss} \quad (4)$$

ensuring (ε,δ)- Differential Privacy (DP) guarantees. Results: >90% data utility [19] [20].

#### 3.2.2.2 Dimension 2: Architecture Performance

| Architecture | Defense Role | Performance | Challenges |
|---|---|---|---|
| DCGAN | Synthetic data, malware | Baseline augmentation | Limited scalability |
| WGAN-GP | Stable IDS training | Accuracy↑ 14-30%, False Negative (FN) ↓ 22% | High computational cost |
| CGAN | Targeted synthesis | F1↑ 10-20% for rare attacks | Requires large labelled data |
| Hybrid GAN-RL | Adaptive defence | Recall↑ 18% in phishing | Training instability |

#### 3.2.2.3 Dimension 3: Cybersecurity Domains

Network Intrusion Detection System (NIDS): WGAN-GP improved detection by 14-22% on CICIDS2017; FN↓ 22% [21], [22]. Malware: CGANs simulate polymorphic variants; >85% detection rate [23], [24]. IoT: Accuracy >90% on IoT-23/Bot-IoT [25]. Phishing: GAN-RL hybrids improved recall by 18% [15].

#### 3.2.2.4 Dimension 4: Threat Models Addressed

| Threat Type | GAN Defense Mechanism | Performance |
|---|---|---|
| Evasion Attacks | CGAN-generated adversarial traffic for retraining | Evasion success↓ 15-20% [26] [23] |
| Poisoning Attacks | GAN-driven data cleansing pipelines | Accuracy↑ 9% under backdoor [16] |

| | | |
|---|---|---|
| Privacy Inference | Federated GANs with DP guarantees | Utility ~90%, Membership Inference Attacks (MIAs) mitigated [19], [20] |

### 3.2.3 Results Synthesis

| Application | GAN Variant | Dataset | Metric | Performance | Ref. |
|---|---|---|---|---|---|
| IDS Augmentation | WGAN-GP | CICIDS2017 | F1-score | ↑ 12-22% | [21] |
| Adversarial Training | CGAN | CICIDS2017 | Robustness | ↑ 25% | [18] |
| Malware Detection | DCGAN/CGAN | EMBER | Detection | >85% | [23], [24] |
| IoT Security | Standard GAN | IoT-23 | Accuracy | >90% | [25] |
| Phishing Detection | GAN-RL | PhishTank | Recall | ↑ 18% | [15] |
| Evasion Defense | CGAN | UNSW-NB15 | ASR | ↓ 15-20% | [26] |
| Privacy-Preserving | Federated GAN | Enterprise | Utility | ~90% | [19] |

**Key Findings:** (1) WGAN-GP dominates due to training stability and mode collapse mitigation. (2) CGANs excel in targeted synthesis for class-specific traffic. (3) Hybrid GAN-RL models show promise but requires stability improvements. (4) NIDS applications report the most consistent efficacy.

### 3.2.4 Technical Challenges

| Challenge | Description | Impact |
|---|---|---|
| Training Instability | Mode collapse, vanishing gradients in WGAN/hybrid | Reduced sample diversity |
| Computational Cost | High Graphics Processing Unit (GPU) requirements; WGAN-GP overhead | Limited IoT/edge deployment |
| Standardization Gap | No benchmark consensus; outdated datasets (NSL-KDD) | Poor reproducibility |
| Dual-Use Risk | GANs enable attacks (deepfakes, adversarial malware) | Governance complexity |
| Explainability | Black-box nature of GAN-based defences | Hindered critical-sector adoption |

### 3.2.5 Future Research Directions

1) **Adaptive Adversary Robustness:** Develop co-evolutionary GAN training where attack and defence models adapt jointly against multi-vector adversaries.
2) **Standardized Benchmarks:** Extend EMBER2024 and IoT-23 with adversarial variants; establish unified metrics (FID, ASR, robust accuracy).
3) **Lightweight Architectures:** Explore quantization, pruning, and knowledge distillation for edge deployment with latency constraints.
4) **Privacy-Utility Optimization:** Integrate differential privacy with federated GAN frameworks; formalize privacy leakage metrics.
5) **Explainable GAN Defences:** Incorporate explainable AI (XAI) techniques (attention visualization, feature attribution) for trust and accountability.
6) **LLM-Driven Threat Mitigation:** Develop adaptive GAN frameworks against emerging LLM-based cyberattacks [27].

## 3.3 Collective Intelligence for Cybersecurity

The increasing scale, diversity, and interconnected nature of current IoT and OT systems make traditional centrally managed cybersecurity solutions inadequate in effectively addressing the evolving and coordinated nature of cybersecurity threats. Collective intelligence for cybersecurity helps address this problem by allowing distributed elements, systems, and organizations to collectively observe, analyse, and respond to cybersecurity-related issues, thereby bringing together evidence, expertise, and capabilities in an integrated manner to formulate an overall cybersecurity strategy. This section discusses collective intelligence mechanisms for cybersecurity, including distributed anomaly detection, data quality verification, trust scoring, security task distribution, structured threat information sharing, and the application of LLMs in cybersecurity, among others. By promoting collaboration between devices, systems, and organizations, these mechanisms are intended to improve detection, enhance response, and reduce risks, thereby creating an opportunity for building an effective cybersecurity strategy in complex, interconnected, and evolving digital ecosystems.

### 3.3.1 Data Quality Verification Methods for Collective Intelligence

Data quality research is evolving and increasingly important. Methodologies for data quality measurement and improvement are evolving in several directions: (1) considering a wider number of data types and data quality dimensions, (2) moving from data quality to information quality, (3) relating data quality issues more closely to business process issues; and (4) considering new types of information systems, specifically Web information systems. In addition to mathematical, statistical and probabilistic methods, AI methods have been researched in recent years, leading to new ways of data quality assessment.

Many different data quality tools are on the market having implemented different of the above-mentioned methodologies helping to streamline and often automate data management activities required to guarantee that data remains fit for analytics, data science, and ML use cases. Table 1 gives an overview of just some of these tools.

**Table 1.** Comparison of data quality tools

| Tool | Release date | OSS | No code | AI/ML based monitoring | On-prem available |
|---|---|---|---|---|---|
| Great Expectations | 2017 | ✅ | ❌ | ❌ | ✅ |
| Deequ | 2018 | ✅ | ❌ | ❌ | ✅ |
| Monte Carlo | 2019 | ❌ | ✅ | ✅ | ❌ |
| Anomalo | 2021 | ❌ | ❌ | ✅ | ✅ |
| Lightup | 2019 | ❌ | ✅ | ✅ | ✅ |
| Bigeye | 2019 | ❌ | ❌ | ✅ | ❌ |
| Acceldata | 2018 | ❌ | ❌ | ❌ | ❌ |
| Observe.ai | 2017 | ❌ | ❌ | ✅ | ❌ |
| Datafold | 2020 | ✅ | ❌ | ❌ | ✅ |
| Collibra | 2008 | ❌ | ✅ | ✅ | ✅ |
| dbt Core | 2021 | ✅ | ✅ | ✅ | ✅ |
| Soda Core | 2022 | ✅ | ❌ | ✅ | ✅ |

While modern data quality tools offer a solid foundation, their capabilities vary significantly. Some excel in AI-powered anomaly detection and metadata management, while others prioritize AI-powered automation or ease of deployment. Table 2 provides a side-by-side comparison of the most critical features across leading AI-powered open-source data quality tools.

**Table 2.** Comparison of the most critical features across leading AI-powered open-source data quality tools

| Feature | Soda Core | Great Expectations | OpenMetadata | Amundsen | DQOps | Datafold | Deequ |
|---|---|---|---|---|---|---|---|
| AI-Powered Anomaly Detection | Partial | No | Yes | No | Yes | No | No |
| Automated Data Profiling | No | Partial | Yes | No | Yes | No | Partial |
| Natural Language Rule Generation | Yes | Partial | No | No | No | No | No |
| Metadata Management | No | No | Yes | Partial | Partial | No | No |
| Data Lineage Tracking | No | No | Yes | Partial | No | Partial | No |
| Built-in Governance & Access Controls | No | Partial | Yes | No | No | No | No |
| Streaming / Real-time Support | No | No | Partial | No | Partial | No | No |
| Integration with CI/CD | Yes | Yes | Yes | Partial | Partial | Yes | Yes |
| Data Diff / Schema Comparison | No | No | No | No | No | Yes | No |
| Ease of Deployment | Yes | Yes | Partial | Yes | Partial | Yes | Partial |

The listed tools enable the assessment of different data quality dimensions which are features that can be evaluated or analysed against a set of criteria to determine data quality. Measuring data quality dimensions helps identify data problems and determine whether data is appropriate to serve its intended purpose. Table 3 lists the most common and relevant data quality dimensions.

**Table 3.** Data quality dimensions

| Data quality dimension | Description | Examples |
|---|---|---|
| Timeliness | Data's readiness within a certain time frame. | A weather app updates its forecast every hour. If the data is delayed by 6 hours, users may make poor decisions based on outdated information. |
| Completeness | The amount of usable or complete data, representative of a typical data sample. | A customer database includes names and email addresses, but 30% of entries are missing phone numbers, making it harder to follow up with clients. |
| Accuracy | Accuracy of the data values based on the agreed-upon source of truth. | A GPS system shows a restaurant's location 2 blocks away from its actual address, leading users to the wrong place. |
| Validity | How much data conforms to acceptable format for any business rules. | A form asks for a birthdate, but someone enters "February 30"—which isn't a real date. The system should reject it as invalid. |
| Consistency | Compares data records from two different datasets. | A product's price is listed as €19.99 on the website but €24.99 in the mobile app. This inconsistency confuses customers and erodes trust. |
| Uniqueness | Tracks the volume of duplicate data in a dataset. | A patient record system has two entries for the same person with identical details. Duplicate records can cause medical errors or billing issues. |

While data quality assessment mostly is performed for economic objectives (e.g., to ensure process and product quality), its application for (cyber)attack detection is a quite new research area, and it is the idea followed in CISSAN.

The Data Quality Verification System developed in CISSAN (and described in Deliverable D4.5 in detail) assesses the data quality dimension "Believability" and aims to identify (cyber)attacks by detecting domain-specific anomalies in OT data, especially in time-series sensor data. The hypothesis is that (cyber)attacks might leave specific patterns, anomalies and/or data quality changes in operational data (e.g., monitoring data, sensor data) that can be detected. The system has been integrated into the CISSAN platform for detecting threats by scoring the believability of OT data at an early stage and triggering counteractions (e.g., the blacklisting of devices, the flagging of sensor data) in case the believability scores exceed defined thresholds. Together with other anomaly detection techniques developed in CISSAN, the method represents an additional component contributing to collective intelligence.

As the occurrence of data quality issues in OT data not necessarily have to be caused by a (cyber)attack but could have operational causes, the system has been based on an empirical/heuristic approach. Its data quality methods/rules have been developed and applied specifically for use case 3 Underground Construction Monitoring and in cooperation with explicit domain experts. While the developed system is based on proven and widely used mathematical/statistical methods, their implementation and application in the given use case goes far beyond the SoA in underground construction industry. The developed system is regarded as an innovation, increasing safety and security of tunnel construction projects.

The system is based on the following concepts:

- **Believability Scores**

  The data quality dimension “Believability” is represented by *Believability Scores* that are defined as calculated numerical values between zero (0) and one (1) for a certain granularity level and a certain believability aspect of the data. They define the degree the data is believed to be free of attacks.

- **Aggregation of Believability Scores, Granularity Levels**

  Believability Scores may be calculated and assigned to data at different granularity levels ranging from single sensor values (= lowest level) to the entire data set available (= highest level). In between, further levels exist (e.g., time series, clusters of sensors) establishing a hierarchical order of Believability Scores. Empirical functions are used to aggregate higher-level scores from lower-level scores.

- **Time Series Windowing**

  To especially enable the analysis and scoring of (sensor) time series in more depth, a Windowing technique has been developed. It allows the calculation of Believability Scores for certain parts (= windows) of time series. Different types of windows can be chosen such as sliding windows, overlapping windows and growing windows. In this way, the Believability of a whole time series can be aggregated from the Believability Scores of Windows of that time series.

- **Data Clustering**

  To allow for the analysis of trend similarity, neighbourhood behaviour, comparison with prediction and other aspects, data can be clustered and Cluster Scores calculated.

- **Data Quality Rules**

  The Believability Scores are calculated by empirically defined Data Quality Rules. They represent and apply the expert knowledge of the domain and are mainly based on statistical and mathematical methods.

- **AI/LLMs**

  The use of AI-based prediction models (Transformers) and LLMs (e.g., Claude AI) have been researched to support anomaly detection.

The above listed concepts have been implemented by using the following technologies and methods:

- **Dynamic Time Warping (DTW)**

  DTW [28] is a mathematical algorithm offering a way to compare the similarity of time series data. It is designed to align, compare and calculate the distances of time series datasets focusing on the shape of the data. The method is used to detect unexpected/unbelievable trends of time series when comparing them with neighbouring or reference time series.

Step-by-Step Execution:

❖ Step 1: Create the Distance Matrix

Suppose we have a reference sequence $X = (x_1, x_2, \dots, x_n)$, and a test sequence $Y = (y_1, y_2, \dots, y_m)$, we build a $n \times m$ matrix where each cell $i \times j$ represents the local distance (usually squared difference) between points $x_i$ and $y_j$.

$$d(i,j) = \left(x_i - y_j\right)^2$$

❖ Step 2: Calculate the Cost Matrix

We want to find the path through this matrix that minimizes the total cumulative distance. We fill a new matrix $D$ using this recursive formula:

$$D(i,j) = d(i,j) + \min\left(D(i-1,j), D(i,j-1), D(i-1,j-1)\right)$$

This ensures that for every step, we choose the "cheapest" path from the previous neighbors (left, bottom, or diagonal).

❖ Step 3: Trace the Warping Path

Starting from the top-right corner $(n,m)$, we trace back to $(1,1)$ by following the lowest values. This path represents the optimal alignment. The total value at $D(n,m)$ is the **DTW Distance**.

❖ Step 4: Using DTW for Anomaly Detection

To detect anomalies, we typically follow this workflow:

- **Establish a Baseline:** Take a "normal" time series (e.g., a healthy machine's vibration profile).
- **Calculate DTW Distance:** Compare new incoming data sequences against this baseline using the DTW algorithm.
- **Set a Threshold:**
  - **Normal:** The DTW distance is low (the shapes are similar, even if timing varies).
  - **Anomalous:** The DTW distance exceeds a predefined threshold (the fundamental shape of the data has changed).

- **Auto-Regressive Integrated Moving Average (ARIMA)**

ARIMA [29] is a statistical technique for analysing and forecasting time series data by combining three components: Autoregression (AR), Integration (I), and Moving Average (MA). It is used to predict time series and to detect unexpected/unbelievable behaviour from comparing them with their predictions.

Step-by-Step Execution:

❖ Step 1: Visualize and Check for Stationarity

ARIMA requires the data to be stationary (mean, variance, and covariance are constant over time). We check this using the **Augmented Dickey-Fuller (ADF) test**. **If the data is non-stationary:** We apply differencing ($d$). This removes trends and seasonality.

❖ Step 2: Identification ($p$ and $q$)

We use two specific plots to determine the $p$ and $q$ values:

- **ACF (Autocorrelation Function):** Helps identify the $q$ (Moving Average) term.

- **PACF (Partial Autocorrelation Function):** Helps identify the $p$ (Autoregressive) term.

❖ Step 3: Train the Model

We fit the ARIMA model on a "clean" historical dataset (data we know is mostly normal). The model learns the underlying coefficients:

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

❖ Step 4: Calculate Residuals (The Anomaly Key)

Once the model is trained, we use it to predict the next value in the series. We then calculate the **Residual** (Error):

$$Error = |ActualValue - \Pr e\,dictedValue|$$

❖ Step 5: Detecting the Anomaly

In ARIMA-based detection, the "Anomaly Score" is essentially the size of the residual.

- **Define a Confidence Interval:** Usually, we calculate the standard deviation ($\sigma$) of the residuals during training.
- **Set a Threshold:** A common threshold is $3\sigma$ (3 standard deviations) from the mean.
- **Flagging:** If the actual value falls outside the predicted confidence interval (the "prediction envelope"), it is marked as an anomaly.

- **Cumulative Sum (CUSUM)**

CUSUM [30] is a statistical quality control technique for monitoring sequential data to detect small, incremental changes in a process's mean. It is used to compare time series and to detect change points therein.

Step-by-Step Execution:

❖ Step 1: Normalize the Data

First, we calculate how far the current point $x_t$ is from the target $\mu$, usually scaled by the standard deviation $\sigma$.

$$z_t = \frac{x_t - \mu}{\sigma}$$

❖ Step 2: Set the "Slack" Parameter ($k$)

We introduce a parameter $k$ (the allowance or slack). This represents the "acceptable" variation. We only start accumulating a sum if the deviation exceeds $k$. Usually, $k$ is set to $0.5$ (representing half a standard deviation).

❖ Step 3: Calculate High and Low Sums

We maintain two running sums, $S_H$ (for upward shifts) and $S_L$ (for downward shifts). They are both initialized at 0.

- Upper CUSUM ($S_{H,t}$):

$$S_{H,t} = \max\left(0, S_{H,t-1} + z_t - k\right)$$

- Lower CUSUM ($S_{L,t}$):

$$S_{L,t} = \max\left(0, S_{L,t-1} - z_t - k\right)$$

The $\max(0, \ldots)$ part is crucial—it ensures that if the process returns to normal, the "memory" of the drift is eventually reset to zero rather than becoming a massive negative number.

- ❖ Step 4: Detecting the Anomaly

  To decide when a shift is an "anomaly" rather than just noise, we set a **Threshold ($h$)**.

  - **Flag Anomaly if:** $S_{H,t} > h$ or $S_{L,t} > h$.
  - Typically, $h$ is set to 4 or 5 times the standard deviation.

- **LLM Claude Sonnet**

  The Large Language Model Claude Sonnet has been researched and used to query for the presence of attack patterns in sensor time series.

- **WebAssembly**

  WebAssembly is a type of code that can be run in modern web browsers. It is a low-level assembly-like language with a compact binary format. It has been used to integrate the developed data quality verification methods into the CISSAN platform

### 3.3.2 Anomaly Detection Mechanisms for Collective Intelligence

**Anomaly detection for smart grids**

Smart-grid security monitoring increasingly treats anomaly detection as a distributed problem: observations originate at heterogeneous endpoints (smart meters, gateways, substations, control-centre systems), and both cyberattacks and faults often appear first as weak local signals that only become convincing when corroborated across multiple vantage points. This has pushed the literature toward collective-intelligence patterns in which local detectors produce lightweight evidence and coordination layers fuse, validate, and contextualize it across advanced monitoring infrastructure (AMI) and wider grid Information and Communications Technology (ICT). The same shift that has motivated threat hunting in Industrial Internet of Things (IIoT) also maps to smart grids: modern adversaries routinely bypass perimeter controls through credential abuse or misuse of legitimate access, making inside-the-environment monitoring necessary even when sessions appear “authorized” [31], [32], [33], [34].

**Previous work**

A practical baseline for collective intelligence in smart grids is hierarchical, fog-based monitoring. [35] propose a hierarchical and distributed IDS spanning AMI tiers (e.g., HAN, neighbourhood aggregation, operation centre) to address the common failure mode of single-tier visibility and to improve containment and tracing under coordinated attacks. [36] similarly argue that fog-based collaboration reduces cloud-induced latency by enabling localized decisions and faster detection in insider-style scenarios but note the management burden and trust implications of relying on intermediate fog nodes.

Alongside these architectures, research continues to refine what each node contributes. Network-centric monitoring remains relevant because it can reduce false positives by understanding industrial protocols and expected traffic patterns; for example, [37] extend deep inspection for industrial traffic to better detect network-borne threats. However, this class of methods can miss intrusions that operate through legitimate Human Machine Interface (HMI)/ Supervisory Control and Data Acquisition (SCADA) channels without obvious network anomalies, which is a serious concern in environments where operator-level manipulation and credential abuse are common [38]. This limitation motivates

CI designs that combine network evidence with device- and process-level indicators rather than relying on one signal type.

A second direction is edge-level monitoring, where devices perform lightweight detection and share only compact evidence. [39] shows that ensemble-style edge analytics (majority voting across simple classifiers) can improve efficiency while maintaining accuracy, suggesting that “many weak checks” can be more deployable than one heavy model. In the smart-grid context, this supports endpoint-level anomaly scoring on meters or gateways, with correlation occurring at higher tiers. At the same time, resource constraints remain central: [40] emphasizes that severe limitations on edge devices make fully local detection hard to operationalize, reinforcing the need for hierarchical splitting of computation.

A third direction embeds “collective reasoning” into the structure of the grid itself. [41] proposes security state awareness by mapping device indices into abstract states and analysing state transitions, enabling anomaly flags when transitions violate learned patterns. Translating to smart grids, this favours sharing state-change evidence and transition violations as hunt-support telemetry rather than raw logs. Complementing this, topology-aware approaches make the collective explicit: [42] proposes temporal, connectivity-informed detection where information is exchanged along grid connectivity to identify false data injection and ramp-style anomalies, demonstrating how the physical network can serve as the substrate for collaborative inference even when the anomaly is gradual rather than abrupt.

Recent CI work also explores “collaboration without raw data sharing,” mainly through federated learning. [43] shows that training locally on smart-meter-class devices and aggregating parameters can achieve performance comparable to centralized training while maintaining privacy, and [44] shows similar collaboration with simpler local models deployed across fog-edge layers to reduce latency. Surveys synthesize these findings by framing federated approaches as a practical CI pattern for large-scale, privacy-sensitive infrastructures, while highlighting unresolved issues around heterogeneity and security of the collaboration itself [45]. Finally, protected collaboration has been explored to mitigate inference and poisoning risks: [46] combines federated updates with homomorphic encryption and client-weighting to resist attacks on the learning process, but at added computational and coordination cost.

**Research gaps**

Across these approaches, three recurring gaps remain. First, feasibility under real resource budgets is still uncertain: many proposals are validated in settings that do not reflect the weakest nodes in the AMI (meters and constrained controllers), and even when local computation is claimed “lightweight,” continuous monitoring, frequent update rounds, encryption, or topology-aware message passing can create CPU, memory, power, and bandwidth burdens that are difficult to justify in outage-intolerant operations [40], [43], [46]. Second, trust and robustness in CI remain incomplete end-to-end: hierarchical and collaborative architectures implicitly assume honest contributors, yet fog nodes, meters, and gateways can be compromised; poisoning, collusion, and insider betrayal are acknowledged but often not addressed with integrated trust-weighting, auditable influence tracking, and resilient fusion logic that remains effective under partial compromise [36], [45], [46]. Third, operational usefulness is often underdeveloped: many systems emphasize detection accuracy but provide limited hunt support, such as explainable triggers, traceable evidence chains, and low-false-positive workflows. This matters because smart-grid incidents frequently involve legitimate channels and slow-evolving manipulations where single-signal alerts are unreliable and operators need justification for actions [38], [42], [31].

Taken together, the literature supports the relevance of CI in smart grids, especially for multi-tier AMI visibility and low-latency detection, while leaving space for architectures that combine lightweight

local anomaly evidence, peer corroboration, and explicit trust and traceability under tight resource constraints. This aligns with CISSAN goals without undermining their value.

### 3.3.3 Blockchain-Based Trust Management

Blockchain-based trust management systems address the disadvantages of centralized trust management systems by eliminating single point of failures by transforming their architectures into distributed ones. While fully distributed trust management systems are beneficial for improving the resilience against the existence of single points of failure, the literature shows that, in resource-constrained and highly dynamic IoT environments, these systems often face challenges related to the global consistency of the trust information, the robustness against manipulation and collusion, and the scalability of the trust dissemination and computation [47].

Recent research has shown a clear trend in the development of hierarchical, lightweight, or semi-centralized blockchain-based trust management systems, which address the trade-offs between decentralization and the limited resources of IoT devices. In this regard, blockchain-based trust management systems provide the required immutability, traceability, and auditability of trust while ensuring the scalability of the system by offloading the computation of trust from the constrained environment of the IoT device. A hierarchical blockchain-based trust management system maintains a blockchain in the intra-organizational environment of the IoT device and another blockchain in the inter-organizational environment of the IoT device, where the trust scores of the device in the inter-organizational environment are aggregated from the trust scores of the device in the intra-organizational environment to improve the scalability of the system and mitigate the effect of the presence of a number of malicious nodes in the system [48]. Similarly, a semi-centralized blockchain-based trust management system has been proposed in [49], which maintains a blockchain-based ledger and a centralized trust computation system to address the trade-offs between the scalability of the system and the energy constraints of the system. The system maintains a blockchain-based ledger and a centralized trust computation system, where the trust computation system takes advantage of the immutability, traceability, and auditability of the blockchain-based ledger while addressing the trade-offs between the scalability of the system and the energy constraints of the system. Lightweight blockchain trust schemes highlight the importance of adjusting the computation, storage, and consensus of trust in constrained environments, with the limitations of energy, storage, and bandwidth remaining the major bottleneck for the development of decentralized systems [50].

In CISSAN, a zero-trust IoT network security framework using distributed blacklisting, trust scoring and smart contracts (ZETROS) [51] was developed. ZETROS secures IoT networks against insider and outsider attacks, relying on blockchain for transparency and trustworthiness of trust computations. Trust computation in ZETROS is based on trust scores that are derived from evidence of past behaviour and interactions and are performed through smart contracts on a blockchain. This is in line with the overall literature, which shows a clear trend toward blockchain-based trust systems being significantly more resilient to trust-based attacks, collusion, and data tampering, as opposed to traditional trust-based systems. The trust model described in ZETROS has been adopted in the CISSAN platform, both conceptually and through simulation. In the CISSAN architecture, global trust scores are computed by the CISSAN management server, and the trust state is stored on the device ledger of the Councilbox blockchain system. It is important to note that the CISSAN project's strategic objective is to facilitate a gradual transition from legacy, centrally managed systems to a more decentralized architecture, without compromising the feasibility, performance predictability, and compatibility with existing industrial environments. In addition, there are several challenges associated with trust score computation, and these challenges have been addressed by the CISSAN architecture, which computes trust off-chain and then anchors it on-chain. In addition, the CISSAN project's approach to decentralization is done at the system level, allowing multiple management servers, each from a different organization, to interoperate via a shared blockchain network. In this case, the approach to decentralization aligns well with the constraints of legacy systems and systems of trust,

allowing for a more decentralized approach to trust and a decentralized architecture, rather than a decentralized application. This is in line with some of the trends identified in the literature, which suggests that a semi-centralized approach, such as a hierarchical trust-based system, is a potential solution to some of the performance, scalability, and resource constraints inherent in IoT systems. However, this also identifies a gap in current research, as many of the current trust-based systems in the literature, including those identified as hierarchical and lightweight, are still primarily validated through simulation, and even those acknowledge that there are still numerous challenges to be overcome.

Recent surveys of blockchain-based trust in IoT and IIoT note various mechanisms for decentralized trust management, highlighting challenges in scalability, energy efficiency, and integration with constrained devices and dynamic environments [52]. The Eventchain system, developed by Councilbox for the CISSAN project, implements a purpose-built blockchain that addresses these gaps.

Eventchain employs a Byzantine Fault Tolerant (BFT) consensus protocol in which a master node creates blocks and all hub nodes independently validate them, with consensus achieved when a majority submits verified votes. Rather than using reputation-weighted or performance-based leader election as explored by [53] and [54], Eventchain implements a zero-trust identity model where node identity is determined through ECDSA signature verification, favouring simplicity and verifiability over throughput optimization—properties desirable in multi-organizational deployments where node reputation cannot be assumed a priori.

A distinctive feature is the integration of anomaly detection directly into consensus. While some systems treat detection and trust scoring as parallel processes, Eventchain embeds anomaly assessments in transactions and requires hub nodes to agree on them during validation. Disagreement causes block rejection, elevating threat detection to a core consensus operation. The deterministic training pipeline—using fixed random seeds, sorted data processing, and parameters derived from the blockchain state—ensures verifiable and reproducible collective anomaly assessment without a centralized model authority.

For efficient state management, Eventchain implements a key-value store approach, providing efficient lookups for current device state without full blockchain traversal, responding to observations that practical systems must offer low-latency queries in time-sensitive OT environments [55]. The complete device lifecycle—registration, trust score updates, anomaly events, and blacklisting—is recorded as native blockchain transactions through a REST API, each signed with ECDSA secp256k1 and validated through BFT consensus [56]. Trust scores are computed off-chain by the CISSAN management server and anchored on-chain, following approaches identified as practical for reconciling decentralization with IoT performance constraints [57]. Periodic blockchain anchoring provides external immutability proof, aligned with emerging standardization efforts for blockchain-based timestamping [58].

The system inherits trade-offs common to blockchain systems for IoT [52] consensus validation scales linearly with hub count, the ECDSA signing scheme is not quantum-resistant (migration to NIST PQC algorithms is planned post-project), and storage grows linearly with transaction volume. Eventchain provides practical evidence that a purpose-built blockchain with BFT consensus, consensus-integrated anomaly detection, and efficient state management can meet the operational requirements of an IoT security platform without introducing prohibitive overhead.

To ensure that anomaly assessments are collectively agreed upon rather than unilaterally declared, the validation protocol requires every hub node to reproduce the anomaly metadata embedded in each transaction before accepting a block. Each node applies the same detection pipeline—comprising statistical and temporal analysis layers—using models trained deterministically from shared blockchain state with fixed parameters and canonical data ordering, an approach consistent with recent work on verifiable computation through controlled determinism [59]. Because all honest nodes derive identical models, any discrepancy in the reproduced scores triggers block rejection under the BFT quorum rule, making the network's anomaly assessment as tamper-resistant as the ledger itself [60]. This consensus-grade verification distinguishes the approach from architectures where

detection runs alongside—but outside—the consensus loop [61], and ensures that no single node can inject or suppress anomaly labels without majority agreement.

From a collective intelligence approach, the literature also points to some potential systemic risks that could impact the outcomes of a project if they are not considered. These include trust systems being subject to trust-based attacks, collusion, and manipulation, as well as being critically dependent on the quality, diversity, and integrity of evidence sources. Blockchain-based trust management also adds trade-offs, such as latency, costs, governance, and scalability, which are important considerations, especially for resource-constrained and real-time OT environments. All of these are potential technical and strategic threats to various collective intelligence-based concepts, including trust scoring, distributed blacklisting, and collaborative defence.

In conclusion, the SoTA is observed to affirm the proposed approach of the CISSAN project to rely on blockchain-supported trust management and trust scoring as a basis for zero-trust-based collective intelligence security. At the same time, the identified gaps regarding scalability, performance, and the cost of fully decentralized trust computation based on smart contracts also affirm the proposed hybrid approach of the project as being more than justified and realistic. It can thus be concluded that the proposed approach to collective intelligence security is viable and can provide the necessary security benefits to the European Union and the stakeholders of the project today while also being able to serve as a basis for the extension of trust computation to more decentralized models in the future.

### 3.3.4 Updates on LLM for Cybersecurity Intelligence and Collaborative Cybersecurity Intelligence

LLMs have transformed our world in recent years. They can understand the context of data and interact with users using natural language. In dynamic IoT environments, LLMs have the potential to analyse heterogeneous data without heavy data-analysis steps such as feature selection and model training. However, cybercriminals can also use LLMs to enhance their attacks. For example, Cisco Talos researchers have identified LLMs that cybercriminals can leverage in cyberattacks [62]. Therefore, it is essential to explore the possibilities of utilizing LLMs in cybersecurity intelligence, as well as to understand their current limitations and the differences between models.

Recently, several organizations have published SoTA LLMs, each claiming that their model outperforms others in one or more tasks. OpenAI's latest model, GPT-5.2, is SoTA across multiple benchmarks [63]. Google's Gemini 3 leads in reasoning [64]. The Mistral 3 series includes both a SoTA open-source model and a SoTA model designed for edge devices [65]. Claude 4 is SoTA in coding [66].

LLMs are rapidly evolving, and previous studies have used them for several cybersecurity-related tasks, such as orchestrating signatures and policies in IoT environments [67], explaining anomalies detected in network traffic [68], [69], [70], suggesting mitigation plans [68], selecting most relevant features and detecting anomalies in wireless communication [71], detecting anomalies in computational workflows [72], acting as honeypots [73], [74], or serving as sparring partner for pen-testers [75]. These works demonstrate that LLMs have the potential to increase the internal intelligence of cybersecurity systems. However, qualitative analysis of LLMs in intrusion detection and LLM agent-powered NIDSs have received only limited attention.

In CISSAN, LLMs have been explored in various ways. The primary focus was intrusion detection from raw IoT network traffic data. First, we prompted LLMs to analyse individual IoT network traffic packets, and we qualitatively evaluated their responses based on answer correctness, reasoning, and evidence identification. Next, we attempted to fine-tune small LLMs to understand hexadecimal-formatted IoT network traffic data. However, the results from this approach were not satisfactory. Finally, we built an LLM agent using five different LLMs and explored its performance in detecting

malicious activity in IoT network traffic captures. Our research showed that LLMs have the capability to understand IoT network traffic data and detect malicious activity without network-specific training or manual data processing. In addition, we proposed an intrusion detection process that utilizes LLM agents as part of NIDS. Future research questions include, for example, how to combine LLM agents and traditional ML models for collaborative cybersecurity intelligence in intrusion detection systems.

Collaborative Cybersecurity Intelligence (CCI) refers to the systematic sharing, fusion, and coordinated exploitation of security-relevant information across multiple entities, platforms, and analytical components to enhance collective situational awareness and defensive capability [76]. In distributed and heterogeneous IoT ecosystems, where individual devices or network segments often possess only partial visibility, collaboration becomes a prerequisite for effective threat detection, attribution, and response.

Recent advances in distributed analytics and federated learning [77] have further reinforced the feasibility of CCI by allowing participants to contribute model insights or behavioural abstractions without exposing sensitive raw data. Such mechanisms are particularly relevant in IoT environments characterized by resource constraints, privacy requirements, and organizational boundaries. Collaborative learning approaches can therefore facilitate the creation of shared detection capabilities while maintaining data sovereignty and regulatory compliance.

In the context of intelligent intrusion detection, CCI can be operationalized through cooperative anomaly detection, shared model adaptation, and coordinated alert enrichment workflows. Detection outputs produced by local agents may be exchanged or summarized into higher-level indicators, enabling peer components to validate findings, refine confidence levels, and contextualize events within broader threat landscapes. This collective reasoning process supports improved detection accuracy, reduced false positives, and faster incident triage.

Within the CISSAN collaboration, a Proof-of-Concept (PoC) implementation of a fully distributed, agent-based anomaly detection system was realized using a distributed autoencoder architecture. In this setup, each agent generates a latent representation of its local IoT data and shares this representation across the network. A shared decoder reconstructs the data locally on each agent, where a dedicated anomaly detection routine evaluates deviations from expected behaviour. This design allows agents to detect anomalies based on reconstructed patterns while minimizing the need to transmit raw data, preserving privacy, and reducing communication overhead.

This approach operationalizes the principles of CCI by enabling agents to collaboratively learn and refine a shared model of normal system behaviour. By exchanging latent representations rather than raw data, the system supports collective situational awareness and coordinates anomaly detection while maintaining autonomy and local decision-making.

### 3.3.5 Collective Intelligence-based threat hunting for cybersecurity

Newer generations of industrial controllers are being designed with additional computational headroom, allowing them to serve as both operational nodes and lightweight security sentinels. This growing onboard capacity creates an opportunity to embed local, device-level anomaly detection that doubles as hunt-support telemetry. In CISSAN, instead of relying exclusively on centralized tools, each device can contribute evidence about abnormal behaviour, support distributed detection logic, and cooperate with peers as part of a collective intelligence driven cybersecurity approach. By sharing locally observed anomalies, these devices support a collective threat-hunting model, where unusual activity identified by one node can trigger heightened vigilance or response across the entire network. This shift represents an important step toward resilient, distributed, and intelligence-driven protection in modern industrial systems, while introducing a negligible strain on the devices themselves.

To support this research direction, this section reviews relevant academic literature to identify early findings, emerging approaches, or recurring observations that may be relevant to the project.

**Previous work**

Threat hunting has gained increasing attention in IIoT security research as modern attacks routinely bypass perimeter controls through compromised credentials or misuse of legitimate access [31], [32], [33], [78]. This recognition has shifted focus toward defensive approaches that operate inside the operational environment, enabling devices to identify abnormal behaviours occurring within authenticated sessions.

A notable research direction involves deploying lightweight threat-hunting or anomaly detection logic directly at the edge. [39] proposes an ensemble-based detection model that aggregates several basic classifiers through majority voting to improve local detection accuracy and efficiency. Their results show strong performance on an IIoT-specific dataset, illustrating the potential of edge-level analytics, although the evaluation relies on high-performance hardware rather than constrained industrial devices. [40] similarly emphasise that while proactive threat hunting is critical for IIoT resilience, the severe resource limitations of edge devices make fully local detection difficult to operationalise in practice.

A complementary direction is found in traditional network-centric approaches. [37] extends the IDS framework with deep inspection capabilities for industrial application-layer protocols. Their enhancements significantly reduce false positives and detect a broader set of network-borne threats, addressing an important limitation in industrial environments where unnecessary alerts can disrupt operations. However, their system cannot detect intrusions performed through legitimate HMI or SCADA channels that do not produce immediate network anomalies. This is a significant gap given modern adversaries' frequent use of credential abuse and operator-level manipulation [38].

To overcome these limitations, research has increasingly turned toward collective intelligence and collaborative defence. [79] demonstrate that distributing Long Short-Term Memory (LSTM)-based intrusion detection across fog nodes can yield very high accuracy, though computational demands restrict deployment to capable fog hardware rather than true IIoT endpoints. [35] similarly proposes a hierarchical fog-based IDS for smart grids, showing that cooperative monitoring reduces detection latency, though their model coordinates across grid layers rather than enabling cooperation among resource-constrained devices within a single segment. [41] advances the edge-intelligence perspective by introducing a security state-awareness model that maps device performance indicators into abstract states and analyses transitions using complex-network theory. Their work demonstrates that meaningful security insights can be generated directly on the device, but the approach remains limited to identifying operational anomalies rather than subtle traces of compromise and presumes computational capacity beyond that available in many industrial controllers.

More autonomous edge-level detection capabilities appear in recent work. [80] presents a Markov-Game-based threat-response architecture capable of classifying local device states and triggering isolation when malicious behaviour is detected. By processing simplified feature sets at the edge, the approach improves response time and reduces false positives. However, the underlying model depends on continuous retraining and high-performance computation, leaving practical deployment on embedded IIoT hardware uncertain. A related direction is explored by [81], who introduces a hierarchical edge-computing anomaly-detection model that distributes analysis across sensor nodes and base stations. They leverage lightweight fuzzy-theory computations, at the sensor end, producing anomaly scores from single-source, single-moment data, while base stations execute more complex temporal and spatial correlation methods across multiple sensors to confirm abnormal trends. This layered approach reduces latency and cloud dependency, while enabling localised triggering of warnings and emergency responses. While effective in structured sensing environments, the model assumes homogeneous sensor behaviour and does not generalise well to heterogeneous industrial

controllers operating diverse processes. Furthermore, the approach is applicable only to data streams and thus lacks device level visibility.

**Research gaps in academic literature**

Across these diverse studies, recurring research gaps become clear. Most detection models are evaluated on high-performance computing systems rather than deployed on the constrained devices they are meant to protect, leaving real-world applicability uncertain. Many approaches rely heavily on curated datasets that fail to capture operational noise, evolving industrial behaviours, or adversarial presence within systems. Detection mechanisms are frequently tailored to specific attack types, rather than addressing the broader challenge of identifying stealthy or unknown intrusion activity within authenticated sessions. ML models often incur computational, and memory demands incompatible with resource-constrained controllers, and industrial environments remain particularly sensitive to false positives due to the operational costs of unnecessary interruptions. Furthermore, research tends to analyse external data streams such as network traffic or aggregate sensor readings, overlooking internal system traces that often contain early indicators of compromise. Finally, although both threat hunting and distributed intelligence appear as emerging research directions, they are rarely integrated into unified architectures that combine lightweight local inspection with cooperative, peer-supported reasoning.

Collectively, these findings show that while distributed, hierarchical, and edge-centric security solutions offer clear promise, there remains a significant gap between academic proposals and deployable systems. This gap highlights the need for lightweight, cooperative, and operationally feasible threat-hunting architectures capable of functioning reliably on constrained IIoT devices, and for approaches that combine local inspection with shared intelligence across the network.

# Conclusions

This report provides a detailed State-of-The-Art (SoTA) analysis of the main research areas of interest to the CISSAN project, with a specific focus on collective intelligence for cybersecurity, the latest AI-based approaches such as generative adversarial models (GANs) and large language models (LLMs), and the implications of the latest advances in these areas for the development of more secure and resilient digital systems. In effect, the SoTA analysis provides an overview of the latest approaches and trends in the relevant areas of interest to the project. This overview highlights the opportunities offered by the latest advances in the field and the challenges they present to the success of the CISSAN approaches and solutions, particularly in terms of long-term security, robustness, and sustainability.

One of the main findings of the SoTA analysis is the identification of the long-term strategic threat to some of the underlying foundations of the project's approaches and solutions posed by the development of quantum computing. While the latest approaches to the development of post-quantum cryptography (PQC) and the integration of the latest cryptographic models into the CISSAN platform and its interfaces have already been assessed, their actual implementation in the CISSAN platform and its interfaces are part of future work post-project. The integration of PQC in collectively intelligent, secure IoT and OT networks is an important step towards future-proofing these infrastructures against quantum-based adversaries. By analysing the potential implications of quantum computing and integration of PQC tools, CISSAN contributes to the development of strategic planning for future critical IoT and OT infrastructures that can maintain confidentiality, integrity, and trustworthiness throughout their entire lifetime, even in the presence of evolving cybersecurity threats. In addition, the SoTA analysis of the latest approaches to AI-based attack and defence models using GANs and LLMs highlights the need for the continuous monitoring of the latest developments in the field in order to ensure the long-term success of the project's approaches and solutions.

# References


[1] K. Tutschku and J. Ding, “Detection and analysis of weak signals, Deliverable D1.1 of the CISSSAN project, available on request.,” 2024.

[2] J. Mascelli and M. Rodden, “Harvest Now, Decrypt Later”: Examining Post-Quantum Cryptography and the Data Privacy Risks for Distributed Ledger Networks,” Finance and Economics Discussion Series 2025-093. Board of Governors of the Federal Reserve System, Washington, 2025.

[3] J. O. d. Moral, A. d. iOlius, G. Vidal, P. M. Crespo and J. E. Martinez, “Cybersecurity in Critical Infrastructures: A Post-Quantum Cryptography Perspective,” *IEEE Internet of Things Journal,* vol. 11, no. 18, pp. 30217-30244.

[4] E. D. Demir, B. Bilgin and M. C. Onbaşlı, “Performance Analysis and Industry Deployment of Post-Quantum Cryptography Algorithms,” *arXiv,* 2025.

[5] National Institute of Standards and Technology (NIST), “Module-Lattice-Based Key-Encapsulation Mechanism Standard (ML-KEM). Federal Information Processing Standards 203,” U.S. Department of Commerce, 2024.

[6] National Institue of Standards and Technology (NIST), “Module-Lattice-Based Digital Signature Standard. Federal Information Processing Standards (FIPS) 204,” U.S. Department of Commerce, Gaithersburg, MD.

[7] National Institute of Standards and Technology (NIST), “Stateless Hash-Based Digital Signature Standard. Federal Information Processing Standards (FIPS) 205,” U.S. Department of Commerce, 2024.

[8] NIS Cooperation Group, “A Coordinated Implementation Roadmap for the Transition to Post-Quantum Cryptography. Part 1, Version: 1.1,” EU PQC Workstream, 2025.

[9] "AI-based quantum-safe cybersecurity automation and orchestration for edge intelligence in future networks (AIQUSEC)," [Online]. Available: https://converis.jyu.fi/converis/portal/detail/Project/184185091.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 2014.

[11] A. Radford, L. Metz and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks.,” *arXiv preprint arXiv:1511.06434,* 2015.

[12] M. Arjovsky, S. Chintala and L. Bottou, “Wasserstein generative adversarial networks.,” in *International conference on machine learning*, 2017.

[13] M. Mirza and S. Osindero, “Conditional generative adversarial nets.,” *arXiv preprint arXiv:1411.1784,* 2014.

[14] D. Zhan, X. Liu, W. Bai, W. Li, S. Guo and Z. Pan, “GAME-RL: Generating Adversarial Malware Examples Against API Call Based Detection via Reinforcement Learning,” *IEEE Transactions on Dependable and Secure Computing,* vol. 22, 2025.

[15] K. Chan, B. Abu-Salih, R. Qaddoura, A. M. Al-Zoubi, V. Palade, D. Pham, J. Del Ser and K. Muhammad, “Deep neural networks in the cloud: Review, applications, challenges and research directions.,” *Neurocomputing,* 2023.

[16] L. Agarwal, B. Jaint and A. K. Mandpura, “Reducing overfitting in deep learning intrusion detection for power systems with CTGAN,” *Chaos, Solitons \& Fractals,* 2024.

[17] I. a. M. Q. H. Ullah, “Design and development of a deep learning-based model for anomaly detection in IoT networks,” *IEEe Access,* vol. 9, 2021.

[18] X. He, Q. Chen, L. Tang, W. Wang and T. Liu, “CGAN-based collaborative intrusion detection for UAV networks: A blockchain-empowered distributed federated learning approach.,” *IEEE Internet of Things,* vol. 10, pp. 120-132, 2022.

[19] B. Liu, Y. Zhu, K. Song and A. Elgammal, “Towards faster and stabilized gan training for high-fidelity few-shot image synthesis.,” in *International conference on learning representations*, 2020.

[20] P. Radanliev, O. Santos and U. Ani, “Generative AI cybersecurity and resilience.,” *Frontiers in Artificial Intelligence,* vol. 8, 2025.

[21] T. Jiang, C. Shen, P. Ding and L. Luo, “Data augmentation based on the WGAN-GP with data block to enhance the prediction of genes associated with RNA methylation pathways.,” *Scientific Reports,* 2024.

[22] V. Kumar and D. Sinha, “Synthetic attack data generation model applying generative adversarial network for intrusion detection.,” *Computers & Security,* 2023.

[23] L. Coppolino, S. D’Antonio, G. Mazzeo and F. Uccello, “The good, the bad, and the algorithm: The impact of generative AI on cybersecurity.,” *Neurocomputing 623,* 2025.

[24] G. Gebrehans, N. Ilyas, K. Eledlebi, W. Lunardi, M. Andreoni, C. Yeun and E. Damiani, “Generative adversarial networks for dynamic malware behavior: A comprehensive review, categorization, and analysis.,” *IEEE Transactions on Artificial Intelligence,* 2025.

[25] S. Bethu, “Malicious attack detection in IoT by generative adversarial networks,” *SN Computer Science,* 2025.

[26] S. Zhao, J. Li, J. Wang, Z. Zhang, L. Zhu and Y. Zhang, “attackgan: Adversarial attack against black-box ids using generative adversarial networks.,” *Procedia Computer Science,* pp. 128-133, 2021.

[27] M. Ferrag, F. Alwahedi, A. Battah, B. Cherif, A. Mechri, N. Tihanyi, T. Bisztray and M. Debbah, “Generative AI in cybersecurity: A comprehensive review of LLM applications and vulnerabilities,” *Internet of Things and Cyber-Physical Systems,* vol. 5, 2025.

[28] D. M. Diab, AsSadhan, Basil, Binsalleeh, Hamad, Lambotharan, Sangarapillai, Kyriakopoulos, Konstantinos G and Ghafir, Ibrahim, “Anomaly detection using dynamic time warping,” in *2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, 2019.

[29] Nelson and Brian K, “Time series analysis using autoregressive integrated moving average (ARIMA) models,” *Academic emergency medicine,* vol. 5, no. 7, pp. 739--744, 1998.

[30] Ellaway and PH, “Cumulative sum technique and its application to the analysis of peristimulus time histograms,” *Electroencephalography and clinical neurophysiology,* vol. 45, no. 2, pp. 302--304, 1978.

[31] S. F. Tan and A. Samsudin, “Recent technologies, security countermeasure and ongoing challenges of industrial Internet of Things (IIoT): A survey,” *Sensors,* vol. 21, no. 19, p. 6647, 2021.

[32] E. Ahmady, A. R. Mojadadi and M. Hakimi, “A comprehensive review of cybersecurity measures in the IoT era,” *Journal of Social Science Utilizing Technology,* vol. 2, no. 1, pp. 28-38, 2024.

[33] Ț. Mircea, “Security and privacy in the IIoT: Threats, possible security countermeasures, and future challenges,” *Computing & AI Connect,* vol. 2, 2025.

[34] M. F. Guato Burgos, J. Morato and F. P. Vizcaíno Imacaña, “A review of smart grid anomaly detection approaches pertaining to artificial intelligence,” *Applied Sciences,* vol. 14, no. 3, 2024.

[35] D. A. Chekired, L. Khoukhi and H. T. Mouftah, “Fog-based distributed intrusion detection system against false metering attacks in smart grid,” in *2019 IEEE International Conference on Communications (ICC)*, Shanghai, 2019.

[36] W. Li, M. H. Au and Y. Wang, “A fog-based collaborative intrusion detection framework for smart grid,” *International Journal of Network Management,* vol. 31, no. 2, 2021.

[37] H. R. Ghaeini and N. O. Tippenhauer, “Hierarchical monitoring intrusion detection system for industrial control systems,” in *2nd ACM Workshop on Cyber-Physical Systems Security and Privacy (CPS-SPC ’16)*, 2016.

[38] N. Abosata, S. Al-Rubaye, G. Inalhan and C. Emmanouilidis, “Internet of Things for system integrity: A comprehensive survey on security, attacks and countermeasures for industrial applications,” *Sensors,* vol. 21, no. 11, p. 3654, 2021.

[39] A. Yazdinejad, B. Zolfaghari, A. Dehghantanha, H. Karimipour, G. Srivastava and R. M. Parizi, “Accurate threat hunting in industrial internet of things edge devices,” *Digital Communications and Networks,* vol. 9, no. 5, pp. 1123-1130, 2023.

[40] S. Ghasemshirazi and G. Shirvani, *Securing the future: Proactive threat hunting for sustainable IoT ecosystems,* arXiv, 2024.

[41] W. Lei, H. Wen, W. Hou and X. Xu, “New security state awareness model for IoT devices with edge intelligence,” *IEEE Access,* vol. 9, p. 69756–69765, 2021.

[42] S. H. Haghshenas, M. A. Hasnat and M. Naeini, “A temporal graph neural network for cyber attack detection and localization in smart grids,” in *2023 IEEE Power & Energy Society Innovative Smart Grid Technologies*, Washingtong, DC, USA, 2023.

[43] J. Jithish, B. Alangot, N. Mahalingam and K. S. Yeo, “Distributed anomaly detection in smart grids: A federated learning-based approach,” *IEEE Access,* 2023.

[44] N. Tariq, A. Alsirhani, M. Humayun, F. Alserhani and M. Shaheen, “A fog-edge-enabled intrusion detection system for smart grids,” *Journal of Cloud Computing,* vol. 13, no. 1, p. 43, 2024.

[45] S. Agrawal, S. Sarkar, O. Aouedi, G. Yenduri, K. Piamrat, S. Bhattacharya, P. Reddy and T. Gadekallu, *Federated learning for intrusion detection system: Concepts, challenges and future directions,* arXiv, 2021.

[46] J. Wang, Z. Xia, Y. Chen, C. Hu and F. Yu, “Intrusion detection framework based on homomorphic encryption in AMI network,” *Frontiers in Physics,* vol. 10, 2022.

[47] Q.-u.-A. Arshad, W. Z. Khan, F. Azam, M. K. Khan and H. Yu, "Blockchain-based decentralized trust management in IoT: systems, requirements and challenges," *Complex & Intelligent Systems,* vol. 9, p. 6155–6176, 2023.

[48] E. Meybodian, S. Mostafavi and M. Ebrahimi, "A Blockchain-based Hierarchical Trust Management Scheme for IoT," in *7th International Conference on Internet of Things and Applications (IoT)*, 2023.

[49] Y. Liu, C. Zhang, Y. Yan, X. Zhou, Z. Tian and J. Zhang, "A Semi-Centralized Trust Management Model Based on Blockchain for Data Exchange in IoT System," *IEEE Transactions on Services Computing,* vol. 16, pp. 858-871, 2023.

[50] M. Deng, Y. Lyu, C. Yang, F. Xu, M. Ahmed, N. Yang, Z. Xu and C. Ke, "Lightweight Trust Management Scheme Based on Blockchain in Resource-Constrained Intelligent IoT Systems," *IEEE Internet of Things Journal,* vol. 11, pp. 25706-25719, 2024.

[51] C. A. Baykara, I. Şafak and K. Kalkan, "ZETROS: A zero-trust IoT network security framework using distributed blacklisting, trust scoring and smart contracts," *Computer Networks,* vol. 271, 2025.

[52] A. Lahbib, K. Toumi, A. Laouiti and S. Martin, “Blockchain based distributed trust management in IoT and IIoT: a survey,” *Journal of Supercomputing,* pp. 21867-21919., 2024.

[53] G. Jia, S. H. Sun, J. Xin and D. Wang, “RWA-BFT: Reputation-Weighted Asynchronous BFT for Large-Scale IoT,” *Sensors,* 2025.

[54] F. Fathi, M. Baghani and M. Bayat, “Light-PerIChain: Using lightweight scalable blockchain based on node performance and improved consensus algorithm in IoT systems.,” *Computer Communications,* p. 246–259, 2024.

[55] Z. Chen, B. Li, X. Cai, Z. Jia, L. Ju, Z. Shao and Z. Shen, “ChainKV: A Semantics-Aware Key-Value Store for Ethereum System,” *Proceedings of the ACM on Management of Data,* vol. 1, 2023.

[56] B. Khayer, S. Mirzaei, H. Alavizadeh and A. Shahraki, “Blockchain for Secure IoT: A Review of Identity Management, Access Control, and Trust Mechanisms,” *IoT,* 2025.

[57] Y. Tsang, C. Lee, K. Zhang, C. Wu and W. Ip, “On-Chain and Off-Chain Data Management for Blockchain-Internet of Things: A Multi-Agent Deep Reinforcement Learning Approach,” *Journal of Grid Computing,* 2024.

[58] A. Brînzea, L. I. Aciobăniței and F. Pop, “Standard-Compliant Blockchain Anchoring for Timestamp Tokens,” *Applied Sciences,* 2025.

[59] M. Srivastava, S. Arora and D. Boneh, “Optimistic verifiable training by controlling hard-ware nondeterminism,” *arXiv,* 2024.

[60] Y. Mirsky, T. Golomb and Y. Elovici, “Lightweight collaborative anomaly detection for the IoT using blockchain,” *Journal of Parallel and Distributed Computing,* no. 145, p. 75–97, 2020.

[61] K. Demertzis, L. Iliadis and N. K. P. Tziritas, “Anomaly detection via blockchained deep learning smart contracts in Industry 4.0,” *Neural Computing and Applications,* no. 32, p. 17361–17378, 2020.

[62] J. Schultz, “Cybercriminal abuse of large language models,” June 2025. [Online]. Available: https://blog.talosintelligence.com/cybercriminal-abuse-of-large-language-models.

[63] OpenAI, “Introducing GPT-5.2,” Dec 2025. [Online]. Available: https://openai.com/index/introducing-gpt-5-2.

[64] Google, “A new era of intelligence with Gemini 3,” Nov 2025. [Online]. Available: https://blog.google/products-and-platforms/products/gemini/gemini-3/#note-from-ceo.

[65] Mistral, “Introducing Mistral 3,” [Online]. Available: https://mistral.ai/news/mistral-3 .

[66] Anthropic, “Introducing Claude 4,” May 2025. [Online]. Available: https://www.anthropic.com/news/claude-4.

[67] B. Karunanayake, I. Khalil, X. Yi and K. Lam, “Toward LLM-Driven Adaptive Policy Orchestration for Host-Based Intrusion Detection Systems in IoT Environments,” *IEEE network, 39(5),* pp. 66-73, 2025.

[68] T. Ali and P. Kostakos, “HuntGPT: Integrating Machine Learning-Based Anomaly Detection and Explainable AI with Large Language Models (LLMs),”,” *doi:10.48550/arxiv.2309.16021,* 2023.

[69] M. Arif Iftakher Mahmood, F. Ashab, M. Saifuzzaman Sohan, M. Hedayetul Islam Chy and M. F. Kader, “LLM-Enhanced Security Framework for IoT Network: Anomaly Detection and Malicious Devices Identification,” *IEEE access,* pp. 168405-168419, 2025.

[70] K. Jerabek, J. Koumar, J. Setinský and J. Pesek, “Explainable Anomaly Detection in Network Traffic Using LLM,” in *NOMS 2025-2025*, Honolulu, HI, USA, 2025.

[71] H. Zhang, A. B. Sediq, A. Afana and M. Erol-Kantarci, “Large Language Models in Wireless Application Design: In-Context Learning-enhanced Automatic Network Intrusion Detection,” *doi:10.48550/arxiv.2405.11002,* 2024.

[72] H. Jin, G. Papadimitriou, K. Raghavan, P. Zuk, P. Balaprakash, C. Wang, A. Mandal and E. Deelman, “Large Language Models for Anomaly Detection in Computational Workflows: From Supervised Fine-Tuning to In-Context Learning,” *doi:10.48550/arxiv.2407.17545,* 2024.

[73] M. Sladić, V. Valeros, C. Catania and S. Garcia, “LLM in the Shell: Generative Honeypots,” in *EuroS&PW*, Vienna, Austria, 2024.

[74] J. Ragsdale and R. V. Boppana, “On Designing Low-Risk Honeypots Using Generative Pre-Trained Transformer Models With Curated Inputs,” *IEEE access,* pp. 117528-117545, 2023.

[75] A. Happe and J. Cito, “Getting pwn’d by AI: Penetration Testing with Large Language Models,”,” in *ESEC/FSE*, San Francisco, CA, USA, 2023.

[76] P. Santos, R. Abreu, M. Reis, C. Serôdio and F. Branco, “A Systematic Review of Cyber Threat Intelligence: The Effectiveness of Technologies, Strategies, and Collaborations in Combating Modern Threats,” *Sensors,* vol. 25, 2025.

[77] C. Papadopoulos, K. Kollias and G. Fragulis, “Recent Advancements in Federated Learning: State of the Art, Fundamentals, Principles, IoT Applications and Future Trends,” *Future Internet,* 2024.

[78] M. J. Assante and R. M. Lee, “The industrial control system cyber kill chain,” SANS Institute, 2015.

[79] A. Diro and N. Chilamkurti, “Leveraging LSTM networks for attack detection in fog-to-things communications,” *IEEE Communications Magazine,* vol. 56, no. 9, pp. 124-130, 2018.

[80] G. Czeczot, I. Rojek and D. Mikołajewski, “Autonomous threat response at the edge processing level in the industrial Internet of Things,” *Electronics,* vol. 13, no. 6, p. 1161, 2024.

[81] Y. Peng, A. Tan, J. Wu and Y. Bi, “Hierarchical edge computing: A novel multi-source multi-dimensional data anomaly detection scheme for industrial Internet of Things,” *IEEE Access,* vol. 7, p. 111257–111270, 2019.