Resilmesh
securing cyber infrastructures

RESILMESH WHITE PAPER

Who Certifies the Detector?

AI-Based Threat Detection and the Resilience Certification Gap in Critical Infrastructure

Prepared by	Branka Stojanović, JOANNEUM RESEARCH, on behalf of the Resilmesh Consortium

Date	October 2026
Document version	Final
Classification	Public / White paper
DOI	To be assigned
Project	Resilmesh – Situation Aware Enabled Cyber Resilience for Dispersed, Heterogenous Cyber Systems. Horizon Europe, Grant Agreement No. 101119681. https://resilmesh.eu



Funded by the European Union

Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Contents

1 The Challenge: A Safety-Relevant Component Nobody Certifies
2 AI-Based Detection and Threat Awareness in Resilmesh
3 What Today's Frameworks Say About AI Detectors
4 Evidence: When the Detector Is the Attack Surface
5 Why the Certificates Do Not Add Up
6 Towards Detectors That Can Be Trusted
7 Recommendations
8 Looking Ahead

Executive Summary

AI-based detection is becoming the eyes of critical infrastructure operators. In the Resilmesh platform, anomaly detectors running at the edge and in the cloud watch IT and OT data around the clock, and their output feeds the correlation, risk-assessment and mitigation functions that decide how to respond. But if the detector is wrong, or is deliberately deceived, nothing in today's certification landscape tells an operator how much that matters.

This white paper makes three points:

1. **Lightweight detection works, and it can be made robust.** Resilmesh detectors reach results at or near the top of published results on the public BATADAL water-distribution benchmark. The lightweight edge detector, one of the project's results, comes within 0.4% of a centralised detector's overall score, detects attacks faster, localises them to the network area concerned and needs a fraction of the energy. This type of detectors in general can be deceived by a determined attacker: in our tests, an attacker who knew the model disguised most attacks from a standard centralised detector, and hardening did not reliably fix this. Our results also show how to deploy edge detection robustly: combined with a system-wide view, as in the Resilmesh architecture, a compromised controller hid fewer anomalies than from edge-only detection.
2. **The standards do not cover the gap.** Security (IEC 62443), functional safety (IEC 61508) and AI and product rules (EU AI Act, Cyber Resilience Act) each certify resilience in their own way, on different assumptions. Taken together, they do not add up to a guarantee for a system in which a detector feeds a safety-relevant decision.
3. **The gap can be managed now.** We recommend treating AI detectors as attackable, safety-relevant components; reporting their robustness next to their accuracy; stating how each certificate was derived; and requiring any combination of certificates to carry an explicit, justified assumption.

The paper closes with recommendations for operators and integrators, technology developers, standardisation bodies and regulators.

1 The Challenge: A Safety-Relevant Component Nobody Certifies

Water, energy and manufacturing are now cyber-physical systems. Pumps, valves and robots are connected, IT and OT networks are converging, and the attack surface has grown with them. Early detection of intrusions is essential, yet attacks are becoming more frequent while skilled security staff are scarce, which makes purely manual response increasingly unsustainable [5].

The 2021 incident at the water-treatment plant in Oldsmar, Florida shows what is at stake. According to the US cybersecurity agency CISA, an intruder used a remote-access tool to raise the sodium hydroxide setpoint from about 100 to 11,100 parts per million [10]. A plant operator noticed the change and reversed it; manual safeguards, not automated protection, averted harm. Cybersecurity controls, plant safety instrumentation and emergency planning each covered one segment of the event. None covered the connections between them.

This is where AI-based detectors come in. They promise to be the automated eye that catches manipulation of a physical process early. But once they are deployed, operators begin to rely on them and automation begins to act on their output. They become part of the safety chain, and, like any software model, they can be fooled. The question

this paper asks is simple: **who certifies the detector, and what does that certificate allow us to conclude about the system around it?**

2 AI-Based Detection and Threat Awareness in Resilmesh

2.1 The platform in brief

Resilmesh builds a Security Orchestration and Analytics Platform Architecture (SOAPA) around one principle: *you can't secure what you don't understand.* Its goals are to make the complexity of digital infrastructure manageable, to detect and predict advanced persistent threats with AI, and to support "zero-trust-ready" security across IT and OT. The platform is organised in six planes (infrastructure, aggregation, collaboration mesh, threat awareness, situation assessment and security operations) [2], and is validated in use cases covering renewable-energy SCADA, smart-manufacturing robotics and regional civil infrastructure, plus further use cases from open calls.

2.2 The role of AI detection

The threat-awareness plane hosts the AI-based anomaly detection module. Models can be placed at the endpoint, at the edge or in the cloud; they work on network traffic, OT sensor readings and logs; and they can be trained in a federated way so that data need not leave the site [2]. Their output flows to an AI correlator and the security-information platform, feeds the assessment of asset dependencies and critical-service risk, and finally reaches the Mitigation Manager, which selects a response and launches it through automated playbooks. Resilmesh deliberately combines centralised detection, with its system-wide view, and lightweight edge detection, which brings speed and local operation and keeps working if higher-level communication is interrupted [2, 11].

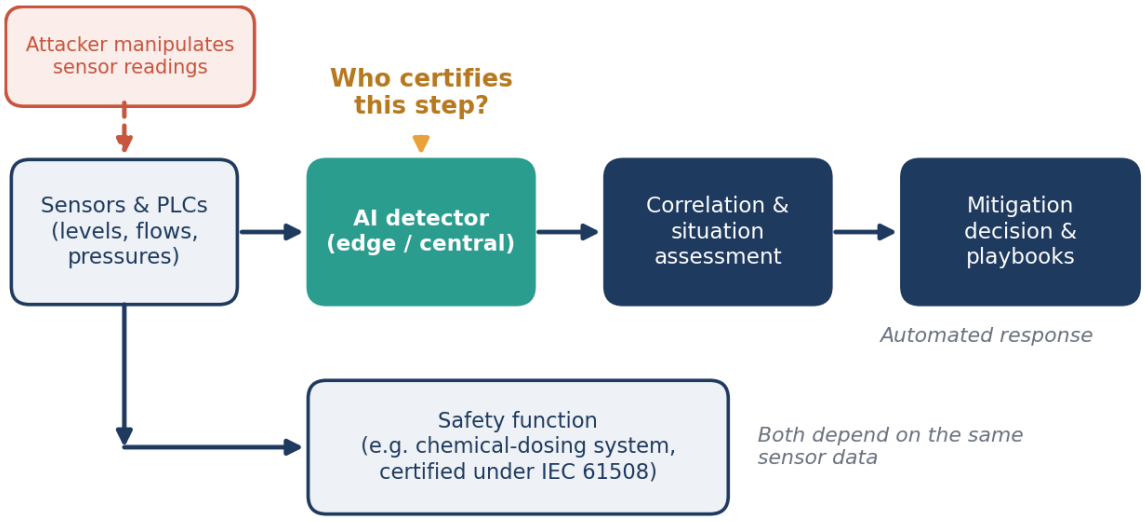


Figure 1. Resilmesh architecture simplified from [2]: the detector sits at the start of an automated chain and shares its data source with safety functions.

In a first proof-of-concept test with synthetic alerts, the average time from suspicious activity to the effect of a mitigation playbook was about 3.2 seconds, with detection itself taking roughly a quarter of that [2]. The chain is fast precisely because it is automated, so an error at its first link travels quickly.

2.3 What the detectors have achieved

Over three generations, Resilmesh partners developed anomaly detectors for critical infrastructures and evaluated them on water-distribution system benchmark BATADAL, a public benchmark that simulates a medium-sized network ("C-Town", with 7 tanks, 11 pumps and 9 controllers) [9]. All three learn what *normal* looks like from attack-free data and flag deviations, so they do not depend on labelled attack examples.

Detector	Key idea	Overall score	Speed score	Highlights
Best published reference	Model-based: detailed physical simulation of the network	0.970	0.965	Needs detailed network model and domain expertise [3]
Enhanced autoencoder (2022)	Time features, robust scaling and smoothing of anomaly scores	0.957	0.956	Best result among methods without a physical model at the time [3]
Edge autoencoders (2024)	Three small models, one per controller area	0.953	0.970	Localises attacks to an edge area; about 75% lower training emissions; runs on a few-watt device [4]
Hybrid physics-inspired autoencoder (2025)	Enforces consistency between a quantity and its rate of change	0.977	0.997	About 3% faster detection; simpler network; classification score slightly below the leader [5]

Table 1. Results on the BATADAL benchmark. Scores follow the challenge definition (0–1, higher is better): the overall score is the mean of a classification score and a detection-speed score. Emission figures are estimates.

The lightweight edge detector is one of the project's main results. Compared with an equivalent centralised model, its overall score is within 0.4% and its detection about 1.5% faster. Because it consists of three small models, it needs roughly a fifth of the computing operations at inference (4,626 against 22,644) and cut training emissions by an estimated 75%. It runs on a device drawing only a few watts, which a small solar panel and battery can power, and it identifies which of three network areas an alarm comes from, something a single centralised model cannot do [4].

3 What Today's Frameworks Say About AI Detectors

Operators of critical infrastructure in the EU face obligations under NIS2 and the Critical Entities Resilience Directive, and must draw on several standards and regulations to show that their systems are resilient. For a detector that feeds safety-relevant decisions, four are particularly relevant.

Framework	What it certifies	Typical evidence	What it leaves open for AI detectors
IEC 62443 (security levels 1–4)	Resistance of a system or zone to an attacker of a defined capability	Audit, penetration testing	Evading an ML detector is not a modelled attack surface; no rule for combining results of different zones
IEC 61508 (safety integrity levels 1–4)	Reliability of a safety function under random hardware faults	Failure-probability bands	Assumes sensor readings are reliable; no equivalent for ML-based components
EU AI Act	Appropriate accuracy, robustness and cybersecurity of high-risk AI systems, where the detector falls in scope	Conformity assessment	Does not specify how resistance to manipulation is measured or certified
EU Cyber Resilience Act	Security of products with digital elements over their lifecycle	CE marking	A CE mark cannot be mapped to a security or safety level in a given integration

Table 2. Frameworks relevant to an AI detector in critical infrastructure. Summarised from [1].

Two observations follow. First, **each framework certifies something different**: resistance to a defined attacker, reliability under random faults, or conformity with a broad requirement. Second, **each rests on its own assumption about what can go wrong.** Functional safety assumes that failures are random and that sensors tell the truth; an attacker who crafts plausible readings is outside its model. Joint guidance exists (IEC TR 63069, NIST SP 800-82), but no formal mapping between a safety integrity level and a security level. In practice, integrators fill the gap with engineering judgement, and that becomes harder as systems grow more heterogeneous.

4 Evidence: When the Detector Is the Attack Surface

Resilmesh partners have tested their detectors against two kinds of manipulation: an attacker who knows the model and alters the data it sees, and an attacker who has compromised one controller and manipulates that area's measurements. The aim was to learn how to deploy detection, including edge detection, as robustly as possible. Both studies use the BATADAL benchmark and autoencoder detectors.

4.1 A knowledgeable attacker can disguise most attacks

An *evasion attack* makes small, deliberate changes to the data a detector sees so that an attack looks normal. The worst case is an attacker who knows the model (a "white-box" attacker). Partners tested exactly this on the centralised autoencoder, using the standard gradient-based attack method (PGD) on the BATADAL test set [6]. For most attacks, the reconstruction error that triggers the alarm could be pushed below the detection threshold, even though the detector performs well against unmodified attacks. This is a property of learned detectors in general, not of edge deployment: the test was run on the centralised model.

Two training methods meant to harden the model helped less than expected. Adversarial training nearly doubled the size of manipulation needed to evade detection when the attacker could change every reading (from 0.012 to 0.022 on the study's normalised measure), yet the share of attacks still caught fell from 22.7% to 18.3% in that scenario. The authors conclude that hardened models are not definitely safer in practice. Figure 2 also shows that the fewer readings an attacker can manipulate, the more attacks remain detectable: even when the attacker was assumed to pick the most influential readings, a single manipulated reading left about 61% of attacks detectable for the baseline model, against about 23% when all readings could be changed.

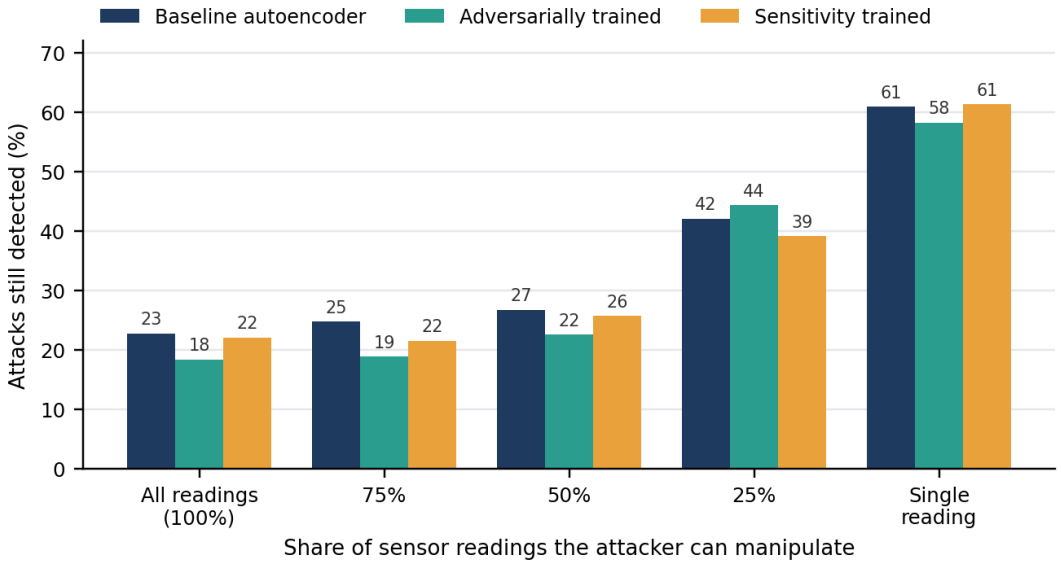


Figure 2. Share of attacks that remained detectable under a white-box evasion attack, by share of sensor readings the attacker can manipulate. Data from [6]; simulated benchmark, one detector architecture.

4.2 Combining edge and system-wide views makes concealment harder

In industrial plants, an attacker is more likely to compromise a single controller (PLC) or edge component than the whole system, and can then alter the measurements that component reports. A second study [11] asks how edge detection should be deployed under this threat. It compares three architectures: an **edge-only** ("localised") setup with one autoencoder per network area, as in the Resilmesh edge detector; a **centralised** one that sees all measurements; and a **vertical federated** one, which keeps the edge encoders and adds a shared decoder that combines the areas' representations, so raw data stay with the data owner. One area is compromised and three manipulations are tried: replacing readings with benign averages, planting anomalous patterns in the training data, and *replaying* earlier normal measurements while an attack is under way. Alarm thresholds were set so that all architectures have the same false-alarm rate.

No architecture was immune, but the edge-only setup was affected most. The clearest case is replay: when one area replayed its measurements from the preceding 24 hours, edge-only detectors lost between 20% and 46% of the

anomalies they had detected before (retaining 54% to 80%), against about 11% to 19% for the centralised and vertical federated detectors (Figure 3). The likely reason is simple: an edge detector sees only its own area's data, which a compromised controller controls, whereas the other architectures also see measurements from areas the attacker cannot reach, where the physical consequences of the attack may still be visible. The practical message is to pair edge detection with a system-wide view, as the Resilmesh architecture does, rather than to rely on the edge alone.

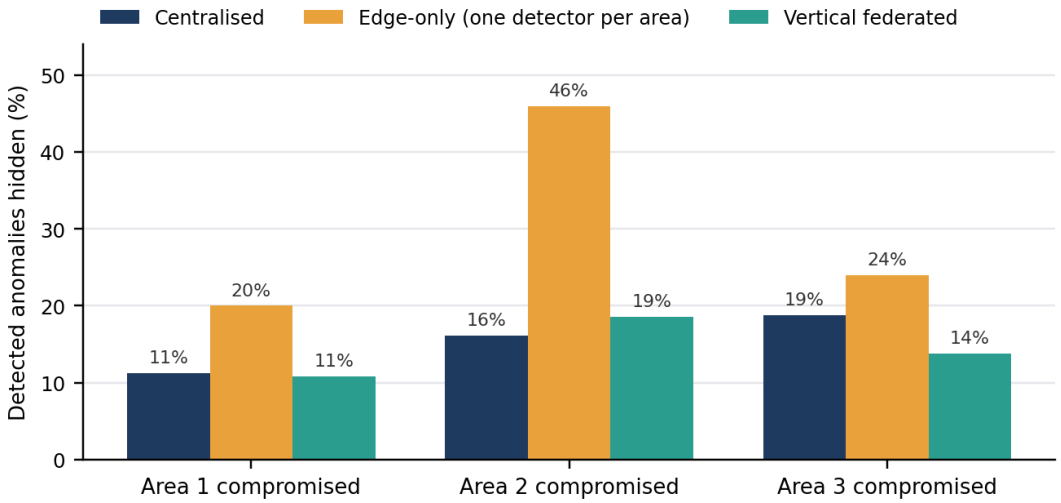


Figure 3. Share of previously detected anomalies hidden by a 24-hour replay attack, by detector architecture and compromised network area. Data from [11]; simulated benchmark.

The authors state several caveats: one benchmark, one fixed partition into three areas, one static attacker, and two exceptions in which local detectors held up relatively well. Even low attack intensities often caused a substantial drop, with little further degradation at higher intensities. The vertical federated variant, which builds directly on the edge detector, looks like a promising compromise between robustness and keeping data local, but its privacy and communication benefits have not yet been quantified.

4.3 Lessons

- **Accuracy is not robustness.** A detector can score well on benchmark attacks and still be easy to evade or conceal.
- **One robustness number can mislead.** Hardening raised one measure while lowering another, so several measures are needed.
- **Architecture is a security property.** Where a detector sits and what data it sees are usually chosen for efficiency, data locality and availability, yet they changed how much an attacker could hide. No current certificate captures this.
- **Edge and system-wide views complement each other.** Edge detectors bring speed, localisation and efficiency; adding information from other parts of the system made concealment harder.

Detectors that test whether readings are physically consistent with each other, such as the 2025 hybrid detector [5, 8], are expected to be harder to evade than purely statistical ones. This is a promising but partial remedy: in the studies underlying this paper, the hybrid detector has not yet been tested against manipulation.

5 Why the Certificates Do Not Add Up

The evasion results matter for certification because of a mismatch between *how* different certificates are derived. Consider a chemical-dosing safety function certified at safety integrity level 2 (SIL 2). That means it is expected to fail on demand between roughly once in 100 and once in 1,000 demands. This is a statement about random equipment failures, **averaged** over many demands.

A determined attacker is not a random fault. Where an attacker can manipulate the data that a detector and a safety function both rely on, he or she picks the scenario that hurts most, so what matters is the **worst case**.

A probabilistic argument does not rescue this. A manipulation may be extremely unlikely under the random-fault model while being fully reachable for an attacker. A certified average cannot rule it out. The result, established formally in [1], is that **the certificates of the individual components do not determine the guarantee for the composed system** unless an explicit bridging assumption is added.

> **In plain words**
>
> A certificate based on averages tells you how a component behaves on a typical day. An attacker does not choose a typical day.
>
> The bridge can take several forms: re-certifying the safety function against a worst-case bound for what the coupled detector can let through; a documented model of which manipulations are reachable; or a hard physical limit that holds whatever the software believes, such as a rate limit on dosing.

Hardening the detector does not substitute for this bridge, as Section 4 showed. Note the scope of the claim: it concerns the boundary between worst-case and average-based certificates; it is a result about *insufficiency of certificates*, not an argument against certification, and a unified framework remains achievable [1].

6 Towards Detectors That Can Be Trusted

The evidence suggests a set of practices that make AI detectors more trustworthy in critical infrastructure. They can be adopted ahead of any change in standards.

1. **Evaluate against an adversary, not only against attack traces.** Add evasion tests to acceptance, and report how many attacks stay detected under stated attacker-access scenarios and how large a manipulation is needed to evade.
2. **Report several measures together:** accuracy, time to detect, localisation, robustness and, where relevant, energy use. The Resilmesh edge detectors show that these trade off against each other.
3. **Check physics, not just statistics.** Consistency checks between related process variables make manipulated readings harder to hide, because they must also be physically plausible.
4. **Combine edge and system-wide views.** Edge detectors give speed, localisation, low energy use and data locality, but a compromised controller can hide anomalies from its own local detector. Pairing them with a cross-area view, centrally or through vertical federated learning that keeps raw data at each site, made concealment harder in our tests.
5. **Keep layers independent.** Safety functions and manual fail-safes should not depend on the same data path or the same model as the detector. Oldsmar was resolved by an attentive operator, not by automation.
6. **Make results explainable and reproducible.** A recent review of time-series analysis in cyber security found low explainability of neural models and that more than a quarter of the reviewed approaches were evaluated only on private data [7]. The same review flags adversarial attacks as a growing threat as models become more complex.

Resilmesh follows several of these principles by design: it combines lightweight edge detection with central detection, correlates multiple sources and detectors rather than trusting any one of them, and adds context about asset dependencies through situation assessment [2]. Nevertheless, none of this removes the need for the certification clarity discussed above.

7 Recommendations

7.1 Operators and system integrators

- List AI detectors as assets in the risk assessment and treat them as safety-relevant components and potential single points of failure.
- For every certificate in the system file, record what it actually promises: against which disturbance (random faults, a defined attacker, none) and in which sense (average, worst case, descriptive). Where a safety function relies on data an attacker could manipulate, require a stated bridging assumption, such as a worst-case bound or a hard interlock, before treating the combined system as certified.

- Pair detection close to the process with a view across the wider system: a compromised component can feed its own local monitor.
- Keep safety layers and manual fail-safes independent of the detection path, and include manipulation of the monitoring itself in exercises and acceptance tests.

7.2 Technology developers and vendors

- Publish robustness evidence next to accuracy: evasion tests under stated attacker-access scenarios, the share of attacks still detected and the size of manipulation needed.
- Do not treat model hardening as proof of safety. Test several measures and repeat the tests after every change to the model.
- Combine lightweight edge detection, which brings speed, localisation and efficiency, with a cross-area view, physics-consistency checks and corroboration across independent sources, and document the assumptions.

7.3 Standardisation and certification bodies

- Require certificates to declare their quantification basis: random-fault average, defined attacker or worst case, or descriptive.
- Develop interface guidance that goes beyond joint-application advice such as IEC TR 63069, and states how safety and security certificates are to be combined where an AI component feeds a safety-relevant function.
- Define how robustness of AI-based monitoring is to be measured and certified, including test protocols and attacker models.

7.4 Regulators and policymakers

- When implementing the AI Act, the Cyber Resilience Act and NIS2, clarify how requirements are to be demonstrated for AI-based security monitoring, including what “robustness” means against deliberate manipulation and how product-level markings relate to system-level levels.
- Support open benchmarks and testbeds that include adversarial scenarios, with results shared publicly.
- Support further research in this domain.

8 Looking Ahead

Several steps would strengthen the evidence base and move from recommendations to practice:

- **Other attackers and detectors.** Black-box attackers, who can only query a detector without knowing its inner workings, are a more realistic threat and remain to be studied. The physics-inspired detector has not yet been tested against manipulation.
- **Beyond one benchmark and one topology.** Further datasets, larger systems, a different number of network areas and real industrial environments are needed to show how far the findings carry; so is extending detectors to network traffic and logs [2, 4, 11].
- **Stronger attackers and defences.** Several compromised areas at once and adaptive attackers are not yet covered. Defences tailored to distributed detection, such as robust aggregation of area-level representations and threshold strategies that separate general degradation from targeted concealment, are a next step [11].
- **Vertical federated learning.** Its privacy and communication benefits need to be quantified, not only argued, before it can be recommended as a standard compromise [11].
- **A formal bridge.** A unified framework would need to cover different disturbance types, multiple levels from component to system of systems, static and dynamic evaluation, domain-specific parameterisation and computational tractability [1]. Supplying these is the subject of further work.

AI detectors will increasingly act as the eyes of critical infrastructure. Making sure that what they report can be trusted, and that the certificates around them say honestly what has and has not been shown, is a task for operators, developers, standards bodies and regulators together.

Acknowledgements and References

The research behind this white paper was mainly funded by the European Union's Horizon Europe programme under the Resilmesh project (Grant Agreement No. 101119681), and in part by the Austrian Security Research Programme KIRAS of the Federal Ministry of Finance (project SeRWas, GA No. FO999905309) the Austrian Research Promotion Agency FFG (project PRESENT, GA No. FO999899544), and the champI4.0ns project (supported by the Austrian BMK, grant no. 891793, and the German BMWK). The detection research summarised here was carried out together with colleagues at JOANNEUM RESEARCH and partner institutions.

[1] Stojanović, B. (2026). The Non-Composability of Resilience Certificates Across Critical Infrastructure Domains. Presented at CRITIS 2026; to appear in Springer Lecture Notes in Computer Science. DOI: 10.5281/zenodo.23317735.

[2] Bernabé, J. B., Lee, B., Husák, M., Sadlek, L., Stojanović, B., Somma, M., et al. (2025). The Resilmesh Architecture: Situation Aware Enabled Cyber Resilience for Dispersed, Heterogenous Cyber Systems. IEEE NetSoft 2025, pp. 621–626. DOI: 10.1109/NetSoft64993.2025.11080622

[3] Stojanović, B., Neuschmied, H., Winter, M., Kleb, U. (2022). Enhanced Anomaly Detection for Cyber-Attack Detection in Smart Water Distribution Systems. ARES 2022. DOI: 10.1145/3538969.3543796

[4] Somma, M., Flatscher, A., Stojanović, B. (2024). Edge-Based Anomaly Detection: Enhancing Performance and Sustainability of Cyber-Attack Detection in Smart Water Distribution Systems. TELFOR 2024. DOI: 10.1109/TELFOR63250.2024.10819044

[5] Somma, M. (2025). Hybrid Temporal Differential Consistency Autoencoder for Efficient and Sustainable Anomaly Detection in Cyber-Physical Systems. CNSM 2025. DOI: 10.23919/CNSM67658.2025.11297553

[6] Flatscher, A., Stojanović, B., Özdenizci, O. (2024). Enhancing Adversarial Robustness of Anomaly Detection-Based IDS in OT Environments. CNSM 2024. DOI: 10.23919/CNSM62983.2024.10814511

[7] Landauer, M., Skopik, F., Stojanović, B., Flatscher, A., Ullrich, T. (2025). A review of time-series analysis for cyber security analytics: from intrusion detection to attack prediction. International Journal of Information Security 24, 3. DOI: 10.1007/s10207-024-00921-0

[8] Somma, M., Gallien, T., Stojanović, B. (2025). Anomaly Detection in Complex Dynamical Systems: A Systematic Framework Using Embedding Theory and Physics-Inspired Consistency. DOI: 10.48550/arXiv.2502.19307.

[9] Taormina, R., et al. (2018). Battle of the Attack Detection Algorithms: Disclosing Cyber Attacks on Water Distribution Networks. Journal of Water Resources Planning and Management 144(8). DOI: 10.1061/(ASCE)WR.1943-5452.0000969

[10] CISA (2021). Alert AA21-042A: Compromise of U.S. Water Treatment Facility.

[11] Flatscher, A., Somma, M., Stojanović, B. (2026). Robustness of Centralised, Localised, and Vertical Federated Anomaly Detection Against Data Manipulation Attacks. Accepted at the 4S-Net Workshop, CNSM 2026 (forthcoming).

Standards and regulations:

IEC 61508:2010; IEC 62443-3-3:2013; IEC TR 63069:2019; NIST SP 800-82 Rev. 3; Regulation (EU) 2024/1689 (AI Act); Regulation (EU) 2024/2847 (Cyber Resilience Act); Directive (EU) 2022/2555 (NIS2); Directive (EU) 2022/2557 (CER).