Resilmesh
securing cyber infrastructures

External Datasets

Deliverable Number	**D7.5**
Deliverable Details: Characterization and formalisation of External Datasets to be used in model training by Resilmesh technical WPs.	
Deliverable Leading:	ALWA, JAMK
Due Date:	May 31, 2024
Submitted Date:	June 26, 2024
Author(s)	Stefano Bianchi (ALWA), Marco Bocca (ALWA), Mirco Antona (ALWA), Eppu Heilimo (JAMK), Marko Angervuo (JAMK), Matti Saarelma (JAMK), Tuomo Sipola (JAMK), Francesco Velásquez Ávila (GMV)
Reviewer(s):	TUS, SLP, JR

Version History

Version	By	Date	Changes
A1	ALWA	20/04/2024	Table of Contents, general parts
A2	ALWA	27/04/2024	Overview and dataset characterization
A3	ALWA	15/05/2024	General descriptive parts
A4	ALWA	31/05/2024	IT/OT dataset tables
A5	ALWA	05/06/2024	Handover for proper completion of deliverable
A6	JAMK	07/06/2024	Minor changes
A7	JAMK	14/06/2024	Chapter 2 Completion
A8	JAMK	17/06/2024	Chapter 3, 4 cleaned up, executive summary added, general cleanup
A9	GMV	19/06/2024	Draft for internal review
A10	JAMK	26/06/2024	Internal review modifications
A11	JAMK	26/06/2024	Formatting
A12	GMV	26/06/2024	Energy OT datasets
A	JAMK	26/06/2024	Release

Disclaimer

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Table of Contents

Funded by
the European Union

Executive summary

This document provides a detailed review of external datasets identified and characterised for model training. It aims to lay a solid foundation for using these datasets to improve AI algorithms and system integration within the Resilmesh project.

The document outlines methodologies for characterising datasets, focusing on External IT and OT datasets. The scope includes components such as the Resource Orchestrator, Event Aggregation & Detection Pipeline, and AI-based Detector.

Three primary use cases are detailed: Civil Regional Infrastructure, Renewable Energy, and Smart Manufacturing Robotics. Each use case addresses specific environmental needs, data collection methods, attack execution, and mitigation strategies.

The milestones include the collaborative efforts of various partners in gathering and refining the datasets. This shows how important these datasets are for improving AI algorithm design and verification activities.

Overall, this deliverable provides a foundational framework for the application of these datasets in securing critical infrastructures through advanced AI methodologies.

1 Introduction

The project has undergone a leadership transition. Previously, ALWA was responsible for leading this initiative. However, due to organisational changes, JAMK has taken over the leadership responsibilities over the T7.5. This document contains the evidence of the first results of task **T7.3 "Management and Generation of Datasets"**, meant to acquire, generate, and curate (external and internal) datasets for use in Artificial Intelligence (AI) algorithm design, system integration and system verification activities in other technical WPs.

In particular, this deliverable focuses on the identification and characterization of External Datasets only.

External Datasets identified and acquired in WP4, 5 and 6 by the design teams (including those from consortium partners), will be stored and curated also as part of related WP7 activities.

It is worth noticing that also Internal Datasets will be generated by the Resilmesh project as a main result of development and testing phases, for example as part of the cyber range activities in WP6 (in this case also using data generation tools from use-case partners ALWA and ALIAS).

Evidence of T7.3 results will be duly included in the following deliverables:

- **D7.5 "External Datasets" Type: DATA — datasets, microdata, etc. – PUBLIC deliverable**
- **D7.6 "Internal Datasets" Type: DATA — datasets, microdata, etc. – PUBLIC deliverable**

This report, D7.5, describes the external datasets which are listed in the accompanying spreadsheet file. The internal datasets listed in the file are not described in this document.

The declared type of both deliverables is "DATA — datasets, microdata, etc." – from this point of view, the actual deliverables are the datasets themselves (e.g., a structured collection of data, presented in a tabular form, where each row corresponds to a unique record and each column corresponds to a specific variable or attribute), whereas these accompanying descriptive documents are meant to provide a general overview of their informative contents and a proof-of-evidence for the Continuous Reporting on the EC portal (deliverable upload).

Funded by
the European Union

1.1 Document Structure

The document consists of 5 sections.

- Section 1 introduces the overall document structure, overview of the project's goals, and the related work packages and deliverables.
- Section 2 provides a brief description of the methodology and the approaches used in the project.
- Section 3 briefly describes the details of internal datasets utilised in the project.
- Section 4 defines an overview of the use case datasets, focusing on the critical datasets and data collection methods relevant to the project.
- Section 5 presents the conclusion, the effectiveness of the datasets, and the future work.

1.2 Related Work Packages and Deliverables

The management and generation of datasets is a pivotal project activity, tightly coupled with the other WPs, tasks, and deliverables: for the sake of comprehension, it is useful to briefly enumerate main references, dependencies, and related documents.

As for relation with other WPs, as stated in the Description of Action (DoA):

- ***Work package WP4 – Threat Awareness***
    - ***T4.1 "Anomaly detection" (Lead: JR, Participants: UMU, ALWA, MONT, M4-M30)*** *This task will develop AI anomaly detection algorithms to detect suspicious events and attacks at both edge and cloud for host and network data including both logs and emitted events (IDS, MMT monitoring framework, etc.). It will multi-view anomaly detection algorithms to enable the blending of heterogeneous data sources to improve cross-domain anomaly detection, especially for mixed IT/OT critical infrastructures. It will use deep stacked networks (DSN), hybrid Stacked Autoencoders (SAE) and Convolutional Neural Networks (CNN) structures, and similar techniques to develop hierarchical feature fusion (e.g. via Autoencoders) and decision fusion (e.g., via ensemble learning) models for edge-based anomaly detection. The specific models to be developed will be guided by the pilot use cases.* Models will be trained using publicly available (i.e., external) datasets in the first place and later with datasets generated within the project.
    - ***T4.2 AI-based correlation (Lead: TUS, Participants: JR, MONT, M4-M30)*** *This task will develop algorithms to correlate security events or raw data with the goal to i) predict attack evolution, ii) determine the root causes of attacks or anomalies, and iii) reduce the number of events (e.g., false positives) submitted to the SOC operator. It will, for this reason, focus*

Funded by
the European Union

> *primarily on sequence/causal-based learning techniques or, when appropriate, hybrid approaches combining sequence and similarity learning or other techniques. It will use both deep learning (including spatio-temporal techniques) and broader data mining (rule mining, time series, Bayesian networks, etc.) approaches. The specific models to be developed will be guided by the pilot use cases.* Models will be trained using publicly available datasets in the first place and later with datasets generated within the project.

Also, according to the Description of action:

- *The **Data Management Plan (DMP)** 📫 **D1.6** will address all issues regarding the use of data within the consortium including the public release, posting, curation, and preservation of data during and after the project's lifetime. It will follow the FAIR principles and will be updated regularly to ensure appropriate data management and a high level of data quality and accessibility. This way we adhere to the principle "as open as possible, as closed as necessary". All project partners will provide input on the data types they will collect and store, the protective measures taken and will follow the stipulated principles and guidelines of the DMP. Specific attention will be given to requirements gathering and analysis activities, which involve acquiring questionnaire, interview, observational study, and user study data, to ensure that these adhere to strict legal and ethical guidelines outlined in the DMP and that users are transparently informed and fully consent to these methods. The ResilMesh solution itself will involve incident reports and cyber threat intelligence exchange between partners and member states. Such data considered can constitute sensitive/personal information and must be treated appropriately to address the data privacy issues.*
- ***Privacy and Data Governance:** No personal data is used for training AI systems in Resilmesh. Datasets are either well-known anomaly datasets or will be generated in the project. Where federated learning may be used as in the civic infrastructure use case, strong cryptographic techniques will be used to achieve data privacy when sharing model parameters. Moreover, we will leverage these techniques to address the call issue of mass surveillance and privacy of personal spaces using federated learning in anomaly detection personal space, such as mobile phones or home networking, in the open call use cases.*

Finally, as for what is included in the Grant Agreement, COMMUNICATION, DISSEMINATION, OPEN SCIENCE AND VISIBILITY (— ARTICLE 17) - Open science: research data management:

- *Metadata of deposited data must be open under a Creative Common Public Domain Dedication (CC 0) or equivalent (to the extent legitimate interests or constraints are safeguarded), in line with the FAIR principles (in particular machine-actionable) and provide information at least about the following: **datasets (description, date of deposit, author(s), venue and embargo);** Horizon*

> *Europe or Euratom funding; grant project name, acronym and number; licensing terms; persistent identifiers for the dataset, the authors involved in the action, and, if possible, for their organisations and the grant. Where applicable, the metadata must include persistent identifiers for related publications and other research outputs.*

Related deliverables, as listed in Table 1.

Table 1: The list of deliverables related to the external datasets.

Deliverable Number	Deliverable Name	Relation to the External Datasets
D2.1	Requirements specification	Provides requirements to test
D2.2	System architecture	Provides the list of components to test
D2.3	Evaluation Strategy	Related task, collaboration
D3.1-2	Agg. & Collab. Tool Selection	Components to test
D4.1-4	Anomaly Detection Models, Event Correlation Models, Federated Learning, CTI Implementation	Components to test
D5.1-4	CSS Metamodel, THF Framework, NSA Forecasting, Attack Mitigation	Components to test
D6.1	Smart Grid Scenario	Related task, collaboration
D6.2	Pen Test Framework	Component to test

Funded by
the European Union

Deliverable Number	Deliverable Name	Relation to the External Datasets
D7.2	Verified Resilmesh System	Realisation of the test plan is its prerequisite of this task
D7.5	External Datasets	Provides external datasets usable in AI-model training and testing
D7.6	Internal Datasets	Provides internal datasets usable in AI-model training and testing
D8.1	Pilots test plan	Test plan for the pilots, successor to this deliverable

Funded by
the European Union

2 External Datasets Characterization

Herein, we provide the overview of the external datasets, including the definition of the scope of testing, setting the approaches to testing, selecting the tools for testing, and setting the levels for testing.

It is worth noting that some related topics, such as the collection of datasets, will be discussed in other documents. Readers are encouraged to see the list provided in the introduction of this document.

2.1 Methodology

2.1.1 External IT Datasets

In Information Technology (IT), datasets serve as a backbone for various analytical, monitoring, and optimization tasks. Among these, network traffic information is a critical component for ensuring the smooth and secure operation of IT infrastructures. Network traffic data encompasses a wide range of features that capture the flow of data across a network, providing insights into usage patterns, potential security threats, and performance bottlenecks. Key features of network traffic datasets include packet sizes, transfer protocols, source and destination IP addresses, port numbers, and timestamps. Understanding and analysing these features generally allows IT professionals to maintain network integrity, optimise performance, and preemptively address issues before they escalate.

In the scope of the Resilmesh project, the identified/suggested external IT datasets are meant to provide actionable information to train models for an AI-based distributed anomaly detection for early attack detection and an AI-based event correlation model for attack prediction. The external IT datasets are also used to develop a federated learning framework to support the training of robust AI models and techniques to improve the robustness of cyber threat intelligence.

The metadata required for the characterisation of external IT datasets are described in Table 2.

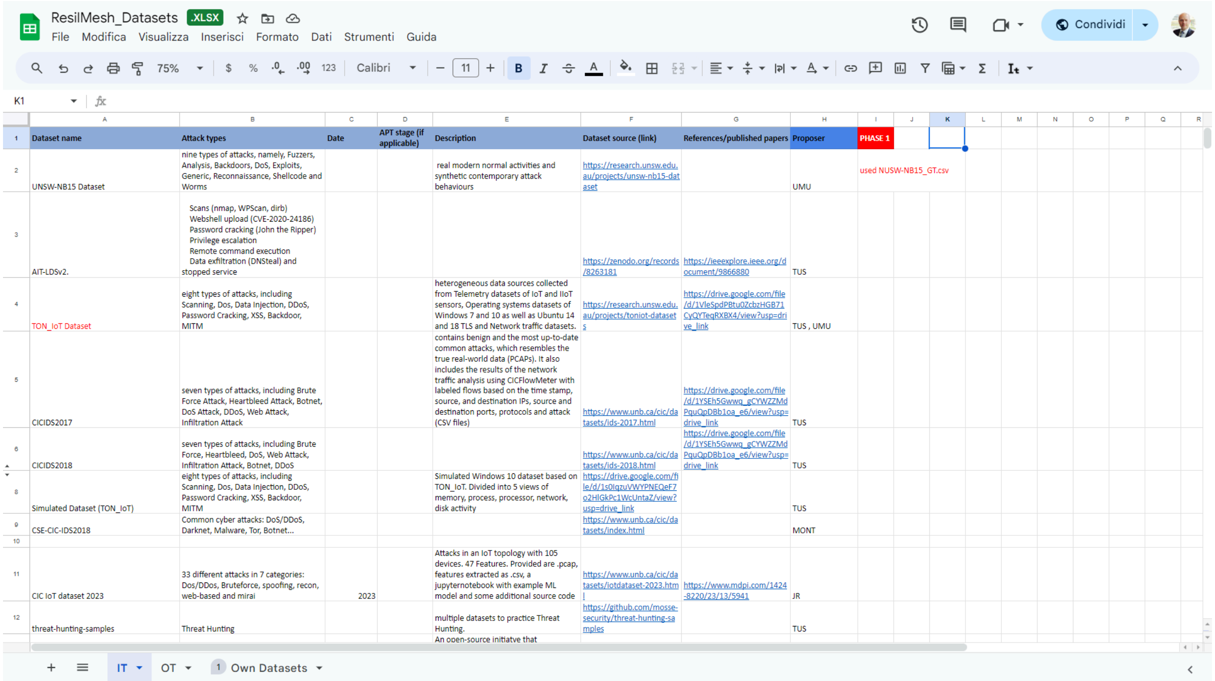
Funded by
the European Union

Metadata field name	Metadata field description
Dataset name	A (possibly) self-explanatory name of the dataset for reference purposes
Attack types	The types of attack that the dataset is about/is able to provide evidence for/can be used to improve the detection of
Date	The date of availability (or generation) of the dataset
APT stage (if applicable)	The specific Advanced Persistent Threat (APT) stage that the dataset is about/is related to
Description	A short meaningful description of the main dataset features (what informative content is represented, the level of granularity, further useful technical explanations/insights etc.)
Dataset source (link)	(If available) The link where the dataset is available for inspection/download (it can be an internal restricted link)
References/published papers	Additional useful information on applicable public references on the dataset or related published papers
Proposer	The Resilmesh Consortium partner proposing the use of the dataset for training models

Table 2: Metadata for the characterisation of external IT datasets.

For External IT Datasets, a simple shared online form (Google Drive Sheet) was provided to involved partners for proper tracking and contributions (see Figure 2).

Funded by
the European Union



Dataset name	Attack types	Date	APT stage (if applicable)	Description	Dataset source (link)	References/published papers	Proposer	PHASE 1
UNSW-NB15 Dataset	nine types of attacks, namely, Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms			real modern normal activities and synthetic contemporary attack behaviours	https://research.unsw.edu.au/projects/unsw-nb15-dataset		UMU	used NUSW-NB15_GT.csv
AIT-LDSv2.	Scans (nmap, WPScan, dirb) Webshell upload (CVE-2020-24186) Password cracking (John the Ripper) Privilege escalation Remote command execution Data exfiltration (DNSteal) and stopped service				https://zenodo.org/records/8263181	https://ieeexplore.ieee.org/document/9866880	TUS	
TON_IoT Dataset	eight types of attacks, including Scanning, Dos, Data Injection, DDoS, Password Cracking, XSS, Backdoor, MITM			heterogeneous data sources collected from Telemetry datasets of IoT and IIoT sensors, Operating systems datasets of Windows 7 and 10 as well as Ubuntu 14 and 18 TLS and Network traffic datasets.	https://research.unsw.edu.au/projects/toniot-datasets	https://drive.google.com/file/d/1VleSpdPBtu0ZcbzHGB71CyQYTeqRXBX4/view?usp=drive_link	TUS , UMU	
CICIDS2017	seven types of attacks, including Brute Force Attack, Heartbleed Attack, Botnet, DoS Attack, DDoS, Web Attack, Infiltration Attack			contains benign and the most up-to-date common attacks, which resembles the true real-world data (PCAPs). It also includes the results of the network traffic analysis using CICFlowMeter with labeled flows based on the time stamp, source, and destination IPs, source and destination ports, protocols and attack (CSV files)	https://www.unb.ca/cic/datasets/ids-2017.html	https://drive.google.com/file/d/1YSEh5Gwwq_gCYWZZMdPquQpDBb1oa_e6/view?usp=drive_link	TUS	
CICIDS2018	seven types of attacks, including Brute Force, Heartbleed, DoS, Web Attack, Infiltration Attack, Botnet, DDoS				https://www.unb.ca/cic/datasets/ids-2018.html	https://drive.google.com/file/d/1YSEh5Gwwq_gCYWZZMdPquQpDBb1oa_e6/view?usp=drive_link	TUS	
Simulated Dataset (TON_IoT)	eight types of attacks, including Scanning, Dos, Data Injection, DDoS, Password Cracking, XSS, Backdoor, MITM			Simulated Windows 10 dataset based on TON_IoT. Divided into 5 views of memory, process, processor, network, disk activity	https://drive.google.com/file/d/1s0IqzuVWYPNEQeF7o2HlGkPc1WcUntaZ/view?usp=drive_link		TUS	
CSE-CIC-IDS2018	Common cyber attacks: DoS/DDoS, Darknet, Malware, Tor, Botnet...				https://www.unb.ca/cic/datasets/index.html		MONT	
CIC IoT dataset 2023	33 different attacks in 7 categories: Dos/DDos, Bruteforce, spoofing, recon, web-based and mirai	2023		Attacks in an IoT topology with 105 devices. 47 Features. Provided are .pcap, features extracted as .csv, a jupyternotebook with example ML model and some additional source code	https://www.unb.ca/cic/datasets/iotdataset-2023.html	https://www.mdpi.com/1424-8220/23/13/5941	JR	
threat-hunting-samples	Threat Hunting			multiple datasets to practice Threat Hunting.	https://github.com/mosse-security/threat-hunting-samples		TUS	
				An open-source initiatve that				

Figure 1. External Information Technology (IT) datasets - online shared metadata collection form.

2.1.2 External OT Datasets

Operational Technology (OT) datasets are pivotal in managing and optimising industrial and infrastructure systems. These datasets contain vital information on the functioning and interaction of hardware and software systems controlling physical processes. Among the crucial elements of OT datasets is (again) network traffic information, which monitors the communication and data exchange within and between devices and control systems. Key features of network traffic in OT datasets include device identifiers, communication protocols, packet sizes, source and destination addresses, and timestamps. By analysing these features, professionals can ensure the reliability, efficiency, and security of industrial operations, enabling proactive maintenance, quick identification of anomalies, and enhanced operational efficiency.

In the scope of the Resilmesh project, the identified/suggested external OT datasets are meant to provide actionable information to train models for an AI-based distributed anomaly detection for early attack detection, an AI-based event correlation model for attack prediction and an innovative AI-penetration testing framework. The external OT datasets are also used to develop a federated learning framework to support the training of robust AI models and techniques to improve the robustness of cyber threat intelligence.

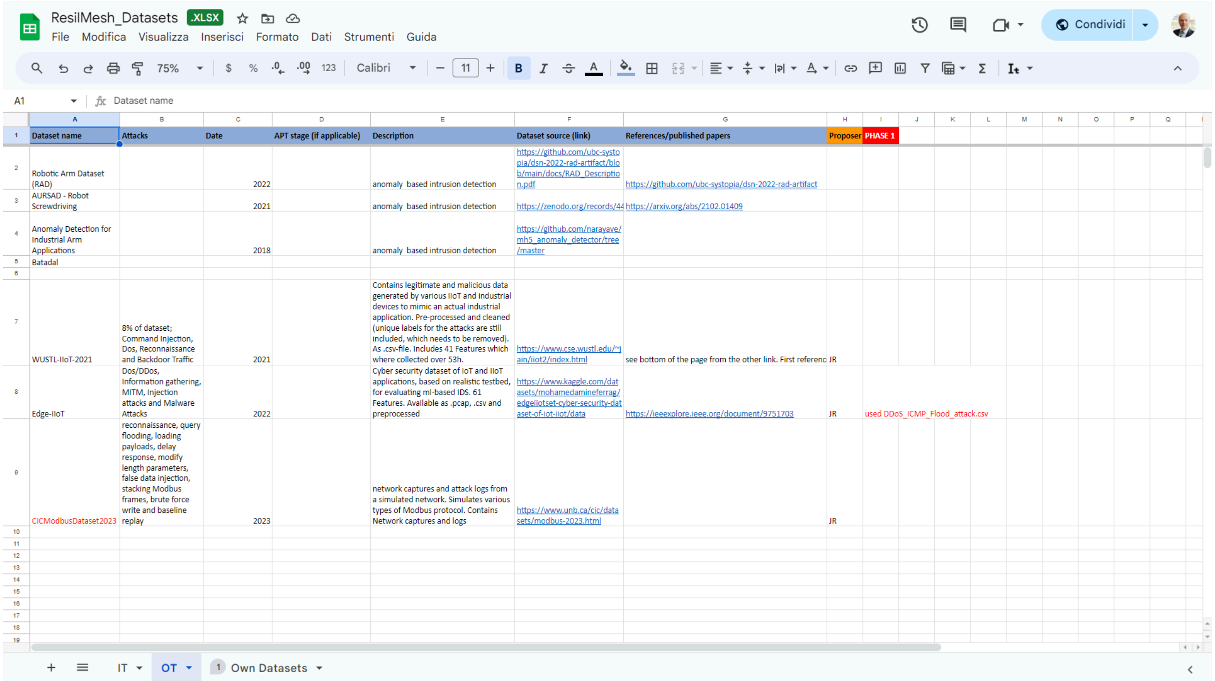
Metadata required for the characterisation of external OT datasets include:

Table 3: Metadata for the characterization of external OT datasets.

Metadata field name	**Metadata field description**
Dataset name	A (possibly) self-explanatory name of the dataset for reference purposes
Attacks	The types of attack that the dataset is about/is able to provide evidence for/can be used to improve the detection of
APT stage (if applicable)	The specific Advanced Persistent Threat (APT) stage that the dataset is about/is related to
Description	A short meaningful description of the main dataset features (what informative content is represented, the level of granularity, further useful technical explanations/insights etc.)
Dataset source (link)	(If available) The link where the dataset is available for inspection/download (it can be an internal restricted link)
References/published papers	Additional useful information on applicable public references on the dataset or related published papers
Proposer	The Resilmesh Consortium partner proposing the use of the dataset for training models

Also for External OT Datasets, a simple shared online form (Google Drive Sheet) was provided to involved partners for proper tracking and contributions (see Figure 2).

Funded by
the European Union



Dataset name	Attacks	Date	APT stage (if applicable)	Description	Dataset source (link)	References/published papers	Proposer	PHASE 1
Robotic Arm Dataset (RAD)		2022		anomaly based intrusion detection	https://github.com/ubc-systopia/dsn-2022-rad-artifact/blob/main/docs/RAD_Description.pdf	https://github.com/ubc-systopia/dsn-2022-rad-artifact		
AURSAD - Robot Screwdriving		2021		anomaly based intrusion detection	https://zenodo.org/records/44	https://arxiv.org/abs/2102.01409		
Anomaly Detection for Industrial Arm Applications		2018		anomaly based intrusion detection	https://github.com/narayave/mh5_anomaly_detector/tree/master			
Batadal								
WUSTL-IIoT-2021	8% of dataset; Command Injection, Dos, Reconnaissance and Backdoor Traffic	2021		Contains legitimate and malicious data generated by various IIoT and industrial devices to mimic an actual industrial application. Pre-processed and cleaned (unique labels for the attacks are still included, which needs to be removed). As .csv-file. Includes 41 Features which where collected over 53h.	https://www.cse.wustl.edu/~jain/iiot2/index.html	see bottom of the page from the other link. First referenc	JR	
Edge-IIoT	Dos/DDos, Information gathering, MITM, Injection attacks and Malware Attacks	2022		Cyber security dataset of IoT and IIoT applications, based on realistic testbed, for evaluating ml-based IDS. 61 Features. Available as .pcap, .csv and preprocessed	https://www.kaggle.com/datasets/mohamedamineferrag/edgeiiotset-cyber-security-dataset-of-iot-iiot/data	https://ieeexplore.ieee.org/document/9751703	JR	used DDoS_ICMP_Flood_attack.csv
CICModbusDataset2023	reconnaissance, query flooding, loading payloads, delay response, modify length parameters, false data injection, stacking Modbus frames, brute force write and baseline replay	2023		network captures and attack logs from a simulated network. Simulates various types of Modbus protocol. Contains Network captures and logs	https://www.unb.ca/cic/datasets/modbus-2023.html		JR	

Figure 2. External Operational Technology (OT) datasets - online shared metadata collection form.

2.2 Scope

The overall system of Resilmesh and its components compose the scope of testing. Only the components developed by the team are considered, external components and services are excluded from testing. The components are to be tested first. The list of components is taken from the Resilmesh project proposal and architecture plan as of Q1, 2024 and might be subject to minor changes. The list of components is presented in Table 5.

Table 4: List of Resilmesh components.

Component	Related Task
Resource Orchestrator	T3.2
Event aggregation & detection pipeline	T3.3
Enrichment	T3.4
Workflow orchestrator	T3.5

Funded by
the European Union

AI-based Detector	T4.1
AI-based Correlator	T4.2
SIEM	T4.3
CTI Sharing	T4.4
CSS Metamodel	T5.1
THF Framework	T5.2
NSSA Forecasting	T5.3
Mitigation Manager	T5.4
Pen Test Framework	T6.2

In the following sections, the use of the provided External Dataset is shortly summarised per component (only for applicable components).

2.2.1 Event aggregation & detection pipeline

This component will be a data processing framework to enable data collection and event aggregation. The component will convert ingested events into a normalised event format and route these using an event streaming platform such as NATS. External datasets may be used in testing the component's capabilities and performance.

2.2.2 AI-based Detector

This component will be an AI anomaly detection algorithm to detect suspicious events and attacks at both edge and cloud for host and network data including both logs and emitted events (IDS, MMT monitoring framework, etc.). It will multi-view anomaly detection algorithms to enable the blending of heterogeneous data sources to improve cross-domain anomaly detection, especially for mixed IT/OT critical infrastructures. It will use deep stacked networks (DSN), hybrid Stacked Autoencoders (SAE) and Convolutional Neural Networks (CNN) structures, and similar techniques to develop hierarchical feature fusion (e.g. via Autoencoders) and decision fusion (e.g., via ensemble learning) models for edge-based anomaly detection. Models will be trained using publicly available datasets in the first place and later with datasets generated within the project. The described deep neural networks require high amounts of data and the models need to be compared against other similar public works, which is why public external datasets are required.

2.2.3 AI-based Correlator

This component will be an algorithm or a set of algorithms to correlate security events or raw data to predict attack evolution, determine the root causes of attacks or anomalies, and reduce the number of events (e.g., false positives) submitted to the

SOC operator. It will use both deep learning (including spatio-temporal techniques) and broader data mining (rule mining, time series, Bayesian networks, etc.) approaches. Models will be trained using publicly available datasets in the first place and later with datasets generated within the project.

2.2.4 Pentest Framework

This component is a reinforcement learning-based testing framework (D6.2). The component will have one or more agents based on the selected algorithms and will incrementally implement and evaluate the framework for several attack scenarios as specified in T6.1. The component will mostly use data gathered during the project, but external datasets may be used to develop the agent training scenarios.

Funded by
the European Union

3 External Datasets Formalization

3.1 External IT Datasets

Table 5: External IT Dataset

Dataset name	UNSW-NB15 Dataset
Attack types	Nine types of attacks, namely, Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms
Date	Not available
APT stage (if applicable)	Not available
Description	Real modern normal activities and synthetic contemporary attack behaviours
Dataset source (link)	https://research.unsw.edu.au/projects/unsw-nb15-dataset
References/published papers	Not available
Proposer	UMU
Dataset IP addresses	private

Funded by
the European Union

Table 6: External IT Dataset

Dataset name	AIT-LDSv2.
Attack types	Scans (nmap, WPScan, dirb) Webshell upload (CVE-2020-24186) Password cracking (John the Ripper) Privilege escalation Remote command execution Data exfiltration (DNSteal) and stopped service
Date	Not available
APT stage (if applicable)	Not available
Description	
Dataset source (link)	https://zenodo.org/records/8263181
References/published papers	https://ieeexplore.ieee.org/document/9866880
Proposer	TUS
Dataset IP addresses	private

Table 7: External IT Dataset

Dataset name	TON_IoT Dataset
Attack types	Eight types of attacks, including Scanning, Dos, Data Injection, DDoS, Password Cracking, XSS, Backdoor, MITM
Date	Not available
APT stage (if applicable)	Not available
Description	Heterogeneous data sources collected from Telemetry datasets of IoT and IIoT sensors, Operating systems datasets of Windows 7 and 10 as well as Ubuntu 14 and 18 TLS and Network traffic datasets.
Dataset source (link)	https://research.unsw.edu.au/projects/toniot-datasets
References/published papers	https://drive.google.com/file/d/1VleSpdPBtu0ZcbzHGB71CyQYTeqRXBX4/view?usp=drive_link
Proposer(s)	TUS, UMU
Dataset IP addresses	private

Funded by
the European Union

Table 8: External IT Dataset

Dataset name	CICIDS2017
Attack types	Seven types of attacks, including Brute Force Attack, Heartbleed Attack, Botnet, DoS Attack, DDoS, Web Attack, Infiltration Attack
Date	Not available
APT stage (if applicable)	Not available
Description	Contains benign and the most up-to-date common attacks, which resembles the true real-world data (PCAPs). It also includes the results of the network traffic analysis using CICFlowMeter with labelled flows based on the timestamp, source, and destination IPs, source and destination ports, protocols and attack (CSV files)
Dataset source (link)	https://www.unb.ca/cic/datasets/ids-2017.html
References/published papers	https://drive.google.com/file/d/1YSEh5Gwwq_gCYWZZMdPquQpDBb1oa_e6/view?usp=drive_link
Proposer	TUS
Dataset IP addresses	private

Table 9: External IT Dataset

Dataset name	CICIDS2018
Attack types	Seven types of attacks, including Brute Force, Heartbleed, DoS, Web Attack, Infiltration Attack, Botnet, DDoS
Date	Not available
APT stage (if applicable)	Not available
Description	
Dataset source (link)	https://www.unb.ca/cic/datasets/ids-2018.html
References/published papers	https://drive.google.com/file/d/1YSEh5Gwwq_gCYWZZMdPquQpDBb1oa_e6/view?usp=drive_link
Proposer	TUS
Dataset IP addresses	public

Funded by
the European Union

Table 10: External IT Dataset

Dataset name	DARPA2000
Attack types	Denial-of-service attack
Date	Not available
APT stage (if applicable)	Not available
Description	
Dataset source (link)	https://www.ll.mit.edu/r-d/datasets/2000-darpa-intrusion-detection-scenario-specific-datasets#:~:text=Off%2Dline%20intrusion%20detection%20datasets,from%20consumers%20of%20these%20data.
References/published papers	https://drive.google.com/file/d/1t9NW1aL8zGRO4Ws4ypB3TG9-QqzYGf3B/view?usp=drive_link
Proposer	TUS
Dataset IP addresses	private

Table 11: External IT Dataset

Dataset name	Simulated Dataset (TON_IoT)
Attack types	Eight types of attacks, including Scanning, Dos, Data Injection, DDoS, Password Cracking, XSS, Backdoor, MITM
Date	Not available
APT stage (if applicable)	Not available
Description	Simulated Windows 10 dataset based on TON_IoT. Divided into 5 views of memory, process, processor, network, disk activity
Dataset source (link)	https://drive.google.com/file/d/1s0IqzuVWYPNEQeF7o2HlGkPc1WcUntaZ/view?usp=drive_link
References/published papers	
Proposer	TUS
Dataset IP addresses	private

Funded by
the European Union

Table 12: External IT Dataset

Dataset name	CSE-CIC-IDS2018
Attack types	Common cyber-attacks: DoS/DDoS, Darknet, Malware, Tor, Botnet... service
Date	Not available
APT stage (if applicable)	Not available
Description	
Dataset source (link)	https://www.unb.ca/cic/datasets/index.html
References/published papers	
Proposer	MONT
Dataset IP addresses	public

Funded by
the European Union

Table 13: External IT Dataset

Dataset name	CIC IoT dataset 2023
Attack types	33 different attacks in 7 categories: Dos/DDos, Bruteforce, spoofing, recon, web-based and mirai
Date	2023
APT stage (if applicable)	Not available
Description	Attacks in an IoT topology with 105 devices. 47 Features. Provided are .pcap, features extracted as .csv, a jupyter notebook with example ML model and some additional source code
Dataset source (link)	https://www.unb.ca/cic/datasets/iotdataset-2023.html
References/published papers	https://www.mdpi.com/1424-8220/23/13/5941
Proposer	JR
Dataset IP addresses	public

Funded by
the European Union

Table 14: External IT Dataset

Dataset name	threat-hunting-samples
Attack types	Threat Hunting
Date	Not available
APT stage (if applicable)	Not available
Description	Multiple datasets to practise Threat Hunting.
Dataset source (link)	https://github.com/mosse-security/threat-hunting-samples
References/published papers	
Proposer	TUS
Dataset IP addresses	public

Funded by
the European Union

Table 15: External IT Dataset

Dataset name	Security Datasets project
Attack types	Threat Hunting
Date	Not available
APT stage (if applicable)	Not available
Description	An open-source initiative that contributes malicious and benign datasets, from different platforms, to the infosec community to expedite data analysis and threat research.
Dataset source (link)	securitydatasets.com
References/published papers	
Proposer	TUS
Dataset IP addresses	n/a

Funded by
the European Union

Table 16: External IT Dataset

Dataset name	ATTACK DB OTX-XFORCE-VT
Attack types	Threat Hunting
Date	Not available
APT stage (if applicable)	Not available
Description	A rich AttackDB that consists of CTI from the MITRE ATT\&CK Enterprise knowledge base, the AlienVault Open Threat Exchange, the IBM X-Force Exchange and VirusTotal.
Dataset source (link)	https://ieee-dataport.org/documents/attack-db-otx-xforce-vt
References/published papers	https://ieeexplore.ieee.org/document/8701505
Proposer	TUS
Dataset IP addresses	n/a

Funded by
the European Union

3.2 External OT Datasets

Table 17: External OT Dataset

Dataset name	AURSAD - Robot Screwdriving
Attack types	
Date	2021
APT stage (if applicable)	Not available
Description	Anomaly based intrusion detection
Dataset source (link)	https://zenodo.org/records/4487073
References/published papers	https://arxiv.org/abs/2102.01409
Proposer	ALIAS
Dataset IP addresses	public

Funded by
the European Union

Table 18: External OT Dataset

Dataset name	Anomaly Detection for Industrial Arm Applications
Attack types	
Date	2018
APT stage (if applicable)	Not available
Description	Anomaly based intrusion detection
Dataset source (link)	https://github.com/narayave/mh5_anomaly_detector/tree/master
References/published papers	
Proposer	ALIAS
Dataset IP addresses	n/a

Funded by
the European Union

Table 19: External OT Dataset

Dataset name	Robotic Arm Dataset (RAD)
Attack types	
Date	2022
APT stage (if applicable)	Not available
Description	Anomaly based intrusion detection
Dataset source (link)	https://github.com/ubc-systopia/dsn-2022-rad-artifact/blob/main/docs/RAD_Description.pdf
References/published papers	https://github.com/ubc-systopia/dsn-2022-rad-artifact
Proposer	TUS
Dataset IP addresses	n/a

Funded by
the European Union

Table 20: External OT Dataset

Dataset name	WUSTL-IIoT-2021
Attack types	8% of dataset; Command Injection, Dos, Reconnaissance and Backdoor Traffic
Date	2021
APT stage (if applicable)	Not available
Description	Contains legitimate and malicious data generated by various IIoT and industrial devices to mimic an actual industrial application. Pre-processed and cleaned (unique labels for the attacks are still included, which needs to be removed). As .csv-file. Includes 41 Features which were collected over 53h.
Dataset source (link)	https://www.cse.wustl.edu/~jain/iiot2/index.html
References/published papers	see bottom of the page from the other link. First reference is the testbed; others are follow up research and case studies
Proposer	JR
Dataset IP addresses	n/a

Funded by
the European Union

Table 21: External OT Dataset

Dataset name	Edge-IIoT
Attack types	Dos/DDos, Information gathering, MITM, Injection attacks and Malware Attacks
Date	2022
APT stage (if applicable)	Not available
Description	Cyber security dataset of IoT and IIoT applications, based on realistic testbed, for evaluating ml-based IDS. 61 Features. Available as .pcap, .csv and preprocessed
Dataset source (link)	https://www.kaggle.com/datasets/mohamedamineferrag/edgeiiotset-cyber-security-dataset-of-iot-iiot/data
References/published papers	https://ieeexplore.ieee.org/document/9751703
Proposer	JR
Dataset IP addresses	n/a

Funded by
the European Union

Table 22: External OT Dataset

Dataset name	CICModbusDataset2023
Attack types	Reconnaissance, query flooding, loading payloads, delay response, modify length parameters, false data injection, stacking Modbus frames, brute force write and baseline replay
Date	2023
APT stage (if applicable)	Not available
Description	Network captures and attack logs from a simulated network. Simulates various types of Modbus protocol. Contains Network captures and logs
Dataset source (link)	https://www.unb.ca/cic/datasets/modbus-2023.html
References/published papers	
Proposer	JR
Dataset IP addresses	n/a

Funded by
the European Union

Table 23: External OT Dataset

Dataset name	Batadal
Attack types	Reconnaissance, Command injection, Denial-of-Service, Data Exfiltration and Stopped Service Attacks
Date	2017
APT stage (if applicable)	Not available
Description	Provides hourly historical SCADA operations data for water distribution systems. It includes Training Dataset 1, Training Dataset 2 and Test Dataset.
Dataset source (link)	https://www.batadal.net/data.html
References/published papers	https://hcis-journal.springeropen.com/articles/10.1186/s13673-019-0175-8
Proposer	ALIAS
Dataset IP addresses	n/a

Funded by
the European Union

Table 24: External OT Dataset

Dataset name	captures1_v2 captures2 captures3
Attack types	nominal state Main-in-the-Middle attack TCP Attacks ICMP Attacks
Date	2019-2021
APT stage (if applicable)	Not available
Description	This dataset was generated on a small-scale process automation scenario using MODBUS/TCP equipment, for research on the application of ML techniques to cybersecurity in Industrial Control Systems. The testbed emulates a CPS process controlled by a SCADA system using the MODBUS/TCP protoco
Dataset source (link)	https://github.com/tjcruz-dei/ICS_PCAPS/releases/tag/MODBUSTCP%231
References/published papers	Not available
Proposer	GMV
Dataset IP addresses	Private

Funded by
the European Union

Table 25: External OT Dataset

Dataset name	MITM Attack Datasets
Attack types	Man-in-the-Middle TCP/IP
Date	2023
APT stage (if applicable)	Not available
Description	This repository contains datasets created to evaluate the detection and classification of man-in-the-middle attacks, operating in eavesdropping mode, targeting MMS and Modbus TCP/IP protocols in the PAN of the smart grid.
Dataset source (link)	https://zenodo.org/records/8375657
References/published papers	Not available
Proposer	GMV
Dataset IP addresses	Private

Funded by
the European Union

Table 26: External OT Dataset

Dataset name	MQTTset
Attack types	Bruteforce Malformed data Flooding DoS attack
Date	2021
APT stage (if applicable)	Not available
Description	The dataset is composed by IoT sensors based on MQTT where each aspect of a real network is defined. In particular, the MQTT broker is instantiated by using Eclipse Mosquitto and the network is composed by 8 sensors.
Dataset source (link)	https://www.kaggle.com/datasets/cnrieiit/mqttset/data
References/published papers	Not available
Proposer	GMV
Dataset IP addresses	Private

Funded by
the European Union

4 Use Cases Datasets Overview

Resilmesh addresses three primary use cases:

1. **Civil regional infrastructure use case** inspired by regional IT infrastructures,
2. **Renewable energy use case** inspired by renewable energy and smart grid sector,
3. **Smart manufacturing robotics use case** inspired by industrial robotics.

The integration of AI in monitoring and controlling distributed IT and OT infrastructures has revolutionised their management, enhancing security against malicious attacks. The success of AI-driven solutions heavily relies on the quality and relevance of the datasets used. These datasets must cater to the specific needs and operational contexts of IT and OT domains to provide accurate, real-time insights and enable automated decision-making.

IT Infrastructure Datasets

Network Traffic Data: Essential for AI algorithms to monitor and manage data flow, including packet sizes, transfer protocols (TCP, UDP), IP addresses, port numbers, and timestamps. AI uses this data to detect anomalies, predict security breaches, and optimise network performance.

System Performance Metrics: Includes CPU and memory usage, disk I/O statistics, and network latency. AI analyses these metrics to predict system failures, automate load balancing, and optimise resource allocation.

Security Logs: Comprising intrusion detection logs, firewall records, and user authentication logs. AI systems use these datasets to identify unusual patterns, flag potential threats, and implement real-time security measures.

OT Infrastructure Datasets

Sensor and Actuator Data: Fundamental for AI applications, including sensor readings (temperature, pressure, humidity) and actuator statuses. AI models use this data to monitor industrial processes, predict equipment malfunctions, and adjust control parameters dynamically.

Equipment Health Metrics: Track equipment health through vibration analysis, thermal imaging, and operational cycles. AI algorithms use these metrics for predictive maintenance, scheduling proactive maintenance, and extending machinery lifespan.

Process Control Data: Control signals, process variables (flow rates, levels, pressures), and alarm logs. AI-based process optimization uses these datasets to enhance efficiency, and ensure product quality making immediate adjustments in response to deviations or alarms.

Data Management Strategies

- **Data Collection**
    - **IT Infrastructure:** Centralised systems aggregate data from various sources using network monitors, performance trackers, and security systems.
    - **OT Infrastructure:** Data collection is distributed, integrating multiple protocols like Modbus and OPC UA for real-time availability.
- **Data Processing**
    - **IT Infrastructure:** High-volume data streams are processed using cloud computing and big data frameworks.
    - **OT Infrastructure:** Edge computing processes data locally to minimise latency, crucial for real-time control.
- **Data Storage**
    - **IT Infrastructure:** Scalable cloud storage and data lakes ensure durability, accessibility, and compliance.
    - **OT Infrastructure:** Time-series databases and industrial data historians maintain historical process data integrity.
- **Data Security**
    - **IT Infrastructure:** Emphasis on encryption, access controls, and compliance with standards (GDPR, ISO 27001…).
    - **OT Infrastructure:** Security measures protect physical processes and ensure operational continuity, adhering to standards (ISA/IEC 62443…).

The test plan outlines key features for each use case:

1. **Environmental needs:** Details the events that occur, the data to be collected, attacks risks, and mitigation strategies
2. **Datasets and data collection:** Describes the sources of data for testing, including tools, formats, and methods to each use case.
3. **Attack execution and mitigation:** Discusses tools and methods for generating attacks and testing mitigation systems.

Funded by
the European Union

4.1 Civil Regional Infrastructure Use Case

This use case focuses on the cybersecurity needs of large IT infrastructures, those of major organisations or regional networks

4.1.1 Environmental needs

Key data includes system logs and network traffic records. Testing includes performance tests with large datasets. Generic IT infrastructure can be easily virtualized or simulated for testing.

4.1.2 Datasets and Data Collection

Public and private datasets contain network traffic or system logs for generic IT infrastructures. Collection tools are widely used. High-speed networks use IDS based on network flow monitoring over deep packet inspection. Testing focuses on flow-based data sets. Information sharing via platforms is common and should be tested for integration with Resilmesh components

4.1.3 Attack Execution and Mitigation

Mitigation strategies and testing includes various attack scenarios in a simulated environment. Plan is to detect attacks with high accuracy, correlate alerts and generate meta-alerts, share threat intelligence, forecast upcoming attacks, conduct penetration testing, selecting and executing mitigation actions.

Specific scenarios focus will be on the collaboration of Resilmesh components. Tests will use large datasets to assess performance, various attacks, and mitigation options. The goal is to verify the functionality of mitigation selectors and workflow tools and ensure successful collaboration of tools.

4.2 Renewable Energy Use Case

This use case focuses on protecting IT, IoT, and OT infrastructures in the energy sector, specifically around renewable energy production. Impacts include property damage, environmental harm, and loss of productivity and revenue. Testing focuses on cyber attacks with major impacts specific to this use case.

4.2.1 Environmental needs

The environment is distributed, with solar plants at multiple locations managed by a centralised SOC. This setup complicates asset management and maintenance. Assets include IT, IoT, and OT, requiring homogenization of diverse data inputs.

The network can be virtualized using simulated IT environments of renewable energy sources like PV plants for scalable and flexible testing.

Funded by
the European Union

4.2.2 Datasets and Data Collection

Testing requires various inputs for asset management during the testing. A combination of IT, IoT, and OT asset inventories should be used and checked whether the Resilmesh system can combine all such inputs.

4.2.3 Attack execution and mitigation

The threat landscape differs from IT use cases, with vulnerability to both cyber and physical attacks. Mitigation requires infrastructure segmentation or human intervention, such as site checks and manual reconfiguration.

A distributed infrastructure includes remote solar plants, control stations, smart metres, and a centralised operations centre. These communications can be intercepted and disrupted.

Plan is to detect and correlate disruptions to identify attacks early, detecting and mitigating attacks with minimal impact on asset availability, localise and isolate attacker's devices, and conduct analysis of such attacks.

This scenario tests environment-specific security monitoring, attack detection, alert correlation, and cyber situational awareness.

4.3 Smart Manufacturing Robotics Use Case

This use case focuses on industrial robotics, specifically OT systems in a single location, such as a factory with industrial robots.

The threat landscape and impacts of cyber attacks are similar to the combined IT, IoT, and OT use case.

4.3.1 Environmental needs

Factories often source robots from one provider, prioritising availability over security. Maintenance is outsourced, leading to delays in patching and hardening.

OT environments can be air-gapped from IT networks, reducing the attack surface, but are still vulnerable to insider attacks.

4.3.2 Datasets and Data Collection

Asset management datasets in this use case need to be collected. The asset inventory could be provided with a high level of detail but might not be available in a machine-readable format or otherwise straightforwardly readable by Resilmesh components.

Network traffic and system logs may differ from one OT system to another and there is a shortage of publicly available datasets. Custom datasets and attack traces should be collected for testing.

Funded by
the European Union

4.3.3 Attack execution and mitigation

Insider and supply chain attacks are key concerns. Different kinds of scenarios need to be created to test the functionality of the application. Mitigation options are limited due to rigid infrastructure.

An air-gapped network operates industrial robots. These networks and robots can be attacked using infection methods.

The plan is to identify and isolate infected assets, stop operations affected by the infection and resume unaffected ones, create infrastructure changes, mitigate the threat and resume normal operations with minimal asset unavailability while analysing the attack.

This scenario assesses incident response and threat mitigation while maintaining OT infrastructure availability.

Funded by
the European Union

5 Conclusions

This document illustrates the External Datasets structure, metadata and contents as identified by Resilmesh partners (use case owners included) for providing input to models and training them. Datasets in fact play a crucial role in training AI models by providing the raw data needed to learn patterns, make predictions, and improve performance over time, in this case as for cybersecurity-related issues in different application domains. The AI model training process involves several key steps – #1 is the one specifically addressed by the activity described in this deliverable D7.5:

1. **Data Collection**: Relevant and high-quality data is gathered from various sources, ensuring it is representative of the problem the AI model aims to solve.

2. **Data Preprocessing**: The collected data is cleaned and prepared for analysis. This involves handling missing values, normalising data, and transforming categorical data into numerical formats.

3. **Data Splitting**: The dataset is divided into three subsets: training, validation, and test sets. The training set is used to teach the model, the validation set tunes hyperparameters, and the test set evaluates the model's performance.

4. **Data Feature Engineering**: Important features are extracted or created from the raw data to enhance the model's ability to learn. This may involve selecting the most relevant variables, creating new features from existing ones, and scaling data appropriately.

5. **Model Training**: The training dataset is fed into the AI model, allowing it to learn the underlying patterns and relationships. This involves iterative processes where the model adjusts its parameters to minimise errors.

6. **Validation**: The model's performance is tested on the validation set to ensure it generalises well to new, unseen data. Adjustments are made based on validation results to avoid overfitting or underfitting.

7. **Evaluation**: After tuning, the final model is evaluated on the test set to assess its accuracy, precision, recall, and other performance metrics.

8. **Deployment**: Once validated, the trained model is deployed into real-world applications where it can make predictions or decisions based on new data.

By using well-structured datasets throughout these steps, AI models can achieve high accuracy and reliability: part of WP7 activities will also be devoted to the verification of quality and consistency of identified datasets for the continuation of technical activities in other WPs.

Overall, IT External Datasets provide data for testing various attack scenarios. The focus on network traffic and system logs allow for evaluation of intrusion detection and prevention systems. OT External Datasets are critical for testing challenges in industrial and energy sectors. Datasets can be used to evaluate managing and securing diverse assets while providing insights into mitigation strategies.

Funded by
the European Union