Public datasets used in Deep learning-based Anomaly Detection methods

YUAN LUO, Wuhan University
YA XIAO, Virginia Tech
LONG CHENG, Clemson University
GUOJUN PENG*, Wuhan University
DANFENG (DAPHNE) YAO*, Virginia Tech

ACM Reference Format:
Yuan Luo, Ya Xiao, Long Cheng, Guojun Peng, and Danfeng (Daphne) Yao. 2020. Public datasets used in Deep learning-based Anomaly Detection methods. *ACM Comput. Surv.* 1, 1 (December 2020), 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PUBLIC DATASETS

DLAD methods usually require a large volume of data to train and test neural models. It is important and essential to collect datasets for DLAD methods. In this section, we present publicly available datasets used in existing work. We summarize the characteristics of these datasets from (1) Systems and devices. The specific systems and devices where data is collected. (2) Period. When and how long the data has been collected. (3) Data types and size. Data types include sensors, actuators, network traffic, control system logs and commands, time series. (4) Attack or fault. We report the characteristics of attack or fault cases (if any). We list all available datasets in Table 1.

1.1 Datasets used in ICSs

SWaT. SWaT [15] is a six-stage scale-down water treatment testbed for research purposes, which implements main functionalities in a real-world water treatment plant. The raw water is pumped into the testbed at the first stage. The following four stages utilize chemical and physical processes (*e.g.*, Ultrafiltration (UF) and Reverse Osmosis (RO) systems) to filter and generate pure water. The final stage is a backwash step to the UF system. The physical devices include pumps, sensors (*e.g.*, the level of water, flow speed), tanks, and chemical/physical treatment devices. The cyber systems consist of a communication network, programmable logic controllers (PLCs) and the SCADA system. The dataset collected 7 days of normal data and 4 days of attack cases. The sensor and actuator values are in time-series form and sampled one data point every second, which is 125MB in normal period and 111MB in the attack period. The dataset also provides 50 network

*Corresponding authors

Authors' addresses: Yuan Luo, School of Cyber Science and Engineering, Wuhan University, Hubei, China, 430072, leonnewton@whu.edu.cn; Ya Xiao, Department of Computer Science, Virginia Tech, Blacksburg, VA, 24060, yax99@vt.edu; Long Cheng, School of Computing, Clemson University, Clemson, SC, 29634, lcheng2@clemson.edu; Guojun Peng, School of Cyber Science and Engineering, Wuhan University, Hubei, China, 430072, guojpeng@whu.edu.cn; Danfeng (Daphne) Yao, Department of Computer Science, Virginia Tech, Blacksburg, VA, 24060, danfeng@vt.edu.

0360-0300/2020/12-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

traces of normal period (300GB) and two network traces of attack period (104GB). 36 attacks (*e.g.*, false control signals, false sensor readings) are designed to simulate real-world attacks. At Aug. 2019, the dataset updated with three hours of normal and one hour of attack data.

Modbus network data. Modbus is one of the communication protocols used in SCADA systems. Lemay *et al.* [19] developed a SCADA sandbox to generate normal and attack Modbus network traffic. The sandbox consists of Master Terminal Units (MTUs), controllers and field devices. For each simulated case, the traffic capturing duration varies from 1 minute to 1 hour. The dataset provides 6 normal and 5 attack network traces, which is configured under different MTU and Remote Terminal Unit (RTU) settings. The size of each trace ranges from 1426 to 305932 entries. The attacks include malware and false control signals. The dataset can be downloaded at [18].

Tennessee Eastman process (TEP) simulation. TEP simulates a realistic setting of a chemical plant, which consists of a reactor, condenser, compressor, separator, and stripper. Totally, 53 measurements are collected from the system, of which 41 are normal values while 12 are manipulated. The normal measurements include temperature, level, pressure, flow, *etc.* The anomalous readings are feed flow, purge valve, steam valve, cooling water flow, *etc.* Since this simulation framework has been used in multiple methods [6, 20, 28, 32], we adopt a well-documented version presented in [25]. The dataset includes the fault-free train (23MB), fault-free test (45MB), fault train (471MB), and fault test (798MB) versions of data. The training and testing datasets run for 25 and 48 hours respectively, which are sampled every 3 minutes. Specifically, twenty-one faults are designed to create anomalies in the system, which includes fixed sensor readings, random variation and the slow drift of sensor values, *etc.* The dataset is available at [26]. The simulator can be obtained at [2].

Gas pipeline testbed. Morris *et al.* [22] built a laboratory-scale gas pipeline system, where Modbus network traffic data in the SCADA system was generated. The testbed consists of a pump, valve, pipeline, fluid flow, and air compressor. A proportional integral derivative (PID) controller is adopted to manage air pressure. Twenty features are captured from traffic data in the dataset, *i.e.*, the length of the packet, the pressure setpoint, PID related information, pressure, *etc.* The dataset is 17MB and contains 214580 normal traffic packets and 60048 packets in the attack period. There are three categories of attacks, *e.g.*, packet injection, DoS, MITM. The dataset is available at [21].

Smart Home Technology (REFIT) dataset. REFIT is a research project that studies buildings, users, energy, communication, and design in UK homes [24]. The project carried out surveys and interviews to understand the perceptions of smart homes and qualitative data on electricity and gas usage. Also, measurement data is collected in real-world households from field sensors and devices. Four datasets focus on different aspects of smart homes in the REFIT project. We report the REFIT smart home dataset [8] that is used in one DLAD method [17]. The devices include thermostats, valves, meters and motion, door, window sensors in 20 homes. The data was collected from October 2014 to April 2015. A description of the location, construction details, energy services of homes is provided. Then, power load, gas usage, temperature, user activity sensors are monitored to form the time-series dataset, which is 94MB. There is no attack or fault in the dataset.

PHM 2015 Challenge. This challenge provides the running status of real industrial plants, which includes time-series sensor measurements, control signals data, and fault events. The devices mainly comprise Heating, ventilation, and air conditioning (HAVC) and some electricity meters. The data sampling frequency is 15 minutes and the collection lasts around three years, which ranges from 2010 to 2012. For each HAVC, sensors 1 to 4 (no details) and control status 1 to 4 are recorded. Meanwhile, the instant power and electricity consumption of each zone are reported. Totally, the dataset contains 70 plants, whose size is about 390MB. Five types of faults are produced in each plant, which covers abnormal temperature, wrong temperature setpoint, wrong cooling zone, *etc.* The dataset is available at [30].

Table 1. Summary of publicly available datasets used in existing work. "●", "-", "◐" means "Yes", "Does not apply", and "Does not clear but inferred to be Yes" respectively. "D", "H" means "Day" and "Hour". "≈" means "approximately".

Name	Domain	Description	When	Period	Data type						Attacks					Faults		
					Sensor/Actuator reading	Size	Network traffic	Size	Control system logs	Size	DoS	MITM	Packet injection	Malware	False control signals	Sensor layer	Network layer	Control system
SWaT[15]	ICS	A scale-down water treatment testbed	2015	11D	●	236MB	●	404GB	-	-	-	●	-	-	●	-	-	-
Modbus[19]	ICS	Simulated Modbus network traffic data	2016	$\approx 7H$	-	-	●	$\approx 912K$ entries	-	-	-	-	-	●	●	-	-	-
TEP[26]	ICS	Simulated chemical plant	2017	146H	●	1.3GB	-	-	-	-	-	-	-	-	-	●	-	-
Gas pipeline[22]	ICS	Gas pipeline testbed	2015	-	-	-	●	17MB	-	-	●	●	●	-	-	-	-	-
REFIT smart home[8]	ICS	Smart home measurements	2015	7Months	●	94MB	-	-	-	-	-	-	-	-	-	-	-	-
PHM 2015 Challenge[30]	ICS	Plant measurements	2015	3Years	●	390MB	-	-	-	-	-	-	-	-	-	●	-	●
NYISO[23]	Smart grid	Power grid pricing, transmission, load data	Present	19Years	-	-	-	-	●	160 KB/day	-	-	-	-	-	-	-	-
IEEE X-bus system [4]	Smart grid	Power grid simulation	-	-	●	-	-	-	-	-	-	-	-	-	-	-	-	-
SPMD [1]	ITS	Safety pilot model deployment program	2014	2Years	●	3.2GB	●	16GB	-	-	-	-	-	-	-	-	-	-
UAH DriveSet[27]	ITS	Driver behaviour data	2016	$\approx 8H$	●	3.3GB	-	-	-	-	-	-	-	-	-	●	-	-
OTIDS [16]	ITS	CAN bus traffic	2017	≈ 42 Minutes	-	-	●	392M	-	-	●	●	●	-	-	-	-	-
SMAP[13]	Aerial systems	Telemetry data of a satellite	-	-	-	-	●	86MB	-	-	-	-	-	-	-	-	●	◐
Curiosity [13]	Aerial systems	Telemetry data of a rover	-	-	-	-	●	86MB	-	-	-	-	-	-	-	-	●	◐
UAV kernel events[29]	Aerial systems	Kernel event traces of a UAV	2016	-	-	-	-	-	●	-	●	-	-	●	-	-	-	-
Flightradar24[10]	Aerial systems	ADS-B messages	-	-	-	-	●	-	-	-	-	-	-	-	-	-	-	-

1.2 Datasets used in smart grid

New York Independent System Operator (NYISO). NYISO is responsible for managing the power grid and marketplace in New York, while it does not operate or own the infrastructure. It publishes the market and operational data (*i.e.*, pricing, power grid transmission, load data) every day. The load data is used by one work [9] to simulate a more real power grid. Researchers could also get pricing and transmission data. The dataset begins in May 2001 and keeps updating daily. Power load data of 11 areas in New York are recorded every five seconds, whose size is about 160KB each day. There is no attack or fault data in the dataset. The dataset is available at [23].

IEEE X-bus system. IEEE X-bus test system [4, 5] is an approximation of the American Electric Power system, which is developed to simulate the power grid system in the U.S. Depending on the bus quantity and network topologies, there are 14, 24, 30, 39, 57, 118-bus systems. The devices include buses, generators, transformers, synchronous condensers, lines. Since it is a simulation platform, researchers can collect simulated data for any period. Voltage, current, and frequency measurements can be recorded in the system. Typically, data from a real power grid can be loaded into the system to generate more realistic scenarios. Though the system does not provide attack or fault cases, users can inject manually created attacks and faults (*e.g.*, FDI) to simulate anomalies.

1.3 Datasets used in ITSs

Safety Pilot Model Deployment (SPMD) program. The SPMD program is to advance vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications with a real environment, equipment, and deployment, which is performed by the University of Michigan. Vehicle awareness devices (VADs) and aftermarket safety devices (ASDs) are installed on over 2500 real passenger vehicles to support safety-ensuring communication messages. From August 2012 to February 2014, the V2V data was collected. Brake events, basic safety messages (BSM), front targets, GPS, radar and network traffic statistics information are published. The sensor data is about 3.2GB (*e.g.*, brake, GPS, radar) and the network traffic is about 16GB (*e.g.*, BSM). There are no attack or fault cases in the dataset. The description of the program is at [1] and the dataset is available at [7].

UAH-DriveSet driver behaviour data. UAH-DriveSet utilizes six different types of passenger vehicles and six different drivers to perform driving behaviors on motorway and secondary road. Three driving strategies (*i.e.*, normal way, drowsy or aggressive mode) are adopted. Real vehicles with multiple sensors are applied to capture data, which are used in the Naturalistic Driving Study [3]. Over 500 minutes of driving performance data are collected in 36 tests. Speed, altitude, acceleration, latitude, and longitude coordinate information are stored in the dataset, which is 3.3GB. Aggressive driving behaviors are considered anomalies, which will cause sensor measurements to be different from the normal driving period. The description is at [27] while the dataset can be downloaded at [31].

CAN Dataset for intrusion detection (OTIDS). OTIDS provides CAN bus traffic that is generated during in-vehicle communication between different nodes. The attack-free dataset includes 2.3 million messages. DoS attack (656K messages), fuzzy attack (591K messages), impersonation attack (1.6 million messages) messages are injected in a real vehicle. The description is at [16] and the dataset can be downloaded at [12].

1.4 Datasets used in aerial systems

Soil Moisture Active Passive (SMAP) satellite. SMAP is a satellite developed to monitor the soil moisture and freeze on Earth. The telemetry data between the satellite and control center is published by [14]. Time information is anonymized and data is scaled between -1 and 1. There are 55 telemetry channels in the dataset and each channel represents one aspect of a spacecraft, *e.g.*, power. For each channel, there can be multiple sensors to measure the status. Totally, the dataset is 86MB. 43 point anomalies and 26 contextual anomalies are also given in the dataset, but the details are not presented. The dataset is available at [13].

Mars Science Laboratory (MSL) rover, Curiosity. The curiosity's mission is to investigate whether there is evidence on Mars that the environment is habitable for humans. Telemetry data is transmitted to send control commands and receive measurement data. In the work [14], this data is published along with the SMAP project. The data is also scaled to (-1,1) and time values are deleted, where 27 telemetry channels are recorded. A telemetry stream consists of several control commands and a telemetry value. Also, 19 point anomalies and 17 contextual anomalies are used as anomalous data. The details of the anomalies are not shown in the paper. Researchers can download the dataset at [13].

Logs from a UAV platform. This dataset offers kernel event logs from QNX RTOS operating system traces on a UAV platform. The UAV is operated in four modes, which are full-while, fifo-ls, hilRF-InFin, and sporadic. Each scenario contains training samples, validation cases, and anomalies. Multiple traces of a scenario are generated when experimenting, each of which contains 50000 samples. Four types of attacks are introduced. The first attack runs a loop to exhaust CPU computing resources. The other two attacks schedule interfering tasks to interrupt normal operations. The last attack runs in a normal mode but deviates from training samples. The description of the dataset is at [29].

Flightradar24. ADS-B messages, which are used by aircraft to broadcast position and running status information, are utilized in the work [11] to build an LSTM-based anomaly detection method.

Aircraft identification, position, velocities, status information can be contained in the message. In the work [11], over 800 flights from 14 airports are adopted, which range from March 2017 to April 2018. No anomaly cases are in the dataset, while the authors manually injected abnormal messages. ADS-B messages can be accessed at Flightradar24 [10].

REFERENCES

[1] National Highway Traffic Safety Administration. 2020. Safety Pilot Model Deployment Test Conductor Team Report. https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/812171-safetypilotmodeldeploydeltestcondrtmrep.pdf. (Accessed on 01/07/2020).

[2] Tennessee Eastman Challenge Archive. 2019. Tennessee Eastman process (TEP) simulation. Retrieved Dec 10, 2019 from http://depts.washington.edu/control/LARRY/TE/download.html

[3] Asher Bender, James R Ward, Stewart Worrall, Marcelo L Moreyra, Santiago Gerling Konrad, Favio Masson, and Eduardo M Nebot. 2016. A flexible system architecture for acquisition and storage of naturalistic driving data. *IEEE Transactions on Intelligent Transportation Systems* 17, 6 (2016), 1748–1761.

[4] Kittavit Buayai, Kittiwut Chinnabutr, Prajuap Intarawong, and Kaan Kerdchuen. 2014. Applied MATPOWER for Power System Optimization Research. *Energy Procedia* 56 (2014), 505–509.

[5] Electric Grid Test Cases. 2020. Retrieved Jan 06, 2020 from https://electricgrids.engr.tamu.edu/electric-grid-test-cases/

[6] Leo H Chiang, Evan L Russell, and Richard D Braatz. 2000. *Fault detection and diagnosis in industrial systems*. Springer Science & Business Media.

[7] Safety Pilot Model Deployment Data. 2020. Retrieved Jan 07, 2020 from https://catalog.data.gov/dataset/safety-pilot-model-deployment-data

[8] REFIT Smart Home dataset. 2017. Retrieved Dec 12, 2019 from https://repository.lboro.ac.uk/articles/REFIT_Smart_Home_dataset/2070091

[9] Qingyu Deng and Jian Sun. 2018. False Data Injection Attack Detection in a Power Grid Using RNN. In *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 5983–5988.

[10] Flightradar24. [n.d.]. Live Flight Tracker - Real-Time Flight Tracker Map. https://www.flightradar24.com/

[11] Edan Habler and Asaf Shabtai. 2018. Using LSTM encoder-decoder algorithm for detecting anomalous ADS-B messages. *Computers & Security* 78 (2018), 155–173.

[12] Hacking and Countermeasure Research Lab. 2020. CAN-intrusion-dataset (OTIDS). Retrieved Jan 07, 2020 from http://ocslab.hksecurity.net/Dataset/CAN-intrusion-dataset

[13] Kyle Hundman. 2020. Data from the Mars Science Laboratory and SMAP missions. Retrieved Jan 06, 2020 from https://github.com/khundman/telemanom

[14] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. 2018. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 387–395.

[15] iTrust Labs. 2019. iTrust Labs_Dataset Info. Retrieved Dec 09,2019 from https://itrust.sutd.edu.sg/itrust_labs_datasets/dataset_info/#swat

[16] Hyunsung Lee, Seong Hoon Jeong, and Huy Kang Kim. 2017. OTIDS: A novel intrusion detection system for in-vehicle network by using remote frame. In *2017 15th Annual Conference on Privacy, Security and Trust (PST)*. IEEE, 57–5709.

[17] Adrien Legrand, Brad Niepceron, Alain Cournier, and Harold Trannois. 2018. Study of Autoencoder Neural Networks for Anomaly Detection in Connected Buildings. In *2018 IEEE Global Conference on Internet of Things (GCIoT)*. IEEE, 1–5.

[18] Antoine Lemay. 2019. Modbus dataset. Retrieved Dec 10, 2019 from https://github.com/antoine-lemay/Modbus_dataset

[19] Antoine Lemay and José M Fernandez. 2016. Providing SCADA Network Data Sets for Intrusion Detection Research. In *9th Workshop on Cyber Security Experimentation and Test (CSET 16)*.

[20] Chao Liu, Kin Gwn Lore, and Soumik Sarkar. 2017. Data-driven root-cause analysis for distributed system anomalies. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 5745–5750.

[21] Tommy Morris. 2019. Industrial Control System (ICS) Cyber Attack Datasets. Retrieved Dec 11, 2019 from https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets

[22] Thomas H Morris, Zach Thornton, and Ian Turnipseed. 2015. Industrial control system simulation and data logging for intrusion detection system research. *7th Annual Southeastern Cyber Security Summit* (2015), 3–4.

[23] New York Independent System Operator. 2020. Energy Load Data. https://www.nyiso.com/load-data

[24] The REFIT project. 2019. REFIT datasets. Retrieved Dec 12, 2019 from https://www.refitsmarthomes.org/datasets/

[25] CA Rieth, BD Amsel, R Tran, and MB Cook. 2017. Additional Tennessee Eastman Process Simulation Data for Anomaly Detection Evaluation. *Harvard Dataverse* (2017).

[26] Cory A. Rieth, Ben D. Amsel, Randy Tran, and Maia B. Cook. 2017. Additional Tennessee Eastman Process Simulation Data for Anomaly Detection Evaluation. (July 2017). https://doi.org/10.7910/DVN/6C3JR1 type: dataset.

[27] Eduardo Romera, Luis M Bergasa, and Roberto Arroyo. 2016. Need data for driver behaviour analysis? Presenting the public UAH-DriveSet. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 387–392.

[28] Evan L Russell, Leo H Chiang, and Richard D Braatz. 2012. *Data-driven methods for fault detection and diagnosis in chemical processes*. Springer Science & Business Media.

[29] Mahmoud Salem, Mark Crowley, and Sebastian Fischmeister. 2016. Anomaly detection using inter-arrival curves for real-time systems. In *2016 28th Euromicro Conference on Real-Time Systems (ECRTS)*. IEEE, 97–106.

[30] PHM Society. 2020. PHM Data Challenge. Retrieved Jan 05, 2020 from https://www.phmsociety.org/events/conference/phm/15/data-challenge

[31] The UAH-DriveSet. 2020. Retrieved Jan 07, 2020 from http://www.robesafe.com/personal/eduardo.romera/uah-driveset/

[32] Shen Yin, Steven X Ding, Adel Haghani, Haiyang Hao, and Ping Zhang. 2012. A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. *Journal of process control* 22, 9 (2012), 1567–1581.