Ital-IA, AI per la Sostenibilità
30.05.2023

Smart Electrical Grids Under the Lens of Adversarial Attacks

Athours: Fatemeh Nazary, Yashar Deldjoo, Tommaso Di Noia, Carmelo Ardito, Eugenio Di Sciascio
Politecnico di Bari

Ital-IA
ITALIA INTELLIGENZA ARTIFICIALE
cini National Lab AIIS

Politecnico di Bari

SISINF LAB

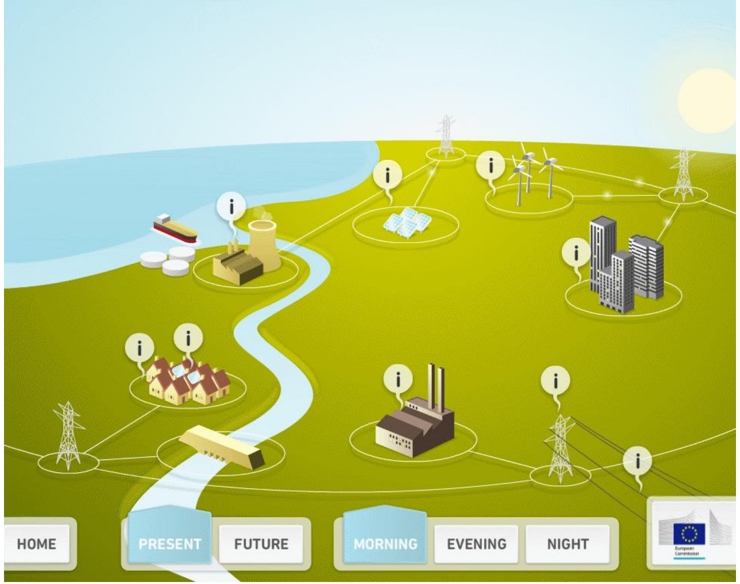
1

Problem Definition

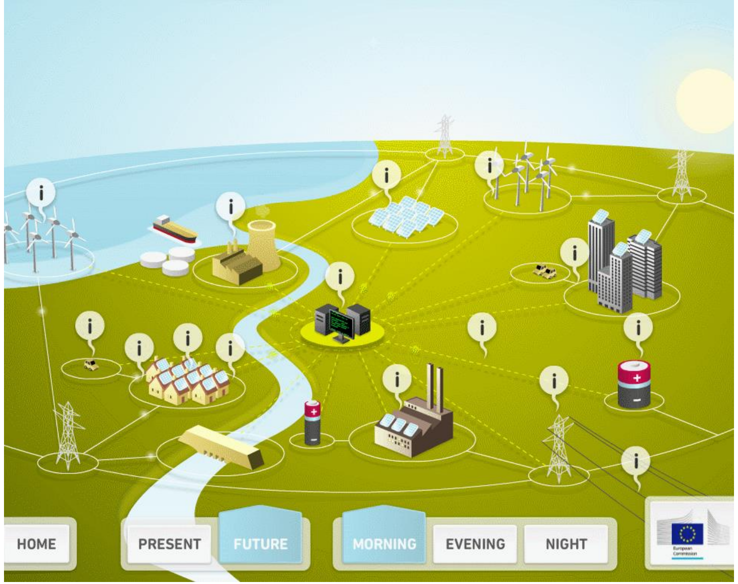
The goal of this research work

What is Smart Electical Grid?

Traditional vs. Smart Grid



Traditional



Smart Grid

https://ses.jrc.ec.europa.eu/smart-grid-interactive-tool-non-flash

Why Smart Grids?

1. Difficulty for the traditional grid to respond to the raising **energy demands**
2. Heterogeneous energy resources and **renewable** ones (winds, solar)
3. A **two-way** dialogue where electricity and information could be exchanged between customers and utilities
4. **Raising security and safety** of SG

Why Smart Grids?

1. Difficulty for the traditional grid to respond to the raising **energy demands**
2. Heterogeneous energy resources and **renewable** ones (winds, solar)
3. A **two-way** dialogue where electricity and information could be exchanged between customers and utilities
4. **Raising security and safety** of SG

self-healing feature

What is Self-Healing Feature in Smart Grids?

General definition

Self-healing ability is a smart network that uses sensing, control, and communication technology to allow for **real-time troubleshooting** for unforeseen events.

What is Self-Healing Feature in Smart Grids?

General definition

Self-healing ability is a smart network that uses sensing, control, and communication technology to allow for **real-time troubleshooting** for unforeseen events.

Unintentional threats

- Natural faults
- Faults created due to human error
- Faults created due to aging of equipment

Intentional adversarial threats

- Adversarial attacks against ML models in SGs
- Adversarial attacks against fault prediction models

What is Self-Healing Feature in Smart Grids?

General definition

Self-healing ability is a smart network that uses sensing, control, and communication technology to allow for **real-time troubleshooting** for unforeseen events.

Unintentional threats

- Natural faults
- Faults created due to human error
- Faults created due to aging of equipments

Intentional adversarial threats

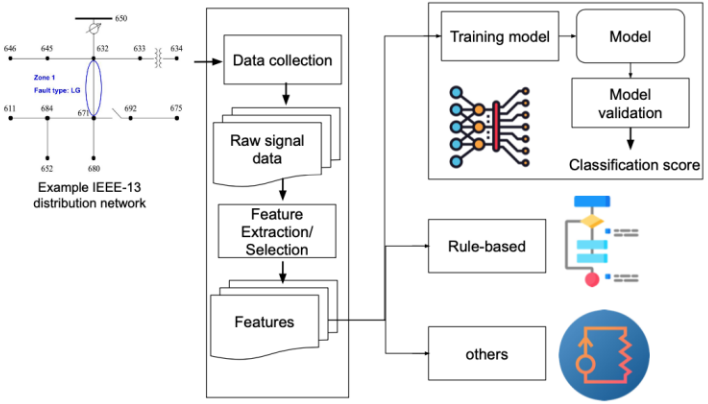
- Adversarial attacks against ML models in SGs
- Adversarial attacks against fault prediction models

Main focus

Trustworthy ML in Smart Grids

Security of smart grids has been studied under different lens

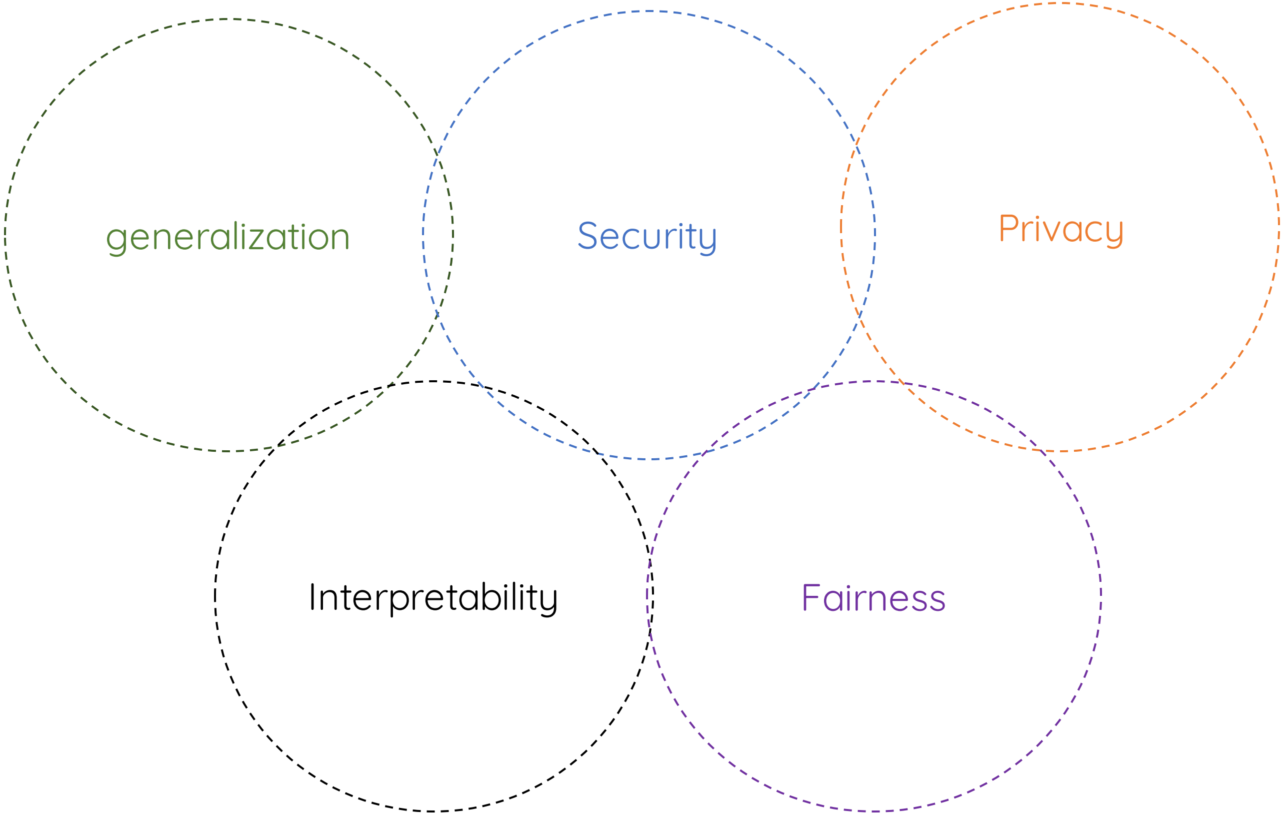
- Electrical engineering
- Signal processing
- Artificial intelligence and ML



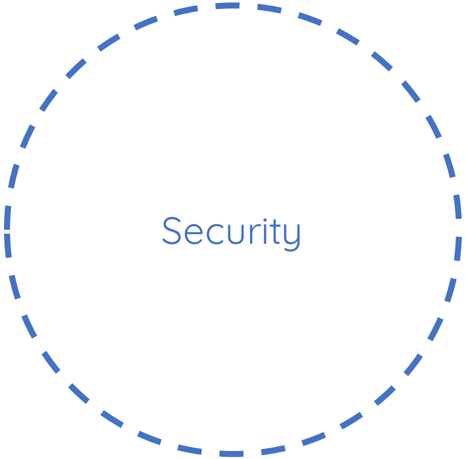
Now we should go for real thing

- Should we change the way we look at trustworthiness of ML models developed for smart grids?

Trustworthy ML in Smart Grids



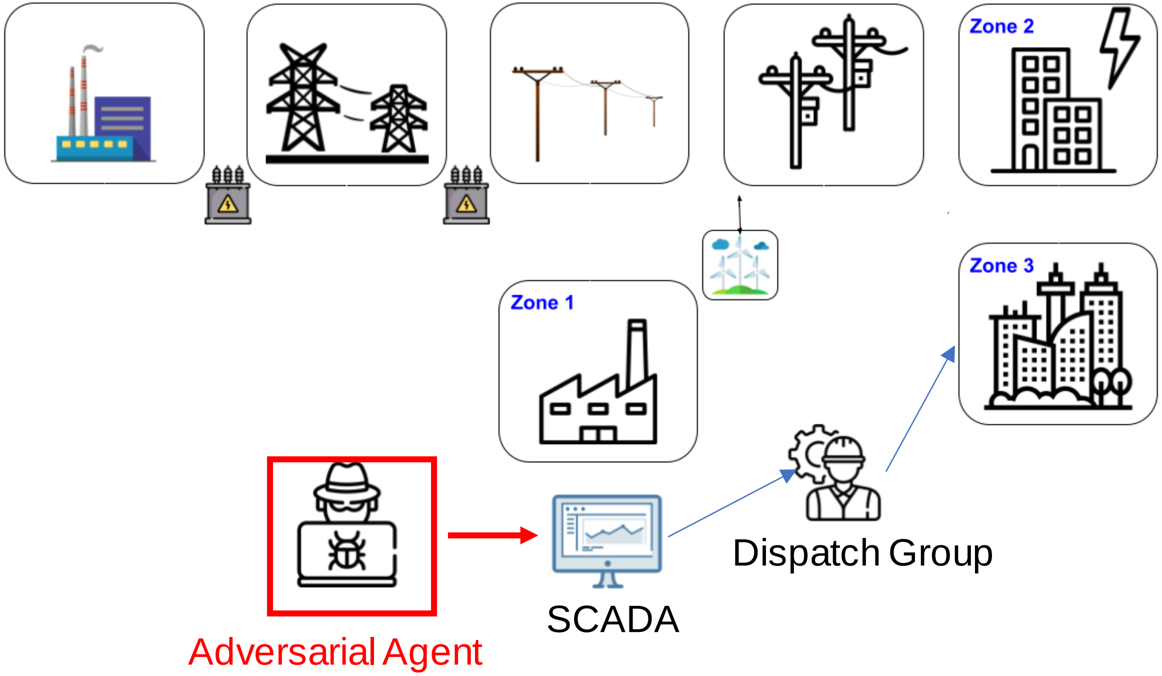
Focus of this Reasearch Work



Security

Development of trustworthy ML solutions for self-healing feature under smart grids by considering their robustness and security

A hypothetical illustration of targeted adversarial attacks against fault zone prediction in smart gids

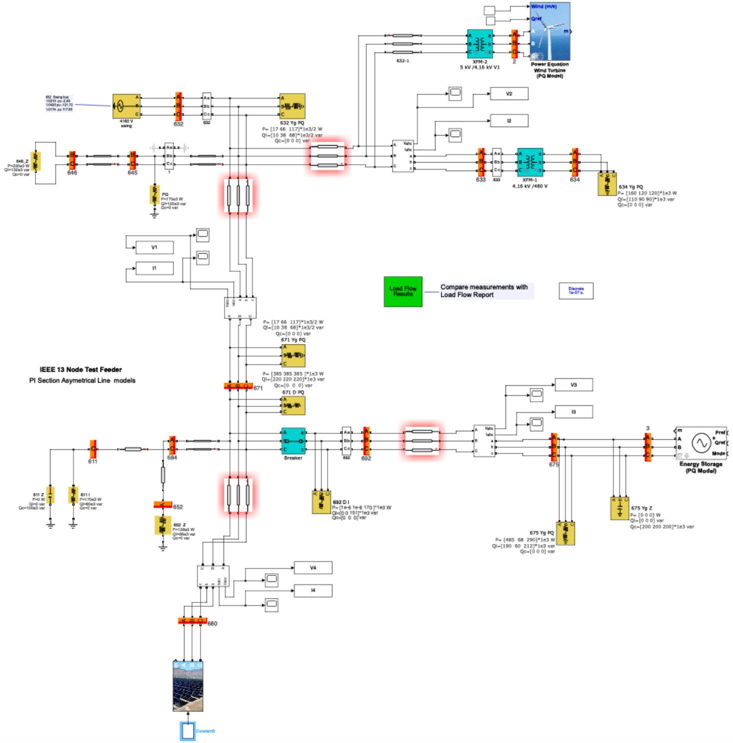


Dataset collected by Simulation environment

Test grids simulate the behavior of an actual distribution feeder that has been established to assess various three-phase grid algorithms.

- IEEE-13

Dataset collected with renewable energies

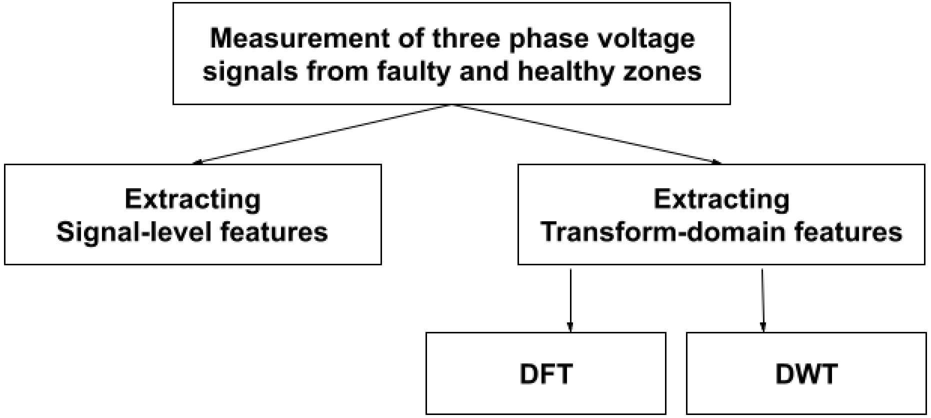


Dataset Characteristics

Duration:
t = [0.0 − 0.02]

Item	Details
Fault type	Phase to ground (AG, BG, CG) Phase to phase (AB, AC, BC) Phase to phase to ground (ABG, ACG, BCG) Three phase (ABC) Three phase to ground (ABCG)
Fault location	zone 1 branch 632-671 zone 2 branch 632-633 zone 3 branch 692-675 zone 4 branch 671-680
Fault resistance	0.0010, 0.0273, 0.0535, 0.0798 0.1061, 0.1323 0.1586, 0.1848 0.2111, 0.2374, 0.2636, 0.2899 0.3162, 0.3424, 0.3687, 0.3949 0.4212, 0.4475, 0.4737, 0.5, 1, 2

Features


Measurement of three phase voltage signals from faulty and healthy zones
Extracting Signal-level features
Extracting Transform-domain features
DFT
DWT

1. Raw time series data
2. three types of features ------> Time-domain features
   Frequency-domain features (DFT)
   Discrete Wavelet transform (DWT)

Six aggregation functions applied to the voltage signal $x\ (t)$ including (mean, standard deviation, skewness, kurtosis, energy, and maximum level of the signals)

Fault Classification in Smart Grid:

A Multi-layer Perceptron (MLP) neural network is trained for different multi-class classification problems pertinent to fault prediction in smart grids with $K \geq 2$ classes.

- Fault location classification (FLC): with $K = 4$ the task aims to classify a given signal into its originating zone
- Fault type classification (FTC): with $K = 11$ the task aims to classify a given signal into one of predefined fault types
- Joint location and type classification (FLC+FTC) $K = 44$ integrating the both fault class labels

Adversarial Attacks against Fault ML model in Smart Grid:

Proposing Adversarial attacks against fault classification

- Attack Scenario: both untargeted and targeted
- Explored attacks: (1) Fast gradient sign method (FGSM)
  (2) Basic iterative method (BIM)
  (3) Carlini and Wagner (C&W)

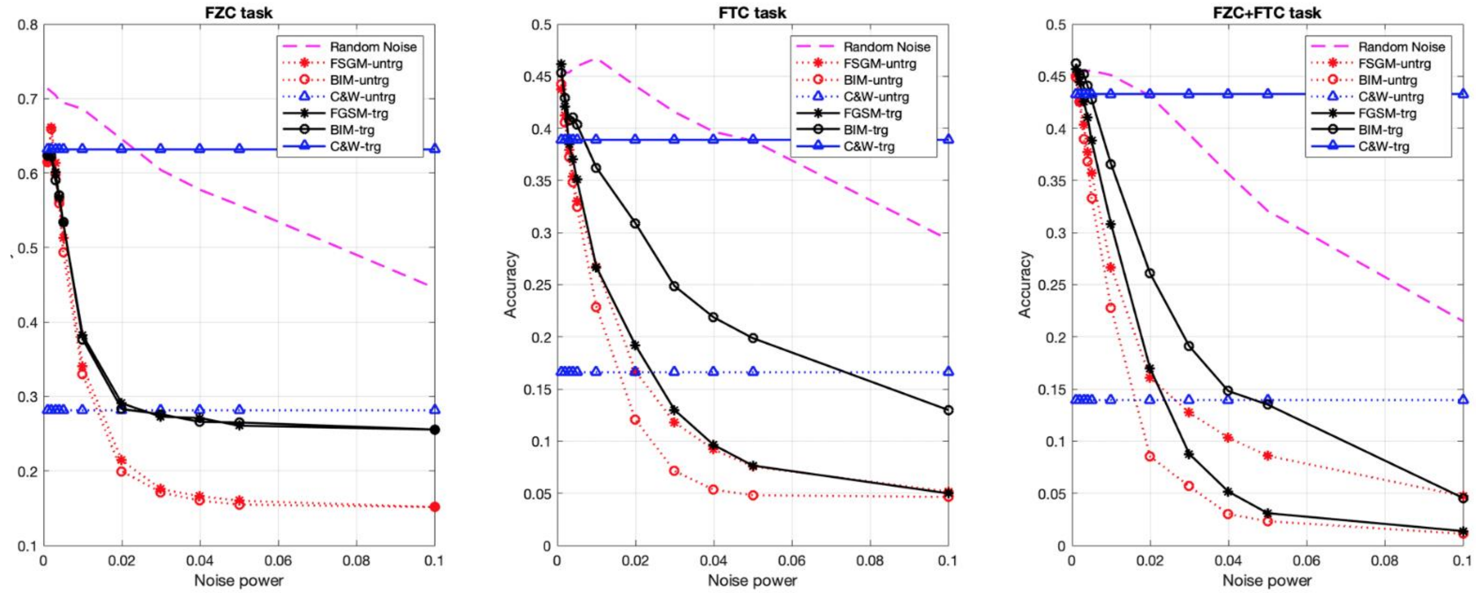
In the untargeted scenario, FGSM aims to generate a perturbation that maximizes the training loss formulated as :

$$\delta = \epsilon \cdot \backslash\mathbf{sign}(\nabla_x \ell(f(x;\theta), y))$$

A targeted FGSM attack is, instead, formulated as:

$$\delta = -\epsilon \cdot \backslash\mathbf{sign}(\nabla_x \ell(f(x;\theta), y_T))$$

Adversarial Attacks against Fault ML model in Smart Grids:



Closing remarks and Future work

- The security and vulnerability of fault classification systems driven in the context of smart electrical grids
- defending against alternative adversarial training and detection techniques would require more in-depth research
- Considering the privacy of fault-prediction systems such that separate zones do not need to exchange their data with a central server (Federated learning)
- Dataset and code are available in: https://bit.ly/3NT5jxG

Thanks!

ANY QUESTIONS?

You can find me at

fatemeh.nazary@poliba.it