SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Machine Learning for Cyber-Physical System Security

David K.Y. Yau

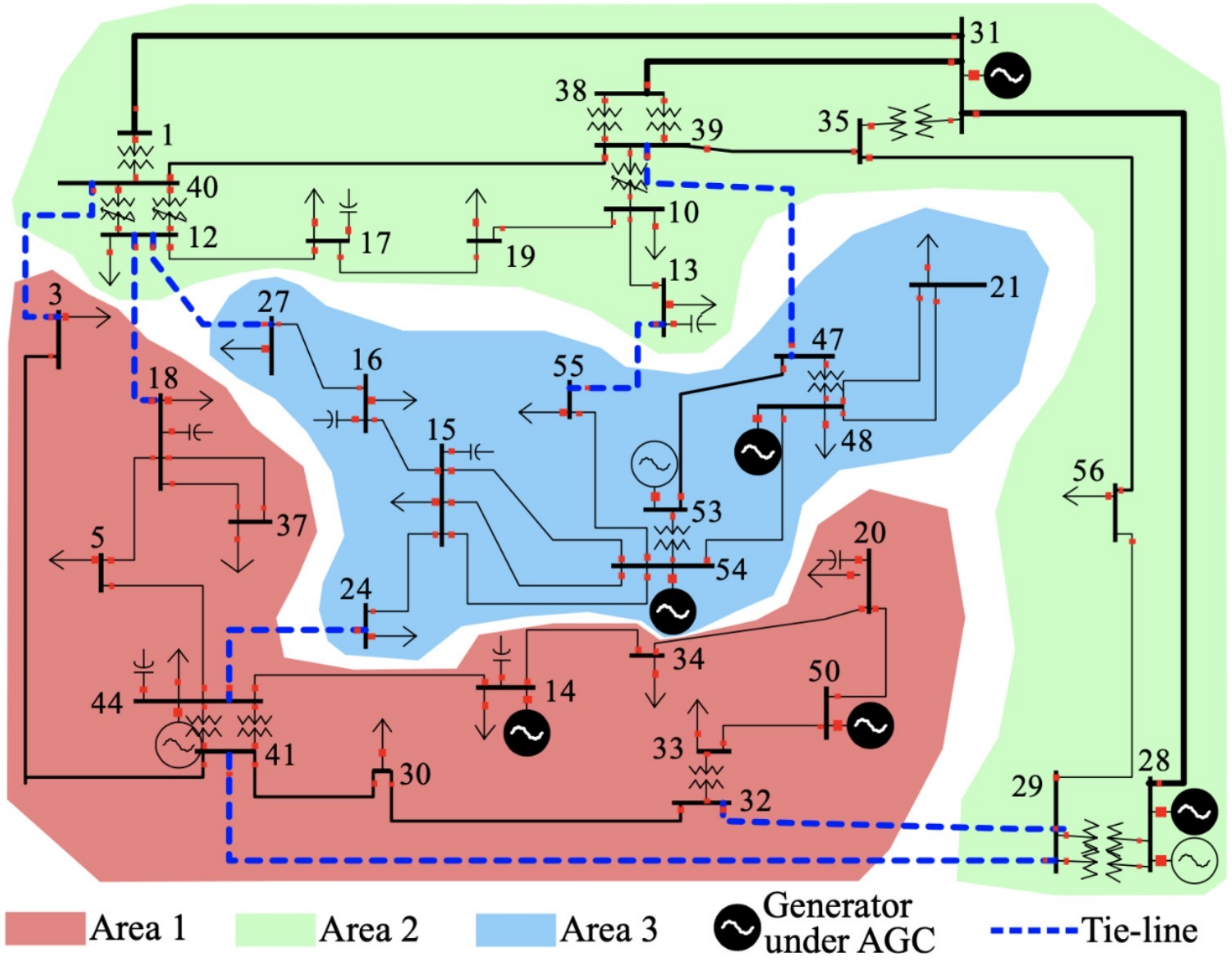
Smart Grid Frequency Control

- Electrical (ac) grids run at a standard nominal frequency (a global property of the grid)
  - E.g., 50Hz in Asia, 60Hz in North America
- Electricity supply should match demand
- If demand increases (exceeds supply), frequency drops
- If deviation from nominal more than 0.5Hz => frequency excursion
- If excursion persists, generators are impacted (e.g., shut down)

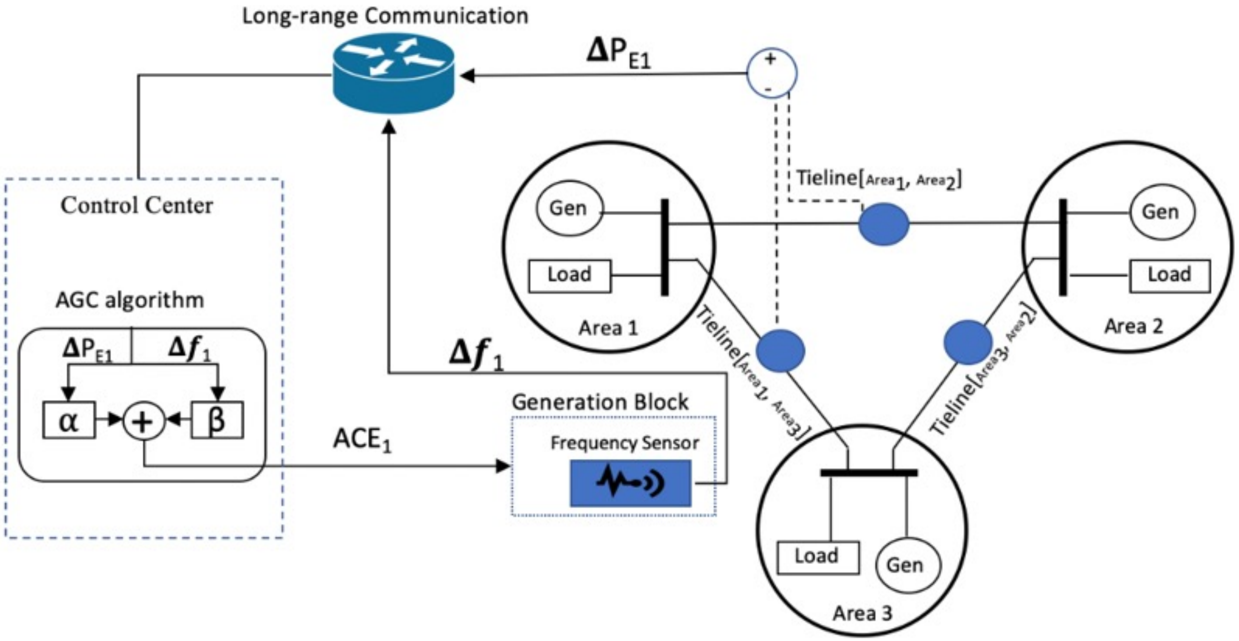
Automatic Generator Control (AGC)

- A fundamental control to maintain grid's nominal frequency
- Aims to adjust supply to match changing demand
  - E.g., when demand rises, ramp up generator speed to supply more
- Works in a feedback control loop under a specifiable *gain* parameter
  - Gain impacts responsiveness and stability
- A large grid may have multiple generator and load buses
  - Organized into multiple (interconnected) *areas*
  - Electricity flows between areas along *tie-lines*, subject to distribution of demand / supply

A multi-area electrical grid

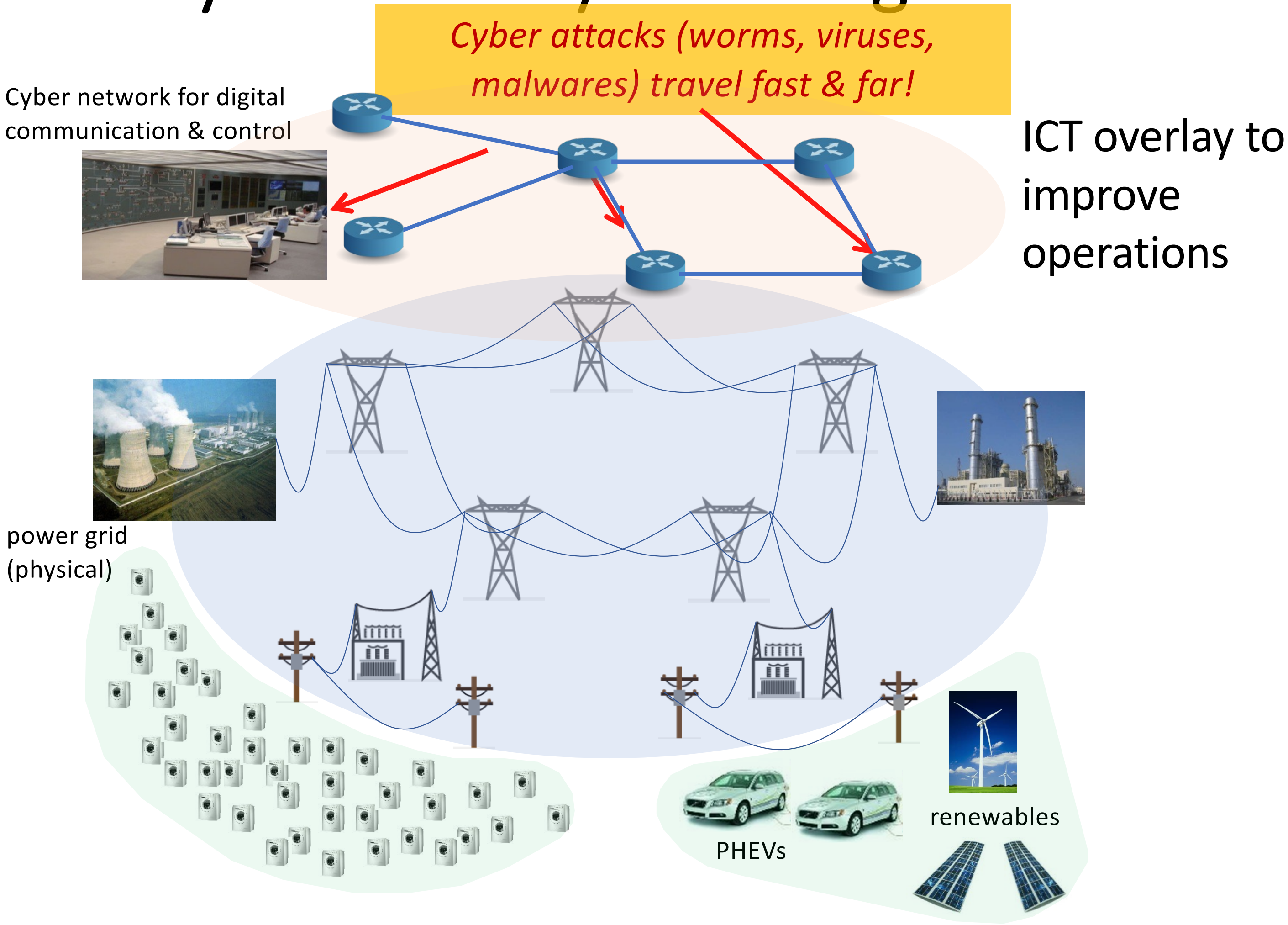


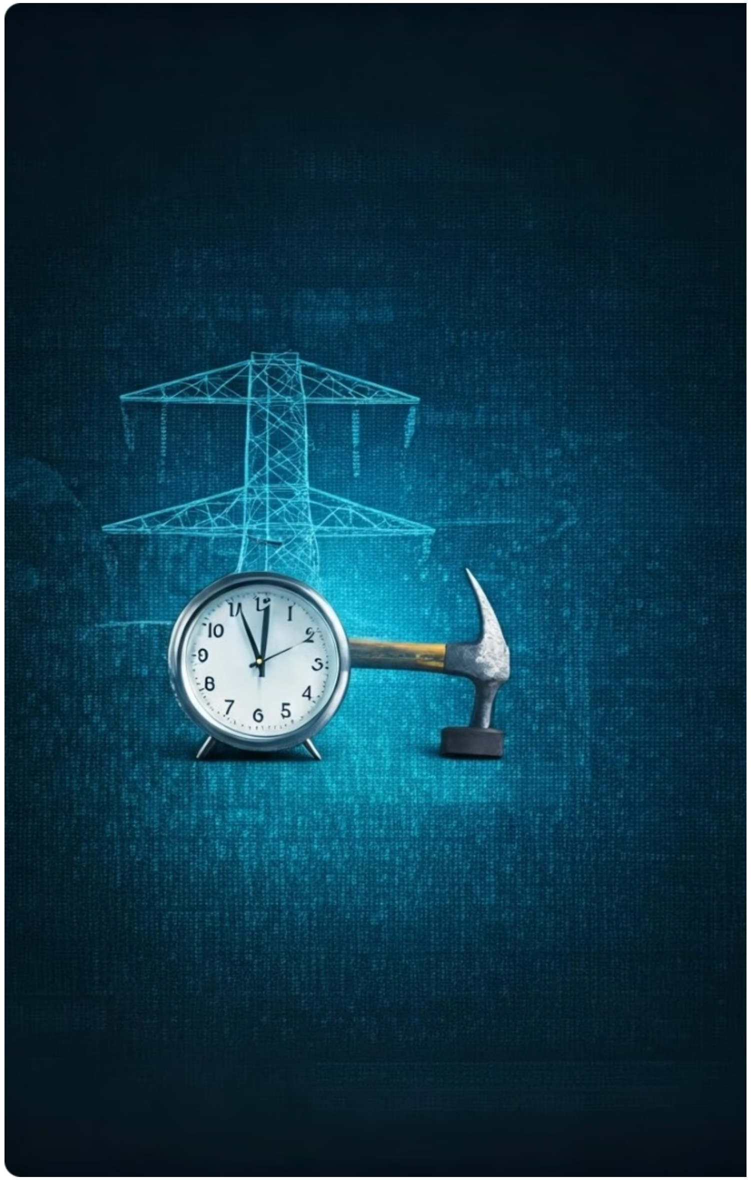
AGC loop in multi-area grid



- Adjustment based on area control error (ACE)
- Aims to correct frequency & power export deviation

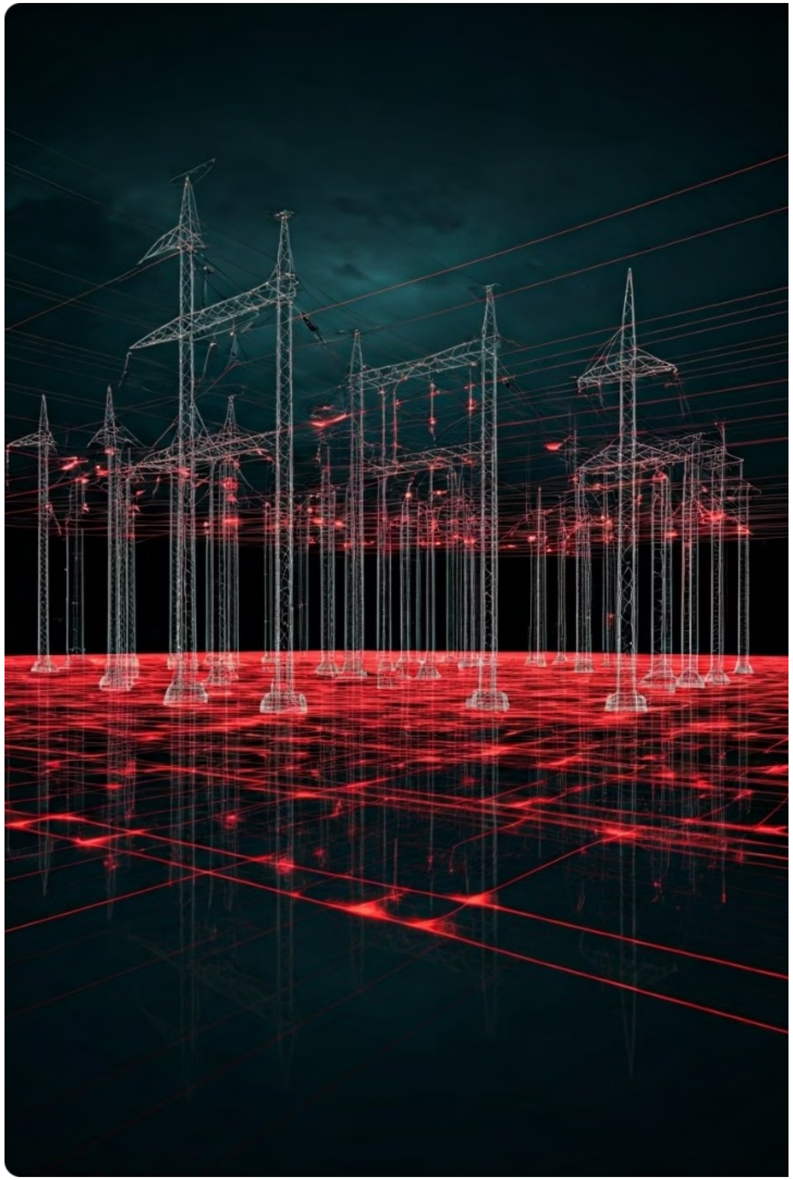
Smart Grid: Cybersecurity Challenges


Cyber attacks (worms, viruses, malwares) travel fast & far!
Cyber network for digital communication & control
ICT overlay to improve operations
power grid (physical)
renewables
PHEVs



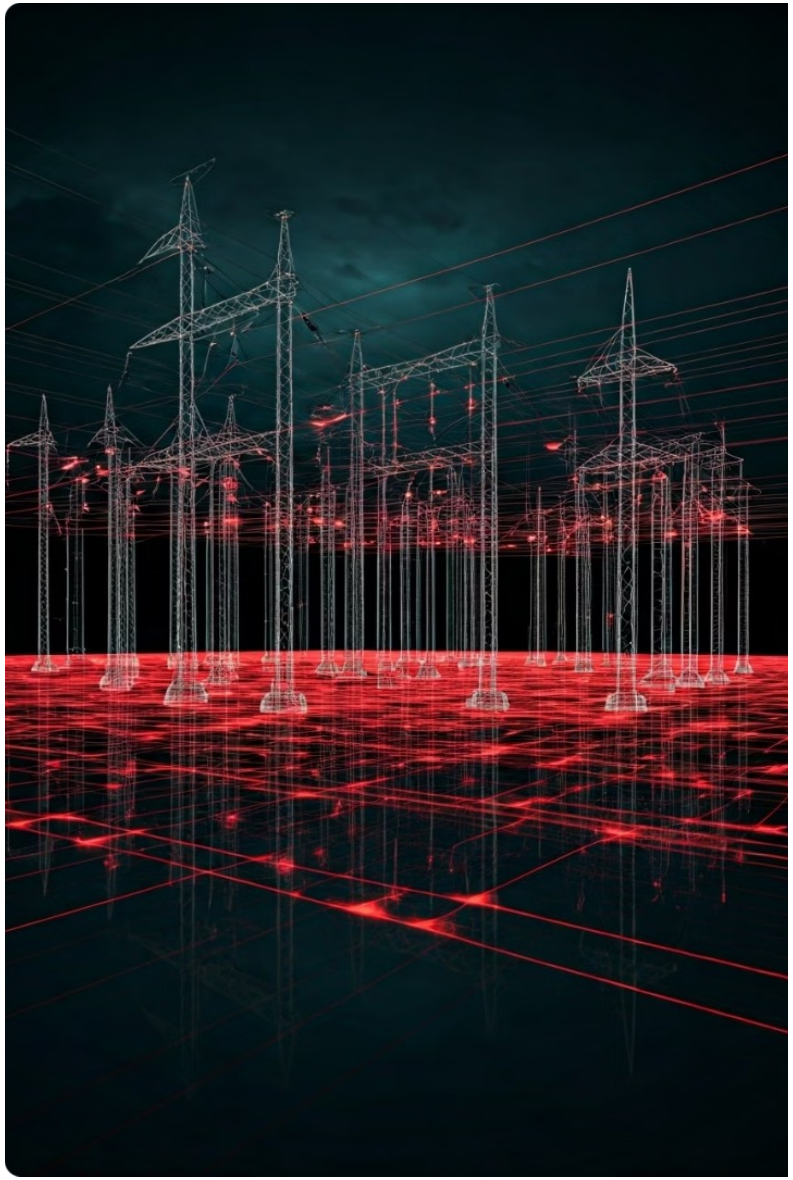
Time Delay Attack (TDA)

- Introduces malicious delays into network communications
  - E.g., MITM buffering of SCADA packets for actuation
- Encrypting packets may not help
- Trustworthy clock synchronization among distributed devices can be challenging



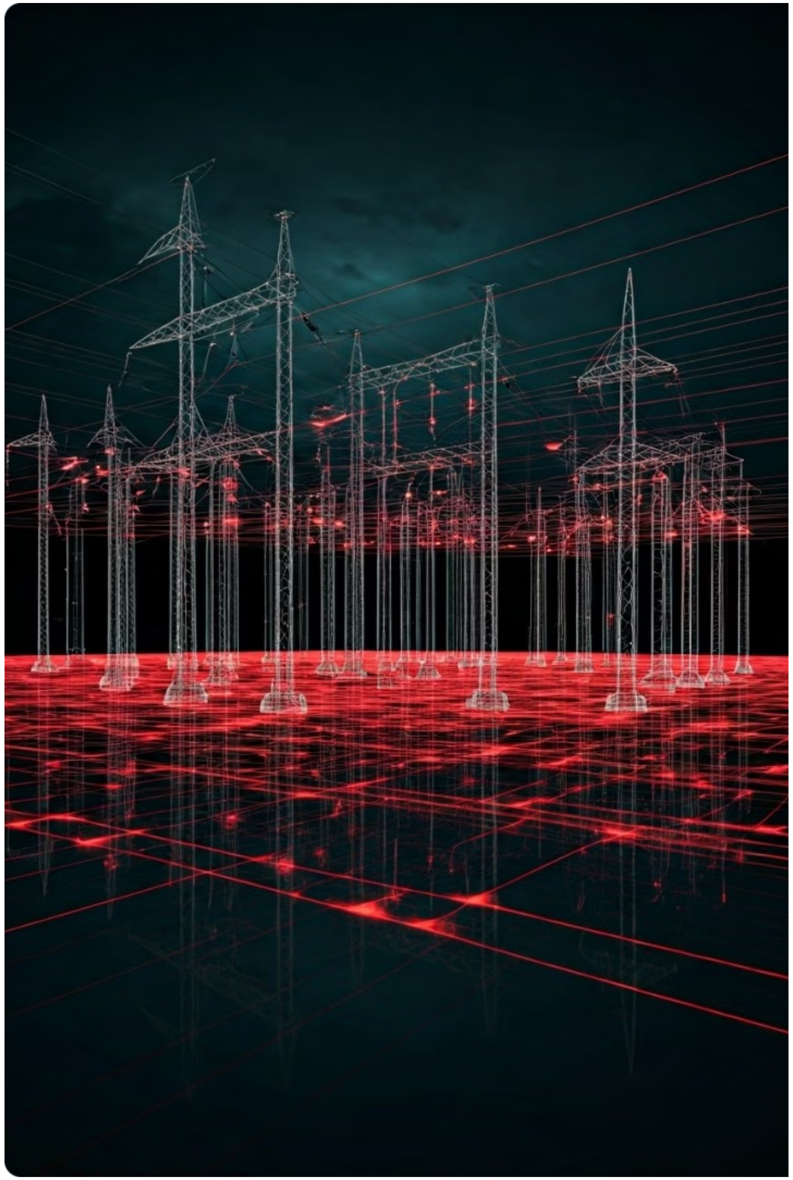
False Data Injection (FDI)

- Tampers with sensing and control content in SCADA packets
- Bypasses operator’s integrity check, e.g., bad data detection (BDD)
- Can take different forms
  - Bias attack, scaling attack, etc.
  - Sophisticated design possible …



Time-optimal FDI (FDI-optimal)

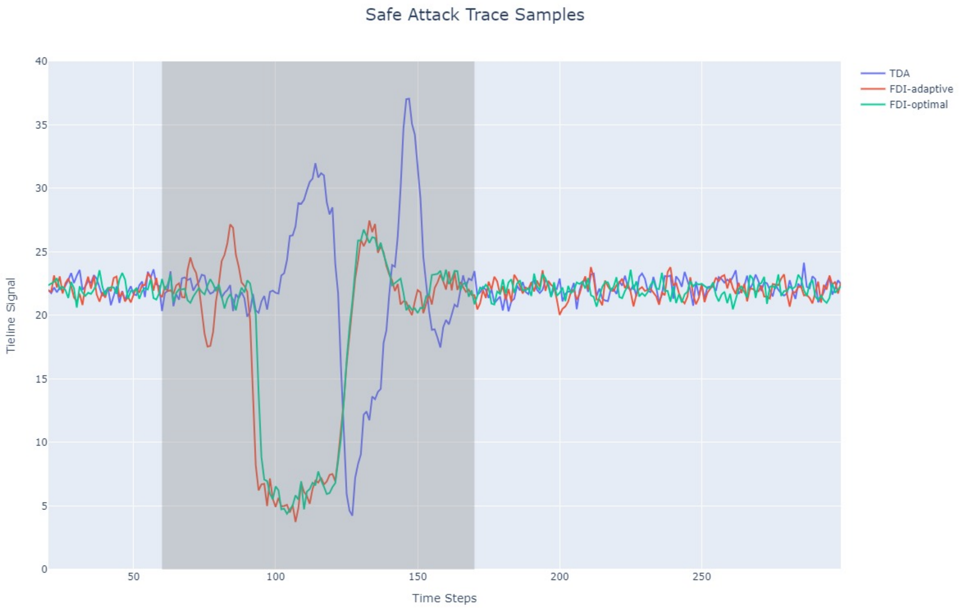
- Minimizes *time-to-emergency* (TTE)
- Causes system damage in the least time (since launch of attack)
- Persists over multiple AGC cycles, while satisfying BDD-bypass constraints



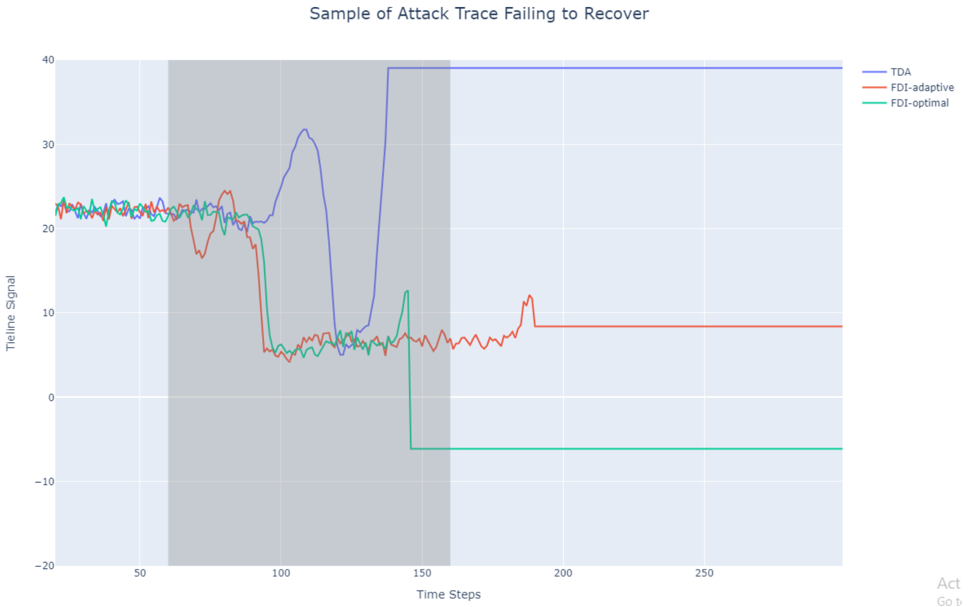
Adaptive FDI to keep stealthy (FDI-adaptive)

- Modifies tie-line measurements while keeping frequency deviations within a specified target
- Phase 1: Learns control model while mimicking normal operation
- Phase 2: Once ready, promptly drives system frequency beyond safe range

Footprint of attacks in tie-line flows



Without system crash



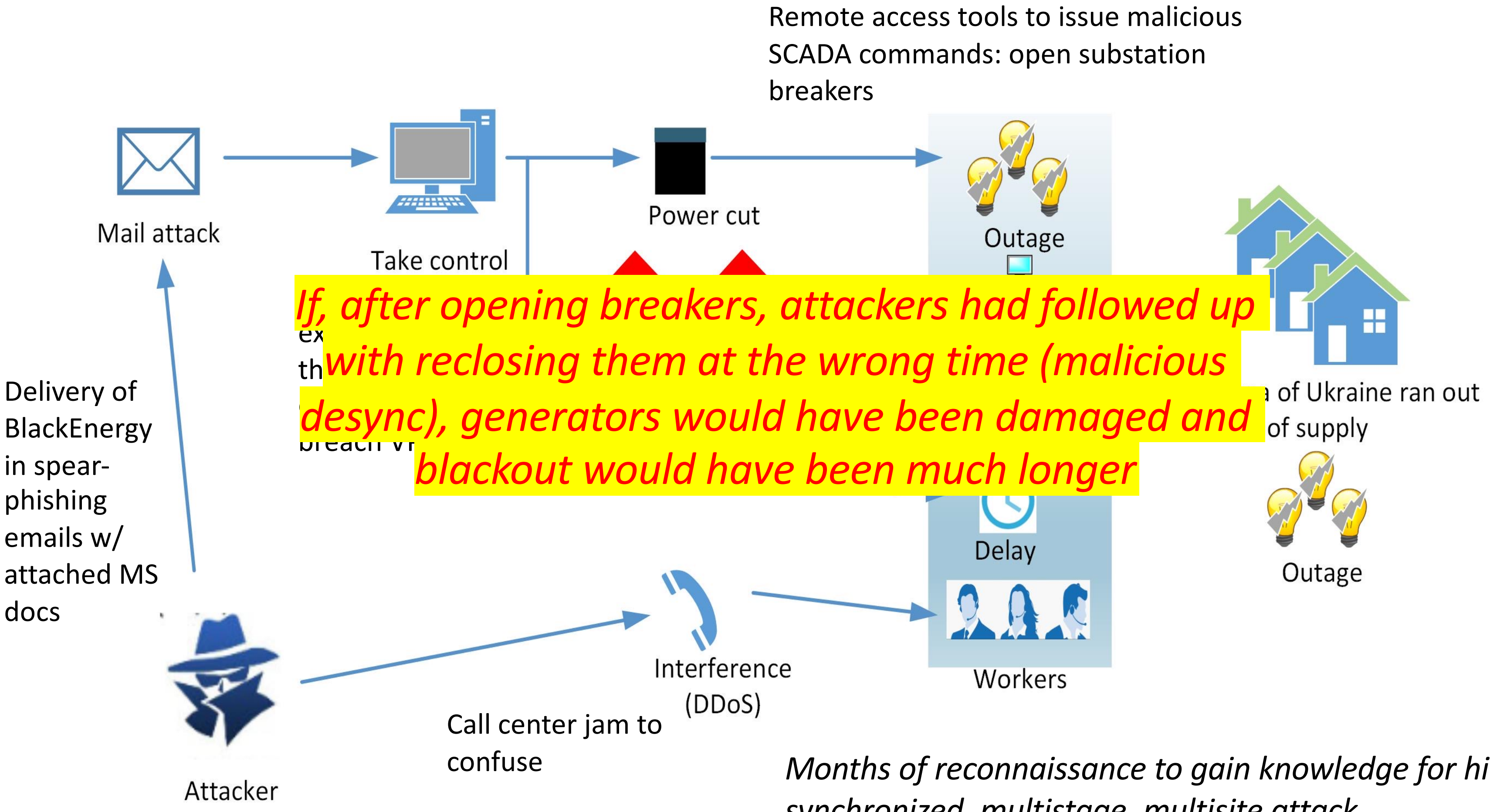
With system crash

Tie-line flows (cf. frequency) give indirect (but earlier) evidence of attacks

Machine Learning for Attack Defense

- Traditionally, OT network is airgapped; now, IT-OT convergence for business analytics, etc
- Perimeter defense (e.g., firewall, DMZ, VPN) can be breached (no lack of real-world incidents)
  - Ukraine power system attack, Colonial Pipeline ransomware attack
- Need resilience against attacks (NIST defense-in-depth)
  - Detect, classify, mitigate attacks
  - E.g., maintain *availability* during attack, forensics afterwards
- Oftentimes, lack of analytical formulas that are sufficiently accurate and complete
  - They also rely on parameters that are changing
- Machine learning provides an alternative *data-driven* approach without a priori detailed system model

The Ukraine attack



Remote access tools to issue malicious SCADA commands: open substation breakers

Mail attack

Take control

Power cut

Outage

ex

th

breach V

of Ukraine ran out of supply

Delivery of BlackEnergy in spear-phishing emails w/ attached MS docs

Delay

Outage

Interference (DDoS)

Workers

Call center jam to confuse

Attacker

If, after opening breakers, attackers had followed up with reclosing them at the wrong time (malicious desync), generators would have been damaged and blackout would have been much longer

Months of reconnaissance to gain knowledge for highly synchronized, multistage, multisite attack

ML/DL challenges

- Attacks do happen in the real world (though only high-profile cases get reported) – system traces will include them
- But hard to label massive data in practice
  - According to SANS survey, many operators suspect they were attacked but can’t tell exactly when / how
- Relative scarcity of attack data itself
  - New types of attack may emerge too (little prior knowledge about them)
- Distribution ICS spans large geographical areas
  - Vastly distributed data sources, rendering massive communications expensive or infeasible
  - Administratively separate data owners (e.g., different utility operators)

Desirable ML/DL features

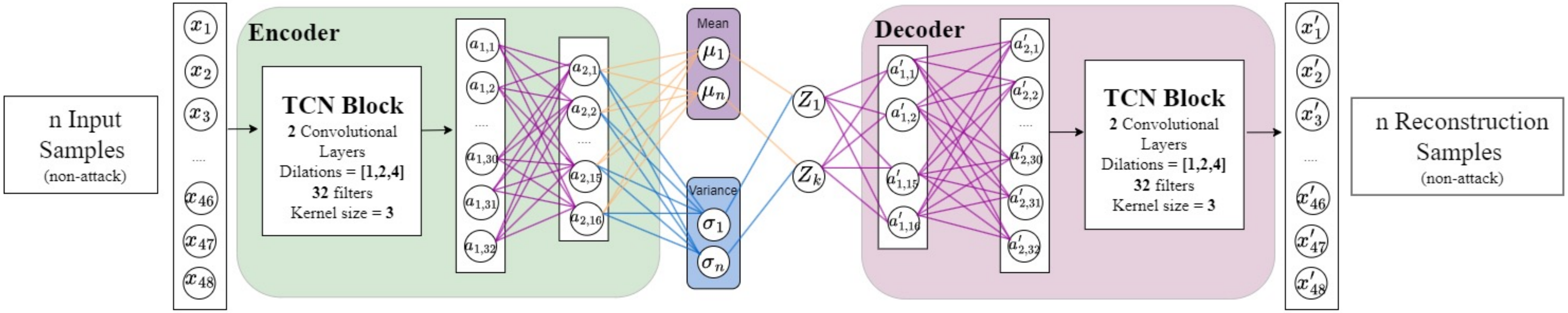
- Techniques that can unravel subtle spatial / temporal correlations in data traces
- Support for finer grained situation awareness, e.g., attack classification beyond detection
- Models trained on (mostly) normal operations
- Unsupervised (or semi-supervised) methods
- Federated learning that is communication-efficient and/or privacy preserving
  - Recent paradigm of learning a latent model of data representation, then fine tuning it for fulfilling different downstream tasks

Unsupervised attack detection and classification based on TCN-VAE …

Variational Auto-encoder (VAE)

- *Encoder* generates variants of input real data in a latent space
- *Decoder* reconstructs data, tracks RMSE of reconstructed data
- Through back propagation optimization based on decoder feedback, encoder minimizes RMSE to make generative samples realistic
- Model obtained depends on data used to train it, e.g., using normal (non-attack) data samples only
- Importance of temporal dimension of data:
  - CNN (convolutional neural network)
  - LSTM (long-short-term memory)
  - TCN (temporal convolutional network)
- Investigation of CNN-VAE *vs.* LSTM-VAE *vs.* TCN-VAE

TCN-VAE architecture



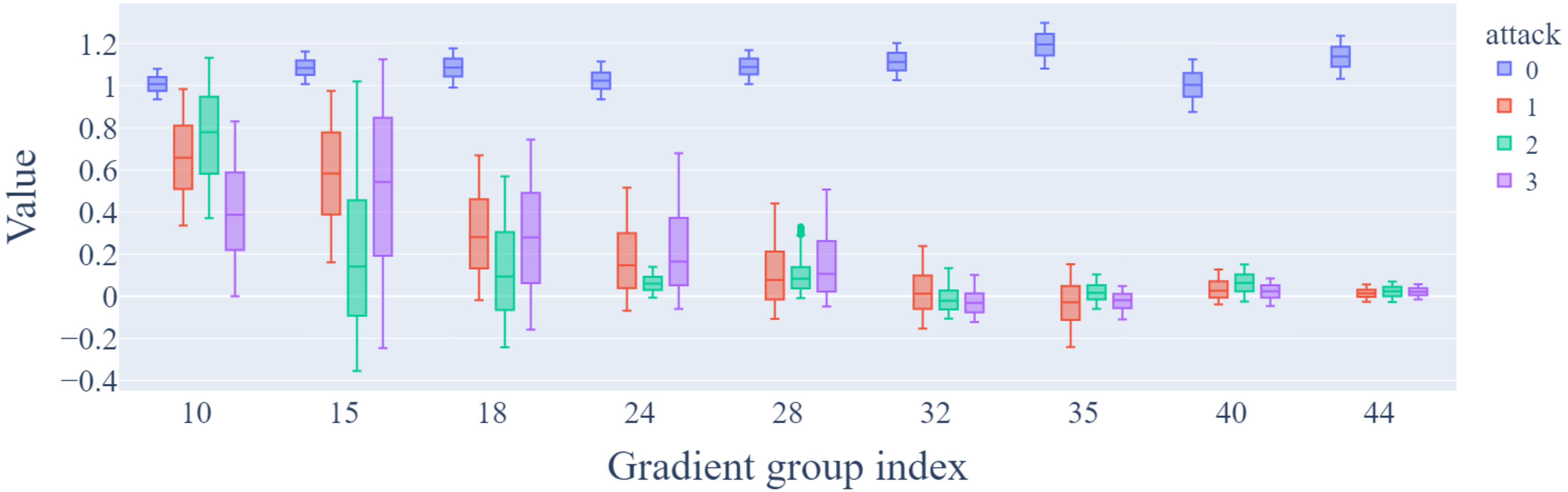
n Input Samples (non-attack)
x_1 x_2 x_3 .... x_{46} x_{47} x_{48}
Encoder
TCN Block
2 Convolutional Layers
Dilations = [1,2,4]
32 filters
Kernel size = 3
$a_{1,1}$ $a_{1,2}$ .... $a_{1,30}$ $a_{1,31}$ $a_{1,32}$
$a_{2,1}$ $a_{2,2}$ .... $a_{2,15}$ $a_{2,16}$
Mean
μ_1 μ_n
Variance
σ_1 σ_n
Z_1 Z_k
Decoder
$a'_{1,1}$ $a'_{1,2}$ $a'_{1,15}$ $a'_{1,16}$
$a'_{2,1}$ $a'_{2,2}$ .... $a'_{2,30}$ $a'_{2,31}$ $a'_{2,32}$
TCN Block
2 Convolutional Layers
Dilations = [1,2,4]
32 filters
Kernel size = 3
x'_1 x'_2 x'_3 .... x'_{46} x'_{47} x'_{48}
n Reconstruction Samples (non-attack)

- Tracks statistics (mean and variance) in latent space
- 48 features as shown

Data sets and training

- Datasets from industry-strength PowerWorld simulator for electrical transmission
  - Transient behaviors in addition to steady state
- Varied loadings subject to short-term randomness
- Normal operation, or under TDA, FDI-optimal, FDI-adaptive attack
- Varied strengths of attack
  - Negligible (not so important), weak (but eventually damaging), moderate, strong
- VAE model trained from normal operation only
  - Attack detected if RMSE deviation from the normal exceeds a (tunable) threshold

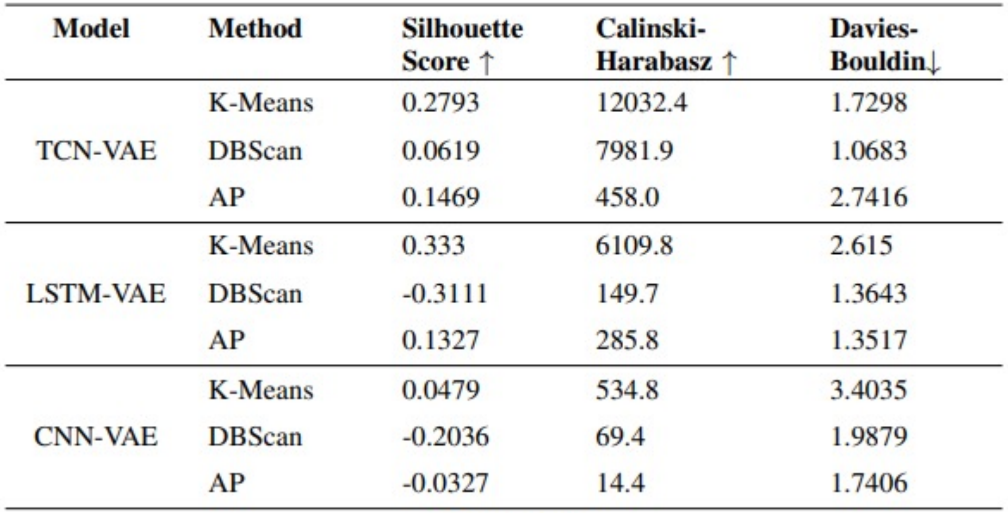
Beyond detection: classification by gradient profiles



- TCN-VAE produces gradients during back-propagation optimization process
- These gradients form a profile (across features in data trace)
- Different classes of attacks (including no attack) can be identified by their gradient profiles

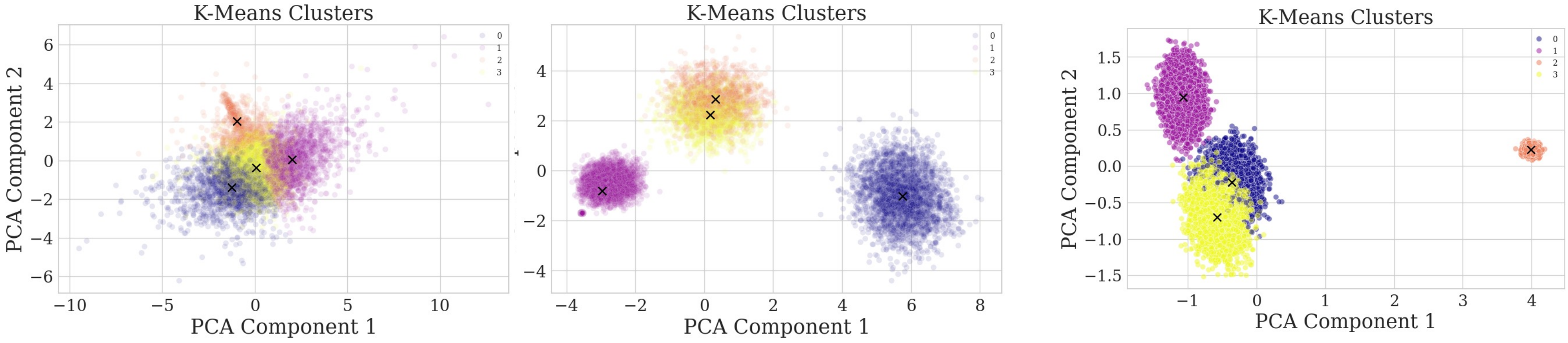
Clustering of gradient profiles

- K-means
  - Based on (multi-dimensional) data distance
- DBScan
  - Based on data density
- Affinity propagation
  - Based on data similarity
- Various metrics of how well the profiles cluster



Model	Method	Silhouette Score ↑	Calinski-Harabasz ↑	Davies-Bouldin↓
TCN-VAE	K-Means	0.2793	12032.4	1.7298
	DBScan	0.0619	7981.9	1.0683
	AP	0.1469	458.0	2.7416
LSTM-VAE	K-Means	0.333	6109.8	2.615
	DBScan	-0.3111	149.7	1.3643
	AP	0.1327	285.8	1.3517
CNN-VAE	K-Means	0.0479	534.8	3.4035
	DBScan	-0.2036	69.4	1.9879
	AP	-0.0327	14.4	1.7406

TABLE I
UNSUPERVISED CLASSIFICATION RESULTS.



CNN-VAE

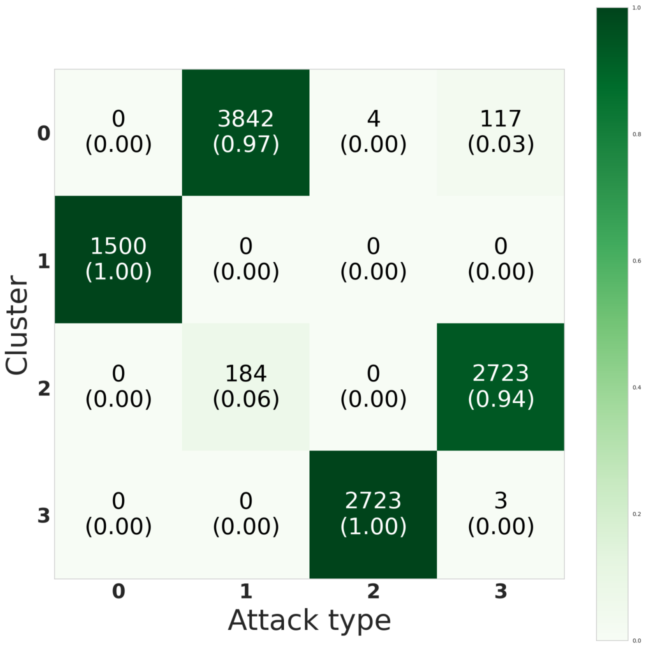
LSTM-VAE

TCN-VAE

2D visualization of K-means clusters

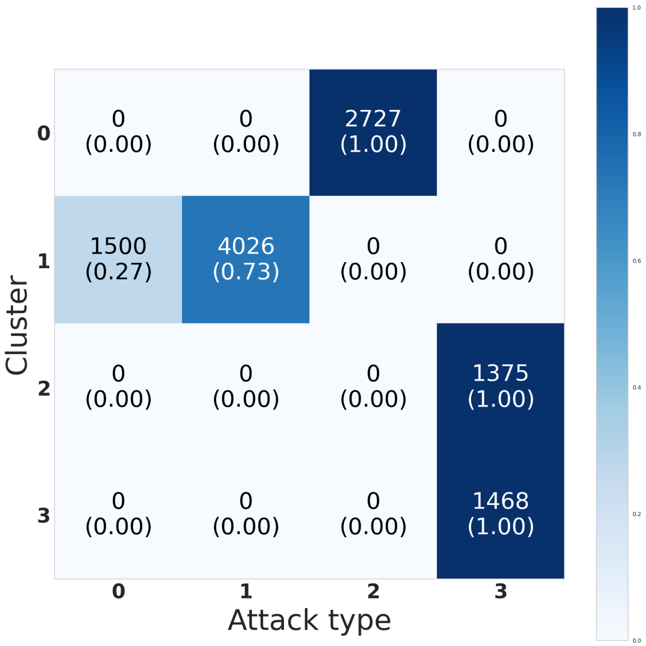
- View of two PCA components
- Results depend on VAE variant, because their back-propagation optimization produces the gradients being clustered
- However, well clustered profiles don’t necessarily agree better with groundtruths (what really matters)

Classification performance



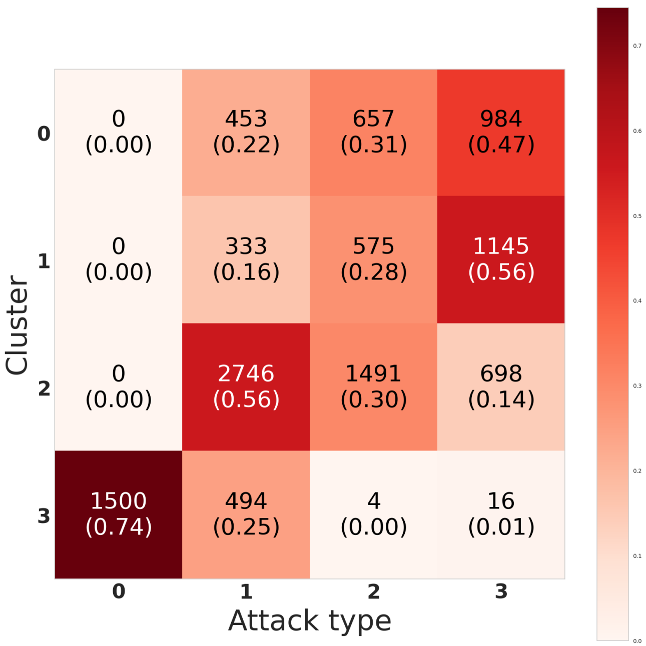
Cluster	0	1	2	3
0	0 (0.00)	3842 (0.97)	4 (0.00)	117 (0.03)
1	1500 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)
2	0 (0.00)	184 (0.06)	0 (0.00)	2723 (0.94)
3	0 (0.00)	0 (0.00)	2723 (1.00)	3 (0.00)

a. TCN-VAE with K-Means



Cluster	0	1	2	3
0	0 (0.00)	0 (0.00)	2727 (1.00)	0 (0.00)
1	1500 (0.27)	4026 (0.73)	0 (0.00)	0 (0.00)
2	0 (0.00)	0 (0.00)	0 (0.00)	1375 (1.00)
3	0 (0.00)	0 (0.00)	0 (0.00)	1468 (1.00)

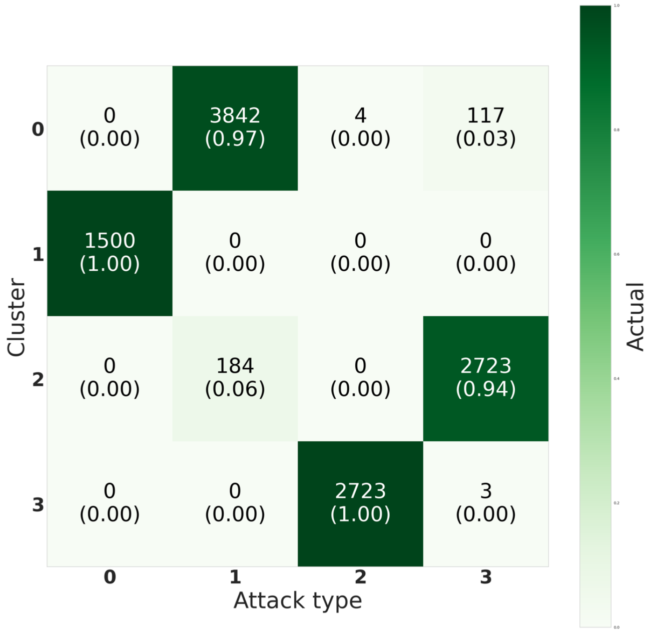
b. LSTM-VAE with K-Means



Cluster	0	1	2	3
0	0 (0.00)	453 (0.22)	657 (0.31)	984 (0.47)
1	0 (0.00)	333 (0.16)	575 (0.28)	1145 (0.56)
2	0 (0.00)	2746 (0.56)	1491 (0.30)	698 (0.14)
3	1500 (0.74)	494 (0.25)	4 (0.00)	16 (0.01)

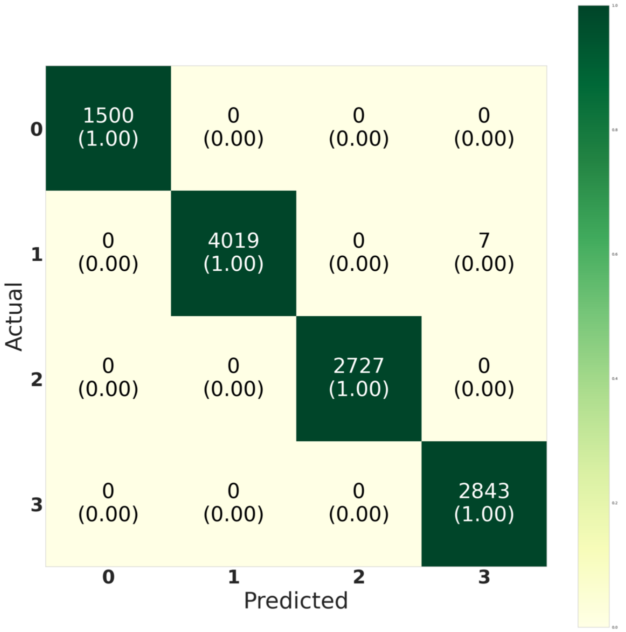
c. CNN-VAE with K-Means

Comparison w/ supervised ML (XGboost)



Cluster \ Attack type	0	1	2	3
0	0 (0.00)	3842 (0.97)	4 (0.00)	117 (0.03)
1	1500 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)
2	0 (0.00)	184 (0.06)	0 (0.00)	2723 (0.94)
3	0 (0.00)	0 (0.00)	2723 (1.00)	3 (0.00)

a. K-Means (TCN-VAE)
Overall Accuracy ~ 97%

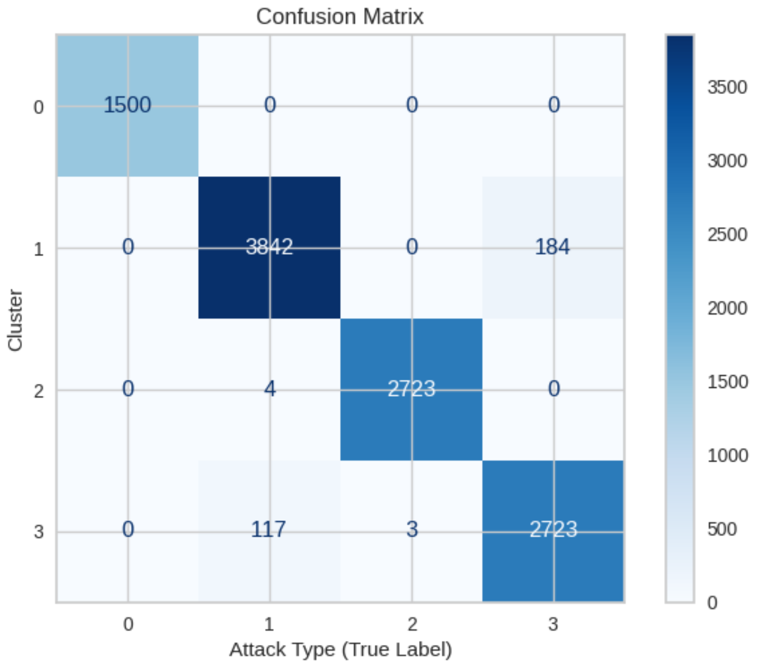


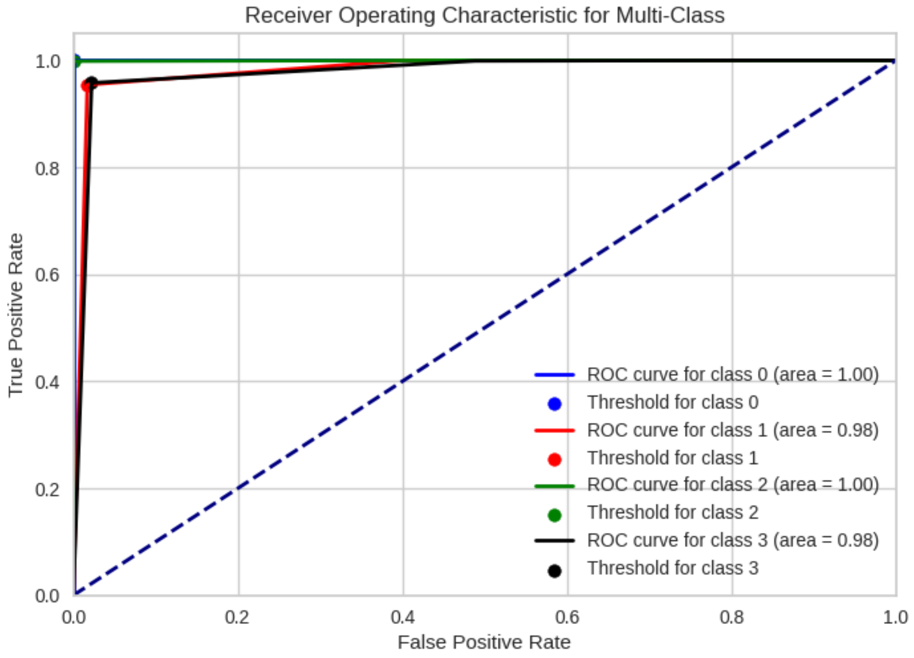
Actual \ Predicted	0	1	2	3
0	1500 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)
1	0 (0.00)	4019 (1.00)	0 (0.00)	7 (0.00)
2	0 (0.00)	0 (0.00)	2727 (1.00)	0 (0.00)
3	0 (0.00)	0 (0.00)	0 (0.00)	2843 (1.00)

b. XGboost
Overall Accuracy > 99%

XGboost has the best performance among several supervised ML alternatives, including SVM and AdaBoost

Confusion matrix and ROC



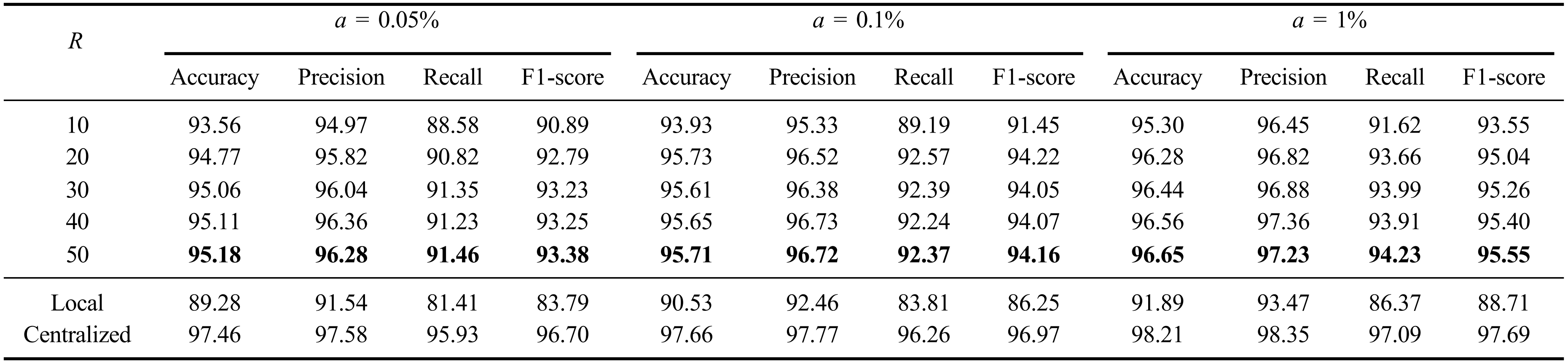


Federated contrastive learning for detecting stealthy attacks with unlabeled data …*

Problem setup

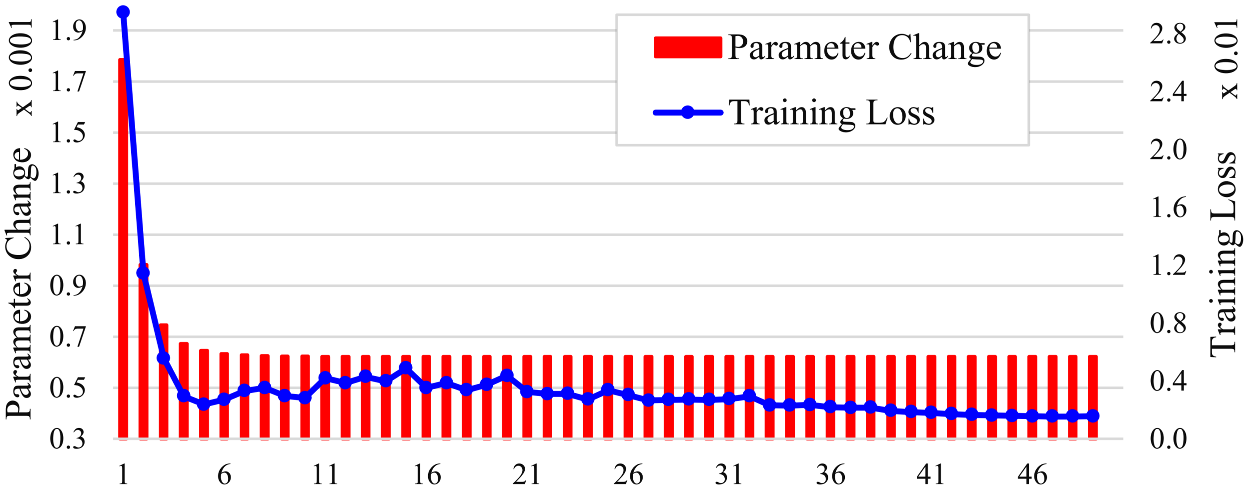
- Detection of BDD-bypassing stealthy FDI attacks
- Data sources are geographically distributed (bandwidth concerns)
- Data owners are administratively separate (privacy concerns)
- Challenge: effective global learning without sharing massive (non-iid) local raw data
- Solution: Federated learning by FedCLD
    - Global control center and local control centers collaborate to learn a latent representation of (mostly) unlabeled grid data, through updates of model parameters only
    - Using learned latent representation, local center runs an online binary classifier to perform downstream task of attack detection

Comparison w/ local and centralized learning

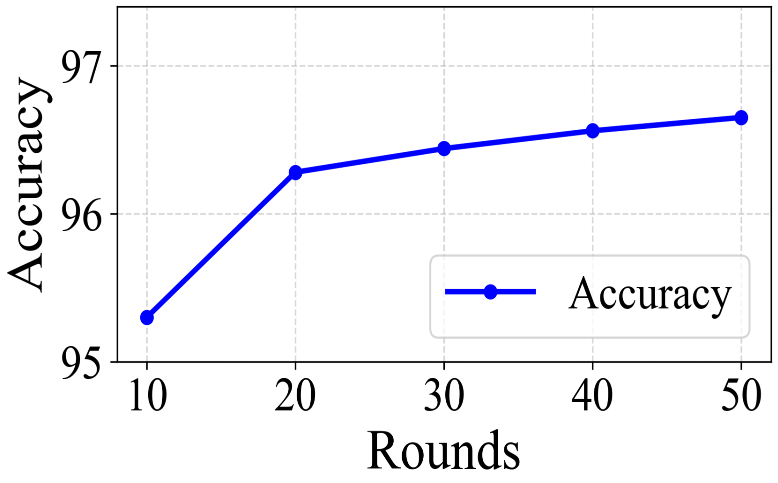


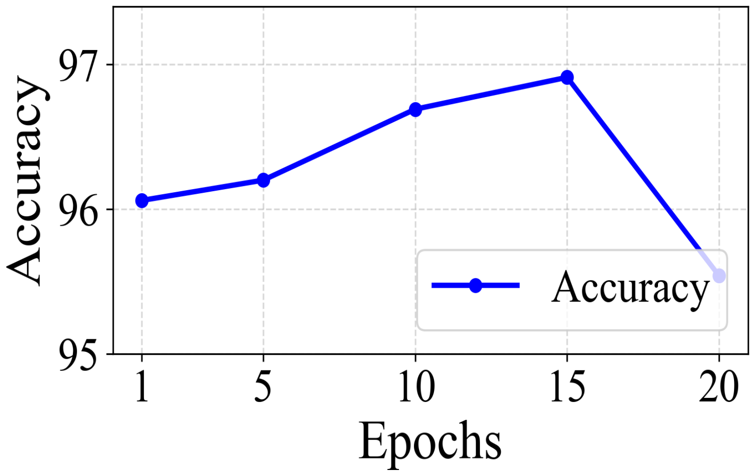
R	a = 0.05%				a = 0.1%				a = 1%			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
10	93.56	94.97	88.58	90.89	93.93	95.33	89.19	91.45	95.30	96.45	91.62	93.55
20	94.77	95.82	90.82	92.79	95.73	96.52	92.57	94.22	96.28	96.82	93.66	95.04
30	95.06	96.04	91.35	93.23	95.61	96.38	92.39	94.05	96.44	96.88	93.99	95.26
40	95.11	96.36	91.23	93.25	95.65	96.73	92.24	94.07	96.56	97.36	93.91	95.40
50	**95.18**	**96.28**	**91.46**	**93.38**	**95.71**	**96.72**	**92.37**	**94.16**	**96.65**	**97.23**	**94.23**	**95.55**
Local	89.28	91.54	81.41	83.79	90.53	92.46	83.81	86.25	91.89	93.47	86.37	88.71
Centralized	97.46	97.58	95.93	96.70	97.66	97.77	96.26	96.97	98.21	98.35	97.09	97.69

Convergence of FedCLD



Impacts of learning rounds and communication frequency





Conclusion

- Next-generation cyber-physical systems (e.g., smart power grids) susceptible to cyberattacks due to reliance on ICT
- Machine learning can be useful for defenses (e.g., attack detection and classification) without good enough analytical models
- Addressed several key challenges in the ML
  - Lack of data labels
  - Lack of attack data
  - Distributed locations of data sources
  - Different ownerships of local data

Acknowledgments

- Singapore National Research Foundation

NRF
SINGAPORE

- Singapore Energy Market Authority

ENERGY
MARKET
AUTHORITY
Our Clean Energy Future

- Guizhou University, China