NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Data-Driven Power Systems Stability Assessment under Adversarial Examples —*Vulnerability Analysis, Robustness Verification and Mitigation*

Dr Yan Xu
Cham Tao Soon Professor in Engineering
Director, Center for Power Engineering
Cluster Director, Energy Research Institute
Nanyang Technological University

OUTLINE

1 **Background**
- What is done and not done?

2 **Problem Description**
- What is the adversarial example?

3 **Proposed Method**
- Adversarial example generation
- Robustness evaluation indices
- Mitigation strategy

4 **Case study**

5 **Conclusions**

Background

Problem Description

Methodology

Adversarial example generation

Robustness evaluation indices

Mitigation strategy

Case Study

Conclusion

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Power System Stability

Definition

"Power system stability is the ability of an electric power system, for a given initial operating condition, to regain a state of operating equilibrium after being subjected to a physical disturbance, with most system variables bounded so that practically the entire system remains intact."

Conventional power grid → "Smart Grid"

- **Generation side**: high-level intermittent renewable energy integration
- **Demand side:** demand response, electric vehicle, distributed energy storage, etc.
- **Device-grid interface:** power-electronics converters

Higher operating uncertainties + Complicated system dynamics

Recent major blackout events

North America Blackout

India Blackout

… 2006 … … 2016 … ?

2003 2012

Very high wind power penetration level (48%)

West Europe Blackout

South Australia Blackout

Classification for Power System Stability

- Rotor Angle Stability (large-disturbance and small-disturbance)
- Voltage Stability (short-term or long-term)
- Frequency Stability (short-term and long-term)

\+

- Resonance stability (electrical and torsional)
- Converter-driven stability (fast and slow interaction)

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{y}, \mathbf{p}, \boldsymbol{\lambda}) \qquad 0 = \mathbf{g}(\mathbf{x}, \mathbf{y}, \mathbf{p}, \boldsymbol{\lambda})$$

Classification for Stability Assessment and Control

On-line Stability Assessment → Preventive Control | Real-time Stability Assessment → Emergency Control

time

Contingency

Steady State (pre-fault)	Dynamic State (post-fault)
↓	↓
Accuracy, Speed, Knowledge	**Accuracy, Earliness, Reliability**
↓	↓
the faster, the more faults can be assessed	**the earlier, the more time is left for control**

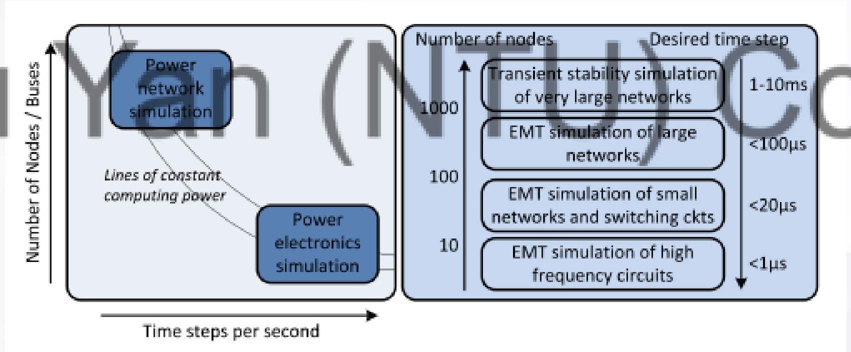
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

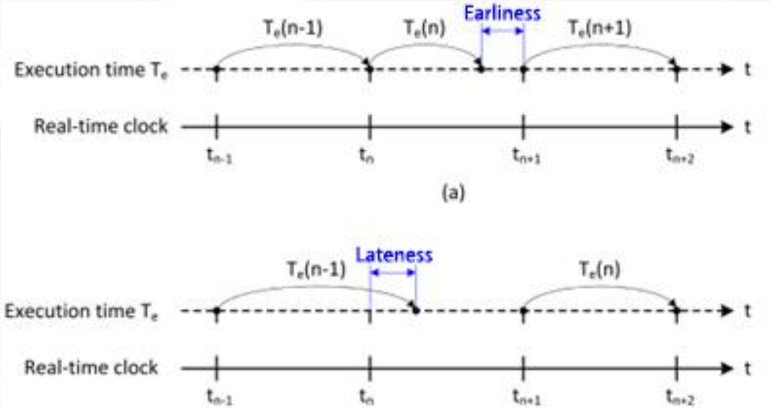
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

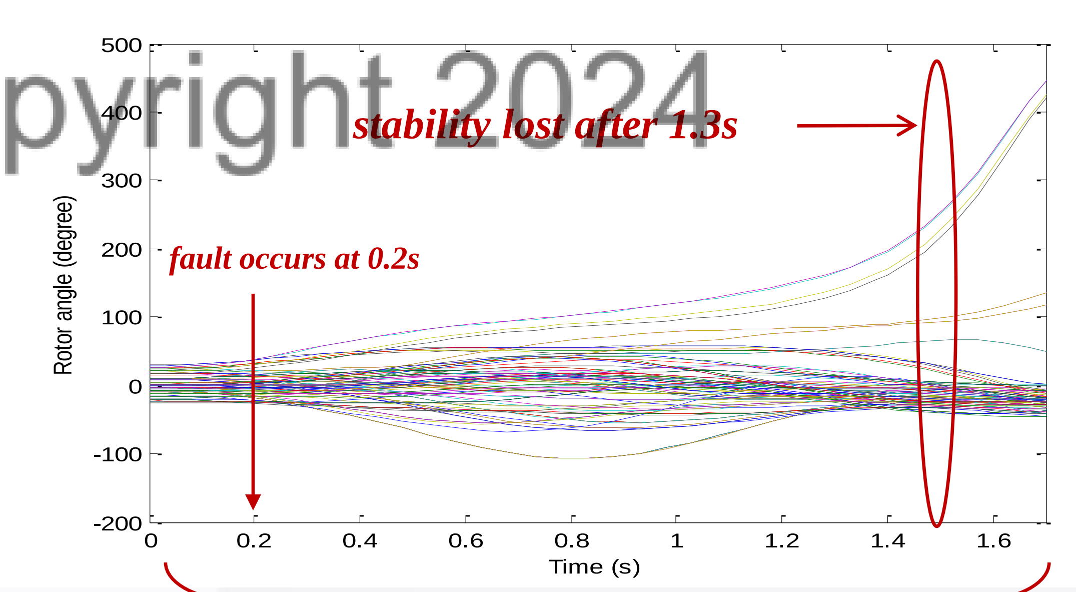
Conventional Methods (Model-based)

- **Time-domain Simulation:** to solve a large-scale differential-algebraic equation (DAE) set
- **Data requirement**: system model (static and dynamic), network topology, state-estimation, fault, etc.
- **Outputs**: system's time-varying trajectories
- **Event-based control:** lookup decision table, contingency indexing

"for a 14,000-bus system, one disturbance analysis could involve a set of 15,000 differential equations and 40,000 nonlinear algebraic equations for a simulation time duration of 10-20s; besides, the number of disturbances to be considered is also enormous, e.g., for the 14,000-bus system, the typical number of postulated disturbances is between 2000 and 3000."

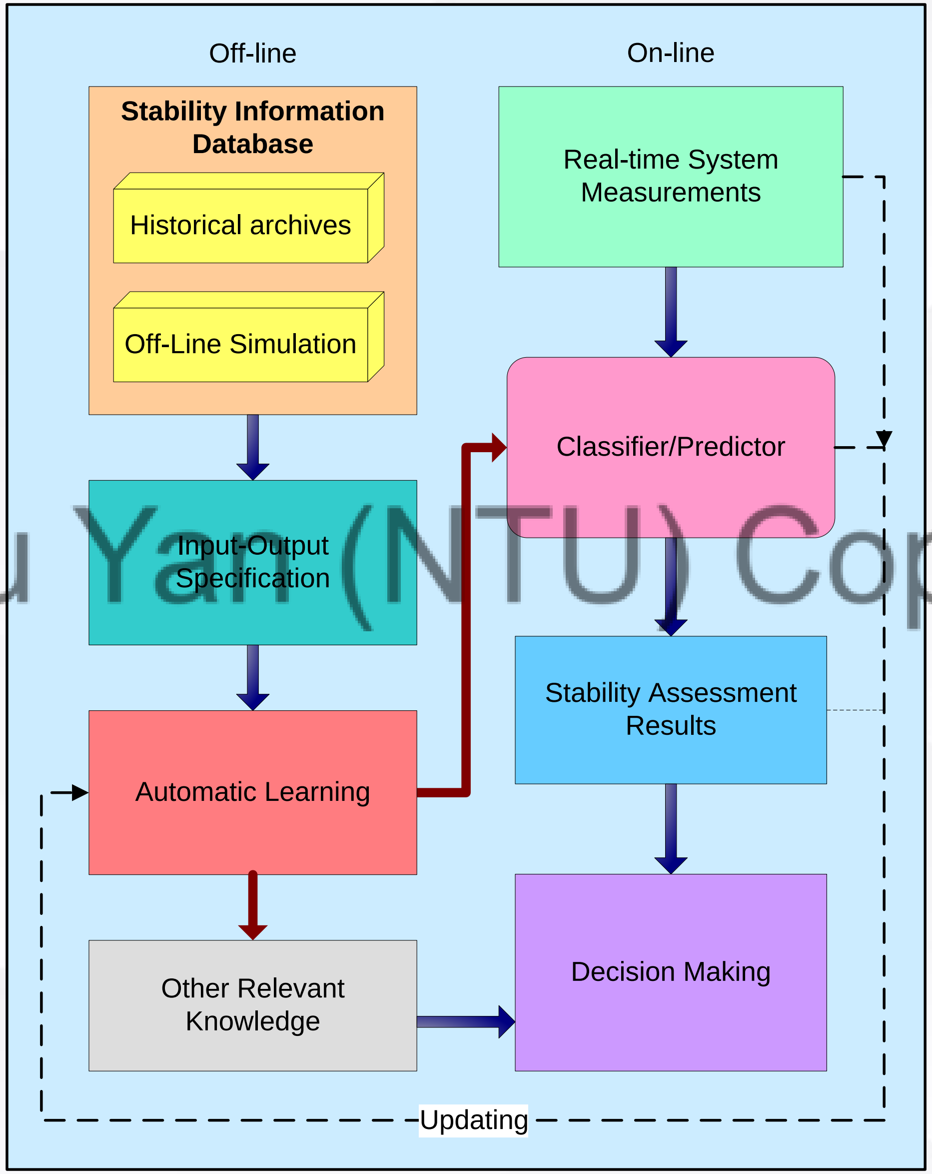


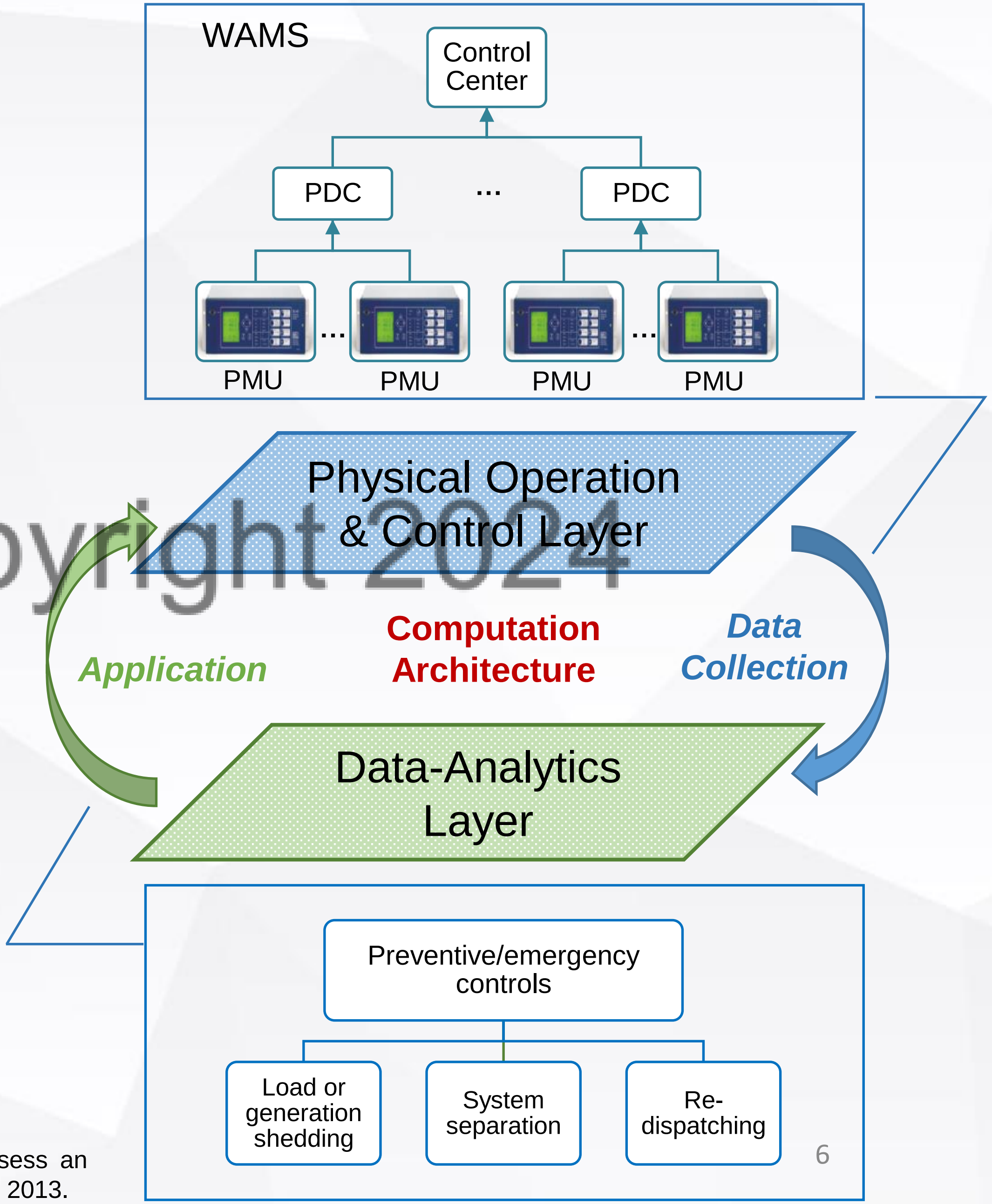




PSS/E simulation costs 2.2s CPU time

Data-Driven Stability Assessment Strategy





Z.Y. Dong, Y. Xu, P. Zhang, and K.P. Wong "Using intelligent system to assess an electric power system's real-time stability," *IEEE Intelligent Systems Magazine*, 2013.

Background

Problem Description

Methodology

Adversarial example generation

Robustness evaluation indices

Mitigation strategy

Case Study

Conclusion

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Key Research Problems: that have been solved by us

1. Generate a comprehensive stability **database**
2. Select/extract **significant features**
3. Evaluate the **credibility** of the model output
4. Improve & tradeoff the **accuracy and speed**
5. Extract **interpretable knowledge** for stability control
6. **Update** the model timely and effectively
7. Mitigate abnormal measurements, such as **missing data, communication delay**
8. Adapt the trained model to unforeseen scenarios, e.g., **unexpected fault, different topologies.**

1. **Y. Xu**, Z.Y. Dong, K. Meng, R. Zhang and K.P. Wong, "Real-time transient stability assessment model using extreme learning machine," *IET Gen. Trans. & Dist*., 2011.
2. **Y. Xu**, Z.Y. Dong, J.H. Zhao, P. Zhang, and K.P. Wong, "A reliable intelligent system for real-time dynamic security assessment of power systems," *IEEE Trans. Power Systems*, 2012.
3. **Y. Xu**, Z.Y. Dong, Z. Xu, K. Meng, and K.P. Wong, "An intelligent dynamic security assessment framework for power systems with wind power," *IEEE Trans. Industrial Informatics*, 2012.
4. R. Zhang, **Y. Xu***, Z.Y. Dong, and K.P. Wong, "Post-disturbance transient stability assessment of power systems by a self-adaptive intelligent system," *IET Gen. Trans. & Dist*., 2015.
5. **Y. Xu**, R. Zhang, J. Zhao, et al, "Assessing short-term voltage stability of electric power systems by a hierarchical intelligent system," *IEEE Trans. Neural Networks and Learning Systems*, 2016.
6. Y. Zhang, **Y. Xu***, Z.Y. Dong, et al, "Intelligent early-warning of power system dynamic insecurity risk towards optimal accuracy-efficiency trade-off," *IEEE Trans. Industrial Informatics*, 2017.
7. Y. Zhang, **Y. Xu***, and Z.Y. Dong. "Robust ensemble data-analytics for incomplete PMU measurement-based power system stability assessment," *IEEE Trans. Power Systems*., 2018.
8. Y. Zhang, **Y. Xu***, Z.Y. Dong, et al, "A Hierarchical Self-Adaptive Data-Analytics Method for Power System Short-term Voltage Stability Assessment," *IEEE Trans. Industrial Informatics*, 2019.

Background

Problem Description

Methodology

Adversarial example generation

Robustness evaluation indices

Mitigation strategy

Case Study

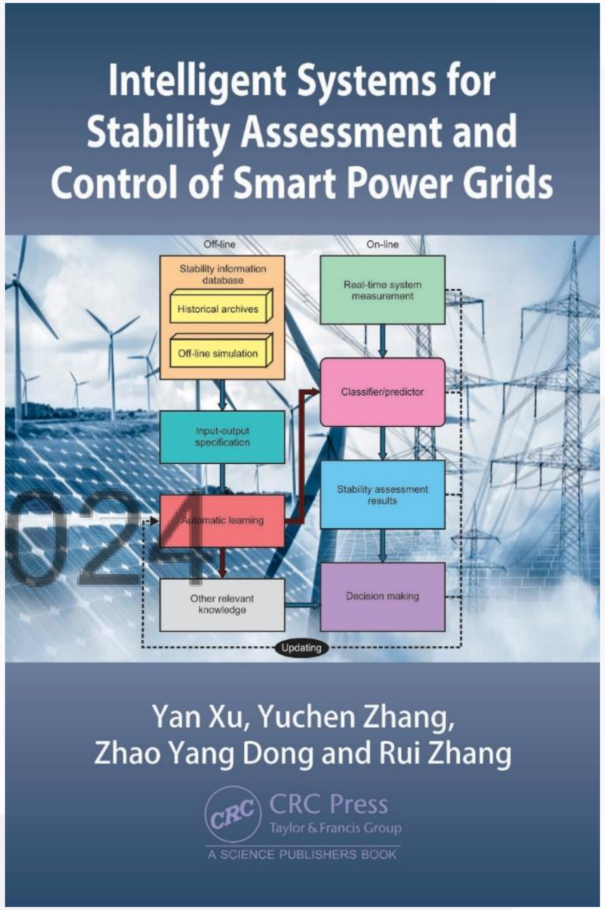
Conclusion

NANYANG TECHNOLOGICAL UNIVERSITY

SINGAPORE

Key Research Problems: that have been solved by us

9. Y. Zhang, **Y. Xu***, Z.Y. Dong, and P. Zhang, “Real-Time Assessment of Fault-Induced Delayed Voltage Recovery: A Probabilistic Self-Adaptive Data-driven Method,” *IEEE Trans. Smart Grid*, 2019.
10. Y. Zhang, **Y. Xu***, Z.Y. Dong, and R. Zhang, “A Missing-Data Tolerant Method for Data-Driven Short-Term Voltage Stability Assessment of Power Systems,” *IEEE Trans. Smart Grid*, 2019.
11. C. Ren and **Y. Xu***, “A Fully Data-Driven Method based on Generative Adversarial Networks for Power System Dynamic Security Assessment with Missing Data,” *IEEE Trans. Power Systems*., 2019.
12. C. Ren and **Y. Xu***, “Transfer Learning-based Power System Online Dynamic Security Assessment: Using One Model to Assess Many Unlearned Faults,” *IEEE Trans. Power Systems*., 2020.
13. C. Ren, **Y. Xu***, Y. Zhang, and R. Zhang, “A Hybrid Randomized Learning System for Temporal-Adaptive Voltage Stability Assessment of Power Systems,” *IEEE Trans. Industrial Informatics*, 2020.
14. C. Ren and **Y. Xu***, “Incremental Broad Learning for Real-Time Updating of Data-Driven Power System Dynamic Security Assessment Models,” *IET Gen. Trans. & Dist*., 2020.
15. C. Ren, **Y. Xu*,** and R. Zhang, “An Interpretable Deep Learning Method for Power System Dynamic Security Assessment via Tree Regularization,” *IEEE Trans. Power Syst*., 2021.
16. C. Ren, **Y. Xu*,** B. Dai, and R. Zhang, “An Integrated Transfer Learning Method for Power System Dynamic Security Assessment for Unlearned Faults with Missing Data,” *IEEE Trans. Power Syst*., 2021.
17. C. Ren, **Y. Xu*** “An Interpretable Deep Learning Method for Power System Dynamic Security Assessment via Tree Regularization,” *IEEE Trans. Power Syst.*, 2022.



Y. Xu, Y. Zhang, Z.Y. Dong, and R. Zhang, "*Intelligent Systems for Stability Assessment and Control of Smart Power Grids*," CRC Press, 2020.

ISBN-13: 978-1138063488

Background

Problem Description

Methodology

Adversarial example generation

Robustness evaluation indices

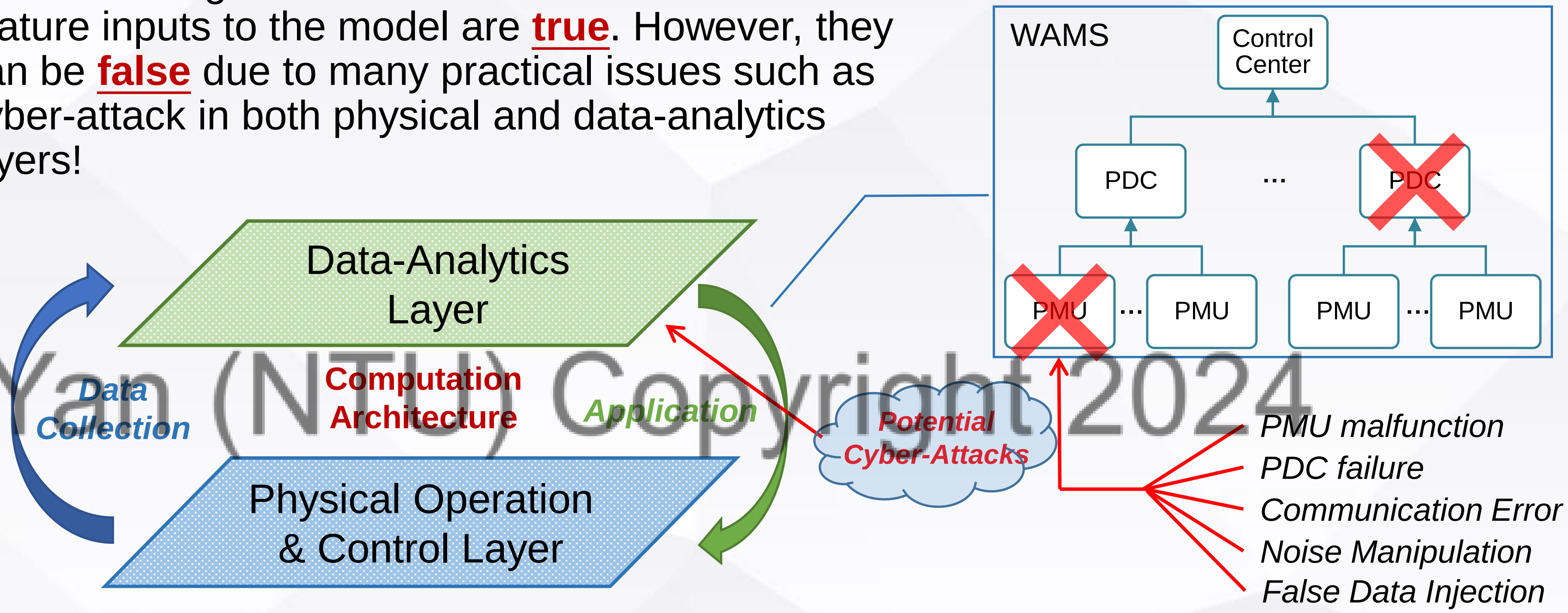
Mitigation strategy

Case Study

Conclusion

Problem Description: Adversarial Examples

All the existing works assume that the values of the feature inputs to the model are **true**. However, they can be **false** due to many practical issues such as cyber-attack in both physical and data-analytics layers!



Adversarial example: *a modified version of the original sample that is intentionally perturbed but retains very close to the original one*. It aims to generate a wrong output.

Mathematical description

$$\min_{\mathbf{x}^{adv}} \left\|\mathbf{x}^{adv} - \mathbf{x}\right\|_p$$

$$s.t. \begin{cases} f_\theta(\mathbf{x}) = y \\ f_\theta\left(\mathbf{x}^{adv}\right) = y' \neq y \end{cases}$$

Background

Problem Description

Methodology

Adversarial example generation

Robustness evaluation indices

Mitigation strategy

Case Study

Conclusion

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Problem Description: Illustration of Adversarial Examples

$$\mathbf{x}^{adv} = \mathbf{x} + \boldsymbol{\varepsilon}_{\mathbf{x}}$$

Small perturbations of the input cause the sound wave and the image to be misclassified.

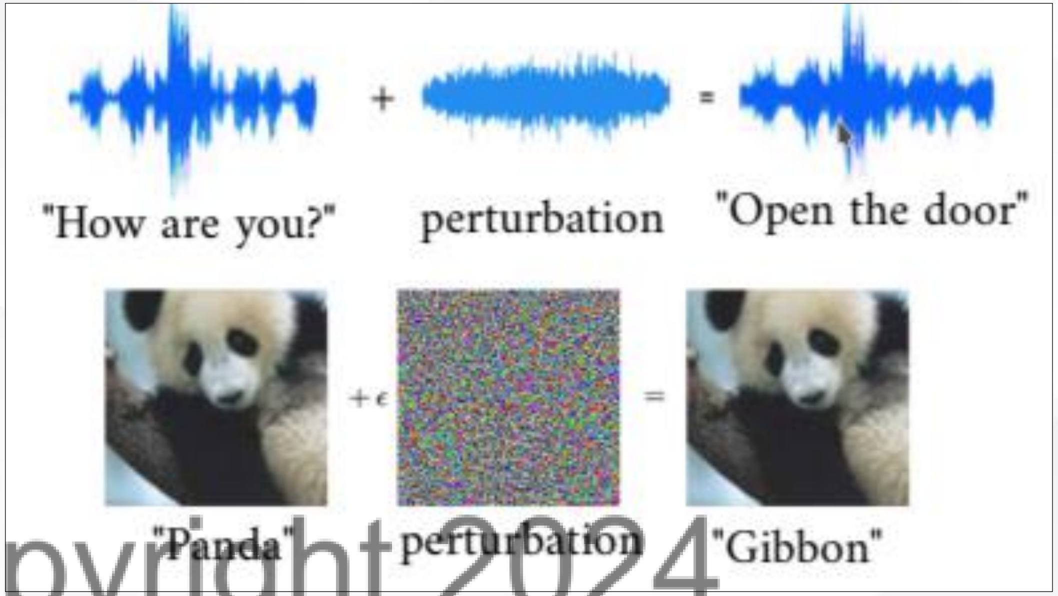
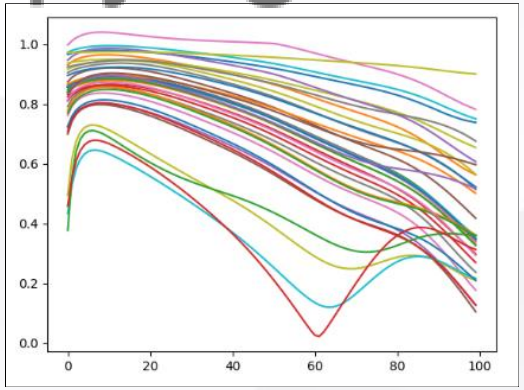
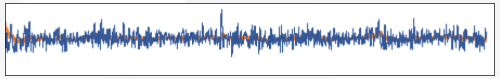


Figure from "G. Anderson, et al. Optimization and Abstraction: A Synergistic Approach for Analyzing Neural Network Robustness. Proc. 40th ACM SIGPLAN (PLDI '19)."

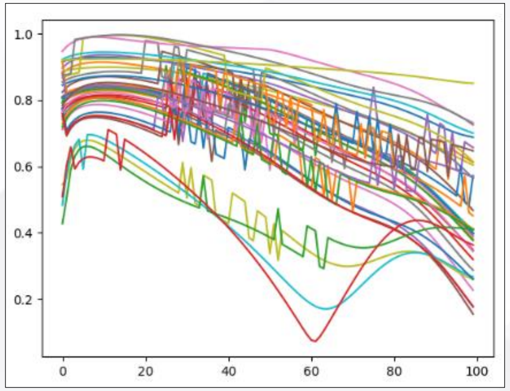
For data-driven stability assessment, a small perturbation to the feature input value that can lead to a different (wrong) stability assessment result.



Unstable voltage trajectories



Adversarial Perturbations



The unstable case is classified as stable, which is wrong.

So, high accuracy is not equal to high robustness under adversarial examples !!!

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Adversarial Example Generation: Fast Calculation Method

Mathematical Model

$$\min_{\mathbf{x}^{adv}} \left\|\mathbf{x}^{adv} - \mathbf{x}\right\|_p$$

$$s.t. \quad \begin{cases} f_\theta(\mathbf{x}) = y \\ f_\theta(\mathbf{x}^{adv}) = y' \neq y \end{cases}$$

y and y' denote the corresponding output label of $\mathbf{x}$ and $\mathbf{x}^{adv}$, respectively; $\|\cdot\|_p$ denotes the distance between $\mathbf{x}$ and $\mathbf{x}^{adv}$, and p measures the magnitude of adversarial perturbation by p-norm distance.

An approximate solution

Fast Gradient Sign Method (FGSM)

$$\mathbf{x}^{adv} = \mathbf{x} + \delta \cdot sign\left(\nabla_{\mathbf{x}} L\left(f_\theta(\mathbf{x}), y\right)\right)$$

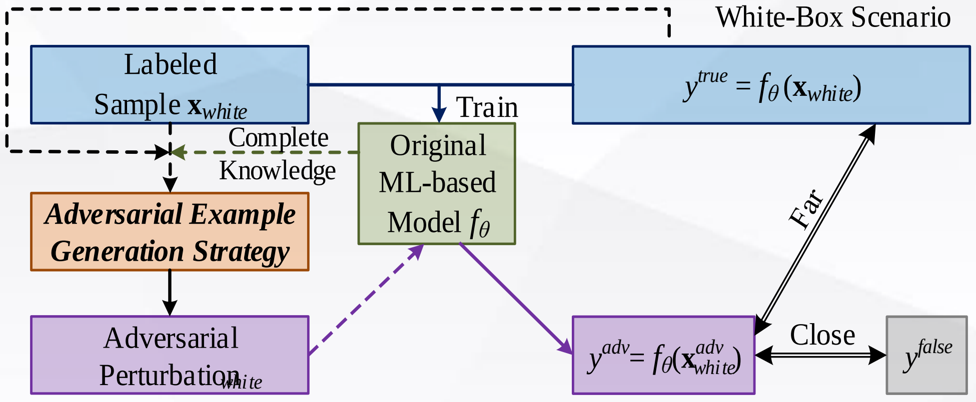
Multi-Step FGSM

$$\mathbf{x}_{i+1}^{adv} = \mathbf{x}_i^{adv} + (\delta/i) \cdot sign\left(\nabla_{\mathbf{x}} L\left(f_\theta\left(\mathbf{x}_i^{adv}\right), y\right)\right)$$

$sign(\cdot)$ is the sign function; δ specifies the boundary of perturbation; $L(\cdot,\cdot)$ represents the loss function of model $f_\theta(\cdot)$.

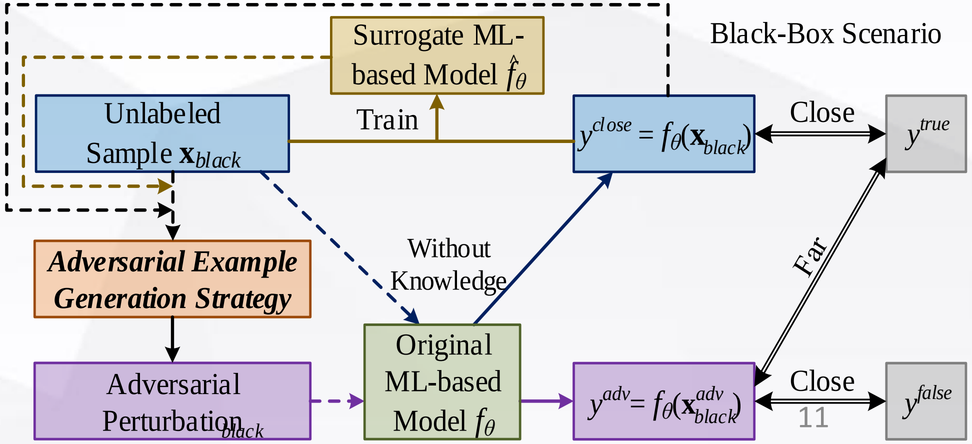
White-box scenario

Complete knowledge of the original trained model (e.g., original training database, training algorithm, model structure, parameter settings, etc.) is available to cyber attackers.



Black-box scenario

None or limited knowledge of the original trained model is available to cyber attackers



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Robustness Evaluation: Principle

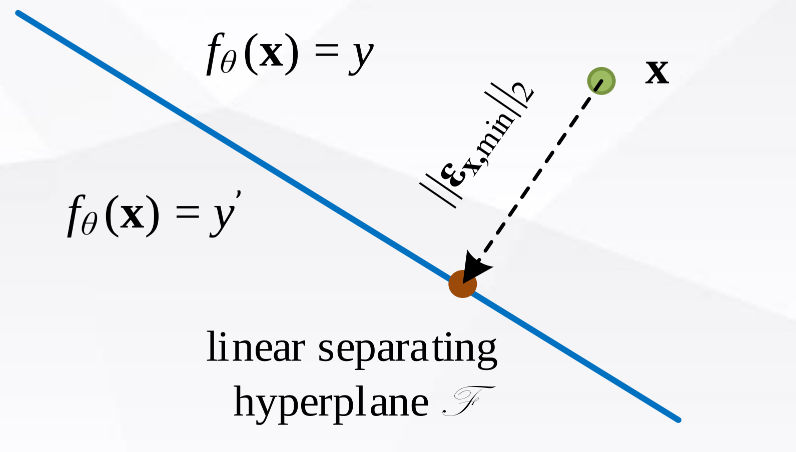
The **robustness** can be evaluated by the average minimal adversarial perturbation for a successful adversarial attack.

$$\min_{\boldsymbol{\varepsilon}_{\mathbf{x}}} \|\boldsymbol{\varepsilon}_{\mathbf{x}}\|_p$$

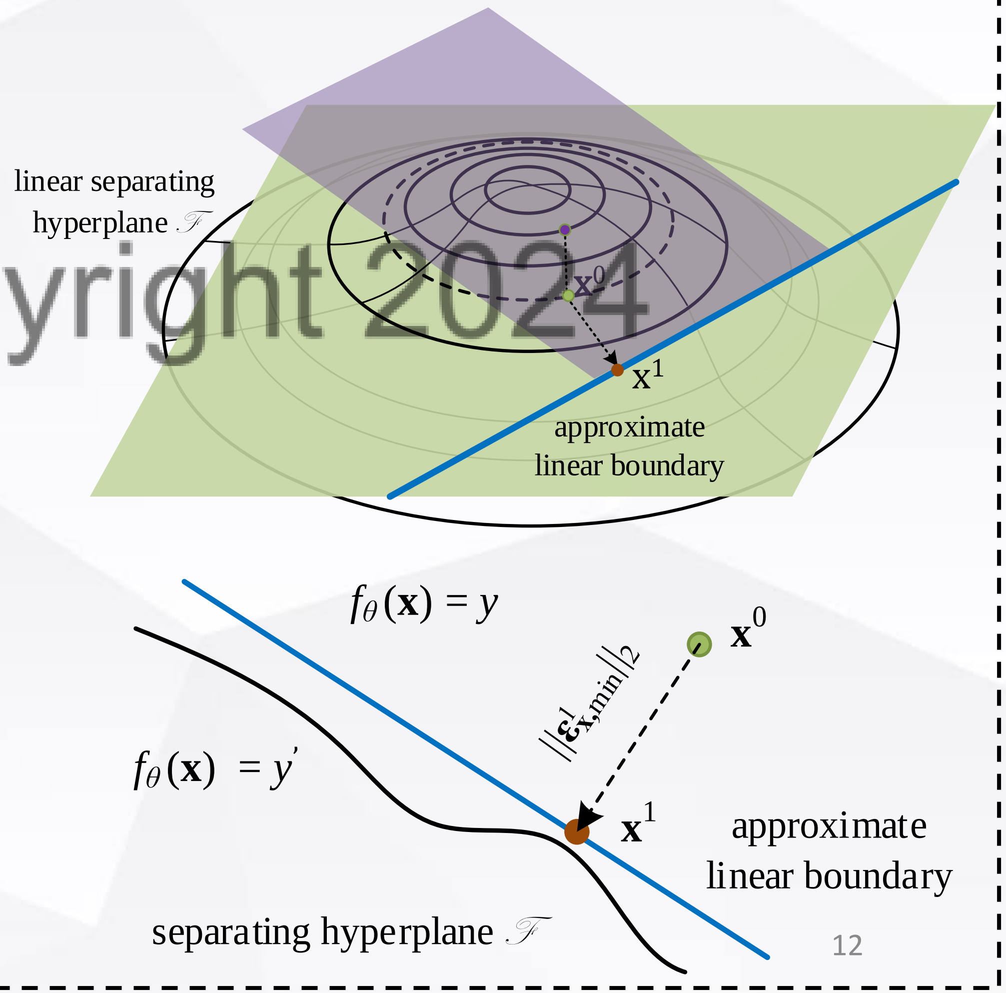
$$s.t. \quad \begin{cases} f_\theta(\mathbf{x}) = y \\ f_\theta(\mathbf{x}) \neq f_\theta(\mathbf{x} + \boldsymbol{\varepsilon}_{\mathbf{x}}) \end{cases}$$

$\boldsymbol{\varepsilon}_{\mathbf{x},\mathbf{min}} = \arg\min_{\boldsymbol{\varepsilon}_{\mathbf{x}}} \|\boldsymbol{\varepsilon}_{\mathbf{x}}\|_2$ represents the minimal adversarial perturbation of the original sample under the classifier $f_\theta(\cdot)$. For the L2-norm distance, $\|\boldsymbol{\varepsilon}_{\mathbf{x},\mathbf{min}}\|_2$ represents the minimal distance from the original sample **x** to the classification boundary.

Adversarial perturbation for a linear binary classifier



Adversarial perturbation for a nonlinear binary classifier



Robustness Evaluation: Proposed Indices

- Adversarial Perturbations for Linear Binary Classifiers

$$\boldsymbol{\varepsilon}_{\mathbf{x},\min} = \arg\min_{\boldsymbol{\varepsilon}_{\mathbf{x}}} \|\boldsymbol{\varepsilon}_{\mathbf{x}}\|_2 = -\frac{f_\theta(\mathbf{x})}{\|\theta\|_2^2}\theta$$

- Adversarial Perturbation for Nonlinear Binary Classifiers

$$\boldsymbol{\varepsilon}^i_{\mathbf{x},\min} = \arg\min_{\boldsymbol{\varepsilon}^i_{\mathbf{x}}} \left\|\boldsymbol{\varepsilon}^i_{\mathbf{x}}\right\|_2 = -\frac{f_\theta(\mathbf{x}^i)}{\|\nabla f_\theta(\mathbf{x}^i)\|_2^2}\nabla f_\theta(\mathbf{x}^i)$$

- The continuous procedure superposes the $\boldsymbol{\varepsilon}^i_{\mathbf{x}}$ of each iteration value as Eq. (13), to obtain the minimal adversarial perturbation $\hat{\boldsymbol{\varepsilon}}_{\mathbf{x},\min}$ until the perturbed instance ($\mathbf{x}+\hat{\boldsymbol{\varepsilon}}_{\mathbf{x},\min}$) makes the different target from the original instance **x**, that is, $f_\theta(\mathbf{x}) \neq f_\theta(\mathbf{x} + \hat{\boldsymbol{\varepsilon}}_{\mathbf{x},\min})$.

$$\hat{\boldsymbol{\varepsilon}}_{\mathbf{x},\min} = \sum_i \boldsymbol{\varepsilon}^i_{\mathbf{x},\min}$$

- ***Robustness index for instance* (RII)** of the classifier $f_\theta(\cdot)$ for original sample **x**:

$$\mathrm{RII}(\mathbf{x}) = \left\|\hat{\boldsymbol{\varepsilon}}_{\mathbf{x},\min}\right\|_2$$

- ***Robustness index for classifier* (RIC)** of the classifier $f_\theta(\cdot)$

$$\mathrm{RIC}(f_\theta(\cdot)) = \frac{1}{|N|}\sum_{\mathbf{x}_n \in \mathcal{D}} \frac{\left\|\hat{\boldsymbol{\varepsilon}}_{\mathbf{x},\min}\right\|_2}{\|\mathbf{x}_n\|_2}$$

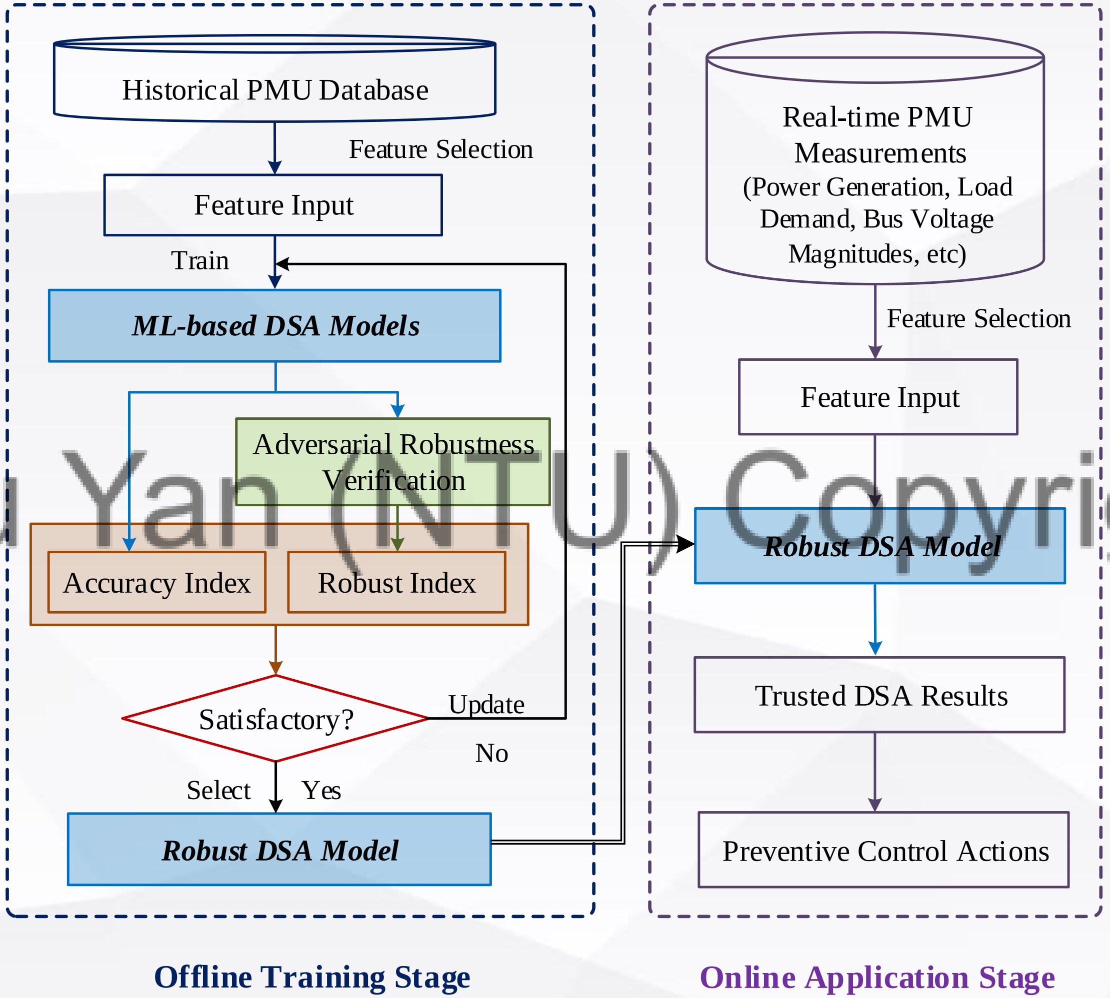
The empirical robustness indices (RII and RIC) can be utilized to measure the robustness of instances and classifiers under the adversarial attack.

The larger RII and RIC values indicate stronger abilities of instances and classifiers against the adversarial perturbations.

C. Ren, X. Du, Y. Xu*, Q. Song, Y. Liu and R. Tan, "Vulnerability Analysis, Robustness Verification and Mitigation Strategy of Machine Learning-based Power Systems Stability Assessment Models under Adversarial Examples," *IEEE Transactions on Smart Grid*, 2021.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Robustness Indices: Application



- *During the offline training stage, in addition to the accuracy, the robustness should also be evaluated to make sure the model is both accurate and robust for practical application.*
- *Besides, if an online instance has a lower robustness value, it should be handled with extra care, e.g., using traditional time-domain simulation method instead.*

Mitigation Strategy against Adversarial Examples

Adversarial training is to use a mixture of adversarial examples and original samples to train the models, rather than using only the original samples.

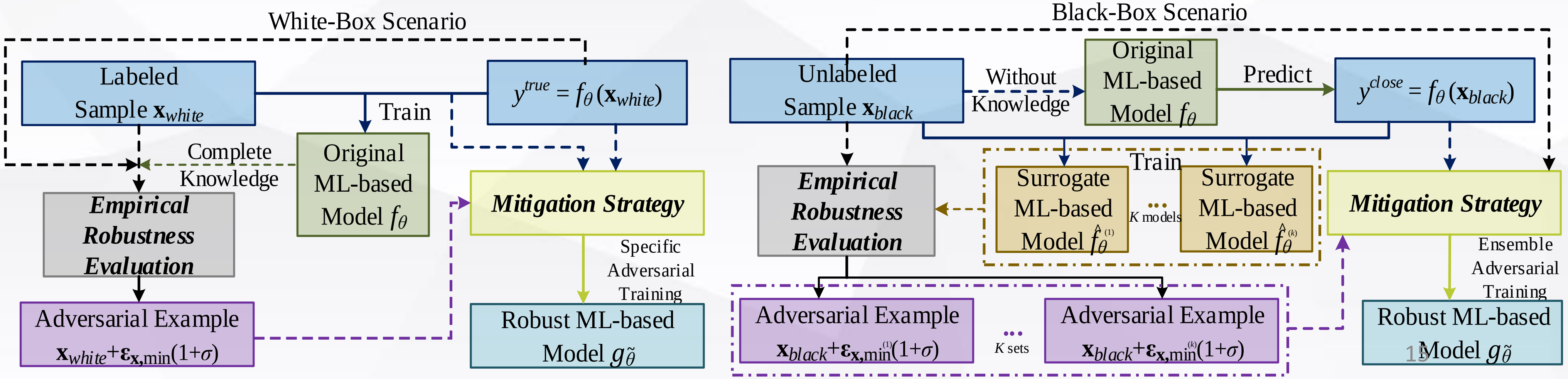
- Loss Function of Specific Single Adversarial Training under the White-Box Scenarios

$$L_g\left(g_{\tilde{\theta}}(\mathbf{x}), y\right) = (1-\alpha)\cdot L_f\left(f(\mathbf{x}), y\right) + \alpha\cdot L_f\left(f\left(\mathbf{x}+\hat{\boldsymbol{\varepsilon}}_{\mathbf{x},\mathbf{min}}(1+\sigma)\right), y\right)$$

- Loss Function of Ensemble Adversarial Training under the Black-Box Scenarios

$$L_g\left(g_{\tilde{\theta}}(\mathbf{x}), y^{close}\right) = \sum_{k=1}^{K}\left\{(1-\alpha)\cdot L_{\hat{f}^{(k)}}\left(\hat{f}^{(k)}(\mathbf{x}), y^{close}\right) + \alpha\cdot L_{\hat{f}^{(k)}}\left(\hat{f}^{(k)}\left(\mathbf{x}+\hat{\boldsymbol{\varepsilon}}_{\mathbf{x},\mathbf{min}}{}^{(k)}(1+\sigma)\right), y^{close}\right)\right\}$$

α represents the ratio of the adversarial examples



Universal Defense Strategy against Adversarial Examples

Randomized Smoothing aims to construct a new smoothed classifier $h(\cdot)$ from any arbitrary base classifier $f(\cdot)$. The smoothed classifier $h(\cdot)$ assigns the most likely class c returned by the base classifier $f(\cdot)$ under the isotropic Gaussian noise perturbation of x to the point x.

$$h(\mathrm{x}) = \arg\max_{c\in Y} \Pr_{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)} (f(\mathrm{x}+\varepsilon) = c)$$

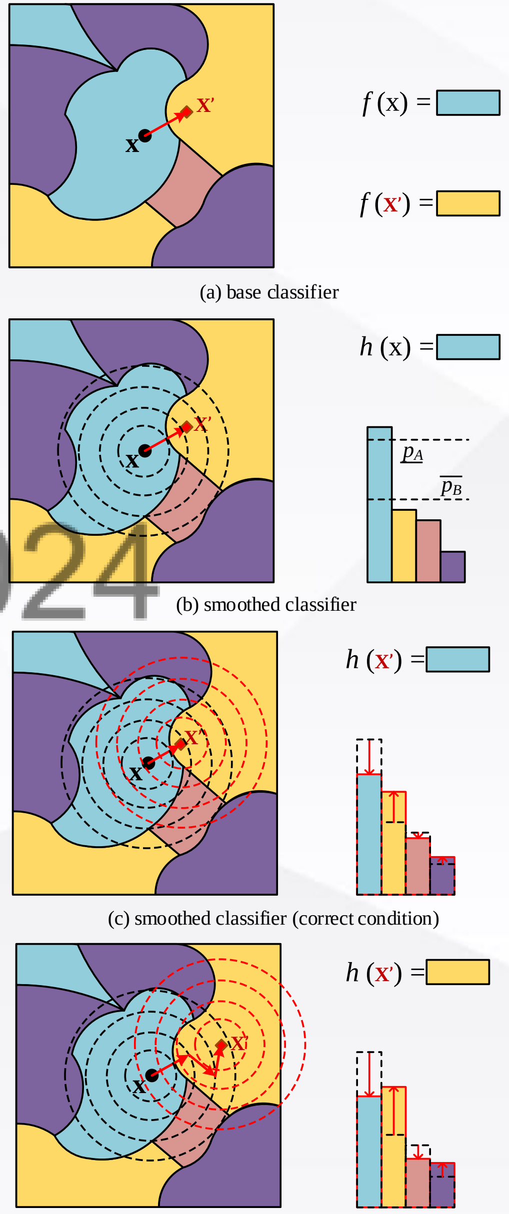
The smoothed classifier $h(\cdot)$ returns the class c with the largest probability value in the decision region $\{f(\hat{\mathrm{x}}) = c | \hat{\mathrm{x}} \in \mathbb{R}^m\}$ under the distribution $\mathcal{N}(x, \sigma^2 I)$.

➢ **Effectiveness Index R** for Universal Defense Strategy

$$\forall \|\delta\|_2 \le R, \text{smoothed classifier } h(\mathrm{x}+\delta) = c_A$$
$$where \quad R = \|\delta\|_2 \le \frac{\sigma}{2}\cdot\left(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})\right)$$

The smoothed classifier $h(\cdot)$ by randomized smoothing are certifiably robust under the l_2-norm ball with the effectiveness index R.

➢ The value of the effectiveness index R is determined by three factors: *1*) noise level σ; *2*) the probability of the most likely class c_A; and *3*) the probability of the other class. The ideal effectiveness index R is under the higher σ, c_A and the lower c_B, but the higher σ may slightly reduce the accuracy.



(a) base classifier
(b) smoothed classifier
(c) smoothed classifier (correct condition)
(d) smoothed classifier (wrong condition)

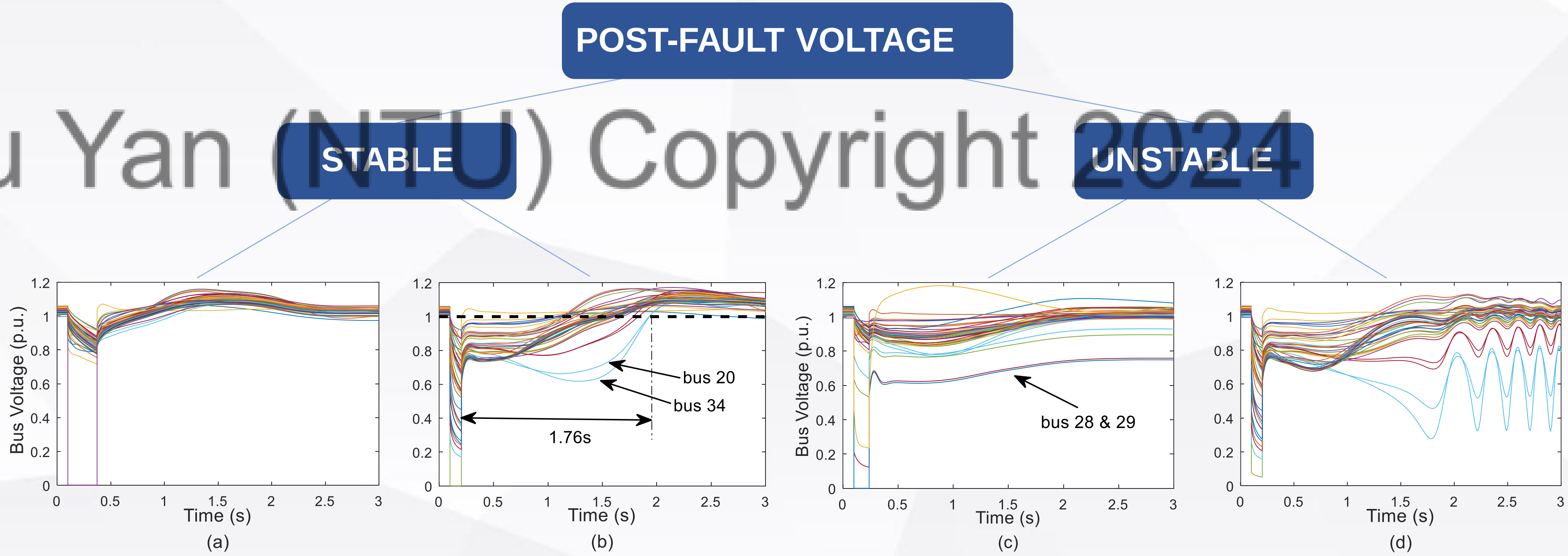
C. Ren and Y. Xu*, "A Universal Defense Strategy for Data-Driven Power System Stability Assessment Models under Adversarial Examples," *IEEE Internet of Things Journal*, 2022.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Case Study: Short-Term Voltage Stability (STVS) Problem

The STVS problem is concerned on:

- Fault-induced delayed voltage recovery (FIDVR) – risk for wind turbine to ride through
- Sustained low voltage without recovery – may lead to voltage collapse in the long-term
- Fast voltage collapse – usually associated with rotor-angle instability



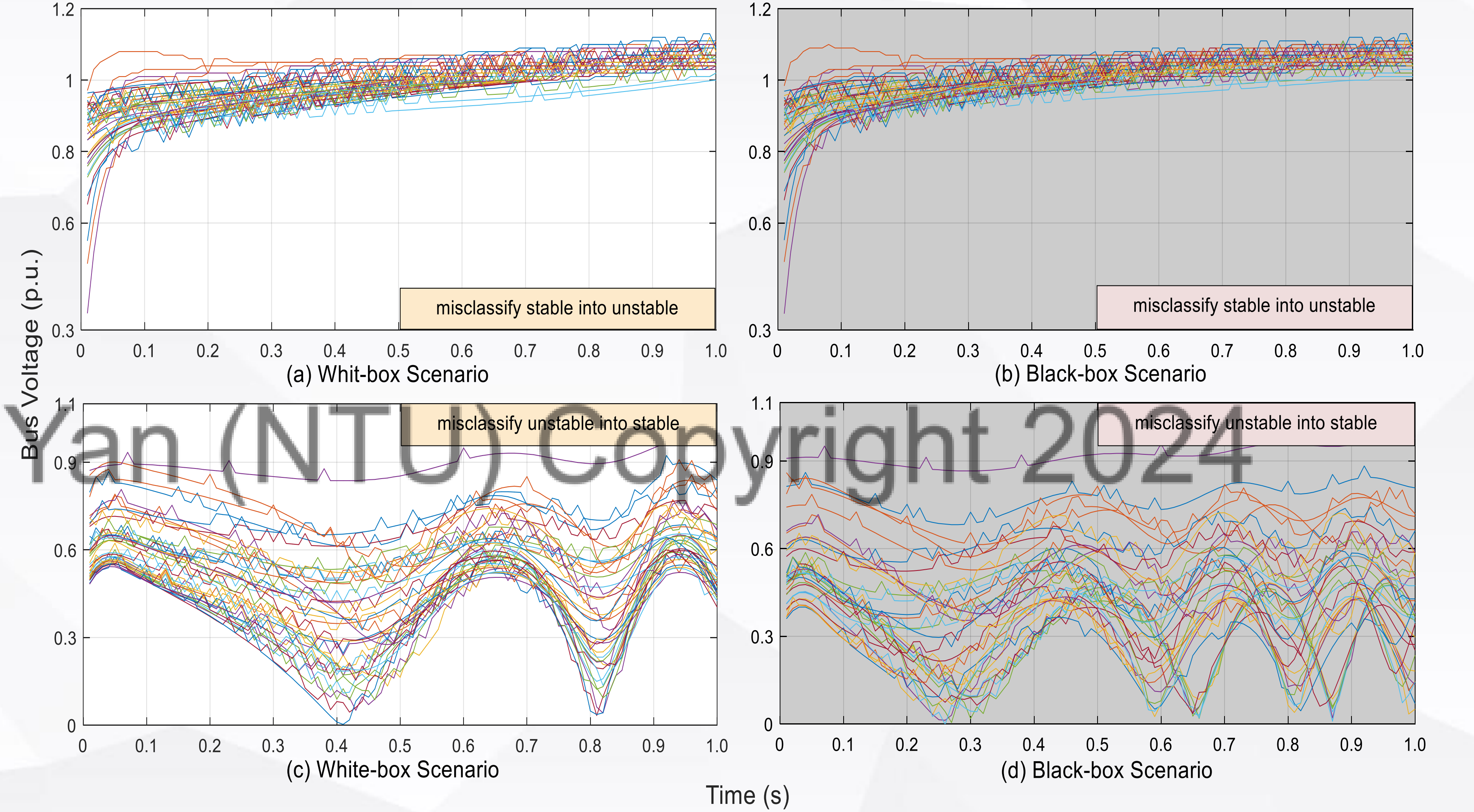
Fast Recovery (satisfactory)

FIDVR (acceptable)

Sustained Low Voltage (unacceptable)

Fast Collapse (unacceptable)

Vulnerability Analysis



Testing results with the adversarial examples

(a) misclassify stable into unstable under the white-box scenarios; (b) misclassify stable into unstable under the black-box scenarios; (c) misclassify unstable into stable under the white-box scenarios; (d) misclassify unstable into stable under the black-box scenarios.

It can be seen that a very small perturbation to the voltage measurement value can lead to a wrong stability assessment result.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Vulnerability Analysis

Testing System and Machine Learning (ML) Models

- Testing system: IEEE New England 10-machine 39-bus system using the industry-standard composite load model "CLOD"
- ML-based STVS assessment models: Long short-term memory (LSTM), fully convolutional neural network (FCNN), and back-propagation neural network (BPNN)
- Testing observation windows: 0.8s, 1.0s, 1.2s after the fault clearance

TABLE I
VULNERABILITY ANALYSIS FOR STVS ASSESSMENT ACCURACY OF ADVERSARIAL EXAMPLES GENERATION STRATEGY

Observation Windows (0.8s, 1.0s, 1.2s)	Without Adversarial Examples			White-Box Scenarios			Black-Box Scenarios (using LSTM)		Black-Box Scenarios (using FCNN)		Black-Box Scenarios (using BPNN)	
	LSTM	FCNN	BPNN	LSTM	FCNN	BPNN	FCNN (Surrogate)	BPNN (Surrogate)	LSTM (Surrogate)	BPNN (Surrogate)	LSTM (Surrogate)	FCNN (Surrogate)
Average	98.78%	97.83%	97.22%	6.37%	6.22%	6.03%	19.02%	17.03%	14.63%	16.62%	13.53%	15.37%

Results and Observations

- For all ML models, the STVS accuracy drops sharply with the generated adversarial examples under both white-box scenario (from 98% down to 6.03% to 6.37%) and black-box scenario (from 98% down to 13.53% to 19.02%).
- The accuracy of the original ML-based STVS models degrades much more significantly in the case of the white-box scenarios than in the black-box scenarios.

Background

Problem Description

Methodology

Adversarial example generation

Robustness evaluation indices

Mitigation strategy

Case Study

Conclusion

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Mitigation Strategy against Adversarial Examples

TABLE II
RIC PERFORMANCE FOR ML-BASED STVS MODELS

Observation Windows (0.8s, 1.0s, 1.2s)	Original ML-based Models without Adversarial Examples			Specific Adversarial Training-based Mitigation Strategy under White-Box Scenarios			Ensemble Adversarial Training-based Mitigation Strategy under the Black-Box Scenarios		
	LSTM	FCNN	BPNN	Specific LSTM (against LSTM)	Specific FCNN (against FCNN)	Specific BPNN (against BPNN)	Ensemble FCNN&BPNN (against LSTM)	Ensemble LSTM&BPNN (against FCNN)	Ensemble LSTM&FCNN (against BPNN)
Average	0.020	0.017	0.016	0.046	0.043	0.041	0.039	0.036	0.035

TABLE III
ACCURACY PERFORMANCE OF ADVERSARIAL TRAINING-BASED MITIGATION STRATEGY AGAINST ADVERSARIAL EXAMPLES

Testing Samples with Observation Windows (0.8s, 1.0s, 1.2s)	White-Box Scenarios			Black-Box Scenarios		
	Specific LSTM (against LSTM)	Specific FCNN (against FCNN)	Specific BPNN (against BPNN)	Ensemble FCNN & BPNN (against LSTM)	Ensemble LSTM & BPNN (against FCNN)	Ensemble LSTM & FCNN (against BPNN)
Clean Samples	98.23%	97.07%	97.05%	96.75%	96.68%	96.22%
Adversarial Examples	97.68%	96.07%	95.69%	95.37%	94.92%	94.70%

- Table II lists the average RIC results of original ML-based STVS model, robust ML-based STVS models after the specific and ensemble adversarial training-based mitigation strategy.
- The RIC value validates the adversarial training-based mitigation strategy for the white-box scenarios is more effective than the black-box scenarios. For RII, the larger the RII value, the greater the adversarial perturbation needed to successfully attack the original sample.
- Table III shows the adversarial training-based mitigation strategy accuracy performances with the original clean samples and adversarial examples for both the white-box and the black-box scenarios

Background

Problem Description

Methodology

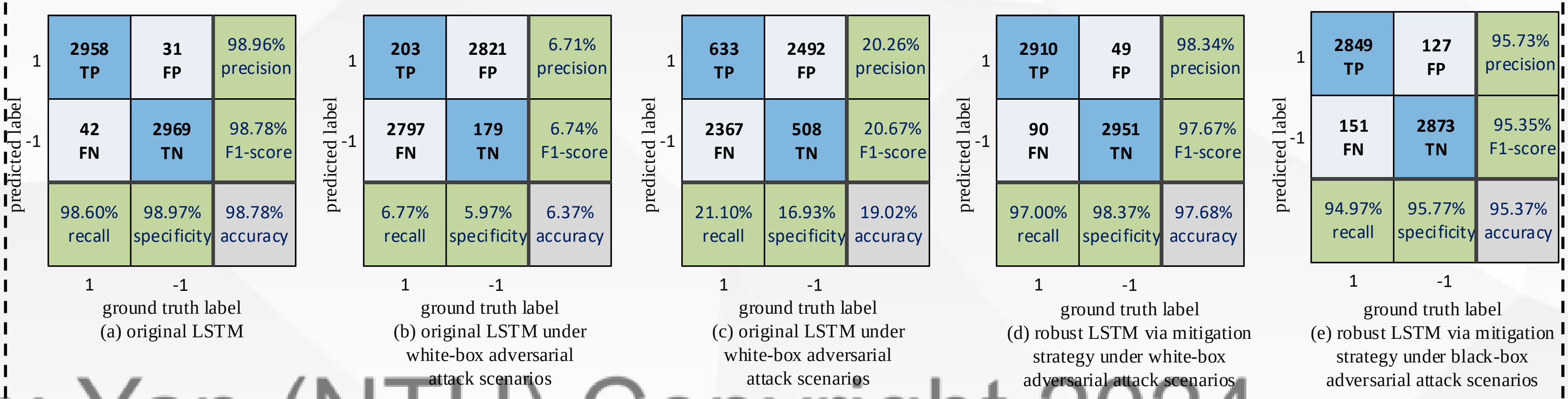
- Adversarial example generation
- Robustness evaluation indices
- Mitigation strategy

Case Study

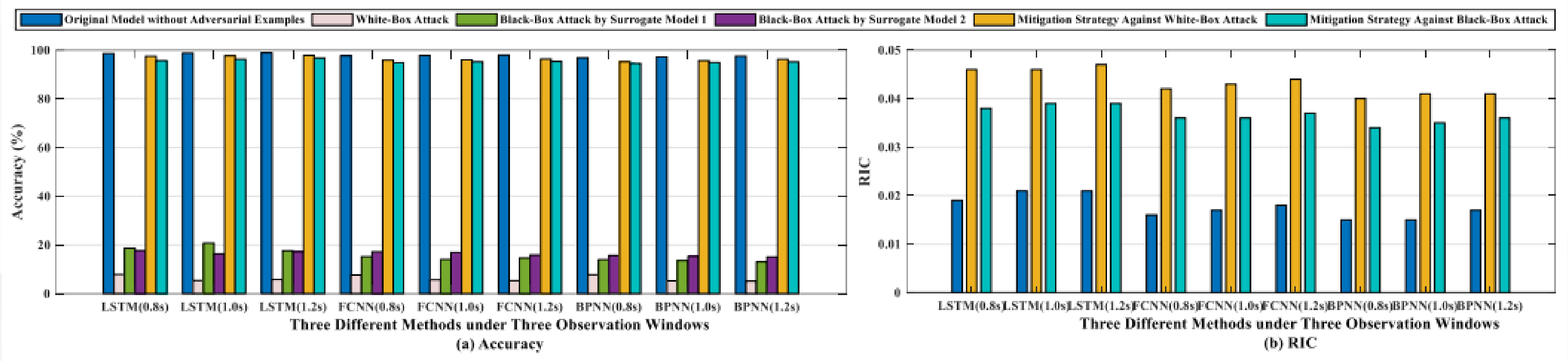
Conclusion

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Mitigation Strategy against Adversarial Examples

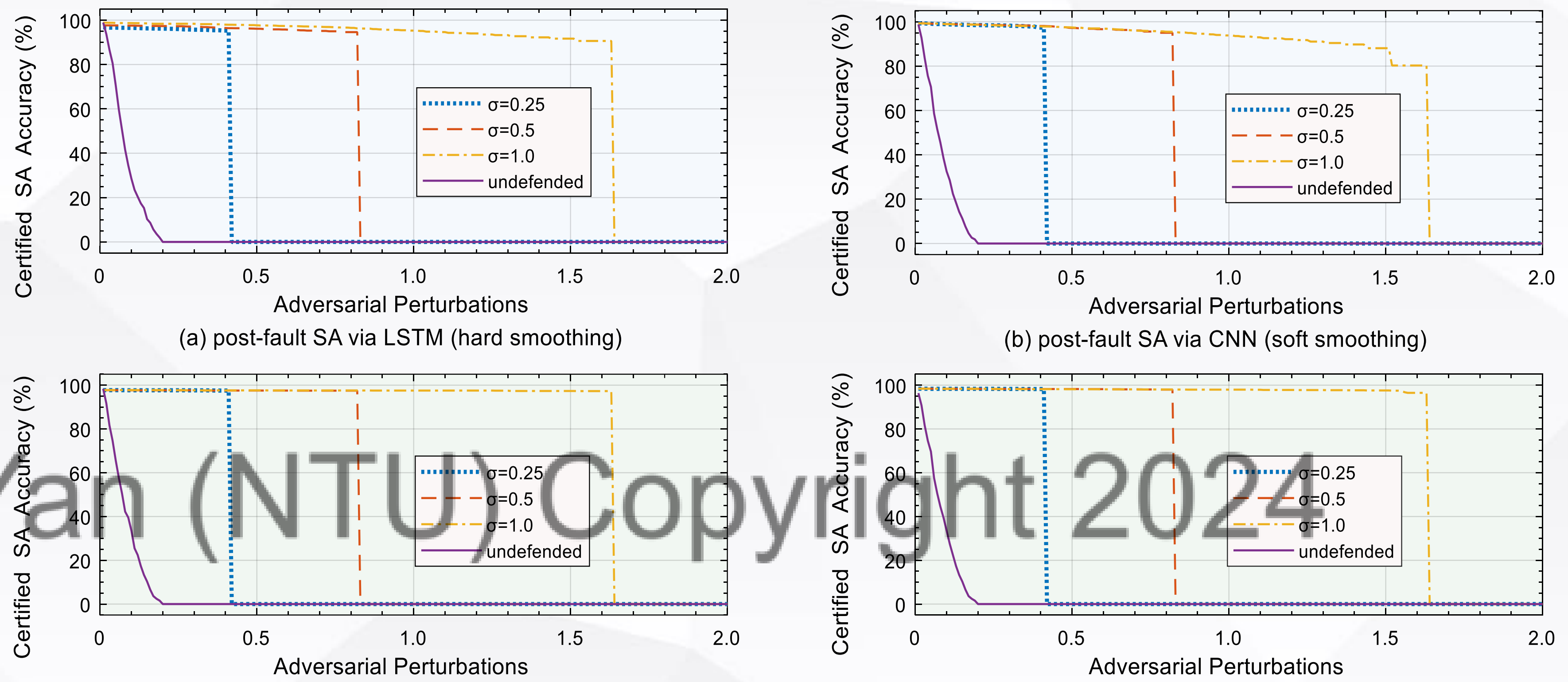


Confusion matrix of original and robust LSTM model with three different observation window under different scenarios.



Accuracy and RIC results for three different methods under the three different observation window (0.8s, 1.0s, 1.2s) (a) STVS accuracy; (b) RIC.

Universal Defense Strategy against Adversarial Examples



(a) post-fault SA via LSTM (hard smoothing)

(b) post-fault SA via CNN (soft smoothing)

(c) pre-fault SA via LSTM (hard smoothing)

(d) pre-fault SA via CNN (soft smoothing)

Testing results of certified SA accuracy under different adversarial perturbations for pre-fault and post-fault SA with different data-driven SA models.

- Above figures show *Tradeoff between Robustness and Accuracy* that the largest noise level σ can only guarantee the largest effectiveness index R, but cannot always achieve the highest certified SA accuracy under all the adversarial perturbations
- Based on such results, we can select the more robust data-driven models, which are trained by different ML algorithms or under the different degree of adversarial attacks.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Conclusions

We firstly reveal the threat of the adversarial examples to the ML-based model, then systematically evaluate the robustness of the ML-based model under the adversarial examples, and finally develop a mitigation strategy against the adversarial examples.

- ✓ The threat of the adversarial examples for the ML-based model under both the white-box and the black-box scenarios is illustrated using an adversarial example generation strategy. It reveals that the adversarial example can obviously lead to ML accuracy degradation.
- ✓ To accurately quantify the vulnerability of the ML-based models and instances, two robust indices are proposed for the empirical robustness evaluation.
- ✓ A mitigation strategy is designed via adversarial training and the empirical robustness evaluation, which can maintain the accuracy and improve the robustness of the ML-based model against the adversarial examples. A defense strategy is proposed to train a smoothed probabilistic classifer.

For more technical details of this work, please refer to our publications:

1. C. Ren, X. Du, **Y. Xu***, Q. Song, Y. Liu and R. Tan, "Vulnerability Analysis, Robustness Verification and Mitigation Strategy of Machine Learning-based Power Systems Stability Assessment Models under Adversarial Examples," ***IEEE Transactions on Smart Grid***, 2021.
2. C. Ren and **Y. Xu***, "Robustness Verification for Machine Learning-based Power System Dynamic Security Assessment," ***IEEE Transactions on Control of Network Systems***, 2022.
3. C. Ren and **Y. Xu***, “A Universal Defense Strategy for Data-Driven Power System Stability Assessment Models under Adversarial Examples,” ***IEEE Internet of Things Journal***, 2022.

THANKS

Dr Yan Xu | Associate Professor
School of Electrical & Electronic Engineering
Nanyang Technological University
Singapore 639798