Perspectives from a Comprehensive Evaluation of Reconstruction-based Anomaly Detection in ICS

Clement Fung, Shreya Srinarasi, Keane Lucas, Hay Bryan Phee, Lujo Bauer

Electrical & Computer ENGINEERING
S3D
Carnegie Mellon University CyLab Security and Privacy Institute
Carnegie Mellon University

Industrial control systems (ICS) govern vital infrastructure

https://samcotech.com/common-industrial-water-treatment-problems-how-to-fix-them/
https://www.britannica.com/technology/chemical-industry/Heavy-inorganic-chemicals
https://itrust.sutd.edu.sg/testbeds/secure-water-treatment-swat/
https://lunesys.com/ukrainian-power-grid-hacked/
https://www.pbs.org/wgbh/nova/article/cyber-attack-german-steel-mill-leads-massive-real-world-damage/

The importance of ICS security is increasing

German Steel Mill (2014)

The importance of ICS security is increasing

German Steel Mill (2014)

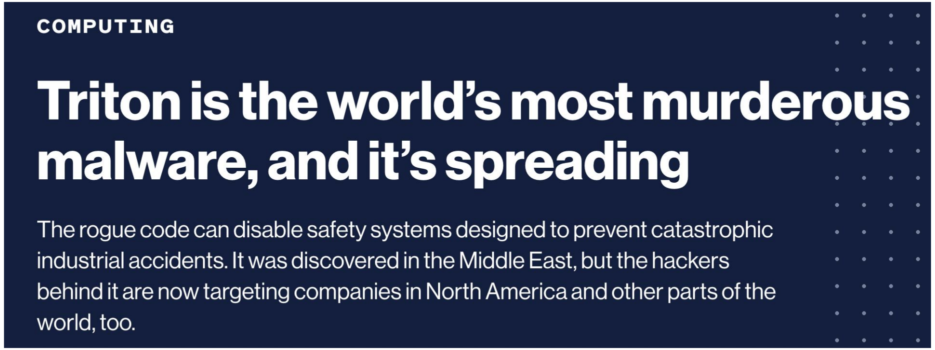
BlackEnergy (2015)
Industroyer (2016)

The importance of ICS security is increasing

German Steel Mill (2014)

BlackEnergy (2015)
Industroyer (2016)

Triton (2017)



COMPUTING

Triton is the world's most murderous malware, and it's spreading

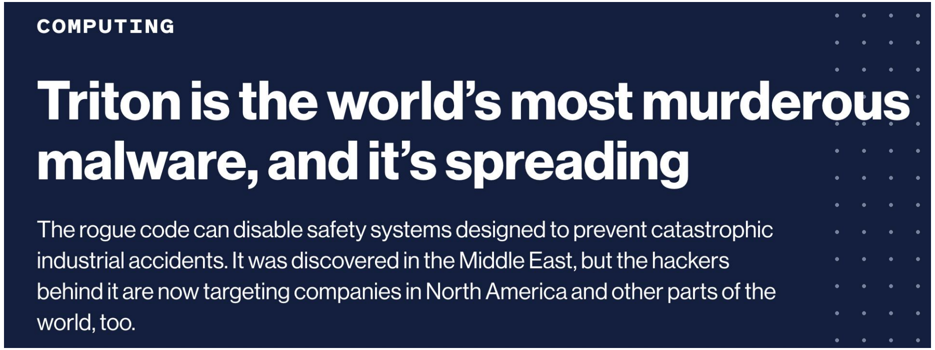
The rogue code can disable safety systems designed to prevent catastrophic industrial accidents. It was discovered in the Middle East, but the hackers behind it are now targeting companies in North America and other parts of the world, too.

The importance of ICS security is increasing

German Steel Mill (2014)

BlackEnergy (2015)
Industroyer (2016)

Triton (2017)

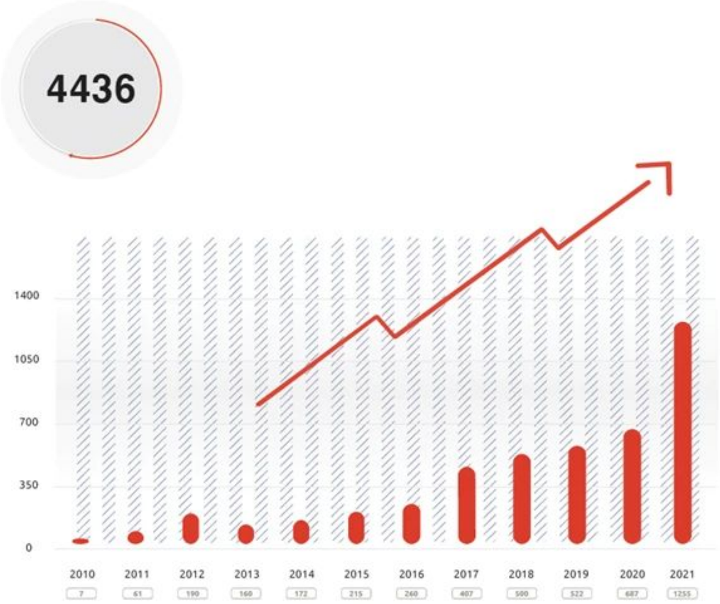


COMPUTING

Triton is the world's most murderous malware, and it's spreading

The rogue code can disable safety systems designed to prevent catastrophic industrial accidents. It was discovered in the Middle East, but the hackers behind it are now targeting companies in North America and other parts of the world, too.

ICS-CERT CVEs (by year)

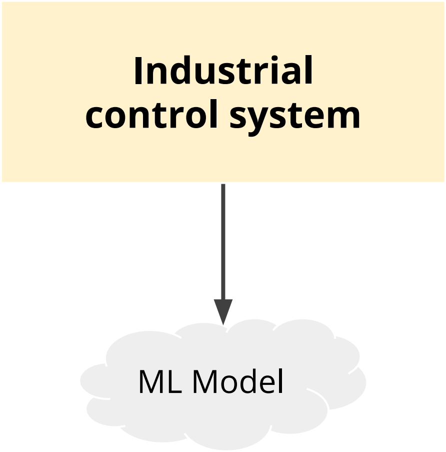


ML-based ICS anomaly detection

Industrial control system

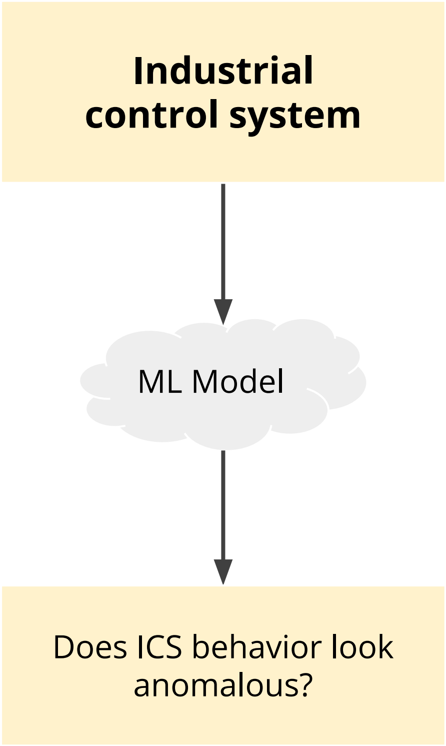
ML-based ICS anomaly detection

- Learn a model of ICS behavior



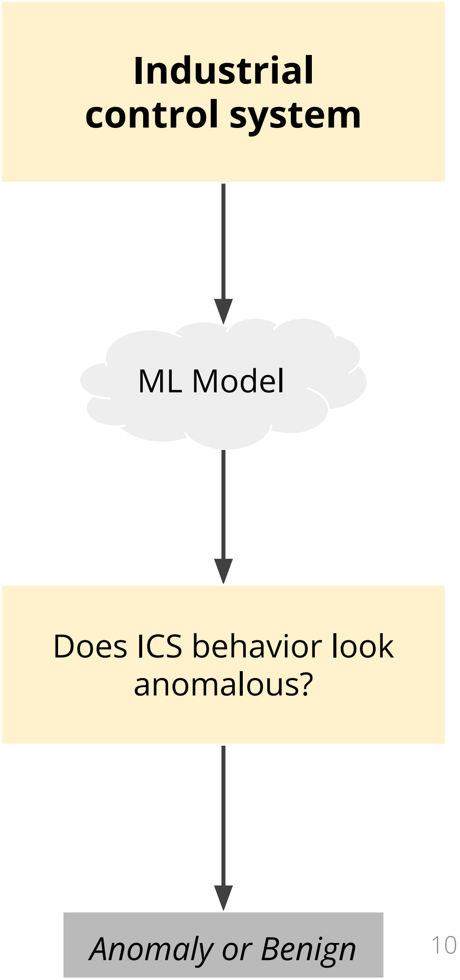
ML-based ICS anomaly detection

- Learn a model of ICS behavior



ML-based ICS anomaly detection

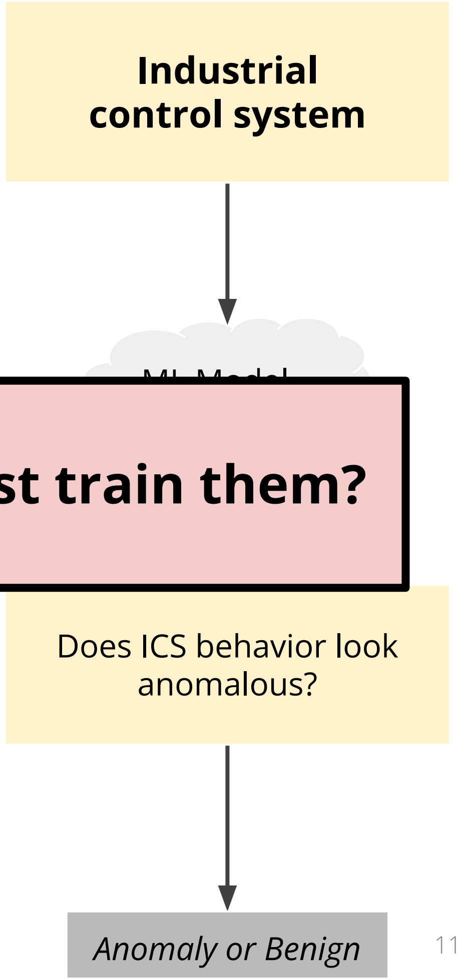
- Learn a model of ICS behavior



ML-based ICS anomaly detection

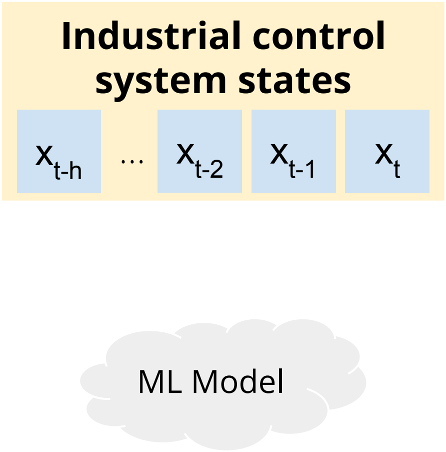
- Learn a model of ICS behavior

We ask: What are the best models and how to best train them?



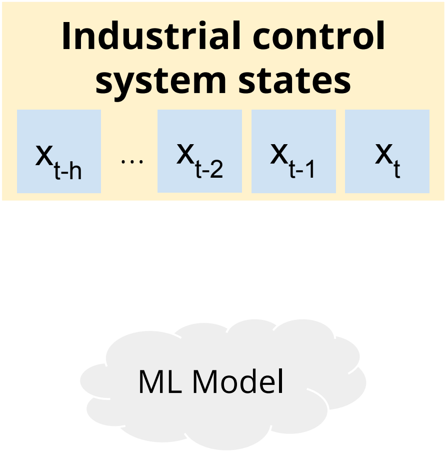
Reconstruction-based anomaly detection

- Learn ICS behavior from system states



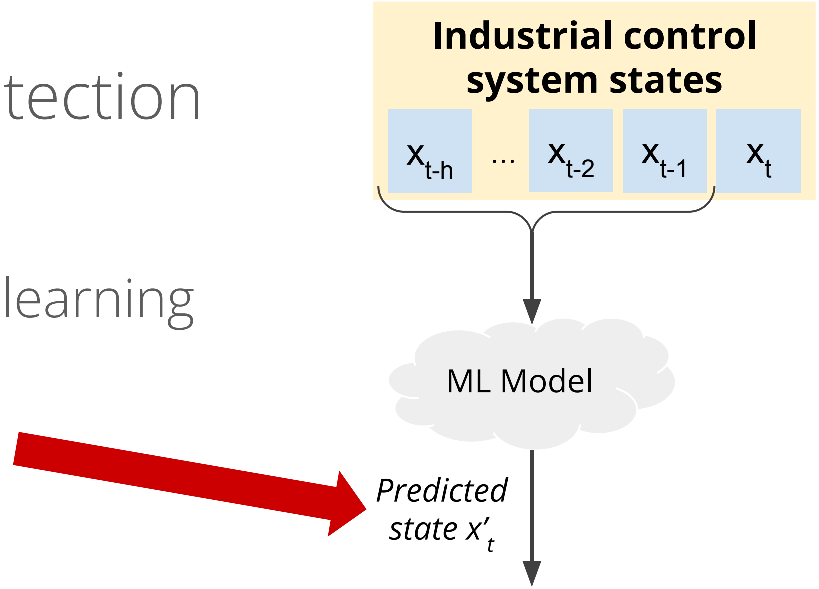
Reconstruction-based anomaly detection

- Learn ICS behavior from system states
- Anomalies are rare: use **unsupervised** learning



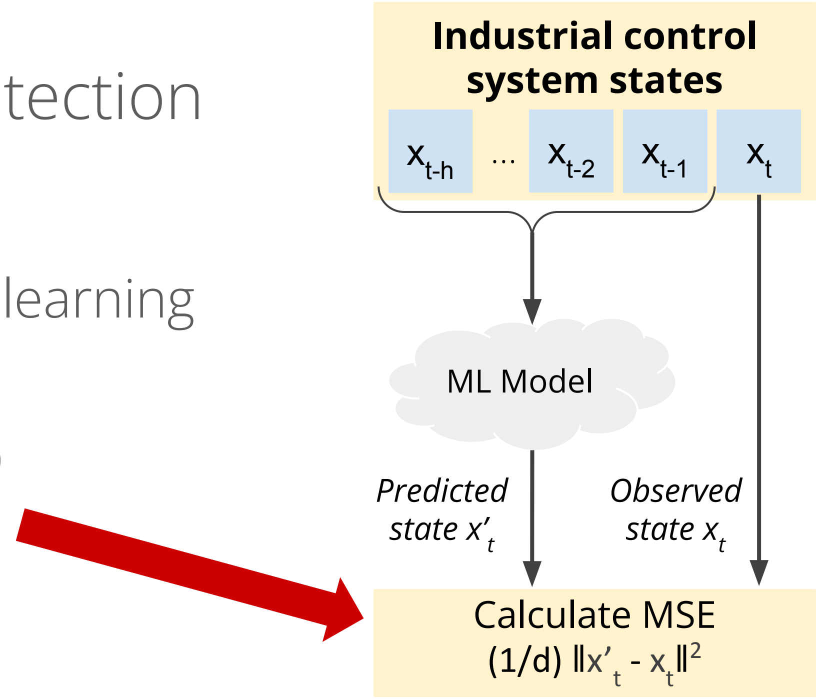
Reconstruction-based anomaly detection

- Learn ICS behavior from system states
- Anomalies are rare: use **unsupervised** learning
  - **Reconstruct** future ICS states



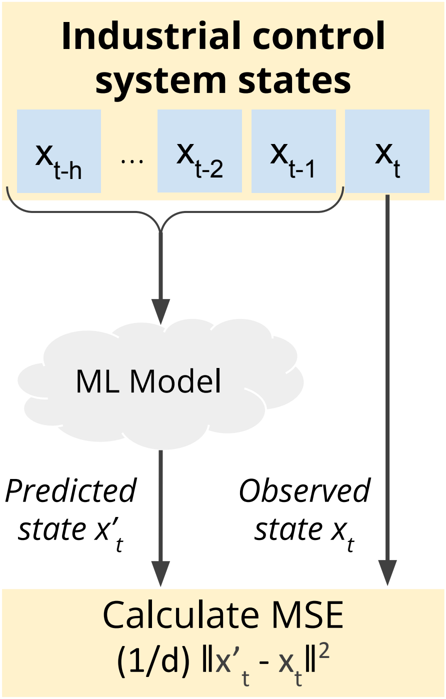
Reconstruction-based anomaly detection

- Learn ICS behavior from system states
- Anomalies are rare: use **unsupervised** learning
  - **Reconstruct** future ICS states
  - Compare with observed states (MSE)

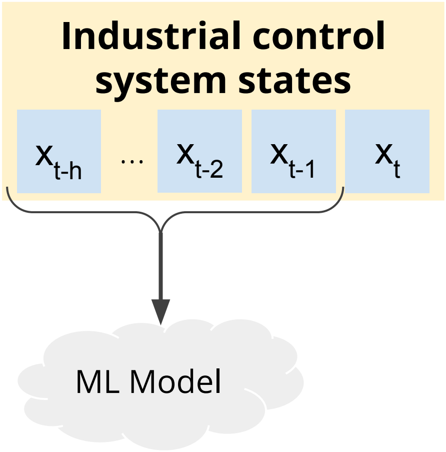


Reconstruction-based anomaly detection

- Learn ICS behavior from system states
- Anomalies are rare: use **unsupervised** learning
  - **Reconstruct** future ICS states
  - Compare with observed states (MSE)
  - Minimize training MSE



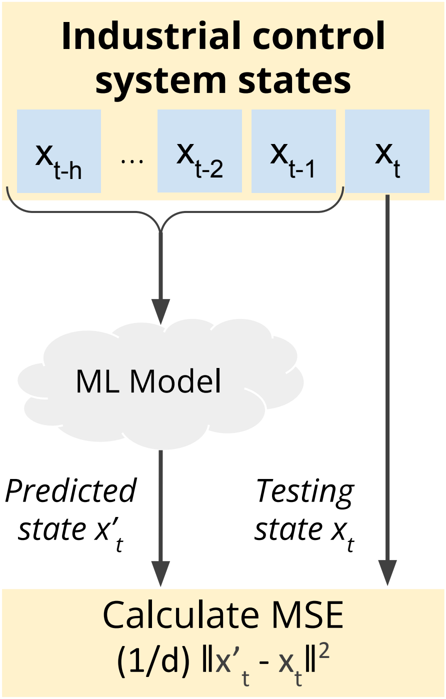
Reconstruction-based anomaly detection



- Learn ICS behavior from system states
- Anomalies are rare: use **unsupervised** learning
  - **Reconstruct** future ICS states
  - Compare with observed states (MSE)
  - Minimize training MSE
- At test time:

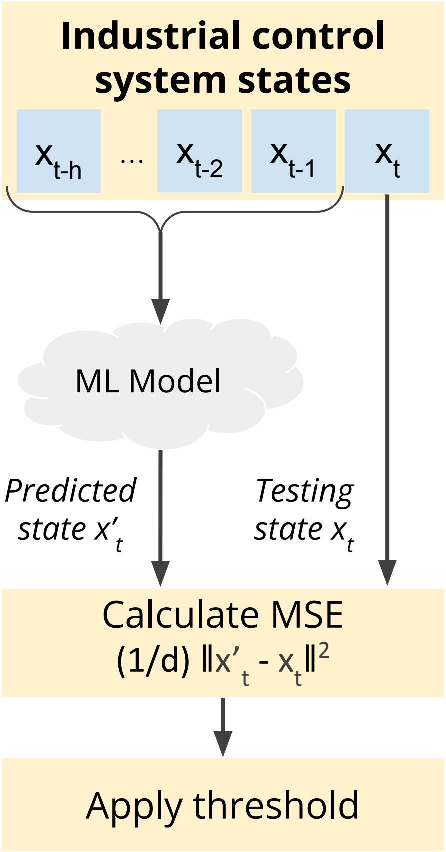
Reconstruction-based anomaly detection

- Learn ICS behavior from system states
- Anomalies are rare: use **unsupervised** learning
  - **Reconstruct** future ICS states
  - Compare with observed states (MSE)
  - Minimize training MSE
- At test time:



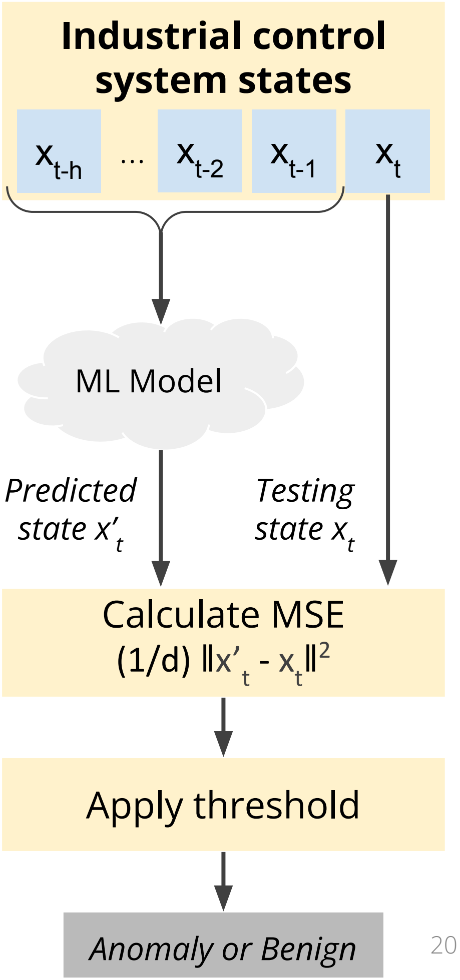
Reconstruction-based anomaly detection

- Learn ICS behavior from system states
- Anomalies are rare: use **unsupervised** learning
  - **Reconstruct** future ICS states
  - Compare with observed states (MSE)
  - Minimize training MSE
- At test time:
  - Apply MSE threshold



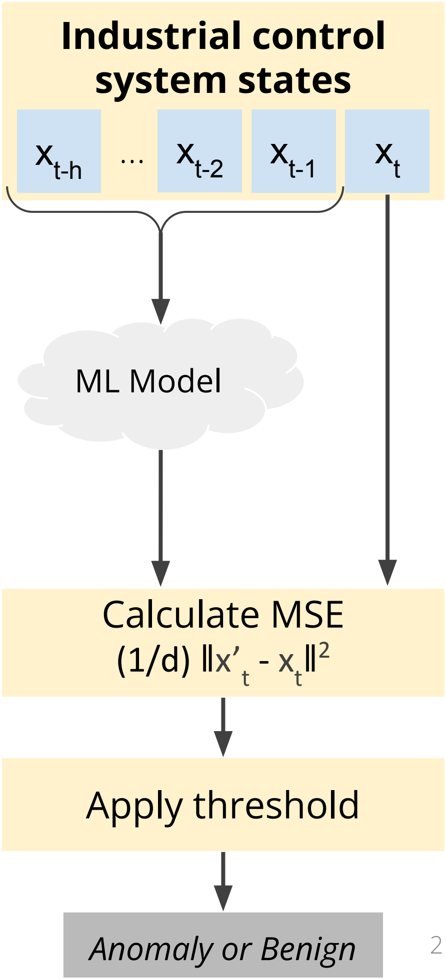
Reconstruction-based anomaly detection

- Learn ICS behavior from system states
- Anomalies are rare: use **unsupervised** learning
  - **Reconstruct** future ICS states
  - Compare with observed states (MSE)
  - Minimize training MSE
- At test time:
  - Apply MSE threshold
  - Raise alarms if exceeded



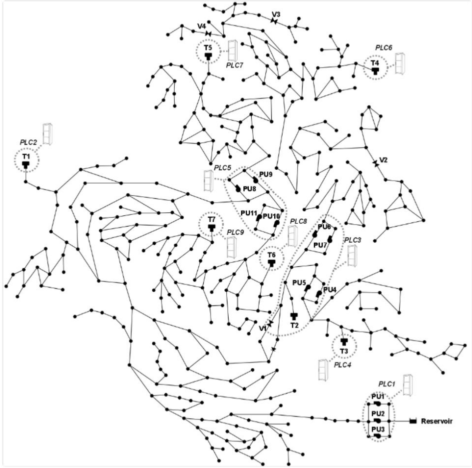
Prior work uses different...

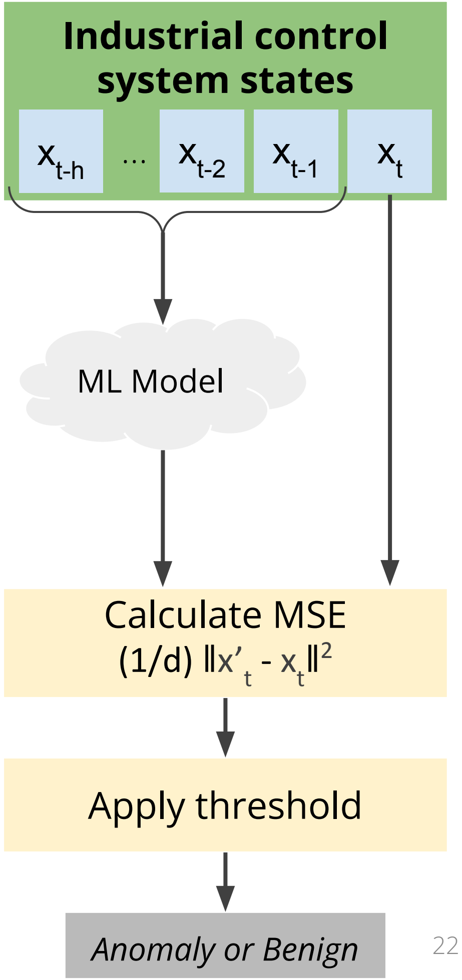
- Datasets
- Architectures
- Techniques



Prior work uses different…

- **Datasets**:
  - SWaT, WADI, BATADAL





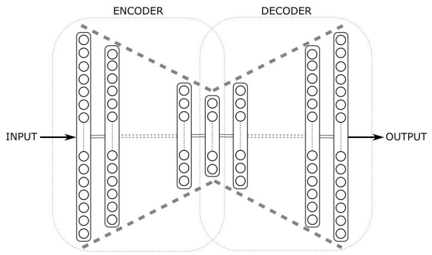
[1] Goh, J., Adepu, S., Junejo, K.N., Mathur, A.: A dataset to support research in the design of secure water treatment systems. *International Conference on Critical Information Infrastructures Security.* 2016.

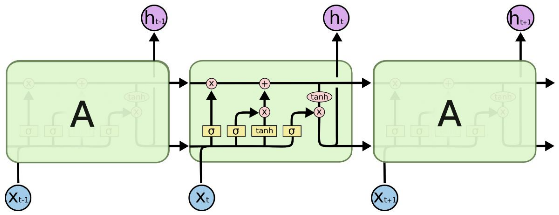
[2] Ahmed, C.M., Palleti, V.R., Mathur, A.: WADI: a water distribution testbed for research in the design of secure cyber physical systems. *3rd International Workshop on Cyber-Physical Systems for Smart Water Networks.* 2017.

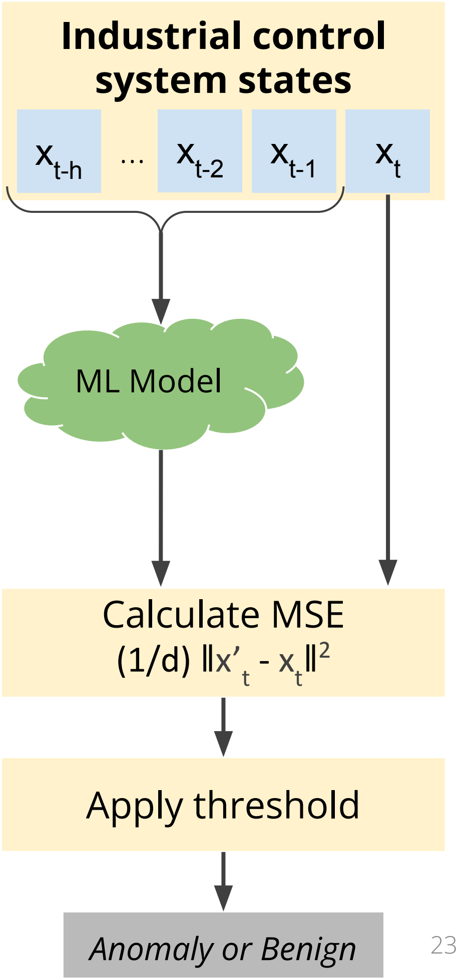
[3] Taormina et al. Battle of the attack detection algorithms: Disclosing cyber attacks on water distribution networks. *Journal of Water Resources Planning and Management 144(8).* 2018.

Prior work uses different…

- Datasets:
  - SWaT, WADI, BATADAL
- **Architectures**:
  - Autoencoders, CNNs, LSTMs



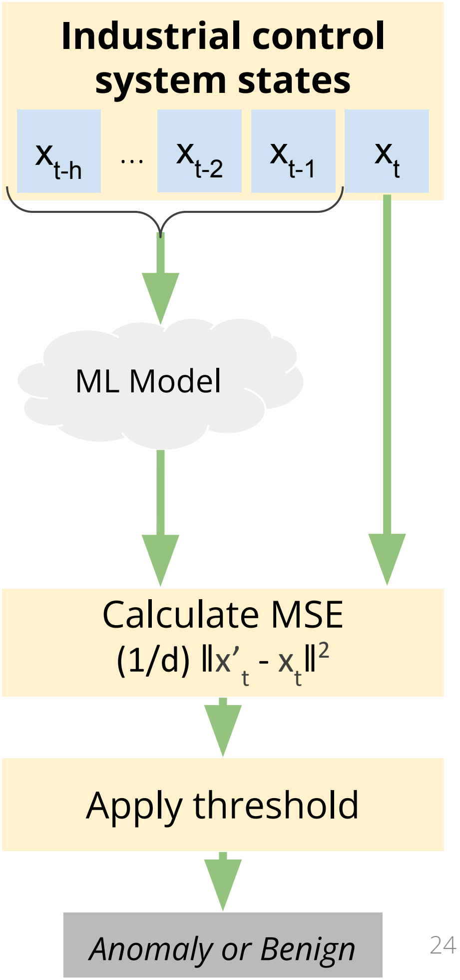




[1] Taormina et al. Deep-Learning Approach to the Detection and Localization of Cyber-Physical Attacks on Water Distribution Systems. *Journal of Water Resources Planning and Management, 144(10).* 2018.
[2] Kravchik et al. Detecting Cyber Attacks in Industrial Control Systems Using Convolutional Neural Networks. CPS-SPC 2018.
[3] Zizzo et al. Intrusion Detection for Industrial Control Systems: Evaluation Analysis and Adversarial Attacks. arXiv 2019.
Images: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Prior work uses different…

- Datasets:
  - SWaT, WADI, BATADAL
- Architectures:
  - Autoencoders, CNNs, LSTMs
- **Techniques**:
  - Early stopping, feature cleaning
  - Various model hyperparameters
  - Various thresholding values

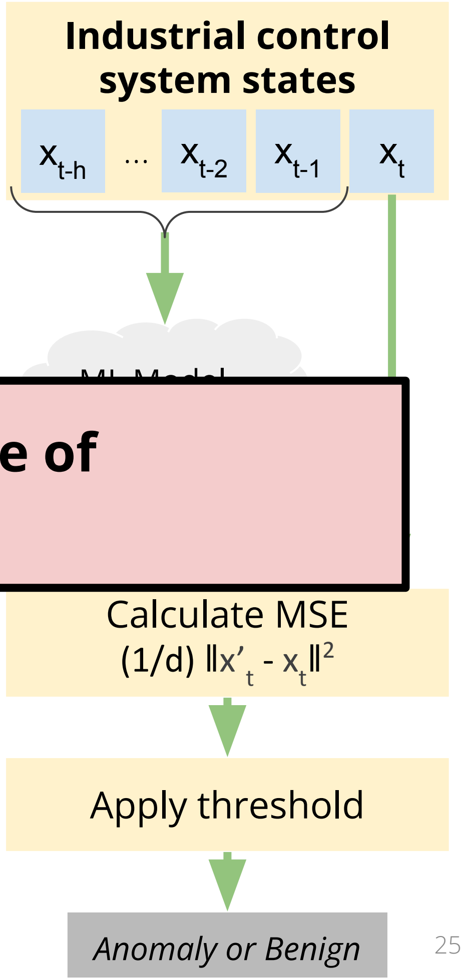


Prior work uses different...

- Datasets:
  - SWaT, WADI, BATADAL
- Architectures:

Difficult to compare prior work because of inconsistent methodology!

  - Early stopping, feature cleaning
  - Various model hyperparameters
  - Various thresholding values



What ML architecture is best?

- **Study A on SWaT**:
  - 8-layer, 32-unit CNN is best

What ML architecture is best?

- **Study A on SWaT**:
  - 8-layer, 32-unit CNN is best

- **Study B on SWaT**:
  - 4-layer, 512-unit LSTM is best

What ML architecture is best?

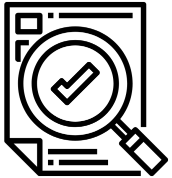
- **Study A on SWaT**:
  - 8-layer, 32-unit CNN is best
- **Study B on SWaT**:
  - 4-layer, 512-unit LSTM is best
- **Study C on SWaT**:
  - 1-layer autoencoder is best

What ML architecture is best?

- **Study A on SWaT**:
  - 8-layer, 32-unit CNN is best
- **Study B on SWaT**:
  - 4-layer, 512-unit LSTM is best
- **Study C on SWaT**:
  - 1-layer autoencoder is best

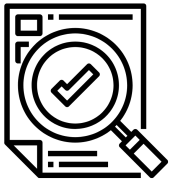
Our Contributions

- We evaluate proposed ICS anomaly detection approaches:
  - Across datasets, model architectures, and hyperparameters
  - With a **common methodology**

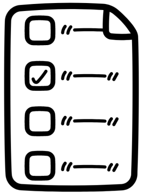


Our Contributions

- We evaluate proposed ICS anomaly detection approaches:
  - Across datasets, model architectures, and hyperparameters
  - With a **common methodology**

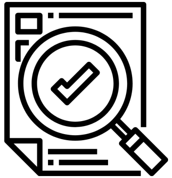


- We identify **four key techniques** in methods:
  - Needed for **reproducible and correct** evaluation!

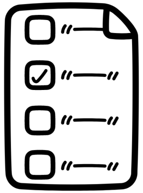


Our Contributions

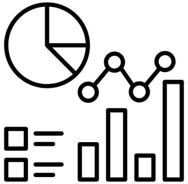
- We evaluate proposed ICS anomaly detection approaches:
  - Across datasets, model architectures, and hyperparameters
  - With a **common methodology**



- We identify **four key techniques** in methods:
  - Needed for **reproducible and correct** evaluation!



- We describe the need for different ICS anomaly-detection metrics
  - Explain why we should **stop using the point-F1 score**
  - **Use range-based metrics** for better tuning and optimization



Part 1

What **models are best** for ICS anomaly detection?

A common training and evaluation methodology

1

Pre-process ICS dataset

Datasets: SWaT, WADI, BATADAL

A common training and evaluation methodology

1 Pre-process ICS dataset

Datasets: SWaT, WADI, BATADAL

2 Train unsupervised ML model

CNN, LSTM: 1-5 layers, 4-256 units, 50-200 history
AE: 1-5 layers, 1.5-4.0 compression

A common training and evaluation methodology

1 Pre-process ICS dataset

Datasets: SWaT, WADI, BATADAL

2 Train unsupervised ML model

CNN, LSTM: 1-5 layers, 4-256 units, 50-200 history
AE: 1-5 layers, 1.5-4.0 compression

3 Tune threshold

MSE threshold **τ**, window length **w**
objective: maximize point-F1 score

A common training and evaluation methodology

1 Pre-process ICS dataset

Datasets: SWaT, WADI, BATADAL

2 Train unsupervised ML model

CNN, LSTM: 1-5 layers, 4-256 units, 50-200 history
AE: 1-5 layers, 1.5-4.0 compression

3 Tune threshold

MSE threshold $\boldsymbol{\tau}$, window length $\mathbf{w}$
objective: maximize point-F1 score

4 Evaluate against attacks at test time

Report final point-F1 score, averaged over 3x random seeds

A common training and evaluation methodology

1

Pre-process ICS dataset

Datasets: SWaT, WADI, BATADAL

Key techniques:

- *Benign data shuffling*
- *Feature selection*
- *Attack cleaning*

2

Train unsupervised ML model

CNN, LSTM: 1-5 layers, 4-256 units, 50-200 history

AE: 1-5 layers, 1.5-4.0 compression

Key technique: *Early stopping*

3

Tune threshold

MSE threshold $\boldsymbol{\tau}$, window length $\mathbf{w}$

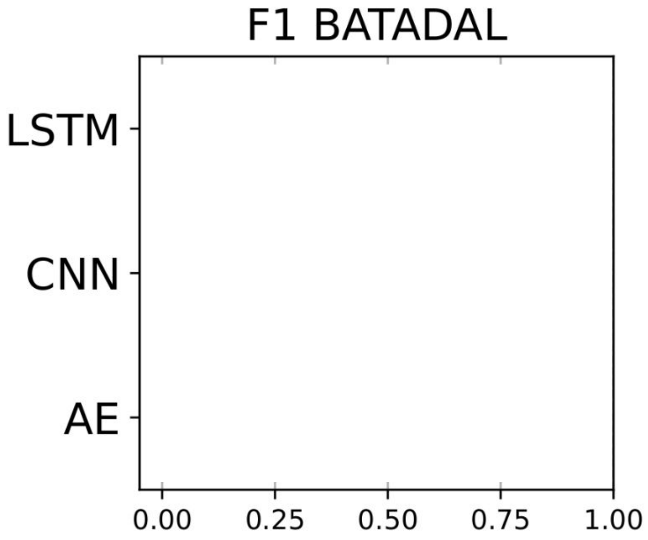
objective: maximize point-F1 score

4

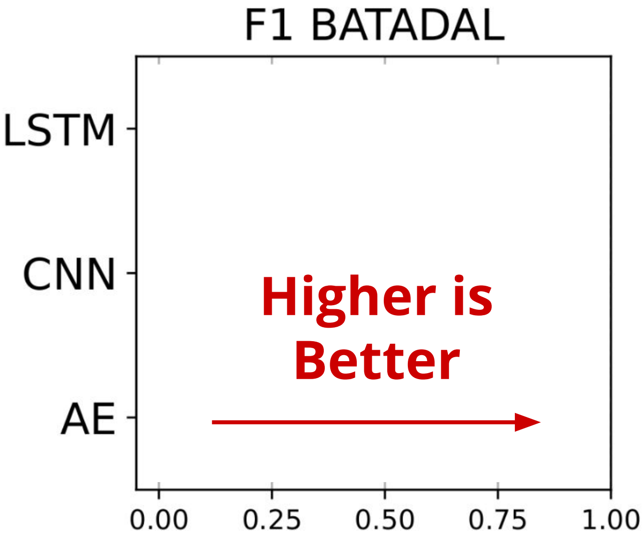
Evaluate against attacks at test time

Report final point-F1 score, averaged over 3x random seeds

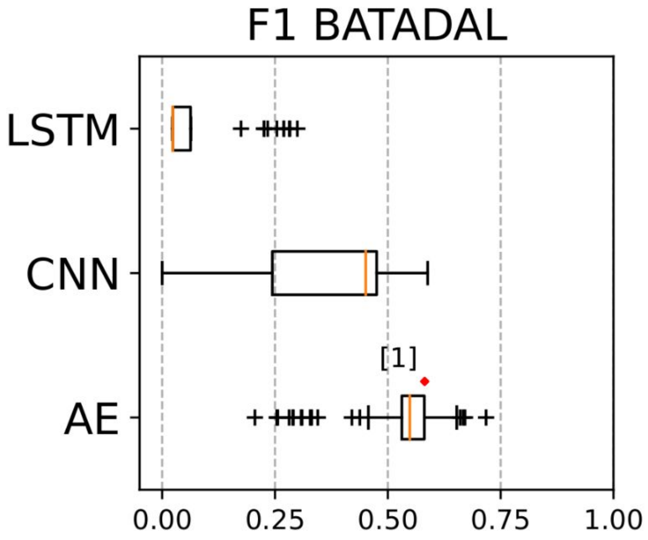
Is there a best model?



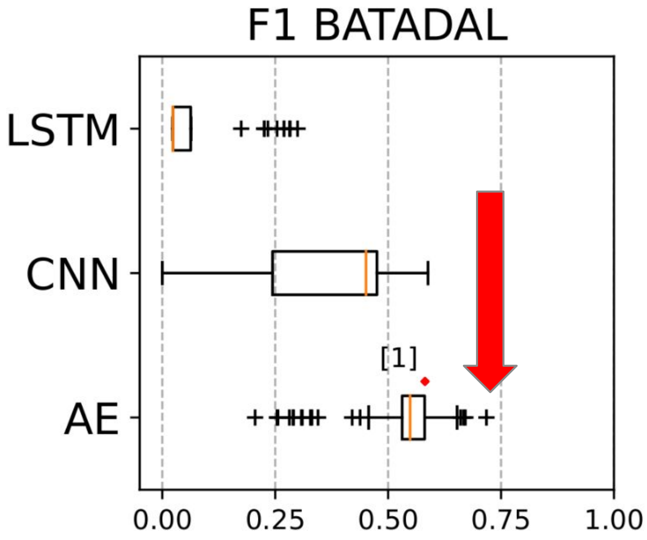
Is there a best model?



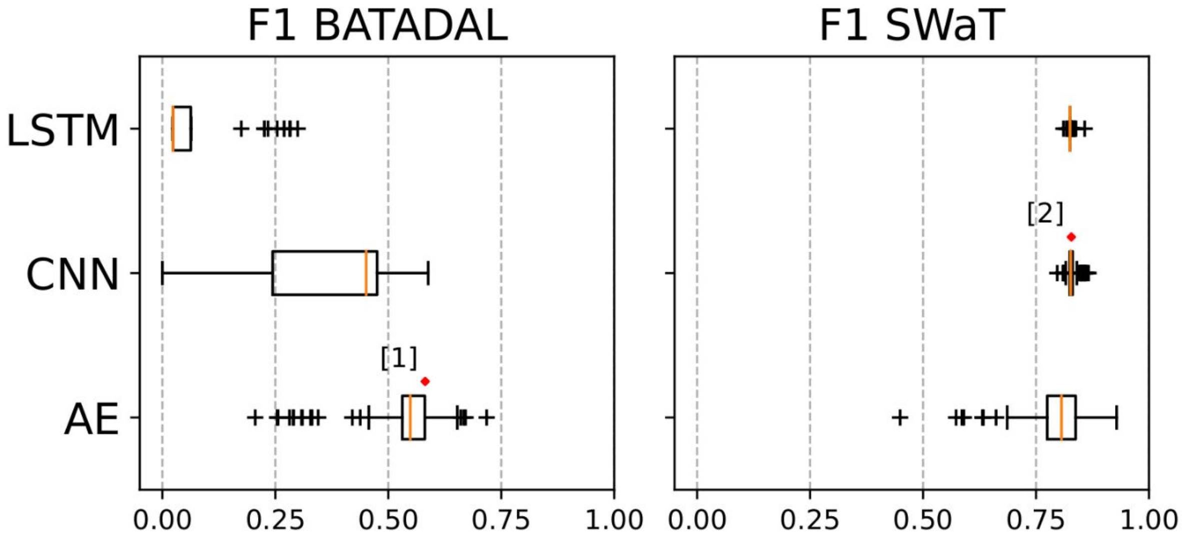
Is there a best model?



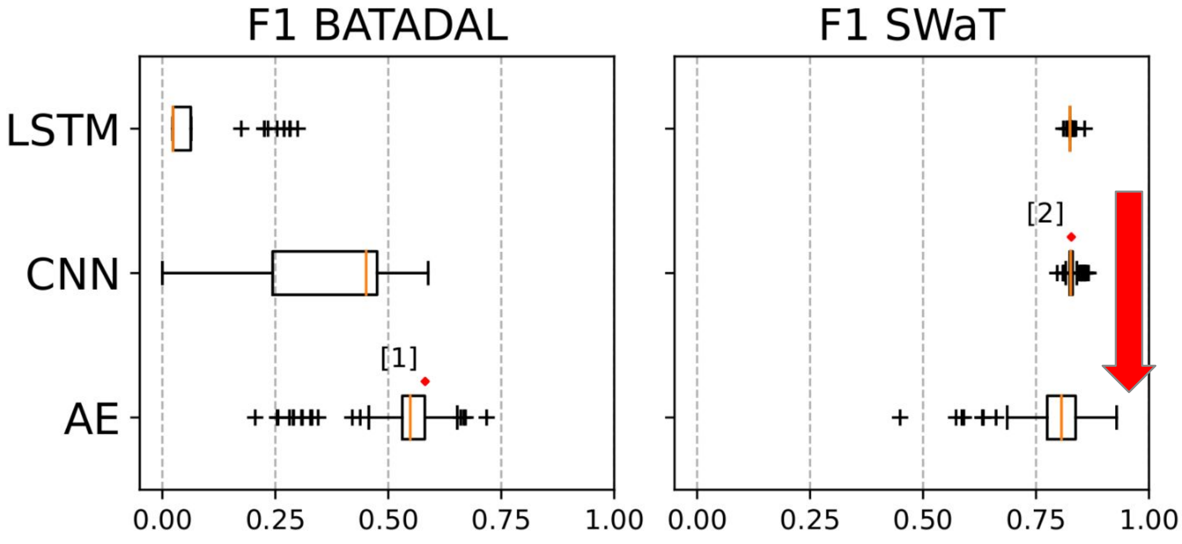
Is there a best model?



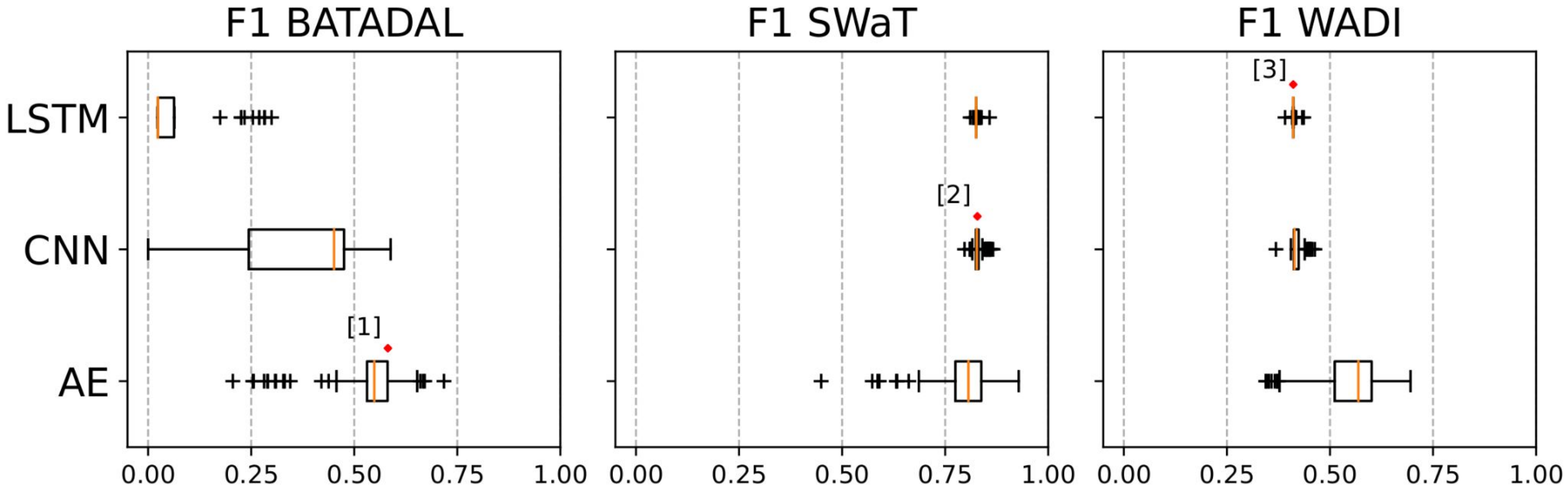
Is there a best model?



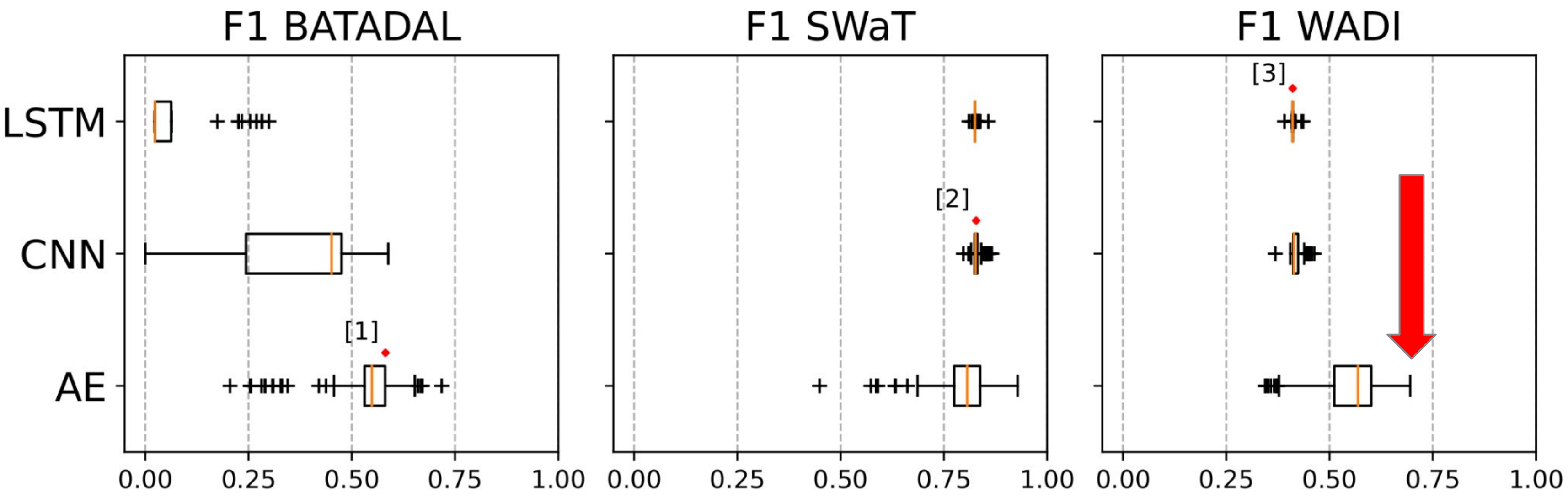
Is there a best model?



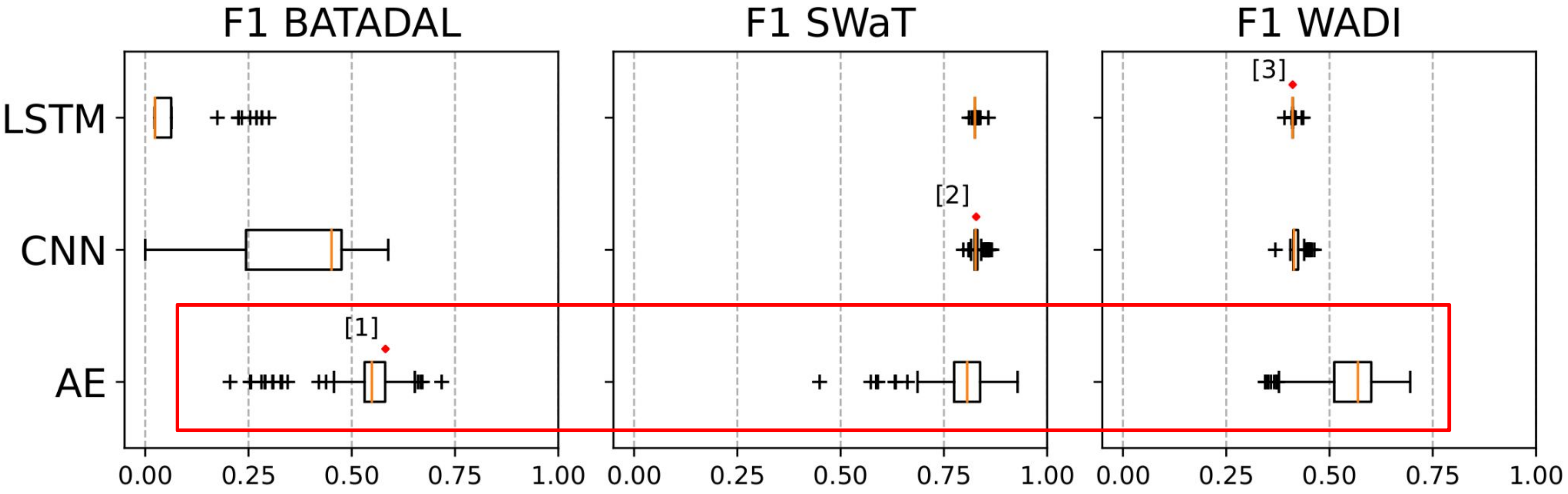
Is there a best model?



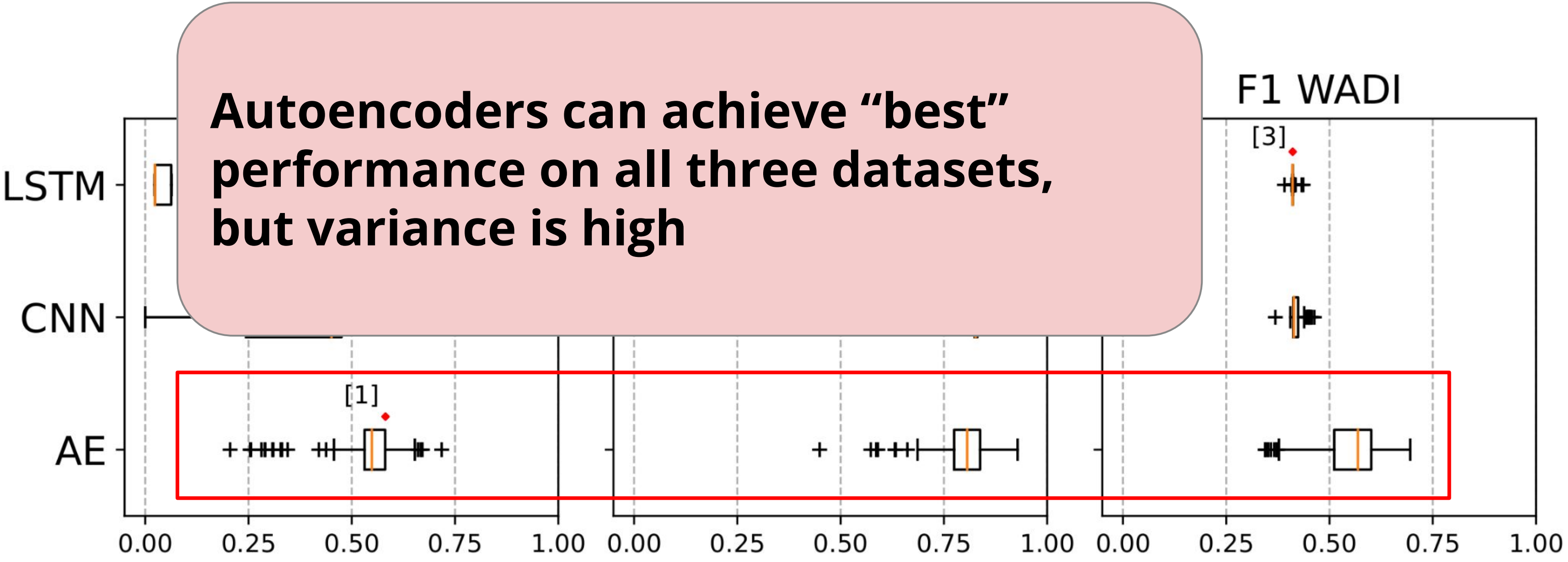
Is there a best model?



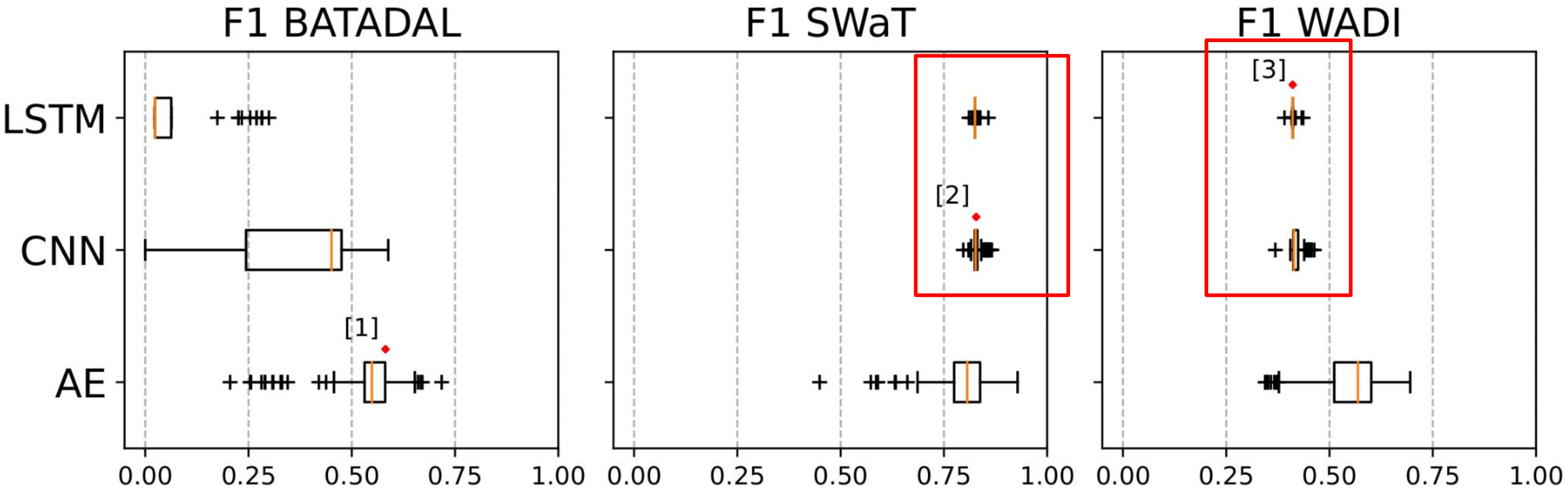
Is there a best model?



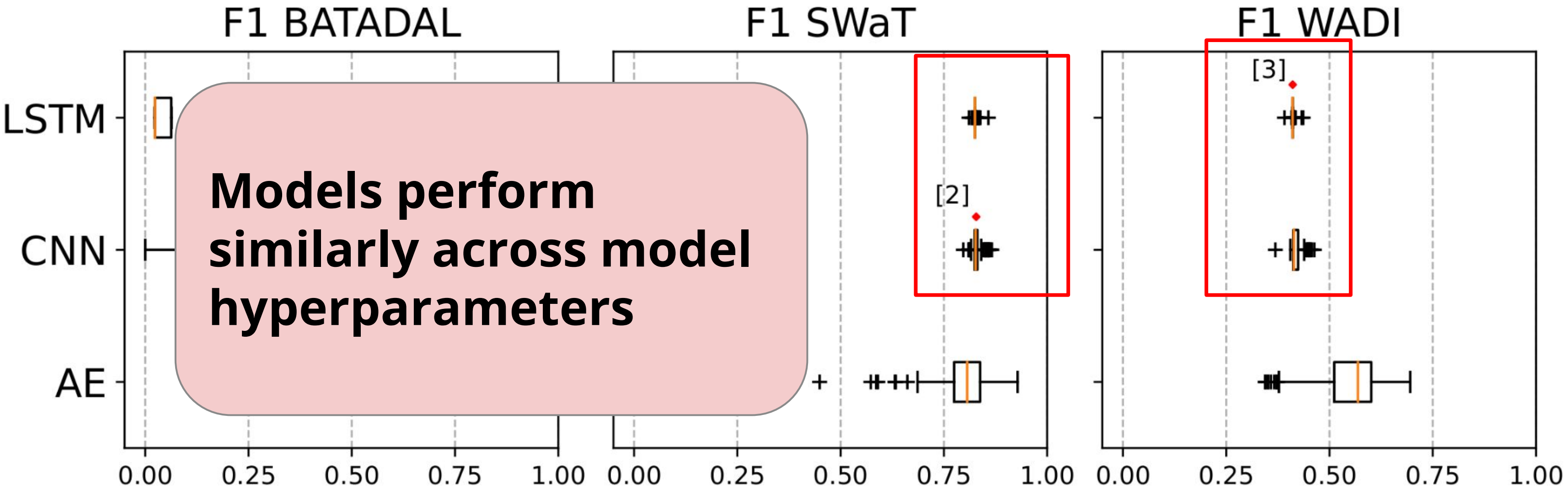
Is there a best model?



Is there a best model?

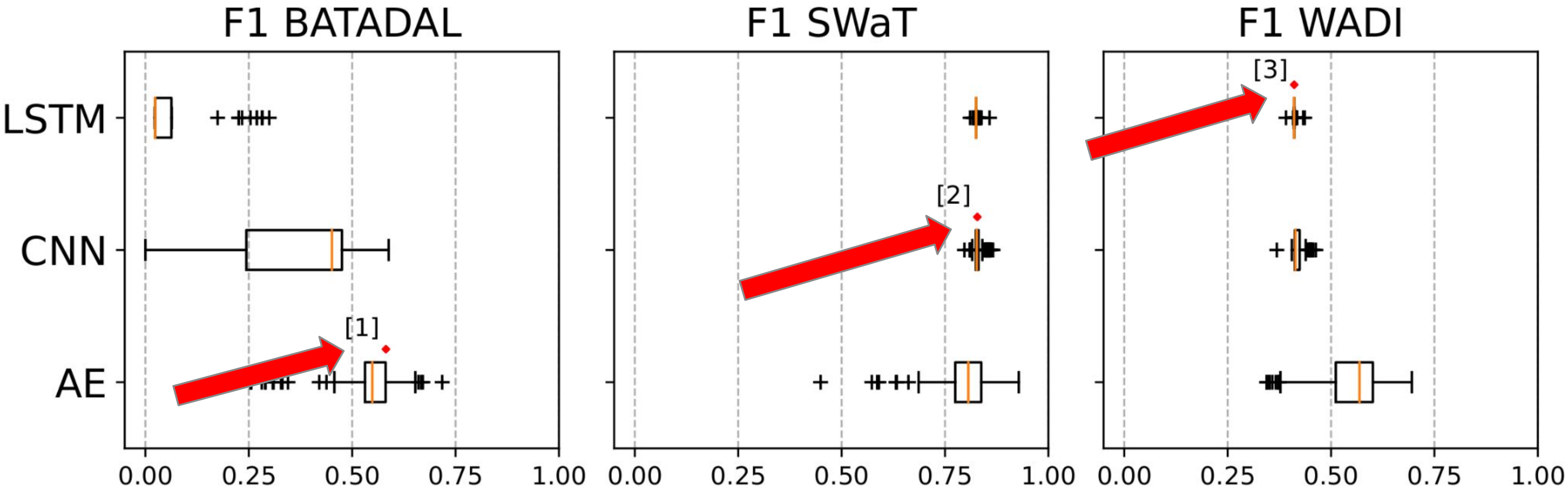


Is there a best model?

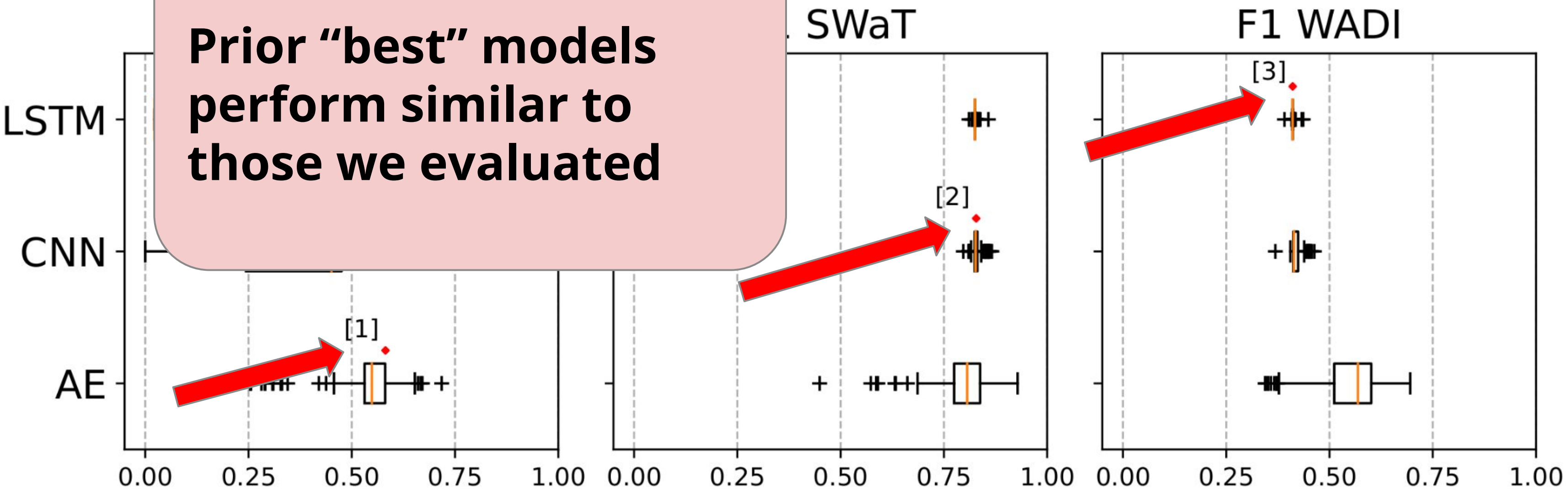


Models perform similarly across model hyperparameters

Is there a best model?

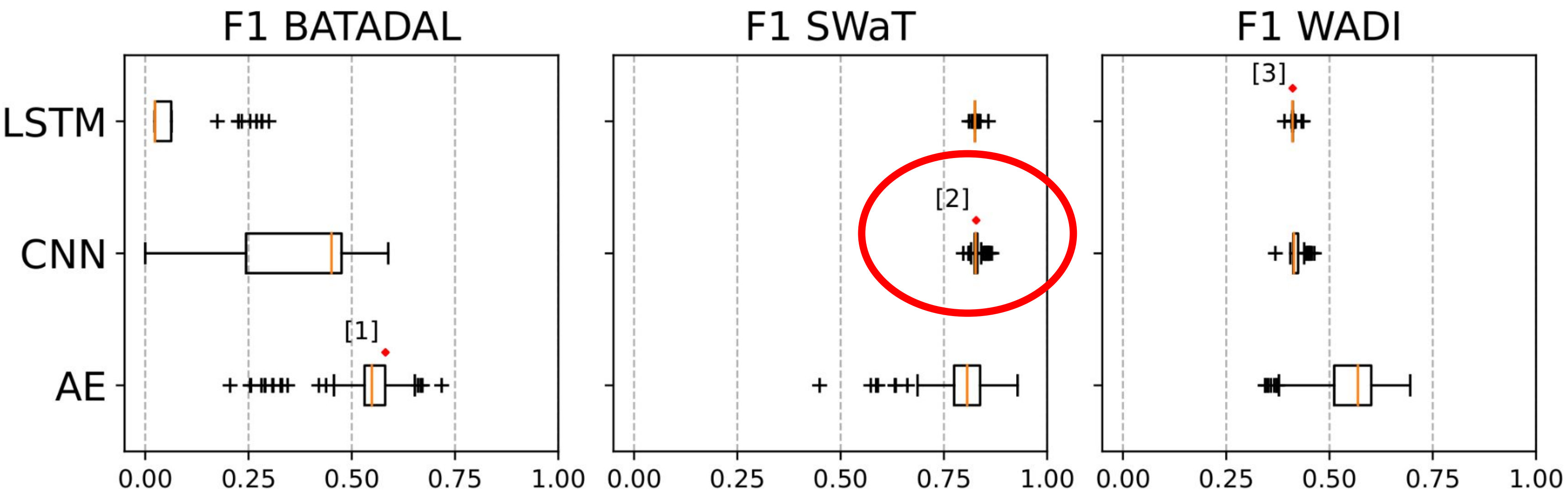


Is there a best model?

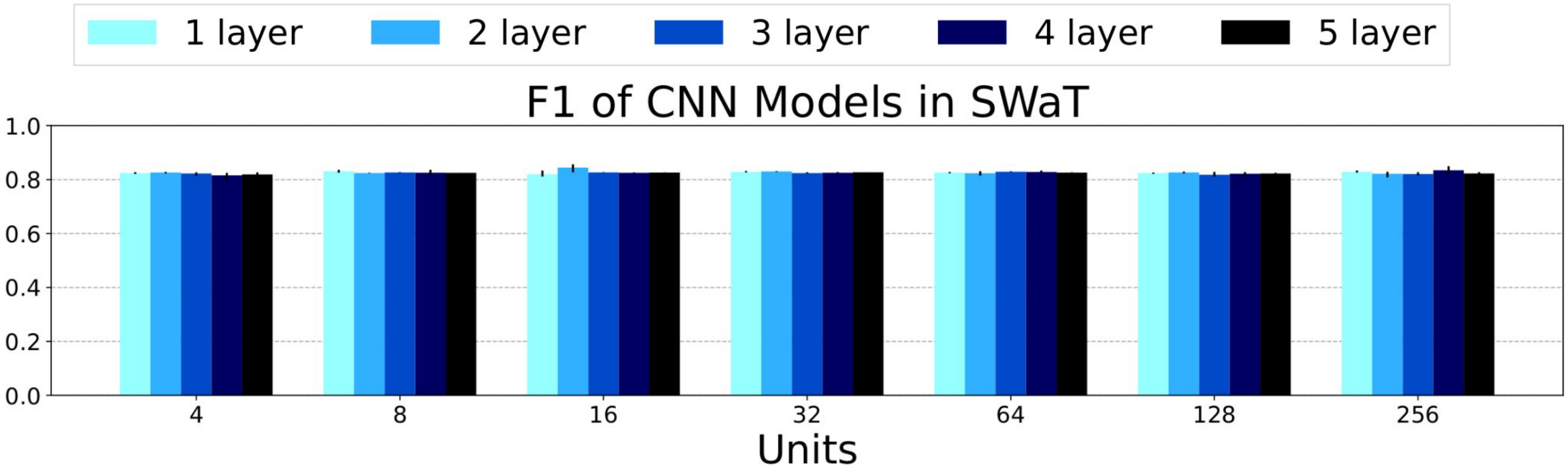


Prior "best" models perform similar to those we evaluated

How important is model hyperparameter tuning?



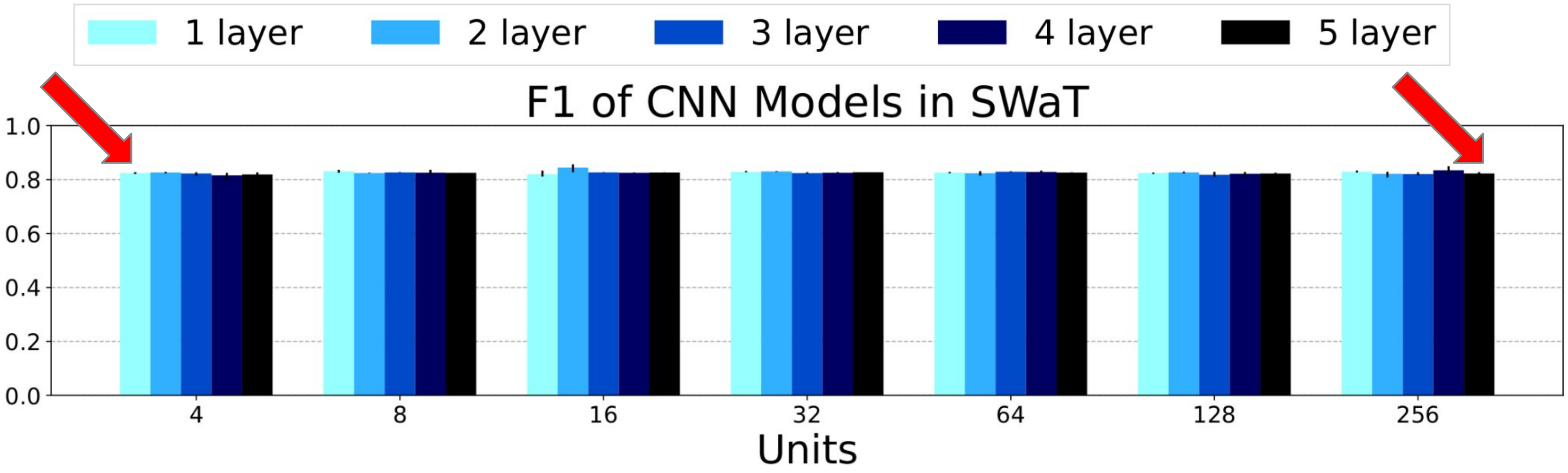
How important is model hyperparameter tuning?



How important is model hyperparameter tuning?

Smallest model: F1 = **0.824**

Largest model: F1 = **0.823**

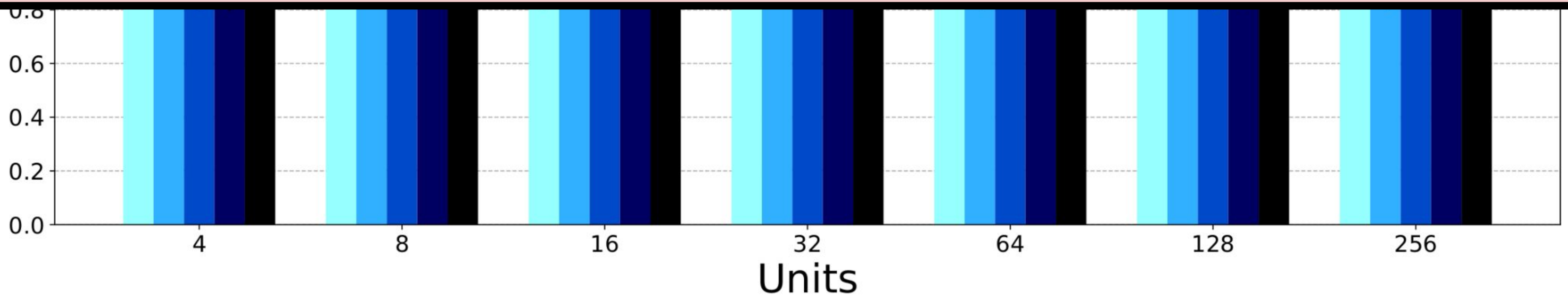


How important is model hyperparameter tuning?

Smallest model: F1 = **0.824**

Largest model: F1 = **0.823**

Model hyperparameter tuning has a limited impact on performance



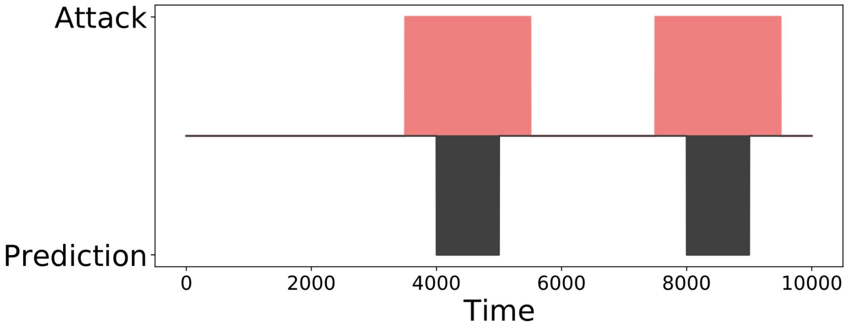
Part 2

How do **range-based metrics** affect tuning and optimization?

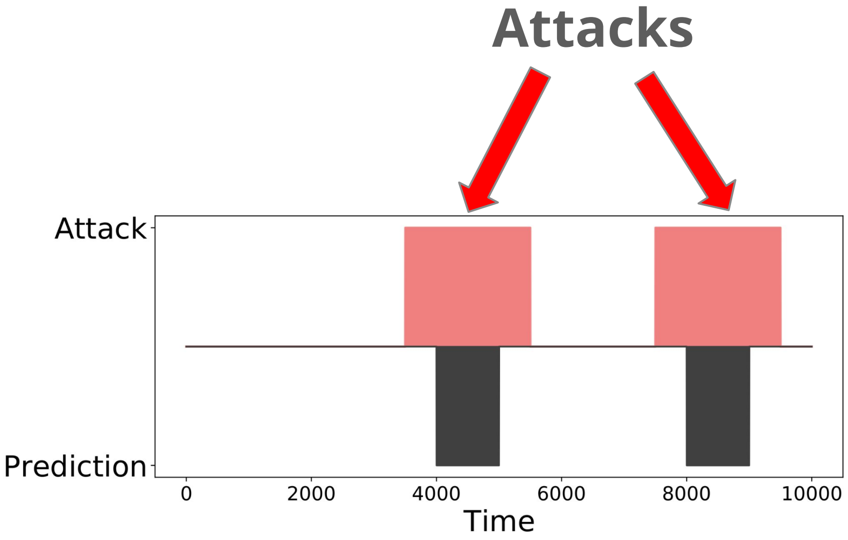
Point-F1: a common metric in ICS anomaly detection

- Point-F1 = Average between precision and recall
  - Each instance in time is equally weighted
- But attacks and predictions are **performed in segments**

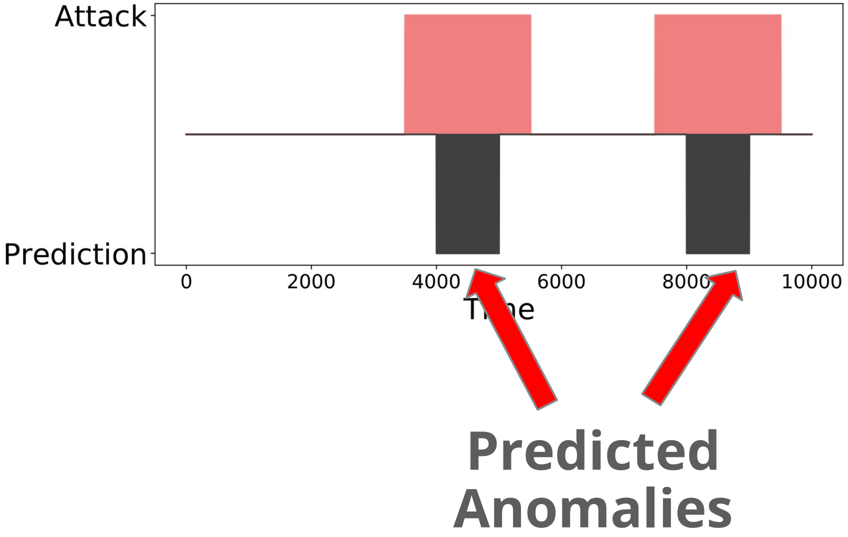
Point-F1 does not capture segment-based objectives



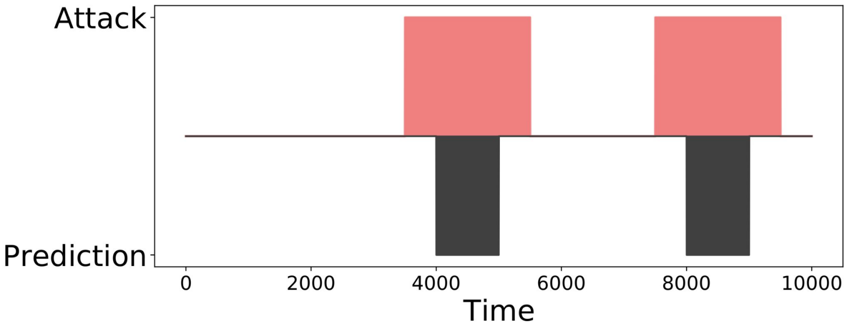
Point-F1 does not capture segment-based objectives



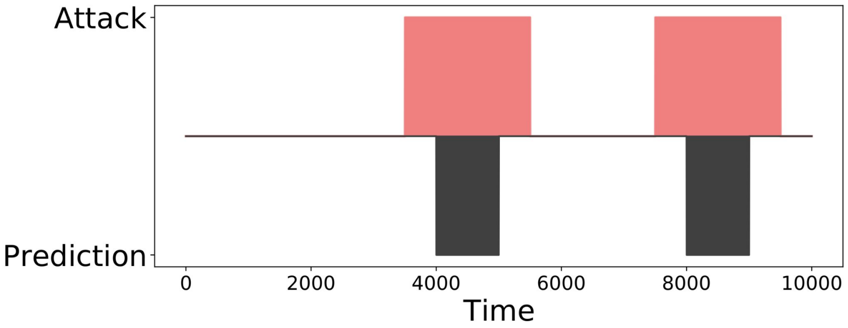
Point-F1 does not capture segment-based objectives



Point-F1 does not capture segment-based objectives

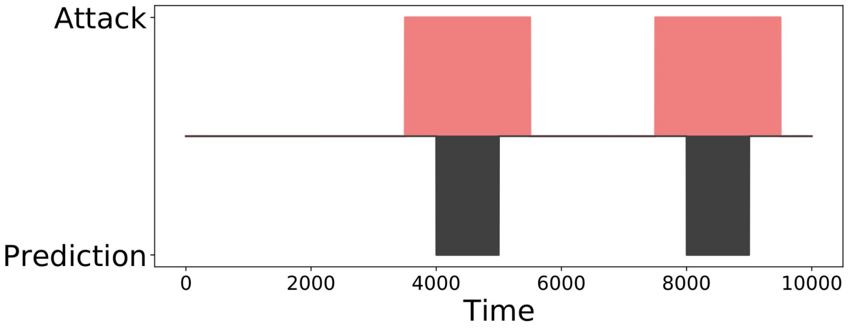


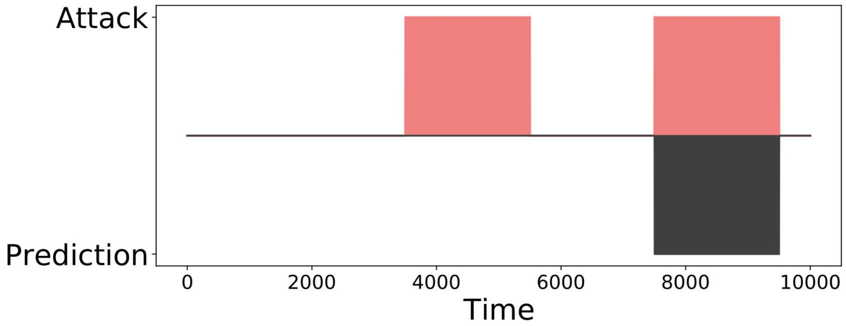
Point-F1 does not capture segment-based objectives



- Both attacks partially detected

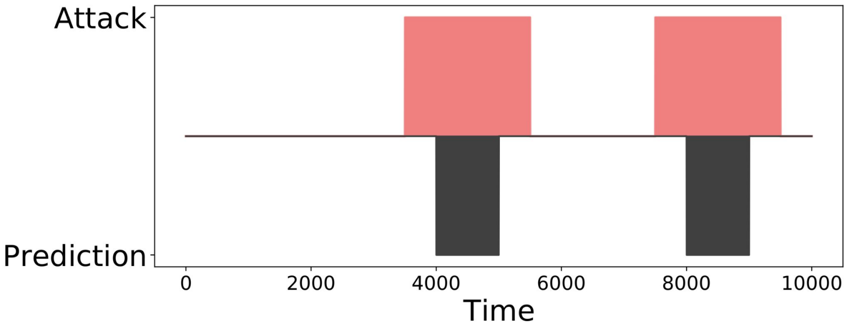
Point-F1 does not capture segment-based objectives



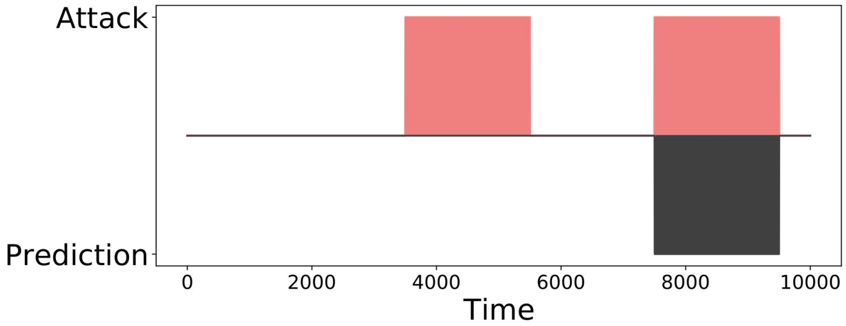


- Both attacks partially detected

Point-F1 does not capture segment-based objectives

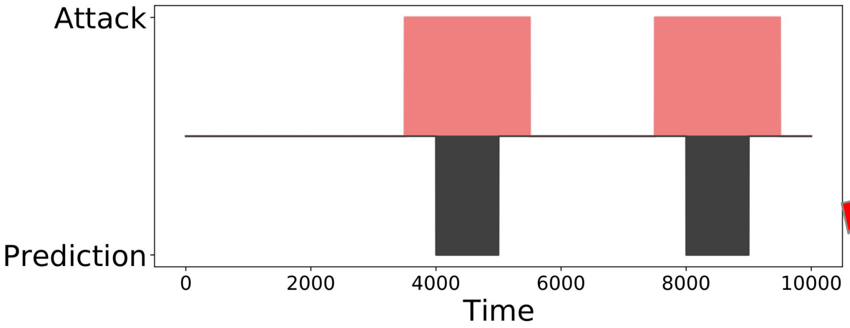


- Both attacks partially detected



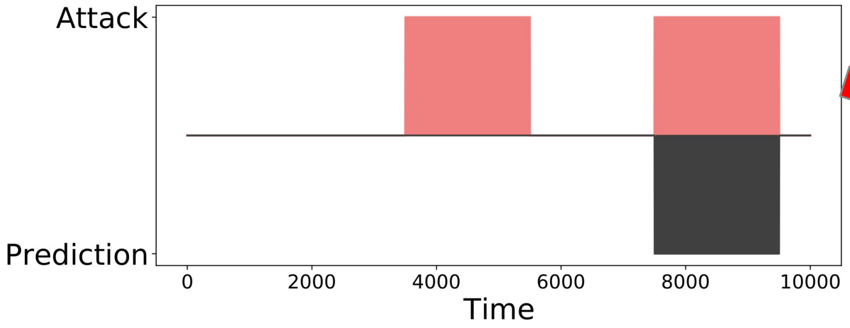
- One attack completely missed
- One attack fully detected

Point-F1 does not capture segment-based objectives



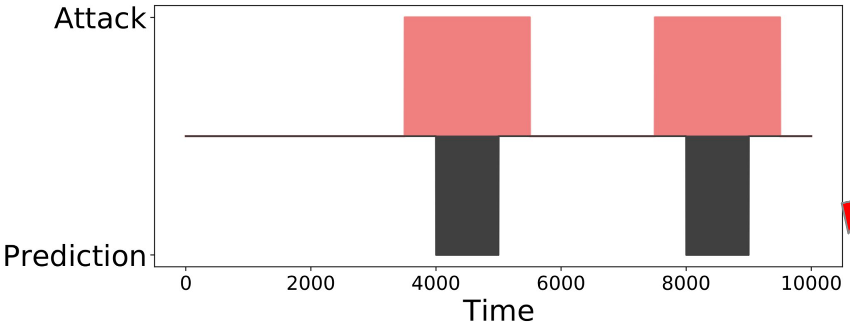
- Both attacks partially detected

Point-F1 = 0.75



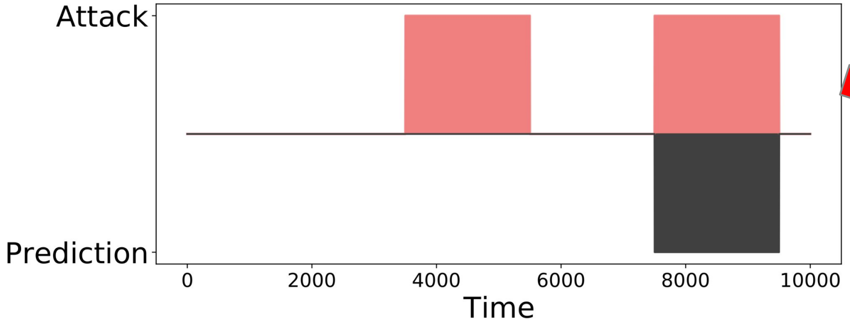
- One attack completely missed
- One attack fully detected

Point-F1 does not capture segment-based objectives



- Both attacks partially detected

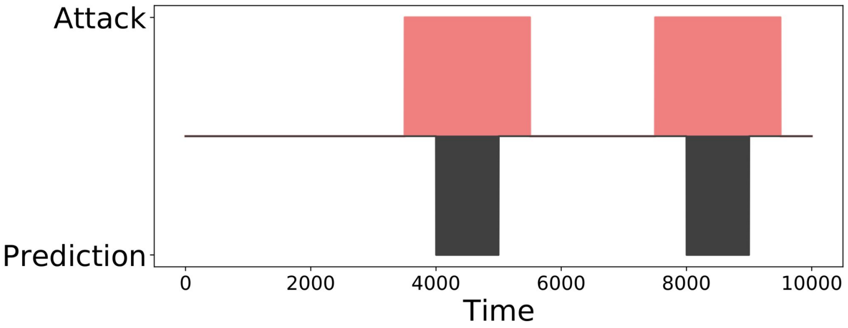
Point-F1 = 0.75



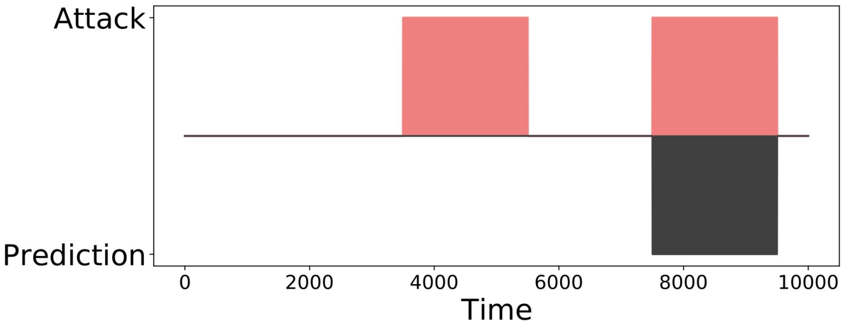
- One attack completely missed
- One attack fully detected

Same point-F1 score, but different outcomes!

Which objectives are important?



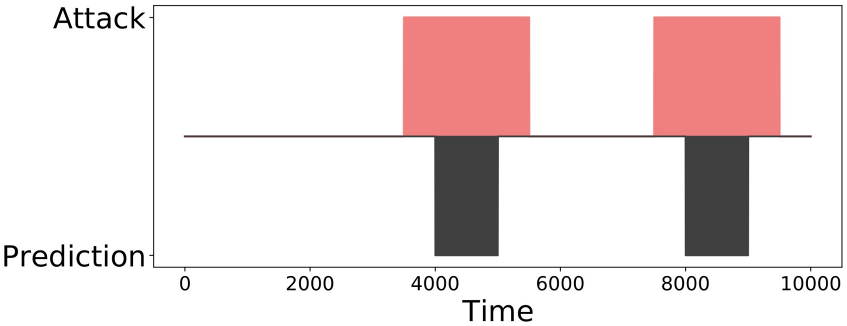
vs.



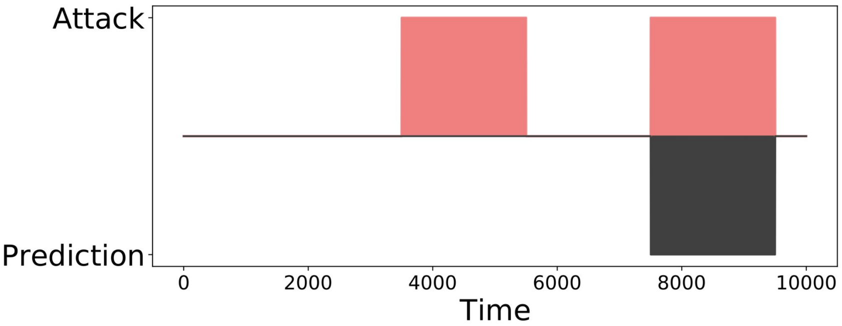
Detect every attack? **Detect all of attack?**

[1] Tatbul, N., Lee, T.J., Zdonik, S., Alam, M., Gottschlich, J.: Precision and recall for time series. *NeurIPS 2018*.

Which objectives are important?



vs.

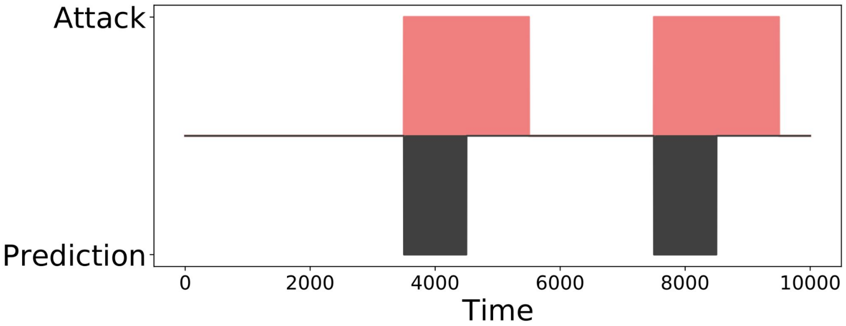


Detect every attack? **Detect all of attack?**

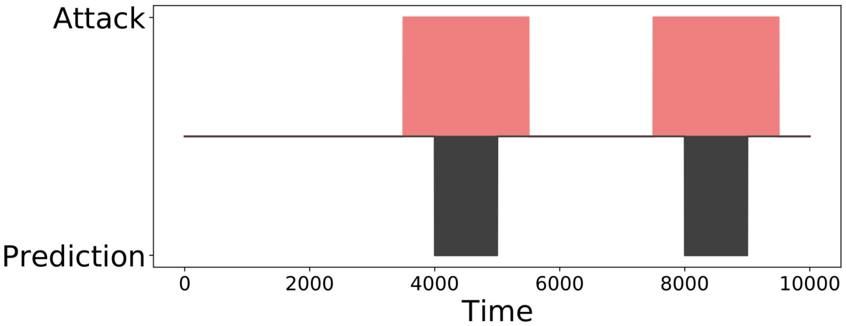
- Detected **segments**, instead of detected timesteps
  - Captured by time-aware precision and recall metric [1]

[1] Tatbul, N., Lee, T.J., Zdonik, S., Alam, M., Gottschlich, J.: Precision and recall for time series. *NeurIPS 2018*.

Which objectives are important?



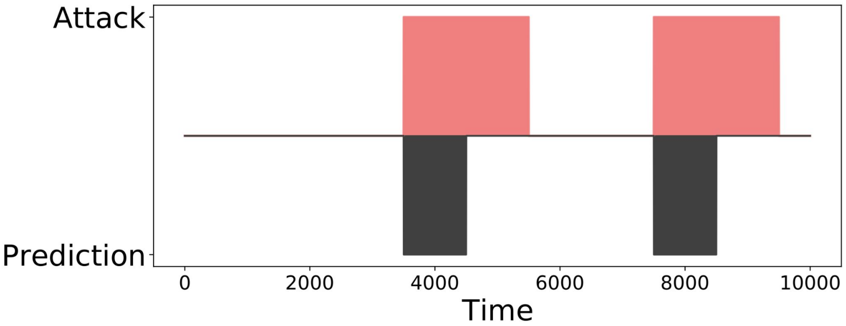
vs.



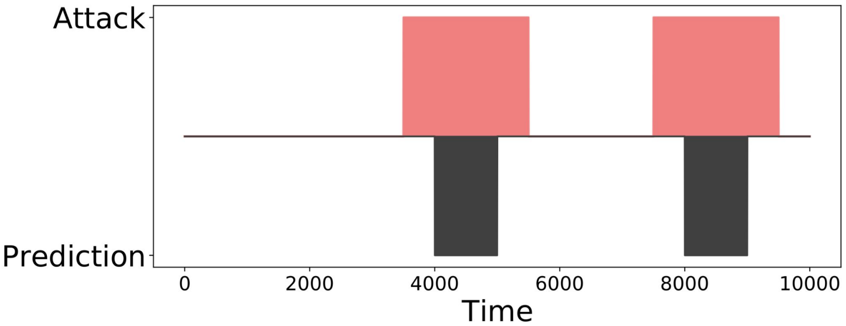
Detect immediately? **Detect later?**

[1] Tatbul, N., Lee, T.J., Zdonik, S., Alam, M., Gottschlich, J.: Precision and recall for time series. *NeurIPS 2018*.
[2] Lavin, A., Ahmad, S.: Evaluating real-time anomaly detection algorithms–the Numenta anomaly benchmark. *14th International Conference on Machine Learning and Applications.* (2015)

Which objectives are important?



vs.



Detect immediately? **Detect later?**

- Detecting attacks **earlier**, rather than later
  - Captured by Numenta metric [2]

[1] Tatbul, N., Lee, T.J., Zdonik, S., Alam, M., Gottschlich, J.: Precision and recall for time series. *NeurIPS 2018*.
[2] Lavin, A., Ahmad, S.: Evaluating real-time anomaly detection algorithms–the Numenta anomaly benchmark. *14th International Conference on Machine Learning and Applications.* (2015)

New training and evaluation methodology

1 Pre-process ICS dataset

Datasets: SWaT, WADI, BATADAL
Key techniques:

- *Benign data shuffling*
- *Feature selection*
- *Attack cleaning*

2 Train unsupervised ML model

CNN, LSTM: 1-5 layers, 4-256 units, 50-200 history
AE: 1-5 layers, 1.5-4.0 compression
Key technique: *Early stopping*

3 Tune threshold

MSE threshold $\boldsymbol{\tau}$, window length **w**
objective: maximize ~~point-F1 score~~ range-F1/Numenta scores

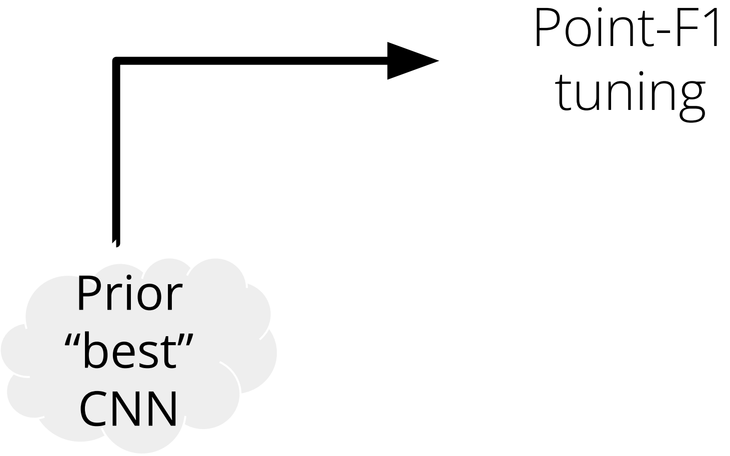
4 Evaluate against attacks at test time

Report ~~final point-F1 score~~: attack precision, attack recall, early detection, range-F1

Flexible tunings with range-based metrics

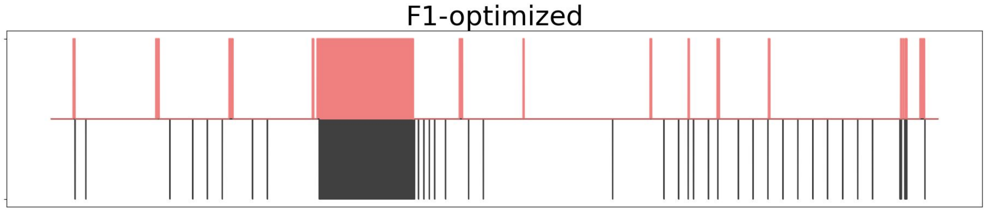
Prior
“best”
CNN

Flexible tunings with range-based metrics



Flexible tunings with range-based metrics

Point-F1
tuning



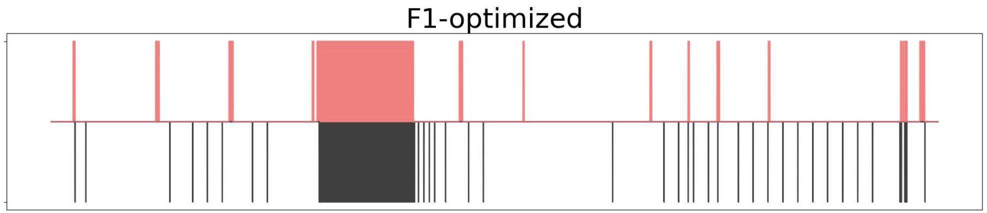
6/18 detected, 32 false alarms

Prior
“best”
CNN

Flexible tunings with range-based metrics

Prior "best" CNN

Point-F1 tuning



6/18 detected, 32 false alarms

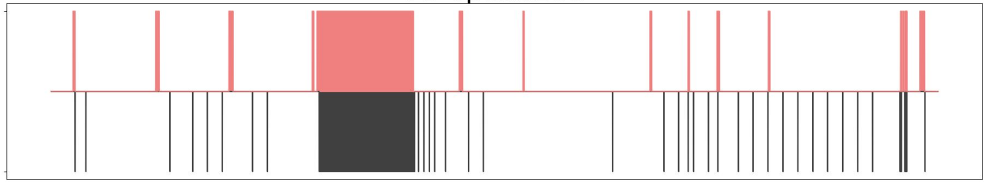
Range-F1 tuning (high precision)

Flexible tunings with range-based metrics

Prior "best" CNN

Point-F1 tuning

F1-optimized



6/18 detected, 32 false alarms

Range-F1 tuning (high precision)

RangeF1 high-precision-optimized

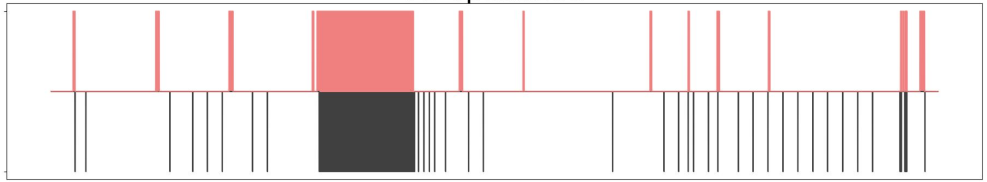
3/18 detected, **no false alarms**

Flexible tunings with range-based metrics

Prior "best" CNN

Point-F1 tuning

F1-optimized



6/18 detected, 32 false alarms

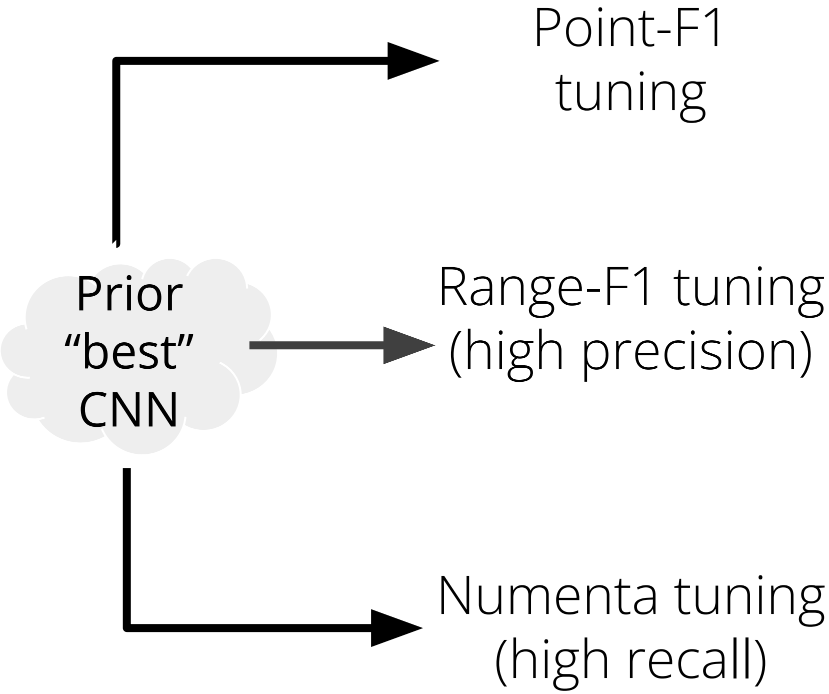
Range-F1 tuning (high precision)

RangeF1 high-precision-optimized

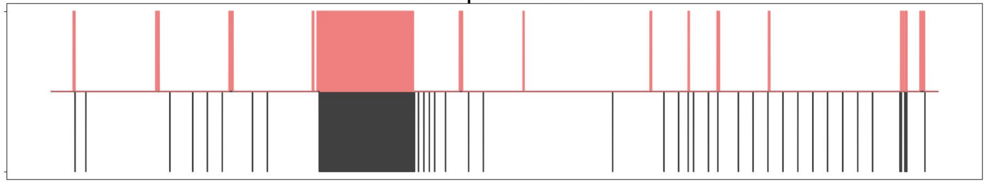
3/18 detected, **no false alarms**

Numenta tuning (high recall)

Flexible tunings with range-based metrics

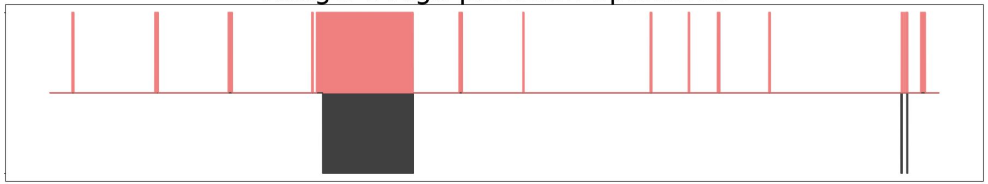


F1-optimized



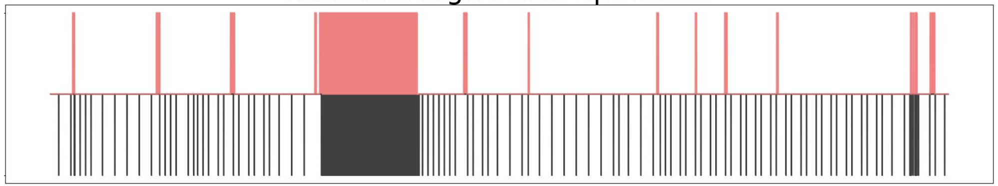
6/18 detected, 32 false alarms

RangeF1 high-precision-optimized



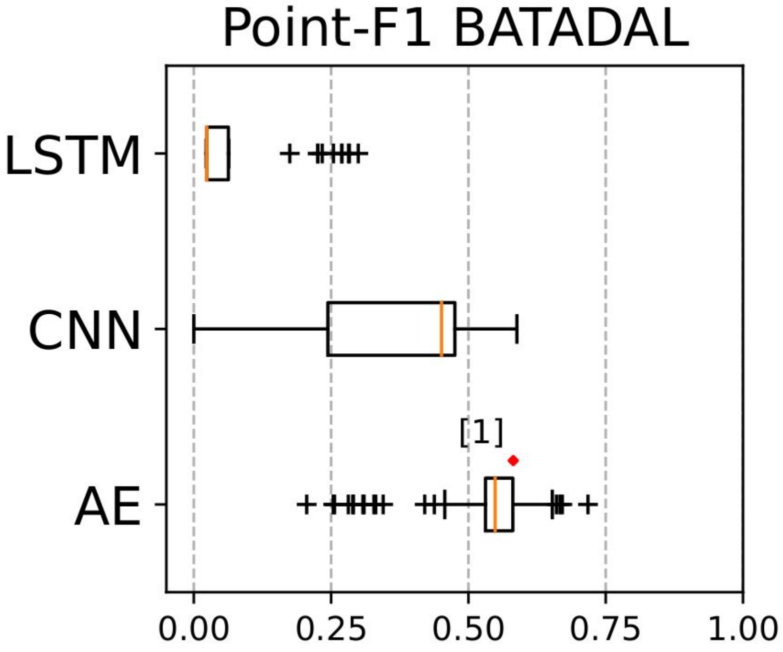
3/18 detected, **no false alarms**

Numenta high-recall-optimized

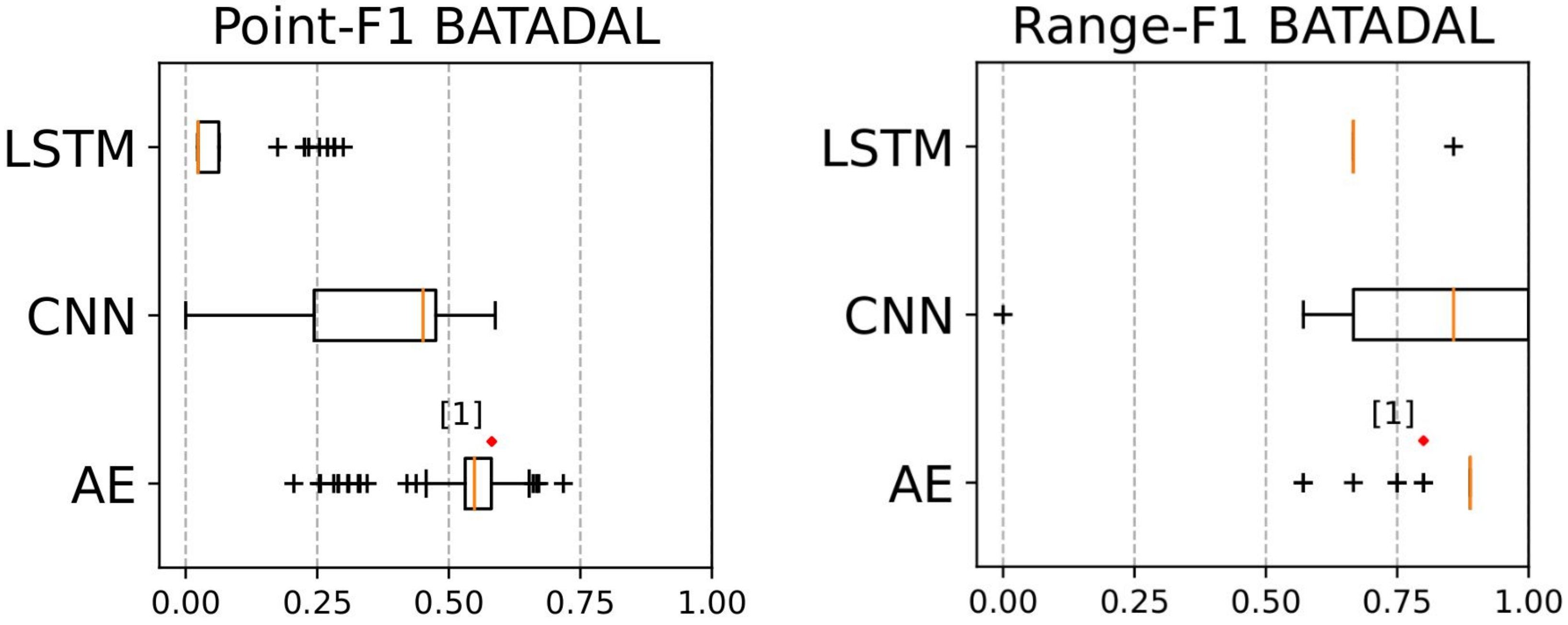


11/18 detected, 89 false alarms
7 early detections

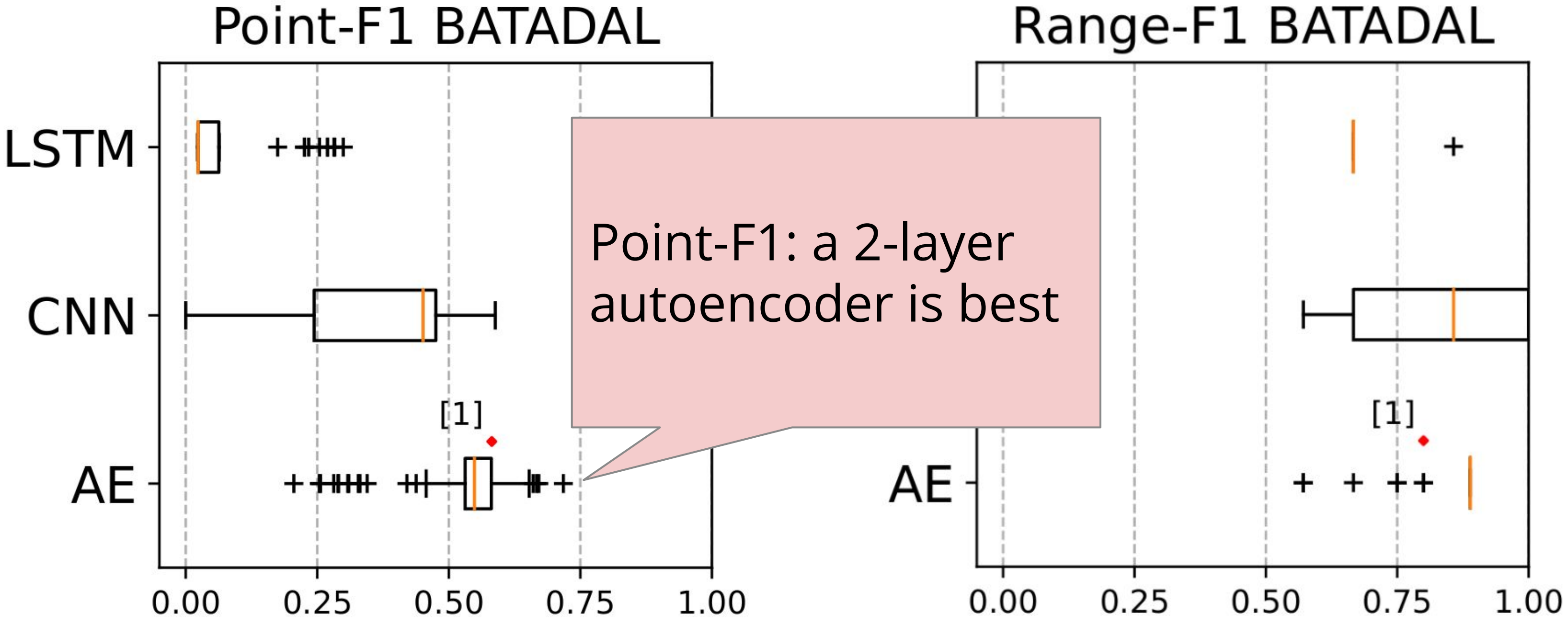
Range-based metrics choose different optimals



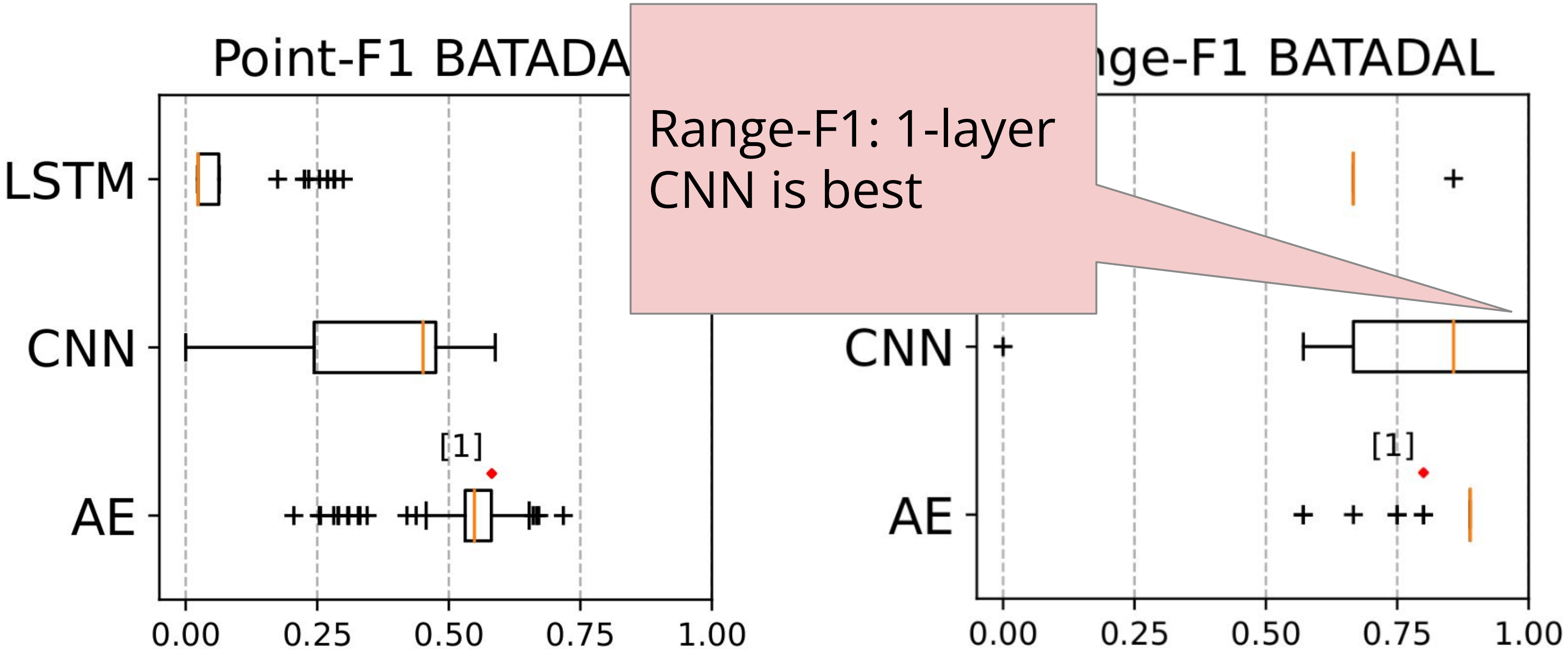
Range-based metrics choose different optimals



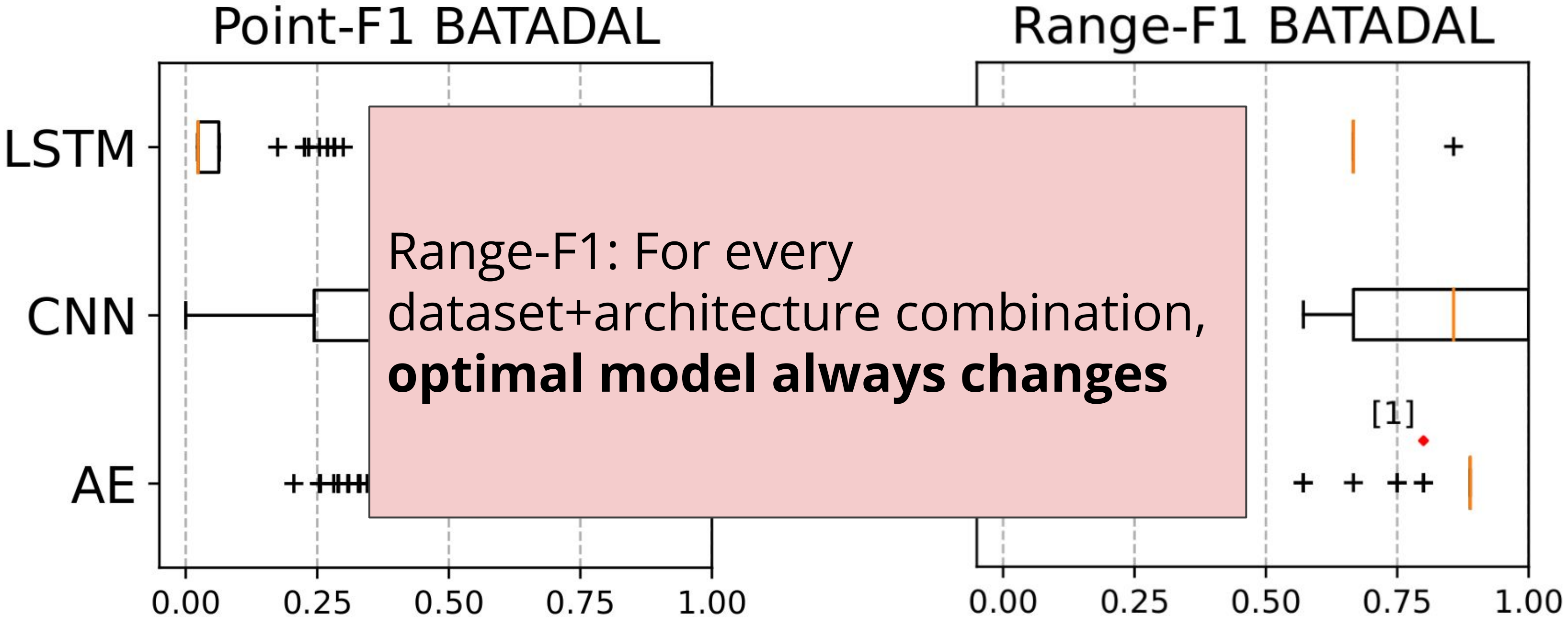
Range-based metrics choose different optimals



Range-based metrics choose different optimals

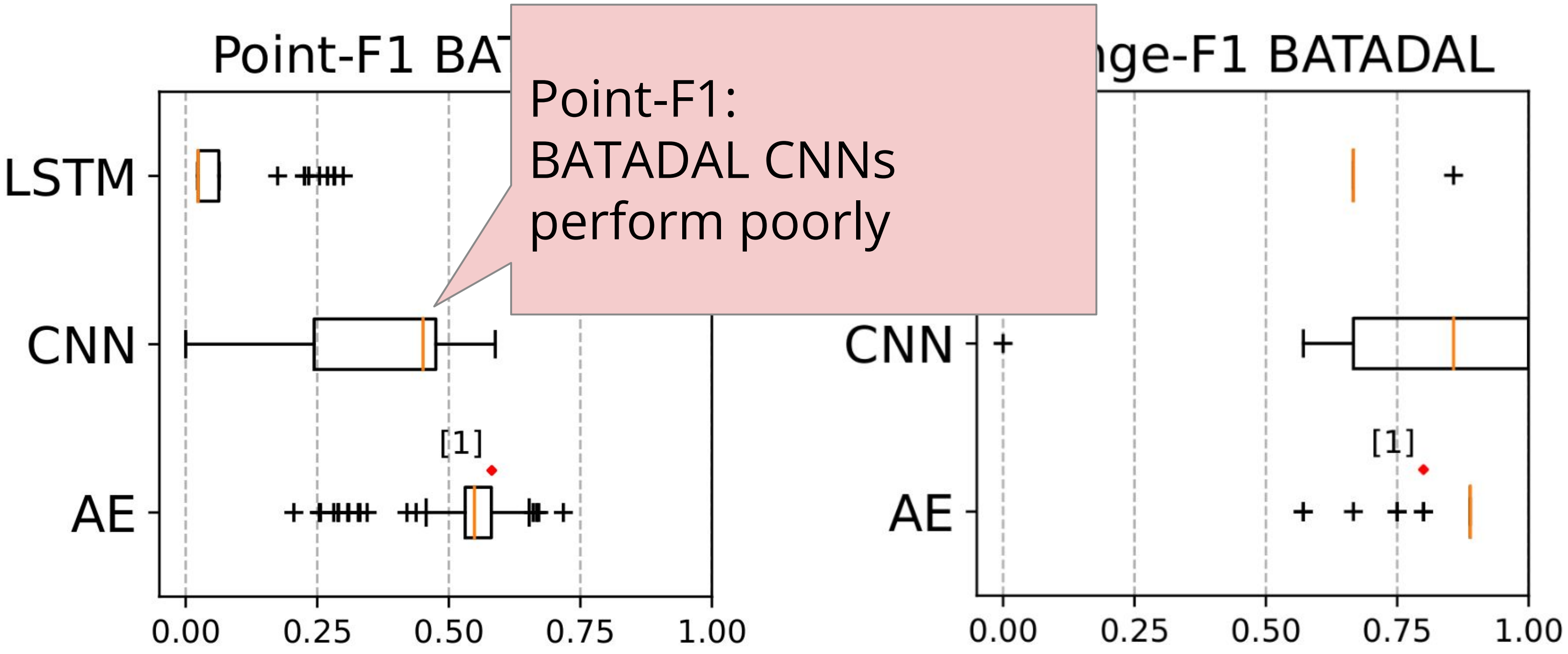


Range-based metrics choose different optimals

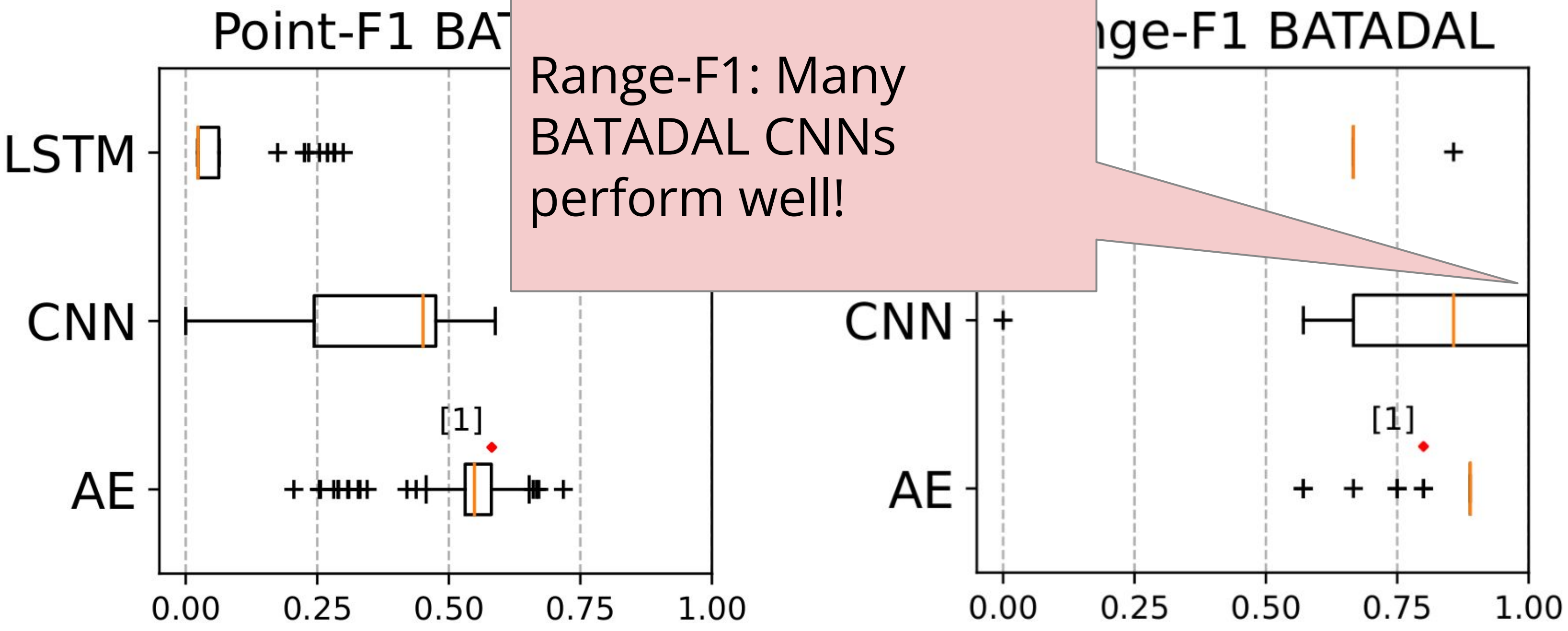


Range-F1: For every dataset+architecture combination, **optimal model always changes**

Range-based metrics choose different optimals

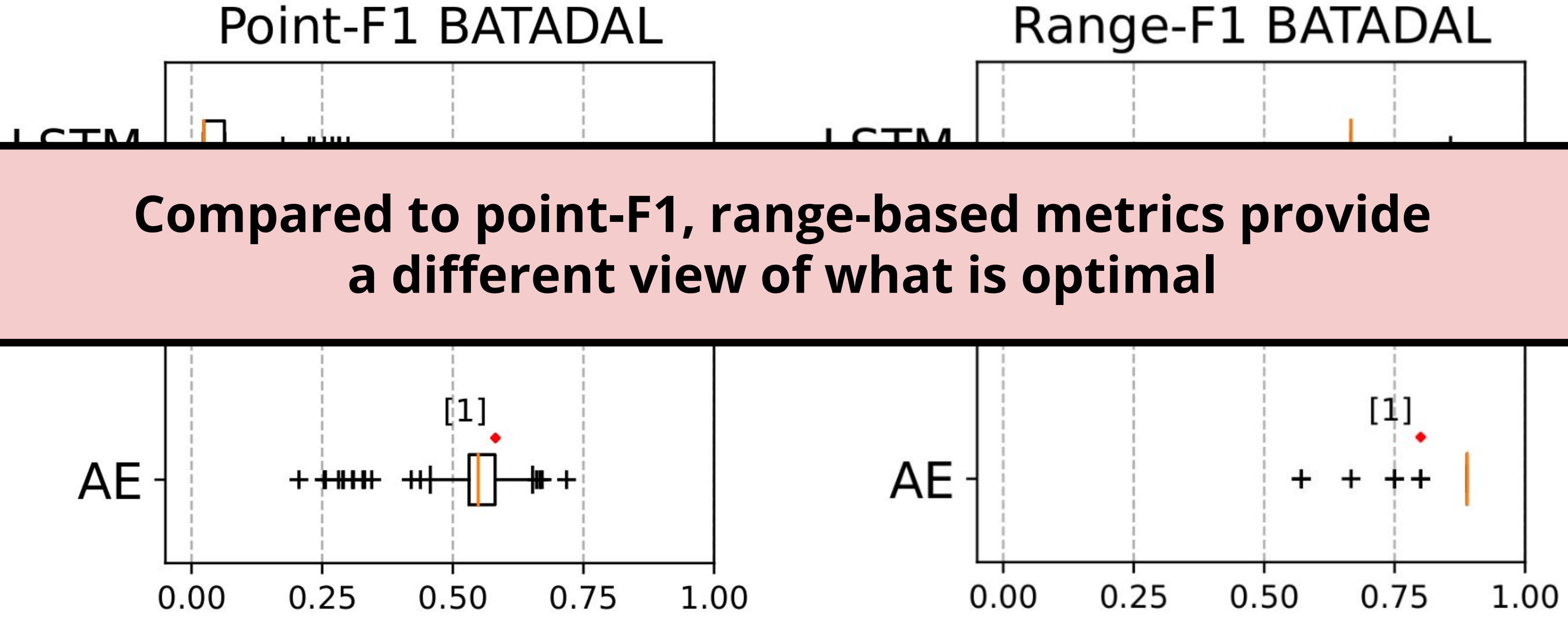


Range-based metrics choose different optimals



Range-F1: Many BATADAL CNNs perform well!

Range-based metrics choose different optimals



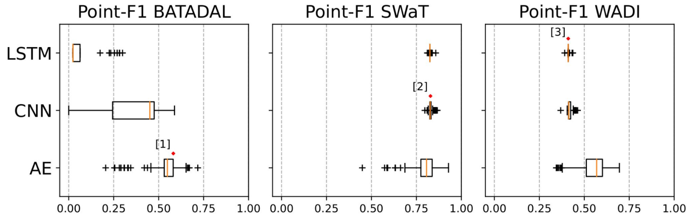
Compared to point-F1, range-based metrics provide a different view of what is optimal

In summary

- We ask: **what are the best models** for ICS anomaly detection?
  - Establish a common methodology for **fair comparison**

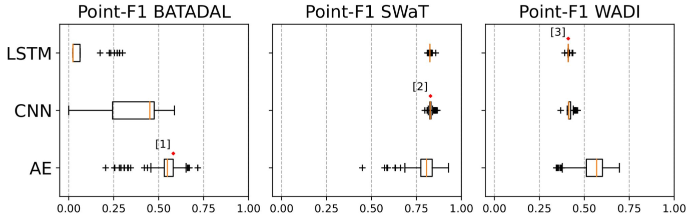
In summary

- We ask: **what are the best models** for ICS anomaly detection?
  - Establish a common methodology for **fair comparison**
  - Identify four key preprocessing/training techniques
  - Find that **small models** are often just as effective!



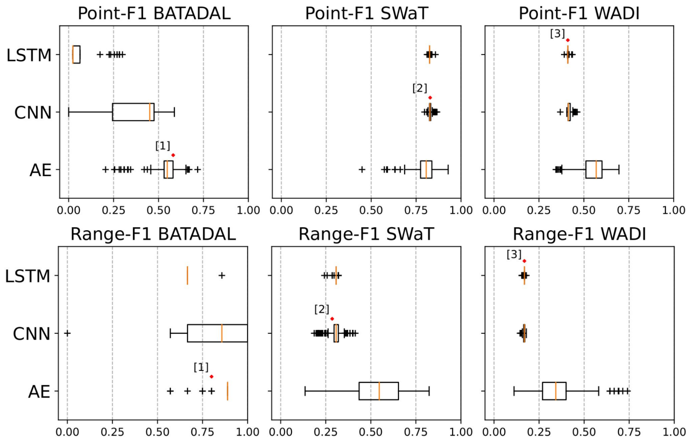
In summary

- We ask: **what are the best models** for ICS anomaly detection?
  - Establish a common methodology for **fair comparison**
  - Identify four key preprocessing/training techniques
  - Find that **small models** are often just as effective!
- **Point-F1 unfit** for time-series data



In summary

- We ask: **what are the best models** for ICS anomaly detection?
  - Establish a common methodology for **fair comparison**
  - Identify four key preprocessing/training techniques
  - Find that **small models** are often just as effective!
- **Point-F1 unfit** for time-series data
- **Range-based metrics** better measure usefulness of ICS anomaly detection



In summary

- We ask: **what are the best models** for ICS anomaly detection?
    - Establish a common methodology for **fair comparison**
    - Identify four key preprocessing/training techniques
    - Find that **small models** are often just as effective!
- **Point-F1 unfit** for time-series data
- **Range-based metrics** better measure usefulness of ICS anomaly detection

Perspectives from a Comprehensive Evaluation of Reconstruction-based Anomaly Detection in ICS

Clement Fung, Shreya Srinarasi, Keane Lucas, Hay Bryan Phee, Lujo Bauer

Carnegie Mellon University

Contact: clementf@cs.cmu.edu

Code: github.com/pwwl/ics-anomaly-detection