A White-Box Adversarial Attack Against a Digital Twin

Wilson Patterson, Ivan A. Fernandez, Subash Neupane, Milan Parmer, Sudip Mittal, Shahram Rahimi

Department of Computer Science & Engineering, Mississippi State University

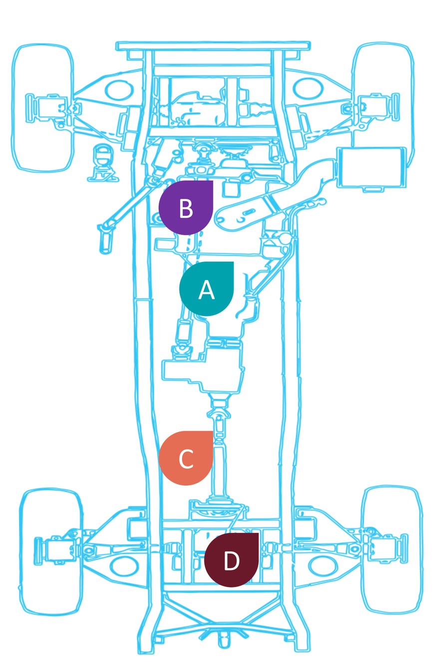
wep104@msstate.edu, iaf28@msstate.edu, sn922@msstate.edu, parmar@cse.msstate.edu, mittal@cse.msstate.edu, rahimi@cse.msstate.edu

MISSISSIPPI STATE UNIVERSITY

Introduction

- Cyber-attacks have evolved to become more effective against **Cyber Physical Systems (CPS).**
- This creates a new area of concern, as cyber-threats can potentially disrupt physical assets digitally. Example include **Stuxnet Attack, Colonial Pipeline attack.**
- Reliance on **Digital Twin (DT)** devices for automotive, military, and medical functions has increased the potential risk of adverse effects, if compromised.
- We demonstrate a white-box adversarial attack against our DT system using a **Machine Learning (ML)** model as a proof-of-concept.

Sensor Channels



A Transmission
- VehSpeedEng
- TransOilTemp
- TrSelGr
- TrOutShaftSp
- TransTorqConvLockup-Engaged

C Fuel
- FuelRate
- InstFuelEco
- InjCtlPres

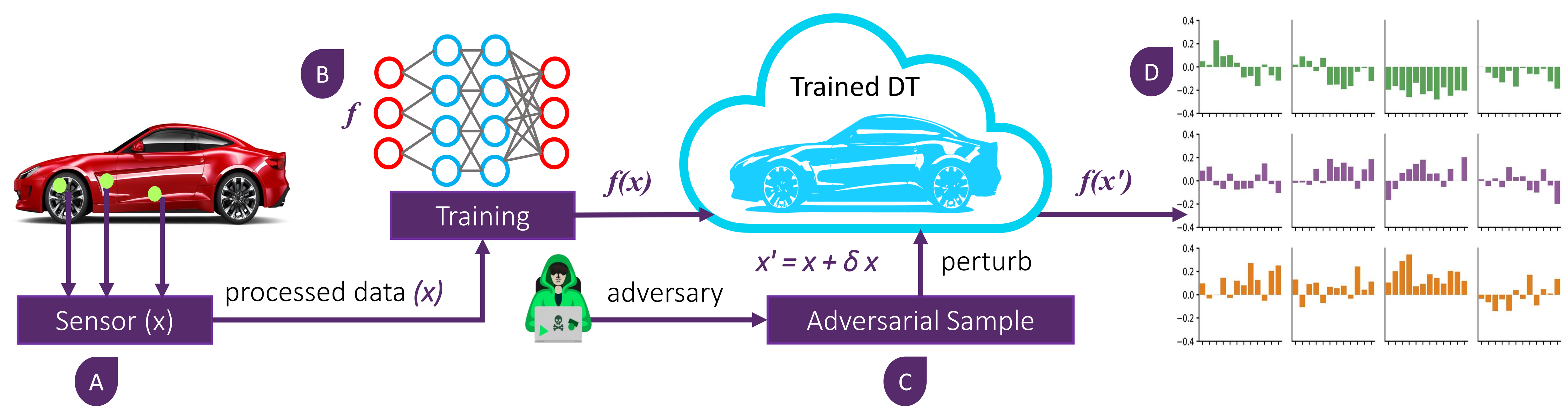
B Engine
- AccelPedalPos
- IntManfTemp
- PctEngLoad
- EngPctTorq
- BoostPres
- EngCoolantTemp

D Brake
- BrakeSwitch

Approach

- Data-driven DT are an important concept in the automotive industry because of their **predictive abilities**.
- The **robustness** of these DT models against **adversarial input samples*** should be thoroughly tested.
- We demonstrate the vulnerability of these systems and how they can be targeted by an adversary in a **white-box attack scenario.**
- We **attack** a trained DT model with an adversarial sample and demonstrate how easily the classifier can be **tricked** into misclassifying normal observation as an anomaly.

Adversarial Attack Architecture Against Trained Digital Twin



A Data Preprocessing

- We preprocess vehicular sensor channel data *(x)*.
- We consider four subsystems such as **Engine**, **Transmission**, **Fuel**, and **Brake**.
- 15 sensor channels are selected for training and modeling. For example, *Engine Coolant Temperature, Accelerator Pedal Position, Transmission Oil Temperature, Transmission Selected Gear, Fuel Rate, Injector Control Pressure, Brake Switch*, etc.
- Processed data *(x)* is then fed to neural network function *f*.

B Training A Digital Twin (DT)

- Deep Learning (DL) model is trained using preprocessed sensor channel data *(x)* as input.
- Resulting DL model is a trained DT.
- DT model can be used to:
  - ✓ Investigate, monitor, and forecast sensor channels behavior.
  - ✓ As well as predictive maintenance step.
- Abnormal sensor channels detected in this phase can be passed to expert for further analysis.

C Adversarial Sample

- The real-world adversary may alter input data *(x)* .
- Adversary can add just enough noise to the input to cause DT model to misclassify* normal observation as anomaly
- Misclassification step:
  - ✓ Input: Adversarial input sample* ($x' = x + \delta x$) is fed to trained model *f*.
  - ✓ Output: $f(x')$ is misclassified.
- An *adversarial sample is the small perturbation to the input *(x)* used by function *f* to predict or forecast that results in misclassified predictions.

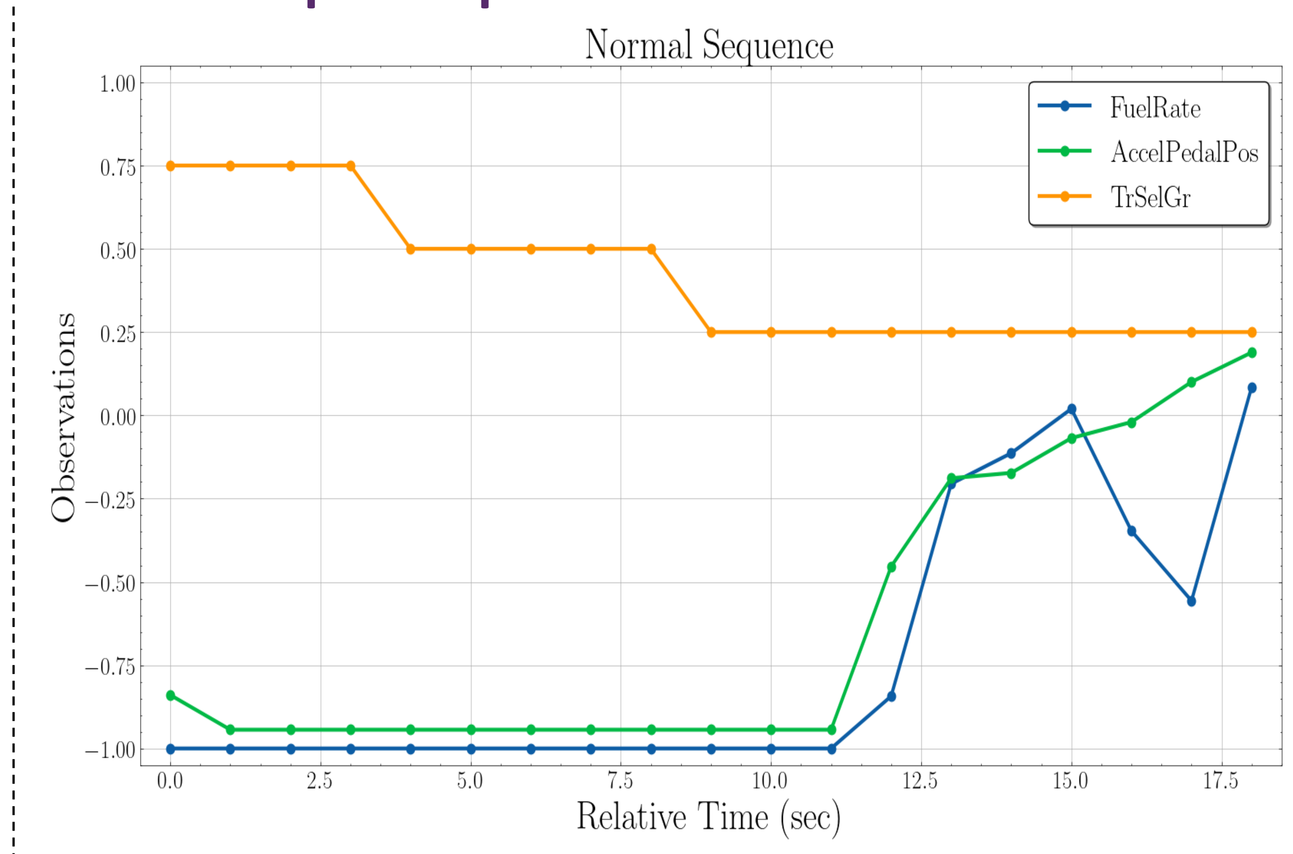
D Proof of Concept Attack

- We randomly perturb the inputs to the trained model.
- Adversarial robustness of DT is checked in this step.
- Perturbation steps:
  - ✓ Calculate standard deviation (σ) for each sensor channel in input sequence *(x)*.
  - ✓ Gaussian white noise* is added to input sequence.
- Gaussian white noise* is determined by sampling from normal distribution centered at zero, $\mathcal{N}(0, \sigma^2)$.

Results

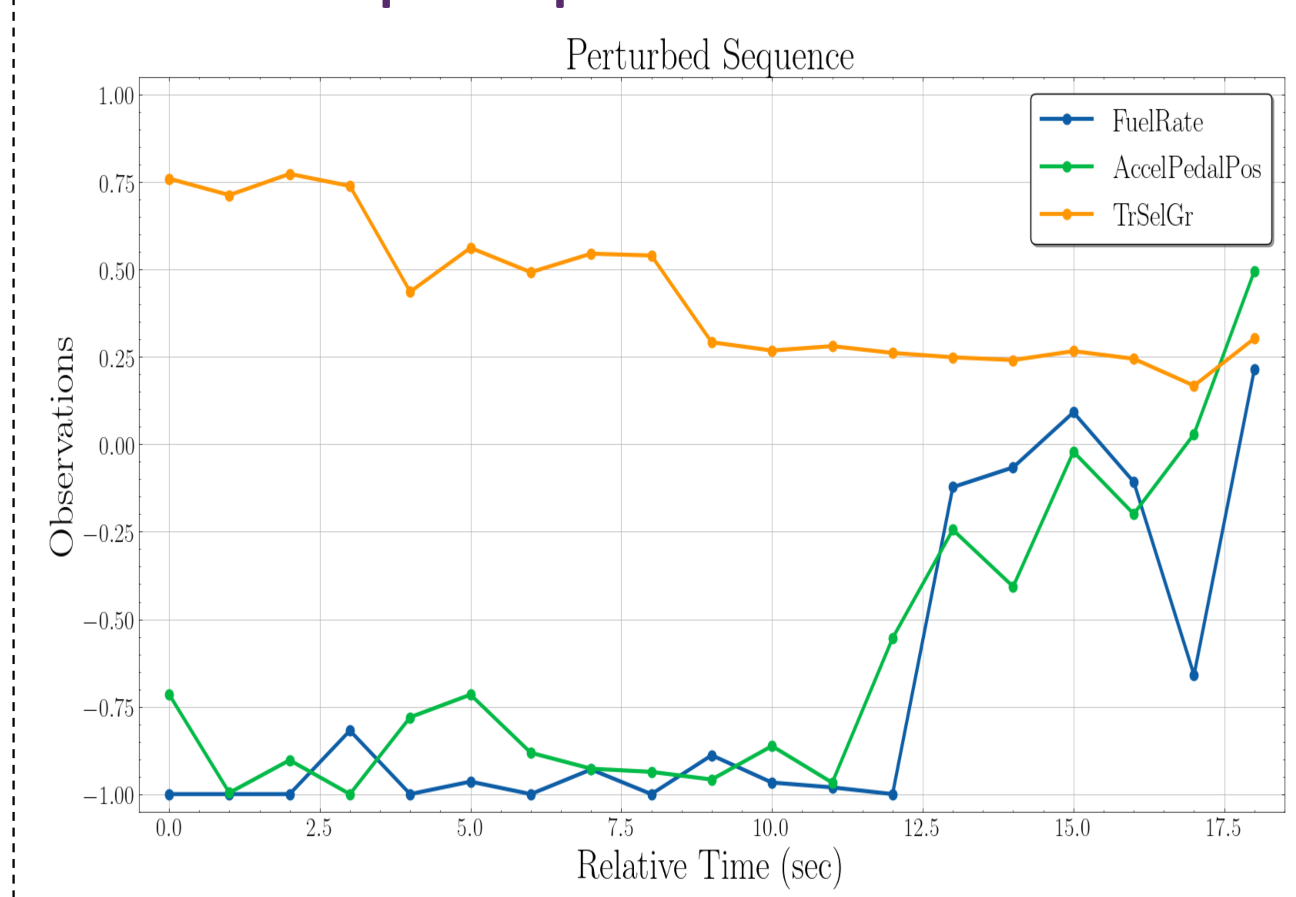
- Robustness of DT is evaluated against test sequences perturbed with white noise.
- Test sequences contain observations from all 15 sensor channels.
- Relationship between Fuel Rate, Accelerator Pedal Position, and Transmission Selected gear is investigated and perturbed.

Normal Input Sequence



- For normal input sequence the model predicts **NORMAL** with a Mahalanobis distance of 5.79.

Perturbed Input Sequence



- For perturbed input sequence the model predicts **ANOMALY** with a Mahalanobis distance of 9.72.

MISSISSIPPI STATE UNIVERSITY JAMES WORTH BAGLEY COLLEGE OF ENGINEERING

Department of Computer Science & Engineering
Bagley College of Engineering
https://cse.msstate.edu/