Securing Cyber-Physical Systems: Physics-Enhanced Adversarial Learning for Autonomous Platoons[*]

Guoxin Sun[1]✉, Tansu Alpcan[1], Benjamin I. P. Rubinstein[1], and Seyit Camtepe[2]

[1] University of Melbourne, Australia
{guoxins@student., tansu.alpcan@, brubinstein@}unimelb.edu.au
[2] CSIRO Data61, Australia, seyit.camtepe@data61.csiro.au

Abstract. The rapid development of cyber-physical systems in high-stakes safety-critical areas requires innovations in protecting them against malicious adversaries. Data-driven attack detection mechanisms based on deep learning (DL) have emerged as powerful tools to fulfil this need. However, it is well-known that adversarial attacks deceive DL models with specifically crafted perturbations added to clean data samples. This work combines cyber-physical system characteristics with DL to develop a hybrid attack detection system. Using knowledge from both physical dynamics and data, we defend against both cyber-physical attacks and adversarial attacks. This approach paves the way to use classical theories from the application domain to mitigate the deficiency of DL, complementing existing adversarial defence methods such as adversarial training. We implement our defence system for an autonomous vehicle platoon test-bed in a sophisticated simulator, where our approach doubles the detection F1 score and increases the minimum inter-vehicle distances compared to existing baselines. Hence, we greatly improve the safety and security of the target system against adversarially-masked cyber-physical attacks.

Keywords: Cyber-physical attacks · Adversarial machine learning · Autonomous platoons.

1 Introduction

Cyber-physical systems, where sensor networks and embedded computing are intertwined with the physical environment, are fast becoming a key driving force of today's economy. Such systems observe and interact with the changes in surrounding environments to achieve high levels of reliability and context-aware autonomy. As a paradigm and example of such a cyber-physical system, *autonomous vehicle platoons* attract attention with potentially improved driving

[*] We gratefully acknowledge support from the DSTG Next Generation Technology Fund and CSIRO Data61 CRP on 'Adversarial Machine Learning for Cyber' and CSIRO Data61 PhD scholarship.

experience and energy efficiency, reduced pollution as well as increased traffic throughput. This concept involves a string of vehicles travelling as a single unit from an origin to a destination. Each platoon member obtains other vehicles' dynamics and manoeuvre-related information through existing and emerging vehicle-to-vehicle communication networks and embedded sensors in order to adapt its own behaviour to maintain a narrow inter-vehicle distance and relative velocity.

The high levels of connectivity and open communication implementations highlight vehicle platoons as appealing targets for cyber-physical attacks causing degradation of their dependability or even catastrophic incidents. The potential impact of security vulnerabilities has motivated the development of attack detection methods. Due to the rapid development of deep learning (DL), researchers have shown an increased interest in applying data-driven techniques, especially in the form of deep neural networks, to study and classify the complex patterns of system behaviour. Although DL-based attack detection demonstrates excellent defence performance against conventional cyber-physical attacks, they are also known to be vulnerable to adversarial attacks, in which specifically crafted perturbations are added on top of clean data with the aim of evading detection.

Motivation and Problem. If a DL-based attack detection system fails to detect cyber-physical attacks masked with adversarial perturbations, then the system is exposed to a much wider range of safety risks, since any conventional attack can be masked to evade detection this way. Traditionally, physical systems have been designed and analysed with classical modelling techniques, which constitute the foundation of control theory. Although data-driven approaches are becoming dominant in many areas, those classical tools still have an important role to play in cyber-physical systems such as vehicle platoons.

In this context, recent work [9] generates adversarial attacks against an anomaly detector for a water treatment problem considering the effects of a 'rule-checker'. Yet, the rules are derived mainly from observations instead of physical laws from first principles. While [1] combine a data-driven algorithm - generalized Extreme Studentized Deviate (ESD) - with the physical laws of kinematics to perform real-time anomaly detection, they have not considered the effects of adversarial attacks in their work. Similarly, model-based approaches alone are not a 'silver bullet' to the cyber-physical security problem either. A well-educated attacker could derive and leverage the underlying system model to increase the level of stealthiness [6].

Novelty and Contributions. This paper presents a novel combination of engineering modelling techniques with DL and proposes a hybrid attack detection system using knowledge from both physical dynamics and data to defend against both cyber-physical attacks and adversarial attacks. This approach paves the way to use classical theories from the application domain to make up for the deficiencies of DL and vice versa. Our approach is also applied in combination with existing adversarial defence techniques such as adversarial training to further improve its robustness. The contributions of this paper include:

(1) We provide a novel physics-enhanced data-driven attack detection system for

cyber-physical systems that leverages knowledge from both data and physics.
(2) We illustrate that classical physics-modelling techniques can help to mitigate the deficiency of deep learning-based approaches, which extends the applicability of many state-of-the-art DL-based approaches for cyber-physical systems.
(3) As a demonstration, we successfully improve the security and dependability of vehicle platoons. Our defence system provides excellent detection performance against an informed white-box attacker.
(4) Our results are demonstrated both analytically and visually using sophisticated, system-level simulations. It outperforms standard baseline attack detection methods and proves the potential to be applied with existing adversarial defence techniques for better performance.

Related Work. Sumra et al. [18] provide a comprehensive survey of the cyber-physical attacks on major security goals, i.e., confidentiality, integrity and availability. For example, data integrity attacks corrupt the legitimacy of transmitted information, which allows malicious or Sybil vehicles to gain the privilege of the road or to cause traffic congestion and even serious collisions [3]. Malicious attackers may conduct eavesdropping attacks to steal and misuse confidential information [21].

In terms of data-driven learning-based attack detection approaches, [11] apply both feed-forward deep neural networks and convolutional neural networks to identify a malicious attacker who tries to cause collisions by altering the controller gains. [22] propose an ensemble model consisting of 4 tree-based algorithms to detect attacks against the Controller Area Network (CAN) bus. To better utilise the embedded temporal information within the time-series data from such systems, several attempts [2] have been made to solve the attack detection problem by examining the deviations of system behaviour and model predictions with machine learning models.

In the past, adversarial attacks have been extensively studied mainly in domains such as image and audio and far less attention has been paid to cyber-physical domains especially systems involved with time-series data. Existing research on the subject has also been mostly restricted to a few pre-generated datasets. For instance, [10] investigate the effects of adversarial attacks against time-series classifiers based on the UCR archive with data generated a posteriori of various types (e.g., motion, sensor etc.). In our work, we investigate the targeted cyber-physical system in various types of fringe and dangerous situations where the data and its corresponding adversarial examples are generated in an online fashion.

The vulnerabilities of DL models have motivated the development of adversarial defences. Adversarial training is a simple yet effective defence approach, which is to include adversarial examples directly as part of the training dataset [8]. Although the improved model is aware of adversarial examples in advance thereby more robust, the defender needs knowledge of the adversarial attacks and efforts to generate those examples a priori. Other defence methods including data distortion [23], defence distillation [14] have been proved to have their own advantages and limitations. Recently, physical knowledge has been exploited to enhance the training procedure or overall performance of neural networks in the targeted domain. Physical models of the underlying system become part of the loss function to bound the space of admissible solutions to the neural network parameters [5]. Nevertheless, few researchers have been able to draw any systematic study on incorporating physical knowledge for adversarial defence.

2 Problem Definition

A typical cyber-physical system (e.g., autonomous vehicles, smart grids, etc.) acquires necessary real-time information via onboard sensors or wireless communication with other parties. Malicious adversaries often target these communication networks and onboard sensors to destabilize or break down such safety-critical systems via cyber-physical attacks. If the system contains machine learning components, a range of adversarial attacks can be utilized by the attacker to perform so-called adversarially-perturbed cyber-physical attacks. The attacker's ultimate goal is to maximize physical damage while remaining stealthy.

This work presents a hybrid defence method that utilizes knowledge both from data and physics to address such security challenges. The data-driven component of our approach learns the complex physical dynamics of a real system purely from data when existing modelling techniques fail to model accurately and reliably. The physics component with a simple system model helps when learning-based methods suffer from adversarial perturbations. Specifically, the underlying system structure can be modelled by physical first principles with differential equations in the form of $\dot{x} = g(x)$, where x contains the states of the system. For example, the motion of autonomous vehicle platoons can be modelled by *the kinematic model* whereas power system dynamics can be modelled by *the swing equation* [20]. As a general defence framework, the physics part of our proposed defence framework could be substituted accordingly based on the underlying cyber-physical system. The deep-learning model could also be replaced by feed-forward neural networks, convolutional neural networks etc. We would utilize the kinematic model for vehicle platoons as a case study in the rest of the work. To generalize from autonomous vehicles to the smart grids, for example, one can replace the kinematic model used by our physics component with the swing equation.

3 Attacker Model

As a paradigm of cyber-physical attacks, false data injection corrupts the content of wirelessly transmitted messages or sensor observations to cause performance degradation or catastrophic failure of safety-critical systems. In the present work, we consider two attack approaches as presented in Sect. 3.1 and Sect. 3.2.

3.1 Conventional Cyber-Physical Attacks

Vanilla False Data Injection Attack. We extend the message falsification attacks [2,19] from only affecting communication messages to attacking both communication and sensor observations [4] in a subtle way and name it vanilla false data injection (v-FDI). In particular, the adversary progressively increases the attack intensity to achieve its malicious objectives (e.g., causing collisions) while evading detection. Take acceleration modification in the vehicle platoon case as an example, the modified acceleration value is similar to the original one at the beginning of the attack. As attack effects progressively build up, it might become too late for the defence system to react since the attack may have already led to limited response time or even collision.

Model-Aware False Data Injection Attack. Model-aware false data injection (m-FDI) can be seen as an evolved version of its vanilla counterpart. Instead of injecting arbitrary modifications, the adversary utilizes the knowledge of the underlying system model to conduct malicious modifications concurrently on a range of observations. Following the acceleration modification example, as the acceleration modification progressively increases, the attacker computes the resulting velocity and position quantities based on the system model and injects velocity and position modifications accordingly. In this way, the modified data is consistent with the underlying system model (i.e., the kinematic model) thereby increasing its stealthiness level and attack strength. We will show in later sections how this type of attack can bypass model-based detection methods but not ours.

3.2 Adversarially-Masked Cyber-Physical Attacks

While attacking the cyber-physical systems with conventional cyber-physical attacks, intelligent adversaries may also create carefully-crafted adversarial perturbations to deceive DL-based attack detection systems. In contrast to conventional adversarial attacks against classifiers, we investigate similar attack methods but applied against regression models in cyber-physical domains. Inspired by the linear behaviour of modern machine learning models, the basic iterative method (BIM) [13] uses the first-order information of the loss function and generates adversarial examples iteratively. It is adopted in our work because of its improved attack performance with an even reduced perturbation level compared to other gradient-based attack methods such as the fast gradient sign method.

3.3 Attacker Capabilities

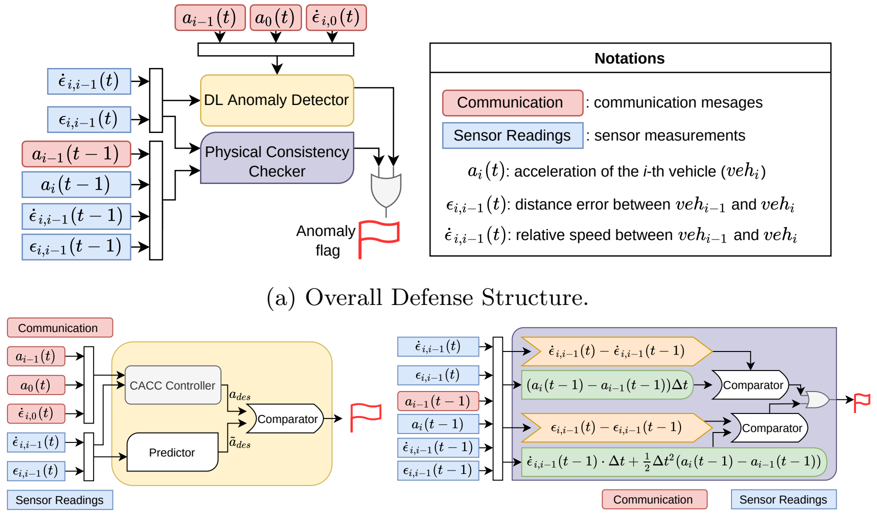
For simplicity, we consider only the dynamic information of a single vehicle (e.g., the preceding vehicle) will be modified by the attacker, which includes the transmitted acceleration messages via wireless communication as well as velocity and position information measured by a rangefinder (e.g., radar). Different levels of a priori knowledge (e.g., white-box knowledge of the controller, the DL-based attack detection system, the underlying system model as well as full access to the onboard memory) are assumed to conduct different types of malicious attacks, which are summarized in Table 1. Attack abbreviation followed by (adv. masked) denotes that adversarial perturbations are added to deceive a machine learning model which in our case is a DL-based anomaly detector.

Table 1: Knowledge required by the attacker to conduct different attacks.

Access to Attack Types	Sensors	Communication	DL Model	System Model	Memory
v-FDI	✓	✓	✗	✗	✗
m-FDI	✓	✓	✗	✓	✗
v-FDI (adv. masked)	✓	✓	✓	✗	✓
m-FDI (adv. masked)	✓	✓	✓	✓	✓

4 Physics-Enhanced Defense Approach

The proposed defence system consists of a data-driven component powered by deep neural networks to detect conventional cyber-physical attacks (e.g., false data injections) and a physics-inspired component to assist in reporting adversarial perturbations to compensate for the deficiency of deep learning models. We apply this general approach to the specific case of autonomous vehicle platoons as an illustrative example. Figure 1a shows the overall structure of our proposed defence system along with the pseudo-code of this new Double-Insured Anomaly Detection (DAD) method presented in Algorithm 1.



$a_{i-1}(t)$
$a_0(t)$
$\dot{\epsilon}_{i,0}(t)$
$\dot{\epsilon}_{i,i-1}(t)$
$\epsilon_{i,i-1}(t)$
$a_{i-1}(t-1)$
$a_i(t-1)$
$\dot{\epsilon}_{i,i-1}(t-1)$
$\epsilon_{i,i-1}(t-1)$
DL Anomaly Detector
Physical Consistency Checker
Anomaly flag
Notations
Communication : communication mesages
Sensor Readings : sensor measurements
$a_i(t)$: acceleration of the i-th vehicle (veh_i)
$\epsilon_{i,i-1}(t)$: distance error between veh_{i-1} and veh_i
$\dot{\epsilon}_{i,i-1}(t)$: relative speed between veh_{i-1} and veh_i
Communication
CACC Controller
Predictor
a_{des}
$\tilde{a}_{des}$
Comparator
Sensor Readings
$\dot{\epsilon}_{i,i-1}(t)-\dot{\epsilon}_{i,i-1}(t-1)$
$(a_i(t-1)-a_{i-1}(t-1))\Delta t$
$\epsilon_{i,i-1}(t)-\epsilon_{i,i-1}(t-1)$
$\dot{\epsilon}_{i,i-1}(t-1)\cdot\Delta t+\frac{1}{2}\Delta t^2(a_i(t-1)-a_{i-1}(t-1))$
Comparator
Comparator
Communication
Sensor Readings

(a) Overall Defense Structure.

(b) Structure of the Deep Learning-based Anomaly Detector.

(c) Structure of the Physical Consistency Checker.

Fig. 1: The defence system applied to the vehicle platoon case study.

Algorithm 1 Double-Insured Anomaly Detection (DAD)

Input: Communication messages S and sensor readings R
Output: Anomaly flag

```
Initialization()
while Destination is not reached do
    Vehicle receives $S$ and measures $R$
    $hist \leftarrow$ Load one-step history data
    $flag1 \leftarrow AnomalyDetector(R, hist)$
    $flag2 \leftarrow PhysicalConsistencyChecker(S, R, hist)$
    if $flag1$ or $flag2$ is TRUE then
        Anomaly flag $\leftarrow$ Anomaly
    else
        Anomaly flag $\leftarrow$ Normal
    end if
end while
```

4.1 Case Study: Autonomous Vehicle Platoons

Platoon Control Policy. We consider a vehicle platoon consisting of N vehicles travelling on a straight highway segment and let veh_i denote the i-th vehicle within the platoon, where $i \in [0, N-1]$. Each vehicle member adopts the predecessor-leader following information flow topology. Specifically, veh_i obtains dynamic information including location, speed, acceleration, etc., from both the platoon leader veh_0 and its immediate preceding vehicle veh_{i-1}. Based on this topology, the vehicle's longitudinal motion is governed by the cooperative adaptive cruise control (CACC) policy, which computes the desired acceleration for veh_i by:

$$\ddot{x}_i = a_{des} = \alpha_1 a_{i-1} + \alpha_2 a_0 + \alpha_3 \dot{\epsilon}_{i,i-1} + \alpha_4 \dot{\epsilon}_{i,0} + \alpha_5 \epsilon_{i,i-1}, \tag{1}$$

$$\epsilon_{i,i-1} = x_i - x_{i-1} + L \ , \quad \dot{\epsilon}_{i,i-1} = v_i - v_{i-1} \ ,$$

where α's are controller gains taken from [16]. a_{i-1} and a_0 are the accelerations of the preceding vehicle and the leader respectively. The distance error $\epsilon_{i,i-1}$ is calculated based on a desired gap distance L and the obtained inter-vehicle distance between veh_i and veh_{i-1}. Similarly, their corresponding relative speed is represented as $\dot{\epsilon}_{i,i-1}$ with v_i denoting the speed of veh_i.

Kinematic Model. The longitudinal motion of each vehicle can be modelled as uniformly accelerated motion along a line by Eq. (2). This motion arises when an object is subjected to a constant acceleration. The acceleration value determines the gradient of the velocity-time function with an initial velocity labelled as $v_i(0)$. Similarly, the steady changing velocity determines the gradient of the position-time function with an arbitrary initial position $x_i(0)$.

$$v_i(t) = v_i(0) + a_i t \ , \quad x_i(t) = x_i(0) + v_i(0)t + \frac{1}{2}a_i t^2 \ , \tag{2}$$

where the acceleration and velocity variables are defined in Eq. (1). This kinematic model can be used to approximate the local behaviour of general longitudinal motions by modeling the object's motion within two consecutive sampling steps $t-1$ and t as uniformly accelerated motion. In general, its approximation strength increases with increased sampling frequency.

4.2 Data-Driven Anomaly Detector

At each time instance t, veh_i obtains the relative speed and distance with respect to its predecessor via a rangefinder (e.g., a radar sensor), which are the inputs to our DL attack detector. As a bonus, although veh_i also receives communication messages from other vehicles, only sensor measurements are used as detector inputs because they are more difficult to modify in practice and inherently immutable to communication-related attacks resulting in high detection success rate when only communication channels are compromised. The overall structure is shown in Figure 1b, which consists of two parts:

1. **Predictor**, trained with data from normal manoeuvre behaviour based on a sliding window, outputs the expected desired acceleration value $\tilde{a}_{des}$ at the current time instance. In our work, we use a multivariate time-series regression model and a sliding window to fully extract the temporal information within the data.
2. **Comparator** computes the difference between the inputs, i.e., the controller output a_{des} and the predicted value $\tilde{a}_{des}$. We use a sliding window to compute the mean absolute error $\bar{e}$ in order to reduce the false alarm rate. Consider an error window of size M, $\bar{e}$ at time t is computed as

$$\bar{e}(t) = \frac{1}{M} \sum_{m=t-M+1}^{i=t} \| a_{des}(m) - \tilde{a}_{des}(m) \| \ . \tag{3}$$

   An anomaly is flagged when $\bar{e}$ is greater than a threshold pre-determined in a benign driving environment.

4.3 Physical Consistency Checker

Corrupted controller inputs may not obey the underlying physical processes of the platoon system. Based on the kinematic model, the physics-based component of the proposed defence method - the physical consistency checker - consists of two components: the distance checker and speed checker.

Distance Checker. The change of inter-vehicle distance ($\Delta\epsilon_{i,i-1}$) within consecutive sampling instances can be directly calculated based on transmitted location information from the preceding vehicle veh_{i-1} and its own location readings. The same quantity ($\Delta\tilde{\epsilon}_{i,i-1}$) can also be computed based on locally measured speed and acceleration information according to the kinematic model.

$$\begin{aligned} \mathit{\Delta}\epsilon_{i,i-1}(t) &= \epsilon_{i,i-1}(t) - \epsilon_{i,i-1}(t-1) \ , \\ \mathit{\Delta}\tilde{\epsilon}_{i,i-1}(t) &= \dot{\epsilon}_{i,i-1} \cdot \mathit{\Delta}t + \frac{1}{2}\mathit{\Delta}t^2 \left(a_i(t-1) - a_{i-1}(t-1)\right) \end{aligned}$$

Speed Checker. Similarly, the change of relative speed can be computed directly by the subtraction of speed measurements or by the kinematic model utilizing acceleration information.

$$\begin{aligned} \mathit{\Delta}\dot{\epsilon}_{i,i-1}(t) &= \dot{\epsilon}_{i,i-1}(t) - \dot{\epsilon}_i(t-1) \ , \\ \mathit{\Delta}\tilde{\dot{\epsilon}}_{i,i-1}(t) &= \left(a_i(t-1) - a_{i-1}(t-1)\right) \mathit{\Delta}t \ . \end{aligned}$$

Both the direct calculation and physical model-based calculation produce similar results when there are no adversarial attacks against the anomaly detector or the proposed defence system in general. However, they deviate in an adversarial environment. If the deviation is greater than a pre-defined threshold, it triggers our physical consistency checker to report anomalies. Note that, these thresholds are domain-specific. In our evaluation, they are determined to balance out the false positive and false negative rates in a benign driving environment, which contains various types of highway driving scenarios. The overall structure is shown in Figure 1c along with the pseudo-code presented in Algorithm 2.

Algorithm 2 Physical Consistency Checker (PCC)

```
Input: Communication messages $S$, sensor readings $R$ and one-step history $hist$
Output: TRUE or FALSE
$\mathit{\Delta}\dot{\epsilon} \leftarrow \dot{\epsilon}_{i,i-1}(t) - \dot{\epsilon}_{i,i-1}(t-1)$ {Speed check}
$\mathit{\Delta}\tilde{\dot{\epsilon}} \leftarrow (a_i(t-1) - a_{i-1}(t-1))\mathit{\Delta}t$
$Anomaly\ flag$ 1 ← $Comparator(\mathit{\Delta}\dot{\epsilon}, \mathit{\Delta}\tilde{\dot{\epsilon}})$
$\mathit{\Delta}\epsilon \leftarrow \epsilon_{i,i-1}(t) - \epsilon_{i,i-1}(t-1)$ {Distance check}
$\mathit{\Delta}\tilde{\epsilon} \leftarrow \dot{\epsilon}_{i,i-1}(t-1) \cdot \mathit{\Delta}t + \frac{1}{2}\mathit{\Delta}t^2(a_i(t-1) - a_{i-1}(t-1))$
$Anomaly\ flag$ 2 ← $Comparator(\mathit{\Delta}\epsilon, \mathit{\Delta}\tilde{\epsilon})$
if $Anomaly\ flag$ 1 or $Anomaly\ flag$ 2 is TRUE then
    $Anomaly\ flag$ ← TRUE
else
    $Anomaly\ flag$ ← FALSE
end if
return $Anomaly\ flag$
```

5 Experimental Results

5.1 Simulation Setup

To provide a comprehensive evaluation of our proposed detection method, we use *Webots* as our simulation platform, which provides a broad range of calibrated vehicle models, sensor modules as well as static objects and materials to realize different simulation scenarios with high physical accuracy. It is a cost-efficient approach to generating adequate training data and constructing different cyber-physical attacks. Our data sets and implementations are available on Github at https://garrisonsun.github.io/Securing-Cyber-Physical-Systems/.
Platoon and Traffic Simulation. We simulate a vehicle platoon of 4 *BMW X5* vehicles driving along a highway segment. Multiple sensors are embedded in each vehicle to measure, transmit and receive critical driving information. For example, veh_i uses a radar sensor to measure the inter-vehicle distance and relative speed with respect to its predecessor veh_{i-1}. Radar noise is calibrated according to the datasheet of a real-world radar (*Delphi ESR 2.5 pulse Doppler cruise control radar*). Other control inputs (e.g., leader's dynamics used in Eq. 1) are obtained via wireless communication. In addition, each vehicle reads its own speed, acceleration, etc. directly from the speedometer and accelerometer respectively.

We generate a large number of vehicles in real-time in *Webots* interfacing with Simulation of Urban MObility (SUMO) [15] in order to construct a more realistic driving environment. Traffic flows involve four types of vehicles (i.e., motorcycles, light-weight vehicles, trucks, and trailers) with various driving characteristics (cooperative or competitive) and intentions to merge, which generates many random situations.

5.2 Double-Insured Anomaly Detection (DAD)

Data-Driven Anomaly Detector. For this work, we train an LSTM network from normal data as our predictor due to its outstanding performance for time-series prediction. It is a many-to-one prediction model, which consists of a normalization layer, and two stacked LSTM layers with 200 and 100 hidden units respectively. Each LSTM layer is followed by a dropout layer (rate=0.3) to avoid overfitting. The last dropout layer is connected with two fully connected layers with 50 and 1 hidden units, respectively. The model takes a sequence of historical sensor measurements (controlled by the sliding window size) and outputs a prediction of the desired acceleration $\tilde{a}_{des}$ for the next time instance. An anomaly is reported if this predicted value significantly deviates from the controller output.

Physical Consistency Checker. Since the physical consistency checker assumes vehicle motion within consecutive sampling time instances as uniformly accelerated motion, high-frequency sensor noise could degrade its anomaly detection performance when noise level and sampling frequency are both high. Therefore, we apply a digital Butterworth low-pass filter to remove the noise and reveal the underlying trend of the residuals between direct and model calculations. Note that, its performance is expected to be improved with low-noise sensors specifically designed for vehicle platoon applications.

5.3 Evaluation Setup

We employ Keras 2.4.3 and Python 3.8.10 to implement DAD and all the baselines on Ubuntu 20.04 operating system with a commodity i7-10510U CPU. The models are trained using an Adam optimizer with a learning rate of 0.001 for up to 500 epochs with early stopping (patience=10). Mean squared error is chosen as the loss function. We found a window size of 20 (input size 20×2) results in the best prediction performance. For the comparator, a window size of 40 and a detection threshold of 2 can effectively smooth out the prediction residuals and reduce the false alarm rate without degrading prediction performance.

Metrics. We prioritize F1 score [24] for this evaluation because a high F1 score indicates a combination of high precision and recall. Missing an attack is often more costly for such safety-critical systems, potentially causing catastrophic damages. Therefore, detection recall (Rec) is also included as our secondary comparison metric.

Attacks. Conventional cyber-physical attacks along with their adversarially-masked versions are examined in our evaluation:

Vanilla false data injection (v-FDI), as described in Sect. 3.1, and progressively modify received/measured dynamics information from the proceeding vehicle. For v-FDI, the modifications can be posed on a single variable such as on acceleration (v-FDI-Acce.) or in combination (e.g., v-FDI-Acce.Speed that alters both acceleration and speed).

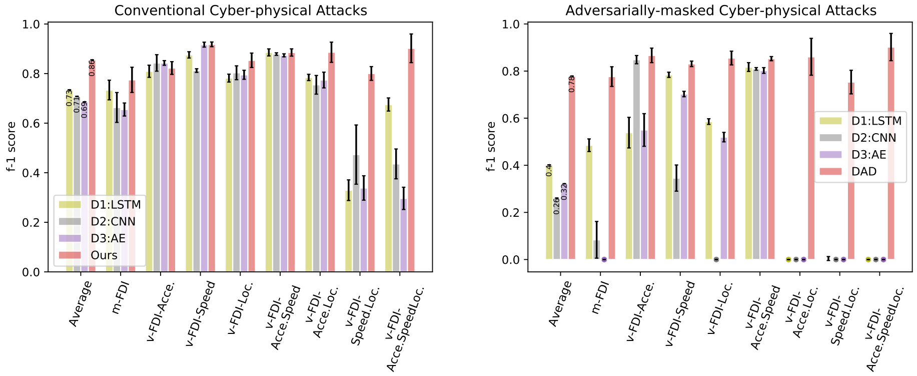
Model-aware false data injection (m-FDI), is seen as an evolved version of v-FDI. We consider the acceleration is modified with the maximum allowable modification as 5 m/s^2 (since higher accelerations are unrealistic in practice) and both the speed and location magnitudes are also modified based on the underlying system model to improve stealthiness.

Adversarial attack, the BIM attack approach [13] in particular, is used to mask these cyber-physical attacks in order to deceive the deployed attack detector (e.g., adversarially masked m-FDI is denoted as m-FDI adv. masked). The max-norm ball ϵ is chosen to be a small value as 0.4. In this way, the attack data is only slightly modified thereby retaining the original attack effects of the cyber-physical attack. Note that, the ϵ value is prefixed in this evaluation and a grid search may be required to find the optimal ϵ for different detectors under different attacks.

Baselines Attack Detectors. To demonstrate the effectiveness and robustness of our proposed defence system, we compare the detection results with pure data-driven and model-based detection approaches and study the impact of each component of our proposed method. In addition, we also consider adversarial training, based on the BIM attack, as an adversarial defence baseline in our evaluation with a robustified LSTM reconstruction-based attack detector (D1: LSTM*). The state-of-the-art data-driven defence baselines include an LSTM reconstruction-based attack detector (D1: LSTM) as in [7] and a CNN reconstruction-based attack detector (D2: CNN) similar to [12]. Some literature [17] also recognizes the effectiveness of autoencoders in performing classification or anomaly detection tasks based on reconstruction errors. For completeness, we also implement a convolutional autoencoder-based detector (D3: AE). Besides, the physics component of our proposed method - physical consistency checker (PCC) - is also used for comparison as a standalone attack detector.

5.4 Attack Detection Results

We demonstrate that our proposed attack detection system (DAD) provides improved attack detection performance against both conventional cyber-physical attacks and their adversarially-masked counterparts. In total, eight conventional cyber-physical attacks are examined including 7 variants of vanilla false data injection and 1 model-aware false data injection. Each type of attack has been performed five times. The complete detection *F1 score* is summarized in Figure 2 along with error bars at the top. The detection results for the model-aware false data injection (m-FDI) are summarized in Table 2.



(a) Conventional cyber-physical attacks. (b) Adv. masked cyber-physical attacks.

Fig. 2: Detection F1 scores comparing our method with alternatives.

Conventional Cyber-Physical Attacks. The detection F1 scores for conventional cyber-physical attacks are summarized in Figure 2a. In general, data-driven methods such as LSTM, CNN and AE reconstruction-based attack detectors perform well against these attacks. Depending on the exact attack type, one may slightly outperform the other. The physical consistency checker (PCC), as seen in the left half of Table 2, misses most of the attack instances when acting alone to detect m-FDI resulting in a recall of 0.18 and an F1 score of 0.29. This highlights the necessity of data-driven approaches to capture complex system behaviour. In comparison, our proposed method takes the advantage of both data-driven and physics-inspired methods achieving an F1 score of 0.86 on average (Figure 2a) over all considered conventional attacks and outperforms other popular data-driven approaches.

Adversarially-Masked Cyber-Physical Attacks. The physics-inspired component of our method, the physical consistency checker, starts to shine when the cyber-physical attacks are masked with adversarial perturbations. The detection performance, as shown in Figure 2b and right half of Table 2, is greatly reduced for all data-driven attack detectors, which exposes the cyber-physical system (i.e., the vehicle platoon) to a wide range of safety risks. Although some attacks require larger perturbations to fully deceive the detector, the average F1 scores are reduced to 0.40, 0.26 and 0.32 from 0.73, 0.71, 0.69 respectively for the LSTM, CNN, and AE based detectors. Because the generated adversarial perturbations are inconsistent with the physics model, our proposed method is able to detect the adversarially-perturbed cyber-physical attacks with an average F1 score of 0.78, which doubles the detection F1 score compared to existing baselines.

Adversarial Training Variants. As seen at the bottom of Table 2, our proposed method can be applied along with existing adversarial defence approaches (e.g., adversarial training). Combining the robustified model with physics knowledge would result in a better detection system DAD*, increasing detection recall and F1 score from 0.77 and 0.78 to 0.84 and 0.81, respectively, against adversarial-perturbed attacks. Although adversarial training might slightly sacrifice detection performance against classical cyber-physical attacks, it is demonstrated that our defence framework has the potential to provide better performance with an improved DL model and/or with other advanced adversarial defence methods against much stronger adversaries.

Table 2: Attack Detection Results against m-FDI with Different Detection Methods. * denotes adversarial training.

Attack	m-FDI		m-FDI (adv. masked)		
Defense	**Rec**	**F1**	**Defense**	**Rec**	**F1**
D1: LSTM	0.70	0.73	D1: LSTM	0.39	0.49
D2: CNN	0.57	0.66	D2: CNN	0.05	0.08
D3: AE	0.56	0.66	D3: AE	0.00	0.00
PCC	0.18	0.29	PCC	0.63	0.75
Ours: DAD	**0.77**	**0.77**	**Ours: DAD**	**0.77**	**0.78**
D1: LSTM*	0.70	0.73	D1: LSTM*	0.48	0.56
Ours: DAD*	**0.75**	**0.76**	**Ours: DAD***	**0.84**	**0.81**

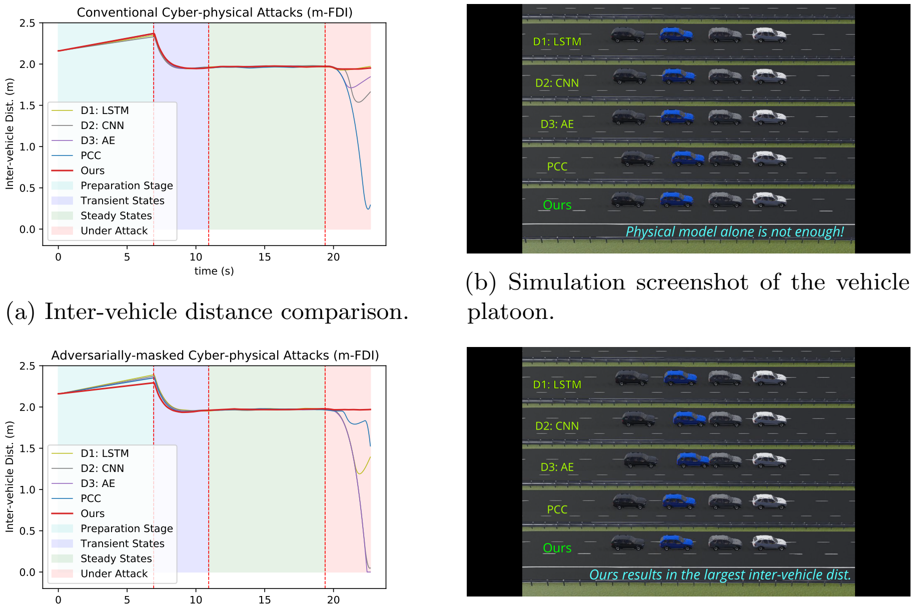
5.5 Simulation Demonstration for the m-FDI attack

In this subsection, we use the inter-vehicle distance as a measuring metric to examine the dangerous level of the compromised vehicle under attack. The entire simulation process can be roughly divided into four stages as shown in Figure 3a and Figure 3c. It starts from the *preparation stage*, where each vehicle starts with zero velocity and accelerates from an arbitrary position with a random inter-vehicle distance. Once the vehicle platoon is established, all platoon members quickly enter the *transient stage* and gradually reach the desired inter-vehicle distance ($2m$ in this case). This distance will be maintained throughout the simulation with only minor fluctuations when traffic condition changes with the power of the CACC platoon controller. The steady stage ends when an attack is initiated and we start to observe the resulting inter-vehicle distances for different defence methods. In this demonstration, we assume the vehicle would request a manual manoeuvre (not affected by data false injection attacks) as a simple mitigation strategy when the defence system reports an attack.

- As indicated in Figure 3a and Figure 3b, our method as well as other data-driven baselines can maintain a relatively safe inter-vehicle distance under conventional cyber-physical attacks (i.e., m-FDI). However, our method results in the best detection performance against m-FDI with nearly unnoticeable fluctuation throughout the entire attack period. In comparison, the model-based detection method PCC leads to a minimum distance of 0.25 meters, which greatly increases safety risks, especially in the highway driving scenario, and highlights the importance of data-driven approaches for cyber-physical systems.
- Figure 3c and Figure 3d indicate that DL-based detectors suffer the most under adversarially-masked cyber-physical attacks with CNN and AE-based detectors leading to catastrophic collisions. Our proposed method again significantly improves system safety and security against such a powerful adversary with white-box knowledge of both the DL and physics models. It is also important to point out that the model-based detector PCC alone cannot detect such attacks accurately during the entire attack period.

6 Conclusions

In this paper, we have presented a novel physics-enhanced attack detection system for autonomous vehicle platoons as a critical cyber-physical system. Our approach and algorithms greatly improve platoon security and dependability against both classical cyber-physical and adversarial attacks. Our methods inherit the advantages of existing data-driven attack detection systems based on recent advances in deep learning as well as utilize physics modelling techniques to improve robustness against adversarial attacks to cyber-physical systems. We consider a powerful white-box attacker and demonstrate that our approach outperforms conventional detection methods with a sophisticated simulator, which highlights its potential to perform even better when dealing with real-world attackers who normally only have limited information about the system. Future research will evaluate the extension of this resiliency architecture to other cyber-physical systems (e.g., smart grids) with various data-driven defence approaches. The scope of adversarial attacks in this work is limited to existing approaches developed mainly in the vision domain. Therefore, a further study could incorporate the physics model with the adversarial example generation process to create a stronger adversarial attack method and investigate its attack



(a) Inter-vehicle distance comparison.

(b) Simulation screenshot of the vehicle platoon.

(c) Inter-vehicle distance comparison.

(d) Simulation screenshot of the vehicle platoon.

Fig. 3: Comparison between different defense methods under attacks: (a)&(b)-Conventional cyber-physical attacks (m-FDI), (c)&(d)-Adversarially-masked cyber-physical attacks (m-FDI (adv. masked)).

effects on the cyber-physical system and evasion strength against the proposed defence method.

References

1. Alotibi, F., Abdelhakim, M.: Anomaly detection for cooperative adaptive cruise control in autonomous vehicles using statistical learning and kinematic model. IEEE Transactions on Intelligent Transportation Systems (2020)
2. Boddupalli, S., Rao, A.S., Ray, S.: Resilient cooperative adaptive cruise control for autonomous vehicles using machine learning. IEEE Transactions on Intelligent Transportation Systems (2022)
3. Boeira, F., Barcellos, M.P., de Freitas, E.P., Vinel, A., Asplund, M.: Effects of colluding sybil nodes in message falsification attacks for vehicular platooning. In: 2017 IEEE Vehicular Networking Conference (VNC). pp. 53–60. IEEE (2017)
4. Cao, Y., Xiao, C., Cyr, B., Zhou, Y., Park, W., Rampazzi, S., Chen, Q.A., Fu, K., Mao, Z.M.: Adversarial sensor attack on lidar-based perception in autonomous driving. In: Proceedings of the 2019 ACM SIGSAC conference on computer and communications security. pp. 2267–2281 (2019)

5. Daw, A., Karpatne, A., Watkins, W., Read, J., Kumar, V.: Physics-guided neural networks (pgnn): An application in lake temperature modeling. arXiv preprint arXiv:1710.11431 (2017)
6. Garcia, L., Brasser, F., Cintuglu, M.H., Sadeghi, A.R., Mohammed, O.A., Zonouz, S.A.: Hey, my malware knows physics! attacking plcs with physical model aware rootkit. In: NDSS (2017)
7. Goh, J., Adepu, S., Tan, M., Lee, Z.S.: Anomaly detection in cyber physical systems using recurrent neural networks. In: 2017 IEEE 18th International Symposium on High Assurance Systems Engineering (HASE). pp. 140–145. IEEE (2017)
8. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
9. Jia, Y., Wang, J., Poskitt, C.M., Chattopadhyay, S., Sun, J., Chen, Y.: Adversarial attacks and mitigation for anomaly detectors of cyber-physical systems. International Journal of Critical Infrastructure Protection p. 100452 (2021)
10. Karim, F., Majumdar, S., Darabi, H.: Adversarial attacks on time series. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
11. Khanapuri, E., Chintalapati, T., Sharma, R., Gerdes, R.: Learning-based adversarial agent detection and identification in cyber physical systems applied to autonomous vehicular platoon. In: 2019 IEEE/ACM 5th International Workshop on Software Engineering for Smart Cyber-Physical Systems (SEsCPS). pp. 39–45. IEEE (2019)
12. Kravchik, M., Shabtai, A.: Detecting cyber attacks in industrial control systems using convolutional neural networks. In: Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and PrivaCy. pp. 72–83 (2018)
13. Kurakin, A., Goodfellow, I., Bengio, S., et al.: Adversarial examples in the physical world (2016)
14. Li, J., Liu, Y., Chen, T., Xiao, Z., Li, Z., Wang, J.: Adversarial attacks and defenses on cyber–physical systems: A survey. IEEE Internet of Things Journal **7**(6), 5103–5115 (2020)
15. Lopez, P.A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P., Wießner, E.: Microscopic traffic simulation using sumo. In: The 21st IEEE International Conference on Intelligent Transportation Systems. IEEE (2018), https://elib.dlr.de/124092/
16. Segata, M., Joerer, S., Bloessl, B., Sommer, C., Dressler, F., Cigno, R.L.: Plexe: A platooning extension for veins. In: 2014 IEEE Vehicular Networking Conference (VNC). pp. 53–60. IEEE (2014)
17. Seyfioğlu, M.S., Özbayoğlu, A.M., Gürbüz, S.Z.: Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities. IEEE Transactions on Aerospace and Electronic Systems **54**(4), 1709–1723 (2018)
18. Sumra, I.A., Hasbullah, H.B., AbManan, J.l.B.: Attacks on security goals (confidentiality, integrity, availability) in vanet: a survey. In: Vehicular Ad-Hoc Networks for Smart Cities, pp. 51–61. Springer (2015)
19. Sun, G., Alpcan, T., Rubinstein, B.I.P., Camtepe, S.: Strategic mitigation against wireless attacks on autonomous platoons. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. ECML-PKDD (2021)
20. Tielens, P., Van Hertem, D.: The relevance of inertia in power systems. Renewable and Sustainable Energy Reviews **55**, 999–1009 (2016)
21. Wiedersheim, B., Ma, Z., Kargl, F., Papadimitratos, P.: Privacy in inter-vehicular networks: Why simple pseudonym change is not enough. In: 2010 Seventh international conference on wireless on-demand network systems and services (WONS). pp. 176–183. IEEE (2010)

22. Yang, L., Moubayed, A., Hamieh, I., Shami, A.: Tree-based intelligent intrusion detection system in internet of vehicles. In: 2019 IEEE Global Communications Conference (GLOBECOM). pp. 1–6. IEEE (2019)
23. Yuan, X., Chen, Y., Zhao, Y., Long, Y., Liu, X., Chen, K., Zhang, S., Huang, H., Wang, X., Gunter, C.A.: Commandersong: A systematic approach for practical adversarial voice recognition. In: 27th USENIX Security Symposium (USENIX Security 18). pp. 49–64 (2018)
24. Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H., Chawla, N.V.: A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 1409–1416 (2019)