On the Effects of Online Black-Box Attacks against Mobile Network Capacity Forecasters under Temporal Constraints

1st Louis Miermont
IMDEA Networks Institute and
Universidad Carlos III de Madrid
Madrid, Spain
louis.miermont@imdea.org

Claudio Fiandrino
IMDEA Networks Institute
Madrid, Spain
claudio.fiandrino@imdea.org

Guillermo Suarez-Tangil
IMDEA Networks Institute
Madrid, Spain
guillermo.suarez-tangil@imdea.org

***Abstract*—Machine Learning (ML) methods are increasingly used to drive resource allocation in mobile networks; however, they are vulnerable to adversarial manipulation. In this work, we investigate attacks on network capacity forecasting that respect temporal constraints inherent to online time series attacks: input sequences overlap, the past is immutable, and future traffic values are unknown. Under these constraints, we propose an attack where an adversary estimates future traffic load, queries the target model, and applies a perturbation that maximizes an evaluation function representing the attacker's objective. We use real-world wireless traffic datasets to validate our method and target well-known ML models for capacity forecasting. Our experiments show that attack effectiveness increases with the adversary's accuracy in forecasting future traffic values and that perturbations impact a sequence of predictions, producing an approximately 2.38-fold relative error increase on targeted forecasts under high forecasting accuracy, with values of the employed capacity forecasting function increasing from 0.13 to 0.31 for perturbations equivalent to 10% of the traffic peak.**

***Index Terms*—Adversarial machine learning, Wireless networks, Time series analysis, Online forecasting**

I. Introduction

The rapid evolution of mobile communication has fundamentally transformed the management of network infrastructures. With the growing complexity of new-generation mobile networks, the use of intelligent mechanisms is becoming essential to enhance their performance [1]. Specifically, the application of ML methods to forecast future traffic demand has garnered considerable attention in recent years [2]. Such techniques leverage historical traffic data by capturing the complex temporal dependencies inherent to time series and anticipating future patterns of traffic load. In this context, time series forecasting models play a crucial role in supporting dynamic resource allocation, performance optimization, and Quality of Service (QoS) provisioning. However, ML models have been shown to be susceptible to adversarial perturbations. Even minimal input modifications can cause a predictor to produce highly inaccurate outputs [3]. Recent studies have extended these findings to models used for time series forecasting, demonstrating that they similarly suffer from adversarial vulnerabilities [4].

The feasibility of adversarial attacks on time series underscores the importance of examining the security of ML-based forecasters before their deployment in wireless network architectures. Specifically, disruptions caused by attacks can lead to degraded service quality, violation of Service Level Agreements (SLAs), and ultimately result in significant financial losses for the operator. However, existing adversarial attacks often overlook essential constraints that frame the problem. Adversaries targeting ML-based traffic forecasting systems rarely have white-box knowledge of the target model and are further constrained by the online nature of the time series forecasting task.

In this paper, we define an attack methodology (§III) that adheres to the following constraints inherent to online time series forecasting: (1) input sequences share values and are therefore interdependent; (2) the past is immutable (i.e., previous observations cannot be altered); and (3) future observations are unknown to the adversary at the time of the attack. We then present an experimental setup that respects these constraints (§IV) and assess the impact of the attack with two real-world datasets (§V and §VI) by addressing the following research questions (RQs):

- RQ1. Can we design a successful attack on network capacity forecasting systems while considering these constraints?
- RQ2. How does an adversary's ability to predict future values influence the effectiveness of the attack?
- RQ3. How does the attack affect subsequent forecasts over time and subsequent model updates?
- RQ4. How consistently does the attack generalize across datasets? What invariant properties persist?

By addressing these RQs, we gain insights into key factors behind attack success and potential mitigation strategies (§VII), filling a gap in prior work that neglects considerations essential to the online nature of traffic forecasting tasks in wireless network architectures (§VIII). Next, we outline the importance of filling this gap.

II. Background

This section introduces the key concepts necessary to understand the practical challenges of implementing an adversarial attack in online time series forecasting settings.

Time series forecasting. Traffic forecasting involves using historical data to predict future traffic demands. This task corresponds to a time series forecasting problem, where a sequence of past values $\mathbf{X}_t = [x_{t-T+1}, \ldots, x_t]$ representing in order the most recent T observations up to time t, is used to forecast future values for the subsequent H time steps by employing a predictive function $f(\cdot)$, which may correspond to a ML model. This forecasting task can be expressed as:

$\hat{\mathbf{X}}_{t+1:t+H} = f(\mathbf{X}_t)$, where $\hat{\mathbf{X}}_{t+1:t+H} = [\hat{x}_{t+1}, \ldots, \hat{x}_{t+H}]$ represents the predictions of the model using $\mathbf{X}_t$ as input. If the objective of the forecasting task is to predict a variable different from the continuation of $\mathbf{X}_t$, we denote the corresponding predictions as $\hat{\mathbf{Y}}_{t+1:t+H}$.

Traffic and Capacity Forecasting. Conventional *traffic forecasting* approaches focus on accurately predicting future traffic demand without considering the economic implications of prediction errors, as we explain next. These approaches optimize generic loss functions that minimize the difference between predicted and observed traffic values, such as Mean Absolute Error (MAE) or Mean Squared Error (MSE). However, these methods treat over-provisioning (allocating more resources than needed) and under-provisioning (allocating fewer) as equivalent, when in reality they have vastly different costs for network operators. While over-provisioning leads to inefficient resource utilization and increased operational expenses, under-provisioning can result in Service Level Agreement (SLA) violations, customer dissatisfaction, and direct monetary penalties.

To address these shortcomings, *capacity forecasting* has emerged as a more practical concept [5]. Instead of predicting traffic demand, it aims to directly estimate the capacity that a network controller must allocate to meet expected demand. Concretely, operators may aim to minimize occurrences of under-provisioning in order to avoid substantial financial penalties associated with violating SLAs [6]. In ML, this shift in objective can be reflected by utilizing a customized loss function that explicitly captures this operational goal. By incorporating the relevant operational criteria into the loss, it guides the training process such that ML models learn parameters that minimize the loss function. For instance, the capacity forecaster *DeepCog* [5] employs α-OMC [6] to specifically perform capacity forecasting tasks.

To implement this objective, the piecewise formulation of α-OMC (eq. 1) introduces an asymmetric cost structure that differentiates between underestimation and overestimation. The relative importance of under- and over-provisioning penalties is controlled by the parameter α, which approximates the amount of resources a network operator is willing to allocate in order to avoid SLA violations. Additionally, the function includes a linear segment controlled by a very small value δ, which introduces a narrow transition region immediately above zero and preserves continuity. Then, when the predicted traffic underestimates the actual demand ($\hat{y} < y$), corresponding to an SLA violation, the model incurs a penalty characterized by a nearly constant linear function. Conversely, when the model overestimates the actual demand ($\hat{y} > y$), the system enters an over-provisioning regime, in which the penalty increases linearly with the extent of over-provisioning.

$$\alpha\text{-OMC}(z) = \begin{cases} \alpha - \delta \cdot z & \text{if } z \leq 0 \\ \alpha - \frac{1}{\delta} z & \text{if } 0 < z \leq \delta\alpha \\ z - \delta\alpha & \text{if } z > \delta\alpha \end{cases} \quad \text{where } z = \hat{y} - y \tag{1}$$

Adversarial Attacks. Adversarial attacks are deliberate strategies that aim to mislead ML models by feeding them carefully crafted data. These attacks can target different stages of the model lifecycle. *Poisoning attacks* occur during training, when an adversary modifies the training set by inserting, removing, or altering examples or their associated labels. In contrast, *evasion attacks* take place after deployment at the inference stage, when model parameters are fixed and the attacker perturbs inputs to induce incorrect predictions [7]. Different attack settings exist depending on the level of knowledge the adversary is assumed to have about the system, with white-box settings being the most permissive. Popular white-box attacks include gradient-based attacks such as Fast Gradient Sign Method (FGSM) [8] and Basic Iterative Method (BIM) [9]. FGSM generates an adversarial example by taking a single step in the direction of the gradient of the loss with respect to the input, while BIM applies the same gradient-based update iteratively with small steps and clipping after each step. In a black-box setting, information available to attackers is highly restricted. They can only interact with the target model through queries or observe its predictions [10].

Although adversarial attacks on time series forecasting have been proposed in prior works [11], [12], many remain impractical in the context of an online scenario as they neglect key temporal constraints intrinsic to this setting. In online forecasting, the target model generates predictions sequentially as new observations become available in a streaming fashion. Consequently, an attacker must make a decision on the spot, without access to ground-truth values, and with no retroactive capability. These characteristics, which are further detailed in the following section §III, impose limitations that constrain adversarial attacks on online time series forecasting systems.

III. Attack Methodology

We detail our attack methodology by introducing temporal restrictions inherent to the problem, followed by a description of the threat model, and our proposed attack.

A. Temporal Restrictions

We identify three key temporal restrictions that must be met for attacks on traffic and capacity forecasting systems to operate in an online setting.

Overlapping Input Sequences. Due to the sequential structure of time series, each input value is typically used for multiple forecasting tasks when employing a sliding-window mechanism. Specifically, one input value is part of the input sequences used for the next T forecasts, where T is the length of the input sequences. This overlap implies that a perturbation to a single value will affect not only the immediate forecast but also subsequent forecasts. This characteristic induces a dependence among different occurrences of an attack, while also giving the adversary an opportunity to increase the impact of perturbations.

Immutable Past. A key characteristic of online attacks is what we refer to as the immutability of the past. Past observations already used by the model cannot be retroactively altered by the adversary. Gong et al. [13] consider this issue by framing the problem as a trade-off between the observation space and the action space. Specifically, they argue that adversarial perturbations can only be applied to unobserved parts of the data. This restriction makes popular unrestricted gradient-based attacks infeasible, as they require the freedom to alter all input values used for prediction, including past observations. Moreover, they treat each input sequence as independent, where in reality they share values, which does not account for the *overlapping input sequences* restriction.

Unknown Future Values. In contrast to other adversarial attack settings, where the ground truth associated with an input is known at the time of the attack [14], this assumption does not hold when the attack is performed online and in real time. In online forecasting settings, the ground truth corresponds to future values that are inherently unavailable to the adversary but can be estimated using a traffic forecasting model trained on previously observed patterns. This introduces a combined challenge given the restriction on *overlapping input sequences*, as evaluating the degradation of performance induced by an adversarial perturbation across all affected forecasts requires knowledge of yet unknown future observations.

B. Threat Model

The adversary's objective is to degrade the model's forecasting performance during inference and in real time (i.e., online). To achieve this, they manipulate inputs at each inference step to maximize an evaluation function $\mathcal{L}(\cdot)$ that reflects their adversarial goal, i.e., the deterioration of a specified forecasting objective. The evaluation function may approximate the loss function used during model training or translate into alternative adversarial goals. For instance, adversaries may seek to cause maximum disruption to network operations by increasing under-provisioning, or to maximize financial losses by combining increased SLA violations and over-provisioning. In our experiments, we use the same function as the one employed during training to assess the attack's degradation of the target models' primary forecasting objective.

Within the constraints outlined above, this scenario assumes the following knowledge and capabilities of the adversary.

1) Attacker's Knowledge: We assume that the adversary has real-time access to traffic-load measurements collected by a cell. Such visibility is practically achievable through the passive sniffing of unencrypted cellular downlink control channels (e.g., PDCCH in 5G and LTE). Recent literature has demonstrated the efficacy of open-source telemetry tools—such as NR-Scope [15], LTESniffer [16], and the framework proposed by Ludant et al. [17] which allow adversaries using commercial off-the-shelf Software Defined Radios (SDRs) to monitor resource allocation messages and accurately reconstruct aggregate traffic volumes at the millisecond level. This over-the-air visibility enables the reconstruction of the target model's input sequences and, if necessary, allows the adversary to construct independent forecasting models, as discussed later in this section.

2) Attacker's Capabilities: The adversary is able to obtain forecasting outputs $f(\mathbf{X}_t)$ for arbitrary inputs in a black-box setting, without knowledge of the model's internals [18], [19]. In cases where access to the target model is unavailable, data collected over an initial time period is used to train a surrogate model for querying. Adversarial perturbations generated using this surrogate may subsequently transfer to the target model, which has already been demonstrated with DNNs for time series prediction [4].

Although future traffic values are not known a priori, as stated in §III-A, they can be estimated. Accordingly, we assume the adversary has access to an external traffic forecasting model $g(\cdot)$, which can be similarly constructed from retrieved traffic load measurements. The corresponding forecasts of this estimator are denoted by $\tilde{\mathbf{X}}_{t+1:t+N}$.

Finally, the adversary can manipulate current traffic values bidirectionally, enabling both injection and removal of traffic. Injection can be achieved by generating additional traffic with devices connected to the network. In contrast, traffic removal is generally considered impractical [2]. Nevertheless, an adversary may artificially reduce the observed traffic load by selectively disconnecting controlled devices that would otherwise remain continuously connected to the network, actively transmitting data. This form of manipulation requires that the target forecasting model operate on traffic data aggregated over sufficiently long time windows (e.g., several minutes), thereby providing the adversary with sufficient time to estimate x_t and apply the desired perturbations within the current time interval.

C. Proposed Attack

Fig. 1 shows a high-level overview of the proposed attack under the previously described threat model. We next outline its problem formulation and the attack description.

1) Problem Definition: The proposed attack aims to identify the direction of perturbation on the currently observed traffic load value that would most significantly degrade the performance of the target forecasting model. Specifically, we seek to determine whether traffic should be injected into or removed from the current observation x_t. In this study, we constrain the perturbation to a fixed magnitude, ε, applied solely to the current observation. The adversary iteratively generates perturbations each time a new observation arrives at the system and constructs the adversarial time series $\mathbf{X}'_\mathbf{t} = [x'_{t-T+1}, \ldots, x'_t]$ so as to maximize the evaluation function $\mathcal{L}(\cdot)$. Due to immutability restrictions, previous observations cannot be modified retroactively. We formalize this situation through the following constrained optimization problem:

$$x_t^* = \arg\max_{x'_t} \mathcal{L}\big(f(\mathbf{X}'_\mathbf{t}), \tilde{\mathbf{X}}_{t+1:t+N}\big) \quad \text{s.t.} \quad \|x'_t - x_t\| \leq \varepsilon. \tag{2}$$

The parameter N denotes the number of future traffic states considered in the attack. Ideally, N should equal $T + H$ to account for all future forecasts affected by the perturbation. However, setting N to such a high value would likely reduce the attack's effectiveness, as the adversary's ability to accurately predict future traffic states typically deteriorates with a longer forecasting horizon.

2) Attack Description: The attack described in Algorithm 1 consists of four distinct steps designed to apply a perturbation to the current observation x_t. This procedure can be iteratively executed as new observations become available (i.e., online). First, in step ①, the adversary utilizes an accessible traffic forecasting model to estimate future traffic loads (`line 1`). In step ②, the adversary queries the target model to retrieve its expected predictions for the subsequent N time steps, and then evaluates these predictions as part of step ③ using the function $\mathcal{L}(\cdot)$ and the previously obtained estimates of future values. Three different evaluations are performed at this stage: one with no perturbation added on x_t (`line 2`), another one by simulating a traffic injection of intensity ε (`line 3`), and a last one (`line 4`) by simulating a removal of traffic of the same intensity With this information, the last step ④ consists in applying the perturbation that most significantly increases the loss value (`lines 5 to 7`). If both values are below the initial evaluation, then no perturbation is introduced (`line 6`). For wireless traffic, negative values are not physically possible;

therefore, the perturbation is constrained to lie above $x_{\min}$ (which depends on the scaling parameters). Due to the clipping induced by the lower bound, removal-based perturbations may have in some cases an effective magnitude smaller than ε.

IV. Experimental Setup

This section describes the experimental setup used to test the proposed attack, including datasets, the architecture, and the model's setup. We also use a random attack as a baseline.

A. Datasets

We use the following datasets in our experiments:

- **Milan dataset** [20]: This dataset, published by Telecom Italia, contains mobile traffic data collected in the Milan metropolitan area from 1 November 2013 to 1 January 2014. The data are partitioned into a grid of 10,000 cells, each corresponding to a specific spatial area. In this study, we extract Internet traffic load measurements for a subset of the 25 most central cells of the grid, aggregated into 10-minute intervals. We then employ the first three weeks of data to train the initial forecasting models, while the subsequent three weeks are reserved for evaluation. All data are normalized using min–max scaling based on statistics computed from the training set.
- **ITU dataset** [21]: This dataset was released in 2024 by the AI for Good initiative of the International Telecommunication Union (ITU) as part of a challenge on spatio-temporal traffic forecasting in 5G wireless networks. It comprises throughput volume, throughput time, physical resource block utilization, and user count. The data span five weeks of hourly measurements from 30 Base Stations (BSs), each divided into 3 cells, with each cell consisting of 32 beams. For our experiments, throughput volume is aggregated at the cell level, and only the first cell of each BS is retained. Due to the smaller size of this dataset compared to the Milan dataset, a higher training-to-test ratio is used: the first four weeks of data are used for model training, while the final week is reserved for evaluation. As with the Milan dataset, values are similarly normalized using min–max scaling.

B. Models

We train ML models tailored for time series forecasting with the following architecture and loss function.

1) Architecture: We implement the *DeepCog* architecture [5], a 3D Convolutional Neural Network (CNN) originally proposed for spatio-temporal applications and to operate with the α-OMC loss function. Therefore, we maintain its objective of capacity forecasting by employing this loss function, but we simplify the problem setting by removing the spatial component of the data while retaining core architecture of the model. In our experiments, we set the parameters of the function to $\alpha = 1$ and $\delta = 0.1$. These values were selected after calibration to reflect a plausible network-operator tradeoff between limiting SLA violations and avoiding excessive overprovisioning.

2) Setup: Separate models are constructed for each cell selected for the experiments, using exclusively the traffic data corresponding to the respective cell. The input sequence length is set to $T = 6$, and the maximum forecasting horizon is $H = 1$. Forecasting is performed in an online setting, where predictions are completed iteratively as each new traffic measurement is observed. This stage may be subject to real-time adversarial manipulation of traffic values by an adversary considering their impact on models' capacity forecasting performance on the subsequent N forecasts, which we set to $N = 1$ in our experiments. Each experiment is performed 25 times, and the results are averaged to reduce variability.

Algorithm 1 Algorithm of the attack at current time t

Inputs: Target forecasting model $f(\cdot)$, external traffic forecasting model $g(\cdot)$, evaluation function $\mathcal{L}(\cdot)$, number N of future traffic states to consider, intensity of the attack ε, input vector of traffic values for current time $\mathbf{X}_t = [x_{t-T+1}, \ldots, x_t]$, minimum possible traffic value $x_{\min}$
Result: Perturbed traffic state for the current time x'_t.

$\tilde{X}_{t+1:t+N} \leftarrow g(X_t)$
$\tilde{L}_o \leftarrow \mathcal{L}\big(f(x_{t-T+1}, \ldots, x_t), \tilde{X}_{t+1:t+N}\big)$
$\tilde{L}_+ \leftarrow \mathcal{L}\big(f(x_{t-T+1}, \ldots, x_t + \varepsilon), \tilde{X}_{t+1:t+N}\big)$
$\tilde{L}_- \leftarrow \mathcal{L}\big(f(x_{t-T+1}, \ldots, max(x_t - \varepsilon, x_{\min})), \tilde{X}_{t+1:t+N}\big)$
if $max(\tilde{L}_-, \tilde{L}_+) < \tilde{L}_0$ **then**
    return x_t ▷ *No perturbation*
end if
if $\tilde{L}_+ > \tilde{L}_-$ **then**
    return $x_t + \varepsilon$ ▷ *Injection*
else if $\tilde{L}_+ \leq \tilde{L}_-$ **then**
    return $max(x_t - \varepsilon, x_{\min})$ ▷ *Removal (bounded)*
end if

C. Estimator Models

The first stage of the attack requires the adversary to estimate future traffic values, which can be achieved using an external traffic forecasting model. The accuracy of the resulting estimates is crucial for the attacker to correctly identify the direction for perturbations. However, forecasting model performance varies significantly depending on the characteristics of the data and on the exact settings of the forecasting task. To enhance generalizability, in our evaluation, we model this model as a estimator parameterized by its accuracy, manipulating the accuracy of estimates for future values. By doing so, we can assess how the quality of the estimator impacts the overall performance and outcome of the attack in various settings. In our implementation, we range between two contrasting scenarios — low- and high-performing estimators.

We simulate a traffic forecasting model by adding controlled noise to the actual future traffic data, with higher noise levels corresponding to lower forecasting accuracy. The probability distribution of the applied noise is modeled as a zero-mean normal distribution as defined in Equation 3. To reflect the heteroscedastic nature of prediction errors observed in our experiments, which tend to increase for higher values, we define the noise variance as a function of the traffic value x. The overall amplitude of noise depends on the parameter a, and the sensitivity of the variance growth with respect to the input is controlled by the parameter b. The simulated prediction of the estimator model is then defined as $\tilde{x} = x + noise$. In our experiments, the value of b is set to 1 and different values are used for a, reflecting varying noise intensities.

$$noise(x) \sim \mathcal{N}\big(0, \sigma^2(x)\big) \quad \text{where } \sigma(x) = a \cdot x^b \tag{3}$$

With this definition, the MSE, representing the desired forecasting performance, is directly related to the parameter a, such as $MSE = a^2 \cdot \mathbb{E}[X^{2b}]$, where $\mathbb{E}[X^{2b}]$ represents the $2b$-th raw moment of the random traffic variable X. In other words, we scale the error based on how large or extreme the input values are. This formulation enables the evaluation of

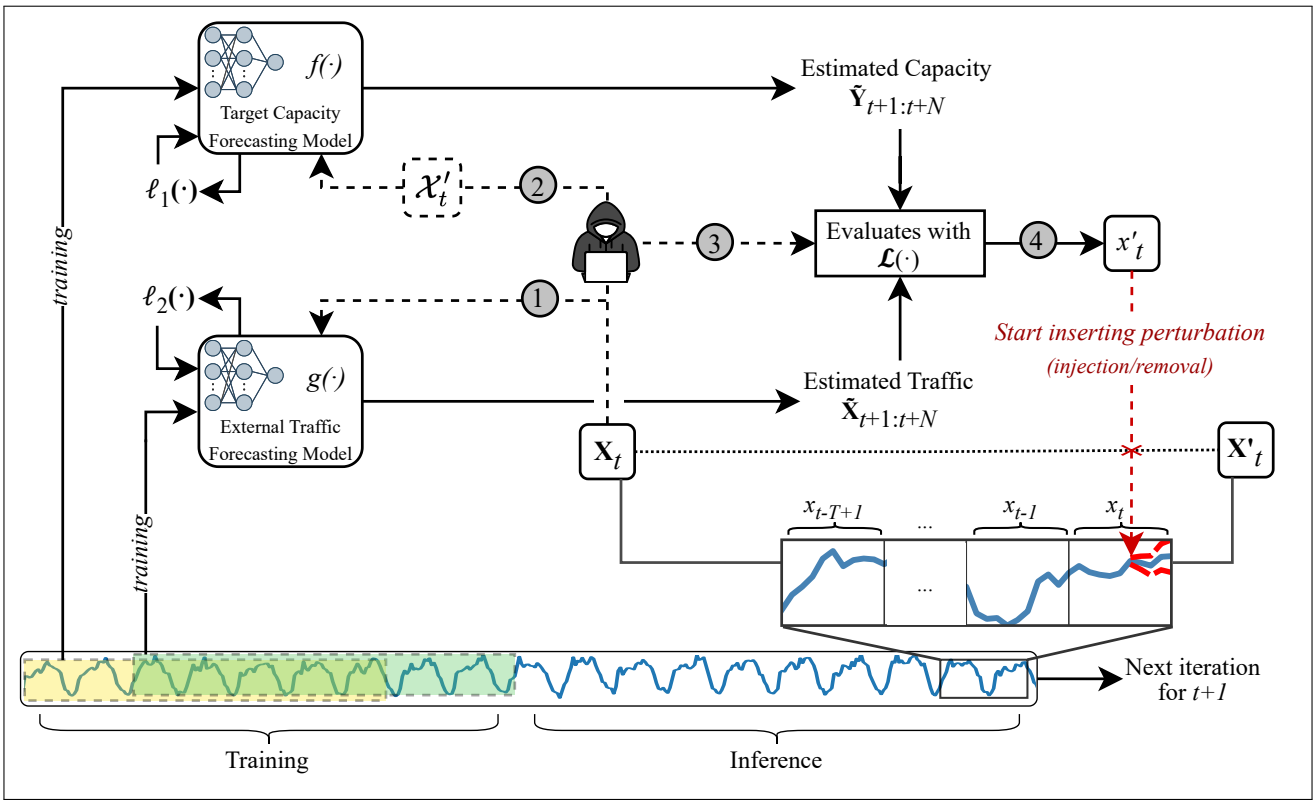

Fig. 1. Attack methodology. Steps are represented in order by ①, ..., and ④. The term $l_1(\cdot)$ represents the target forecasting model loss function, which may be designed for capacity forecasting (i.e., α-OMC), and $l_2(\cdot)$ represents the external traffic forecasting and is typically equal to the MAE or MSE function. The two models do not require training on the same set of data. $\mathcal{X}'_t$ denotes the set of applied perturbations simulated by the adversary, crafted prior to evaluation. The time series values are aggregates of traffic load values measured within their corresponding time windows (e.g., x_t is the aggregate of measurements over the time interval $(t-1, t]$).

the attack with a simulated traffic forecasting model at desired MSE levels. Because the resulting estimator is unbiased with zero-mean error (i.e., its output expectation is equal to the mean traffic value), it represents a theoretical upper bound on the attack's performance for the selected MSE values.

D. Random Attack

Perturbing traffic values is expected to reduce the performances of the target models. To ensure a meaningful evaluation of the proposed attack, we compare our method not only against a baseline scenario without any perturbation but also against a *random attack*. In the random attack setting, the direction for the perturbations (whether to increase or decrease values) is chosen uniformly at random. This comparison enables us to distinguish the effect of deliberately crafted perturbations from the degradation that could occur due to arbitrary noise.

V. Attack Performance in the Milan Dataset

In what follows, we employ the Milan dataset to evaluate the impact of the proposed attack while answering RQ1–RQ3: (i) verifying whether an attack meeting our temporal restrictions can meaningfully degrade the models' capacity forecasting objective, (ii) assessing how the accuracy in future traffic load values modulates attack effectiveness, and (iii) measuring how perturbations propagate over subsequent forecasts and model updates, representing the longer-term effects of adversarial manipulations.

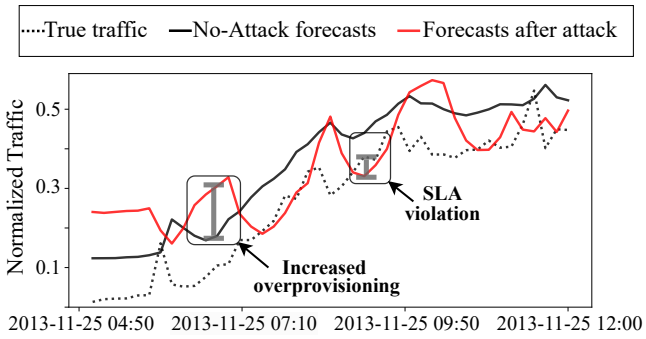

Fig. 2. Illustration of local impact of the attack on capacity forecasts for cell 5062 using *low-noise* estimates and $\varepsilon = 0.1$.

A. Attack Performance under Temporal Restrictions

We begin our evaluation by assessing whether the proposed attack design can effectively degrade the performance of target models (RQ1). Specifically, we examine whether the predictor's behavior after the attack matches the attack objective. Fig. 2 portrays a representative example of how the attack degrades the model's capacity forecasting ability. Prior to the attack, the predictor's forecasts remain above actual traffic, preventing SLA breaks while avoiding excessive overprovisioning. Under attack, the forecasts exhibit increased overprovisioning and, when reachable, induce SLA violations. The magnitude of these two types of degradation scales strongly with the attack intensity. In the following subsection, we summarize the results for different parameters across all evaluated cells.

B. Importance of Future Estimates' Quality

To evaluate the importance of accurately forecasting future traffic values on attack performance (RQ2), we consider several estimation settings that represent different levels of traffic forecasting capability available to the adversary.

Noise-based Estimates. We simulate external traffic forecasting models by following the method described in §IV-C. To ensure that their performance aligns with that of existing traffic forecasting models, we draw values of a comparable scale from the ones reported in [22], which also uses inputs with a 10-minute granularity. Specifically, we assume a well-performing predictor achieves a MSE of approximately 0.025 on normalized data for the first forecast step $t+1$. However, the conditions under which such performance is achieved might not be accessible to an adversary. To account for this limitation, we also consider another forecaster with a MSE of 0.10, i.e., four times larger. From these two values, we simulate a *low-noise* and a *high-noise* estimator, representing scenarios with differing forecasting accuracy.

Optimal Scenario. We additionally evaluate the attack under an "optimal scenario", in which we assume an oracle provides the adversary with the true future traffic observations. This corresponds to the assumption that the estimator is a perfect traffic forecasting model (i.e., with an MSE of 0), thereby violating inaccessibility of future observations described in section §III-A. Although theoretically impractical, as we previously stated, an attack can still operate in an online fashion under this assumption. However, it is expected to strongly overestimate its impact relative to scenarios based on realistic traffic forecasting capabilities. Evaluating under this upper-bound setting serves two purposes: to quantify the maximum theoretical impact of the attack and to provide insight into the extent of performance overestimation associated with approaches that do not respect the aforementioned restriction.

Fig. 3 presents the cell-wise differences in the α-OMC loss between different attack configurations and the random-attack baseline. The results indicate that the proposed attack consistently outperforms the random baseline. They further show that increasing the accuracy of the estimates (high-noise vs. low-noise) improves the impact of the attack, while this effect remains less pronounced than the gain observed when increasing the attack intensity ε. Notably, the "optimal scenario" setting demonstrates the potential impact of an adversary employing a near-perfect traffic-forecasting model. The higher α-OMC disruption observed in this setting compared to the others also highlights that attack scenarios assuming knowledge of future values may overestimate what is realistically achievable in practice.

C. Impact of the Attack on Future Predictions

To assess future effects of the attack (RQ3), we first evaluate the degradation caused by the adversarial perturbations on subsequent impacted forecasts, as mentioned in §III-A. Specifically, we simulate a scenario in which the current time step t represents the last time window during which the attack was executed, targeting the capacity forecast $\hat{y}_{t+1}$. Then, for each newly observed traffic load value x, the system performs another forecast, continuing until the input sequences are no longer affected by the prior perturbations. In our experiment, 6 sequences are impacted, each containing progressively fewer manipulated traffic values.

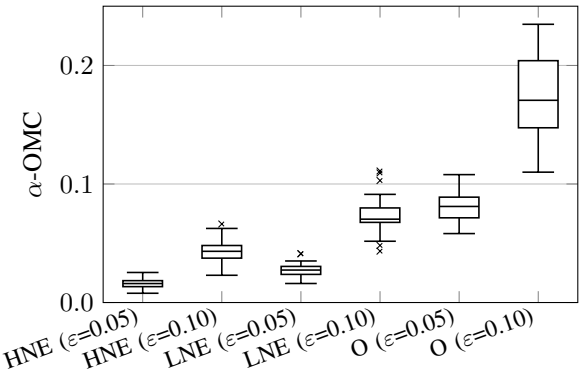

Fig. 3. Box plot of α-OMC difference with the random attack, on 25 cells. HNE and LNE stand respectively for High- and Low-Noise Estimates; O stands for Optimal.

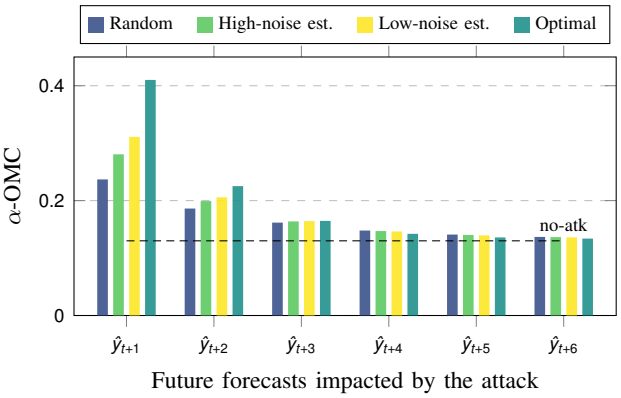

Fig. 4. α-OMC for different attack settings with $\varepsilon = 0.10$ on the Milan dataset.

The results displayed in Fig. 4 show that the attack has its strongest effect on the one-step-ahead forecast $\hat{y}_{t+1}$. Specifically, the low-noise estimator setting shows a 2.4-fold increase in average α-OMC values relative to performance prior to the attack for perturbations equivalent to 10% of the traffic peak ($\varepsilon = 0.1$), and optimal settings indicate that the degradation may reach up to a maximum potential 3.1 times of average base performance, assuming the adversary has access to an even more accurate traffic forecaster. Beyond this point, the impact is still noticeable on $\hat{y}_{t+2}$, then progressively decreases towards performance prior to the attack. This aligns with the expected behavior, as the input sequence contains progressively fewer perturbed traffic values. However, the pronounced impact on $\hat{y}_{t+1}$ is also partially explained by the attack settings, which optimize perturbations solely with respect to this forecast. We can expect that, when targeting multiple future steps (i.e., $N > 1$), the impact would be distributed more evenly across the affected forecasts. Overall, the results reveal a clear ordering of attack effectiveness: the optimal setting yields the most significant impact, followed by the low-noise estimator, the high-noise estimator, and finally the random attack. This trend displays again the dependence of the attack's success on the accuracy of estimates.

D. Updates on Adversarial Perturbations

In real-world settings, data-stream distributions evolve over time, requiring periodic model updates. Such adaptations open the possibility of training on adversarially manipulated data. We therefore decide to evaluate the effect of training

TABLE I
AVERAGE α-OMC GAIN ON MODELS' UPDATE.

ε	Random	High-Noise	Low-Noise	Optimal
0.05	+0.02	+0.01	+0.01	+0.00
0.10	+0.10	+0.08	+0.06	+0.01
0.20	+0.25	+0.11	+0.02	+0.01

models on the adversarial perturbations generated during attacks as part of our investigation of RQ3. Specifically, we compare models trained solely on adversarial data with those trained on clean data. For testing, we use the remaining 2.5 weeks of clean data from the Milan dataset that were not employed in previous experiments. The results of this analysis are displayed in Table I. They display a clear trend: adversarial perturbations derived from more accurate estimates lead to smaller performance degradation for models trained on them. One possible explanation is that, when the attack consistently picks the best directions for the perturbation (i.e., with estimates closer to those used in the optimal scenario), the resulting adversarial examples exhibit structured patterns that can be used for training without degrading the model's forecasting accuracy on clean data. As the estimates become less precise, the perturbations increasingly resemble random noise, which shows the most disruptive impact on the updates.

This finding leads to a counterintuitive implication: perturbations that closely follow the model's loss gradient produce patterns that can be partially learned from, whereas less precise, random-like perturbations cause stronger degradation after retraining. In other words, an attacker aiming for long-term disruption may achieve greater impact by injecting high-variance random noise into the online update process rather than relying on the secondary effects of a carefully crafted attack targeting the performances of the currently deployed model. Additionally, these results suggest that the architecture of the models exhibits a degree of robustness when trained on adversarial perturbations. This indicates that adversarial perturbations crafted to target deployed models do not directly translate into effective poisoning attacks during retraining. It also suggests that purposefully incorporating adversarial examples into the training set, a process referred to as adversarial training, might be feasible without causing substantial deterioration of the models' original performances.

VI. ATTACK PERFORMANCE IN THE ITU DATASET

In the following section, we evaluate the attack's domain generalization by targeting models trained on an alternative data source, the ITU dataset. This analysis shows which invariant properties of the attack persist and further provides insights into the characteristics of the data and models that contribute to the attack's success, thereby answering RQ4.

A. Overall Performance

Fig. 5 presents the cell-wise difference in the degradation of α-OMC values between different attack settings relative to the random attack for the ITU dataset. The results confirm the previously established conclusions: the attack's impact increases with its intensity (represented by ε) and with the quality of estimates of future traffic values. Although the observed trend remains consistent with that reported in Fig. 3, the magnitude of the degradation is noticeably smaller than for the Milan dataset, indicating that the attack has a

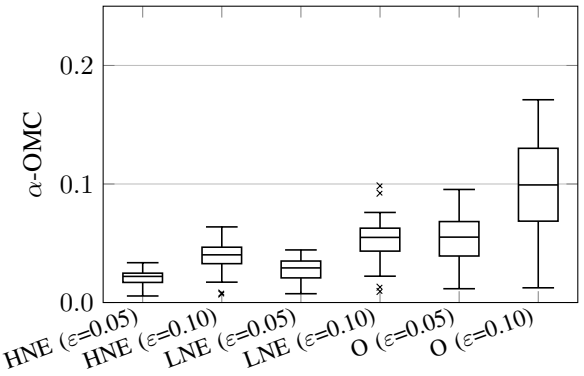

Fig. 5. Box plot of α-OMC difference with the random attack for the 30 cells of the ITU dataset.

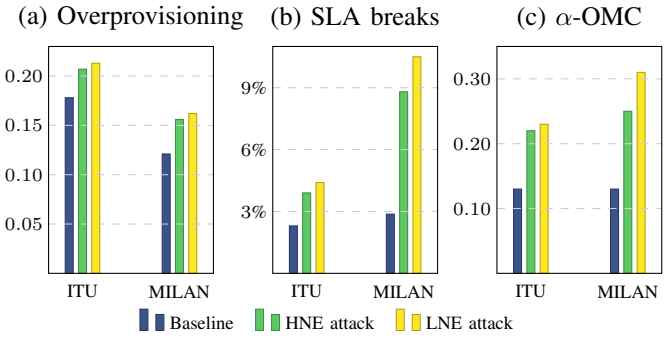

Fig. 6. Comparison of average performance metrics of target models on the two datasets with the baseline (no attack), HNE and LNE attack settings ($\varepsilon = 0.1$). *Overprovisioning* is computed as $\max(0, \hat{y}_i - y_i)$ for each data point.

smaller impact on the ITU dataset despite comparable baseline performance. In the following section, we investigate potential factors contributing to this divergence.

B. Cross-Dataset Result Analysis

Models trained using the DeepCog architecture on both datasets achieve the same average α-OMC value of 0.13 across clean test sets corresponding to their respective cells. Nevertheless, the attack produces a smaller degradation of the capacity forecasts of models trained on the ITU dataset. For instance, with $\varepsilon = 0.1$ and using the low-noise traffic estimates, the mean α-OMC value increases to 0.23 after the attack. Under the same conditions, the corresponding value reaches 0.31 for the Milan dataset. This discrepancy is primarily due to different increases in the SLA violations caused by the attack, as illustrated in Fig. 6, which substantially raises the average α-OMC value. With the aforementioned settings, the attack increases the proportion of SLA breaks by a factor of 3.6 (from 2.9% to 10.5%) for the Milan dataset, whereas this factor is only 1.9 for the ITU dataset (from 2.3% to 4.4%).

This discrepancy is not primarily attributable to the models trained on the ITU dataset being more robust to adversarial perturbations. Indeed, the average variation in forecasts induced by the attack is comparable across datasets: 17.16% for the ITU dataset and 20.83% for the Milan dataset. This difference in magnitude is relatively small and insufficient to account for the observed performance gap. Instead, the discrepancy is mainly explained by the combined effects of the definition of the α-OMC loss function and the characteristics of the initial forecasts produced by targeted models. Fig. 7

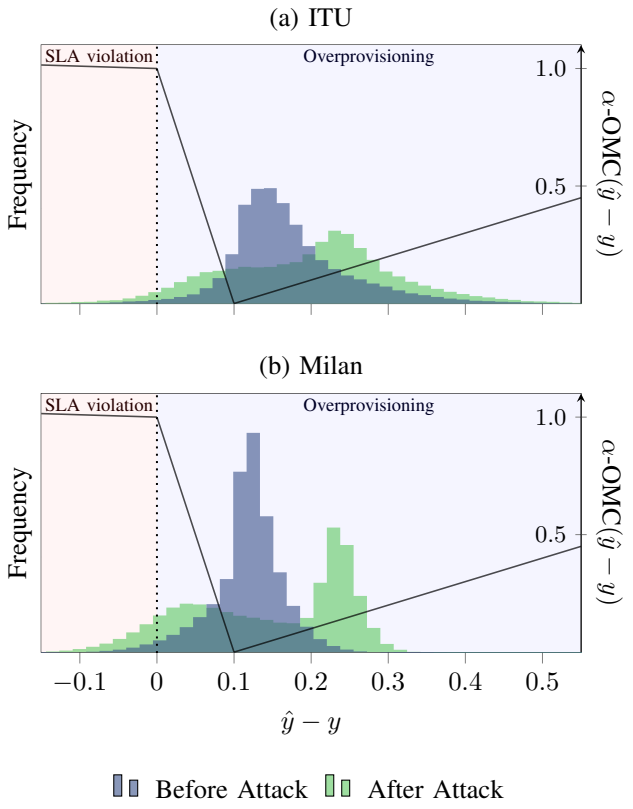

Fig. 7. Distribution of differences between predictions and traffic values on the two datasets and associated α-OMC values for the parameters in use $\alpha = 1$ and $\delta = 0.1$

presents the distribution of the difference between forecasts and observed traffic values for models trained on both datasets. Although the two distributions differ, the average α-OMC value across all predictions is identical before the attack for both datasets, due to the specific shape of the function. More specifically, we observe that the distribution corresponding to the ITU dataset is flatter and more right-skewed that the one corresponding to the Milan dataset. This indicates stronger over-provisioning in the forecasts (i.e., when $\hat{y} - y > 0$) and a smaller proportion of forecasts that violate SLAs (i.e., when $\hat{y} - y < 0$) or lie close to, but above, the violation boundary. In contrast, the Milan dataset exhibits a higher concentration of values in this latter region, which accounts for the larger increase in SLA violations under attack, as these forecasts can be more easily shifted into the associated area (i.e., moved towards the left side of the distribution). This interpretation is supported by the observed shifts in distributions induced by the attack, which moved a higher proportion of samples toward or below the SLA threshold ($\hat{y} - y = 0$) in the case of the Milan dataset. Furthermore, these shifts are consistent with the attack behavior previously described in §V-A. Specifically, forecasts initially well above the SLA violation boundary are further displaced toward higher overprovisioning (i.e., shifted to the right side of the distribution).

These results indicate that when evaluated on the α-OMC function, models with a high proportion of forecasts located just above the SLA threshold are more vulnerable to adversarial perturbations.

VII. Discussion

Temporal restrictions inherent to online time series forecasting substantially narrow the space of feasible adversarial manipulations targeting traffic forecasters, yet they must be accounted for when designing an attack. Although existing work on adversarial attacks against time series forecasters often overlooks these constraints, our results show that, even under such limitations, it remains possible to design attacks that increase over-provisioning and induce SLA violations in capacity forecasting systems. We further show that invariant properties persist during the attack across datasets. We therefore emphasize the need to develop and evaluate mitigations for these attacks, exploring next some preliminary ideas.

Mitigations. Because the effectiveness of the attack scales with the adversary's ability to predict short-term traffic, a potential high-leverage mitigation could aim to reduce this ability while preserving performance in system forecasting objectives. For instance, an Internet Traffic Provider can make it more difficult for an adversary to observe the true sequence of traffic load values associated with a cell by limiting telemetry access or injecting controlled noise into the exposed data stream, thereby increasing the difficulty of crafting appropriate input queries. Strategies limiting the observation of traffic load values may additionally hinder the construction of surrogate models, complicating transfer-based attacks. Additional mitigation may be achieved through Adversarial Defense techniques designed to prevent or reduce the impact of crafted perturbations. For instance, including adversarial examples during model training can improve robustness, in a manner similar to techniques employed in [23] and [24]. Additional defense strategies could involve active monitoring mechanisms that detect data manipulation in real-time, allowing the system to respond upon attack detection. Initial efforts on detecting adversarial perturbations in time series data have primarily focused on the FGSM and BIM attacks [25]. Subsequent research includes the detection of adversarial attacks against time series classifiers [26], as well as more recent investigations regarding time series LLMs [27]. While prior studies have examined Adversarial Defense methods, further research is needed to understand how these approaches can be effectively applied in real-time, online, and on a regression task, as the combination of these characteristics is representative of traffic forecasting systems.

Limitation & Future work. Validating our approach with a broader empirical evaluation—particularly using real deployed forecasting models—will be important for understanding practical behavior under realistic conditions. However, gaining access to such operational systems is challenging, and the number of publicly available datasets for this type of evaluation remains limited, which constrains external validation. Our work is also limited by the scope of our forecast horizon, and it remains unclear how the attack scales when targeting further forecasts for degradation. Therefore, another promising direction is to evaluate the method on long-term forecasting architectures, which would clarify how long input sequences affect both the feasibility and the impact of the attack. Finally, further improving the practicability of online attacks under realistic conditions requires a systematic analysis of how attack effectiveness depends on the adversary's knowledge of the target system. Because direct access to the model or its internal information is often difficult to obtain, attackers may rely on surrogate models or loss-function approximations that only partially capture the target system's behavior, which may lower the attack's impact. Similarly, advancing toward

realistic deployment requires moving beyond noise-based estimates of future values and instead adopting established traffic forecasting models trained under varying levels of data accessibility, thereby reflecting scenarios in which the adversary must independently retrieve the observations associated with a given cell. Despite these limitations, our evaluation provides a controlled setting in which to analyze the core attack mechanism and its impact, allowing us to isolate the fundamental behaviors that future evaluations can build upon.

VIII. Related Work

The study of adversarial machine learning has become a focal point of AI research [3]. As these techniques are increasingly used to drive resource allocation in mobile networks, we first review relevant attacks on time series data and then examine adversarial ML approaches within the broader context of wireless communications.

Existing attacks on time series. Adversarial vulnerabilities in time series prediction have only recently begun to receive attention [28]. Early work adapted image-based methods such as FGSM and BIM to time series classification, demonstrating that perturbations can significantly degrade the accuracy of classifiers [11]. As access to model internals is often limited in practice, several studies employ surrogate models to enable black-box attacks [23], [29], [30]. For example, Liu et al. [19] attack spatiotemporal traffic forecasters by training a surrogate model and applying an iterative gradient-guided method to identify a suitable set of victim nodes. In their work, the authors consider an important characteristic that distinguishes attacking temporal forecasters from conventional classification tasks, which is the inaccessibility of ground truth at the time of attack, highlighting the challenging nature of operating under temporal constraints. Similarly, Gong et al. [13] emphasize the existence of a trade-off between the observation and action space: perturbations are crafted from the observed parts of a sample but can only be applied to the unobserved parts. They show that conventional methods such as FGSM or DeepFool [31] cannot satisfy this requirement. In addition, Su et al. [32] propose the Timestamp-wise Gradient Accumulation Method (TGAM), a gradient-based attack that operates while ensuring identical perturbation values for each timestamp, therefore respecting the temporal interdependency of input sequences caused by their overlap. Similar limitations arise in online adversarial settings [33], highlighting the need for attack strategies tailored to sequential and partially observable inputs.

Adversarial ML applied to Wireless Communications. The demonstrated ability of adversarial perturbations to disrupt wireless communication has spurred growing interest in both attacks and defenses in the field [3]. In the context of traffic forecasting, Zheng et al. [30] developed a gradient-based poisoning attack that leverages a surrogate model to degrade predictors trained on the Telecom Italia dataset. Beyond poisoning, evasion attacks have also been explored across tasks such as semantic communications [34] and RF signal classification [23]. In parallel, defenses are emerging. Detection-oriented approaches aim to identify perturbed inputs [35] for cooperative spectrum sensing. To strengthen intrinsic robustness, Li et al. [36] propose an adversarially robust modulation recognition system capable of maintaining accuracy under attack.

IX. Conclusion

In this paper, we examined adversarial attacks on capacity forecasters under temporal restrictions anchored to online time series forecasting. We demonstrated that, even when the capacities of the adversary are further limited by these restrictions, perturbations can increase over-provisioning and provoke SLA violations. Our evaluation using the DeepCog architecture and real-world datasets further reveals that the effectiveness of such attacks is largely driven by the attacker's ability to estimate future traffic. Moreover, the potential impact of an attack depends on the tendency of the target capacity forecaster to produce predictions slightly above the SLA threshold on natural data. These findings highlight that temporal constraints restrict but do not eliminate adversarial threats to network forecasting systems. Potential extensions of this work are clearly identified and include, among others, considering more than a single forecast when selecting the perturbation direction, investigating how attack effectiveness depends on the fidelity of the attacker's approximation of the target system, as well as implementing an actual traffic forecasting model controlled by the adversary rather than relying on simulated estimators.

Acknowledgment

L. Miermont's work has been funded by Comunidad de Madrid predoctoral grant PIPF-2023/COM-31147. His work was supported by project PID2022-143304OB-I00 (PARASITE), funded by MICIU/AEI /10.13039/501100011033/ and by the ERDF, EU. C. Fiandrino is a Ramón y Cajal fellow (RYC2022-036375-I), funded by MCIU/AEI/10.13039/501100011033 and ESF+. His work was also partially supported by the TUCAN6-CM project (TEC-2024/COM-460), funded by the Madrid Regional Government (ORDEN 5696/2024). G. Suárez-Tangil is a Ramón y Cajal fellow (RYC2020-029401-I), funded by MCIU/AEI/10.13039/501100011033 and the ESF Investing in your future. His work was supported by a 2025 Leonardo Grant for Scientific Research and Cultural Creation from the BBVA Foundation (SHIFT, reference LEO25-1-19720). The BBVA Foundation accepts no responsibility for the opinions, statements, and contents included in the project and/or the results thereof, which are entirely the responsibility of the authors.

References

[1] V. e. a. Gudepu, "Adaptive Retraining of AI/ML Model for Beyond 5G Networks: A Predictive Approach," in *Proc. of IEEE NetSoft*, 2023, pp. 282–286.

[2] S. M. Gholian *et al.*, "DeExp: Revealing Model Vulnerabilities for Spatio-Temporal Mobile Traffic Forecasting with Explainable AI," *IEEE Trans. on Mobile Computing*, pp. 1–18, 2025.

[3] D. Adesina *et al.*, "Adversarial machine learning in wireless communications using RF data: A review," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 77–100, 2022.

[4] T. Wu *et al.*, "Small perturbations are enough: Adversarial attacks on time series prediction," *Information Sciences*, vol. 587, pp. 794–812, 2022.

[5] D. Bega *et al.*, "DeepCog: Optimizing Resource Provisioning in Network Slicing With AI-Based Capacity Forecasting," *IEEE Journal on Selected Areas in Comms.*, 2020.

[6] ——, "α-omc: Cost-aware deep learning for mobile network resource orchestration," in *Proc. of IEEE INFOCOM WKSHPS*, 2019, pp. 423–428.

[7] A. Chakraborty *et al.*, "A survey on adversarial attacks and defences," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 25–45, 2021.

[8] I. J. Goodfellow *et al.*, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[9] A. Kurakin *et al.*, "Adversarial examples in the physical world," in *Artificial int. safety and security*. Chapman and Hall/CRC, 2018.

[10] C. Wang *et al.*, "Black-box adversarial attacks on deep neural networks: A survey," in *Proc. of IEEE ICDIS*, 2022, pp. 88–93.

[11] H. I. Fawaz *et al.*, "Adversarial attacks on deep neural networks for time series classification," in *Proc. of IEEE IJCNN*, 2019, pp. 1–8.

[12] F. Karim *et al.*, "Adversarial Attacks on Time Series," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3309–3320, 2021-10.

[13] Y. Gong *et al.*, "Real-time adversarial attacks," in *IJCAI*, 2019, pp. 4672–4680.

[14] J. Cortellazzi *et al.*, "Intriguing properties of adversarial ml attacks in the problem space," *ACM Trans. on Privacy and Security*, 2025.

[15] H. Wan *et al.*, "NR-Scope: A Practical 5G Standalone Telemetry Tool," in *Proc. of ACM CoNEXT*, 2024.

[16] T. D. e. a. Hoang, "LTESniffer: An Open-source LTE Downlink/Uplink Eavesdropper," in *Proc. of ACM WiSec*, 2023, pp. 43–48.

[17] N. Ludant *et al.*, "From 5G sniffing to harvesting leakages of privacy-preserving messengers," in *Proc. of IEEE S&P*, 2023, pp. 3146–3161.

[18] B. Poudel *et al.*, "Black-box adversarial attacks on network-wide multi-step traffic state prediction models," in *Proc. of IEEE ITSC*, 2021, pp. 3652–3658.

[19] F. Liu *et al.*, "Practical adversarial attacks on spatiotemporal traffic forecasting models," *Proc. of NeurIPS*, vol. 35, pp. 19 035–19 047, 2022.

[20] G. Barlacchi *et al.*, "A multi-source dataset of urban life in the city of milan and the province of trentino," *Scientific data*, vol. 2, no. 1, pp. 1–15, 2015.

[21] Zindi. Spatio-Temporal Beam-Level Traffic Forecasting Challenge by ITU. [Online]. Available: https://zindi.africa/competitions/spatio-temporal-beam-level-traffic-forecasting-challenge

[22] P. F. Pérez *et al.*, "An in-depth analysis of advanced time series forecasting models for the Open RAN," in *Proc. of IEEE INFOCOM WKSHPS*, 2024, pp. 1–6.

[23] W. Zhang *et al.*, "Stealthy Adversarial Attacks on Machine Learning-Based Classifiers of Wireless Signals," *IEEE Trans. on Machine Learning in Communications and Networking*, vol. 2, pp. 261–279, 2024.

[24] L. Liu *et al.*, "Robust multivariate time-series forecasting: Adversarial attacks and defense mechanisms," *ICLR*, 2023.

[25] M. G. Abdu-Aguye *et al.*, "Detecting adversarial attacks in time-series data," in *Proc. of IEEE ICASSP*, 2020, pp. 3092–3096.

[26] ——, "Recurrence-based disentanglement for detecting adversarial attacks on timeseries classifiers," in *EUSIPCO*. IEEE, 2023, pp. 625–629.

[27] H. Ma *et al.*, "Keep the lights on, keep the lengths in check: Plug-in adversarial detection for time-series llms in energy forecasting," *arXiv preprint arXiv:2512.12154*, 2025.

[28] H. Wu *et al.*, "Timesnet: Temporal 2d-variation modeling for general time series analysis," in *Proc. of ICLR*, 2023.

[29] J. Hutchins *et al.*, "Black-Box Adversarial Attacks on Spiking Neural Network for Time Series Data," in *ICONS*, 2024.

[30] T. Zheng *et al.*, "Poisoning Attacks on Deep Learning based Wireless Traffic Prediction," in *Proc. of IEEE INFOCOM*, 2022, pp. 660–669.

[31] S.-M. Moosavi-Dezfooli *et al.*, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proc. of IEEE CVPR*, 2016, pp. 2574–2582.

[32] R. Su *et al.*, "Temporally Unified Adversarial Perturbations for Time Series Forecasting," *arXiv preprint arXiv:2602.11940*, 2026.

[33] A. Mladenovic *et al.*, "Online Adversarial Attacks," *arXiv preprint arXiv:2103.02014*, 2022.

[34] V.-T. Hoang *et al.*, "Adversarial Attacks Against Shared Knowledge Interpretation in Semantic Communications," *IEEE Trans. on Cognitive Communications and Networking*, vol. 11, no. 2, pp. 1024–1040, 2025.

[35] W. Zhao *et al.*, "Detecting Adversarial Spectrum Attacks via Distance to Decision Boundary Statistics," in *Proc. of IEEE INFOCOM*, 2024, pp. 691–700.

[36] G. Li *et al.*, "Adversarial Robust ViT-Based Automatic Modulation Recognition in Practical Deep Learning-Based Wireless Systems," in *Proc. of IEEE S&P*, 2025, pp. 3672–3690.